A MODULAR HEAD/EYE PLATFORM
FOR REAL-TIME REACTIVE VISION

P. M. SHARKEY, D. W. MURRAY, S. VANDEVELDE*, I. D. REID, and
P. F. McLAUCHLAN

Robotics Research Group, Department of Engineering Science,
University of Oxford, Parks Road, Oxford, OX1 3PJ, UK
*Department of Mechanical Engineering,
Katholieke Universiteit, Leuven, Belgium

Abstract — This paper describes the design, implementation and testing of a high speed controlled stereo “head/eye” platform which facilitates the rapid redirection of gaze in response to visual input. It details the mechanical device, which is based around geared DC motors, and describes hardware aspects of the controller and vision system, which are implemented on a reconfigurable network of general purpose parallel processors. The servo-controller is described in detail and higher level gaze and vision constructs outlined. The paper gives performance figures gained both from mechanical tests on the platform alone, and from closed loop tests on the entire system using visual feedback from a feature detector.

1. INTRODUCTION

It is evident that sensing and perception are essential precursors to intelligent robotic action in dynamic and unstructured environments. It is therefore surprising to note that the principal emphasis of research into computational visual sensing and perception over the last quarter century has not been the enabling of actions per se, but rather the creation of an over-arching percept of what is in the scene and where it is relative to the observer. This approach, espoused most forcibly in the writings of Marr [1], has, many would argue, reached something of an impasse. Creating and maintaining the generally useful percept proves too hard in terms of computational theory, representational complexity and computational cost.

The last few years have seen the emergence of a new paradigm in machine vision, that of active vision. Also described variously as animate, reactive, purposive or behavioural [2, 3, 4, 5], this paradigm seeks to remarry sensing and perception to the specific robotic actions that are being executed. One can regard the vision system as a collection of specialists tuned to specific functions, rather than a monolithic but undirected hoarder of percepts. This has several far reaching consequences. First, visual data can be used in all its richness — no longer does it have to be of sufficient quality to drive 3D reconstruction. Secondly, each vision specialist need only represent that required to perform its particular action. Thirdly, if the sensor can compute on the fly considerable faster than the world is changing, it can use the world itself as its own
best memory. There is no necessity to construct and maintain an internal description of the entire surroundings.

The obvious outward manifestation of an active vision system is a camera steering system which allows gaze to be directed at an area of interest in the scene under the control of the vision system itself. At one level, this hardware focussing of attention lightens the computational burden, but it is becoming clear that, at a deeper level, problems in perception and acting are made easier by the ability to maintain fixation on a specific point in the scene.

To date within the active vision paradigm there has been emphasis on (i) the execution of defined camera movements, and (ii) the ability to fixate on a scene point and to track it smoothly. Both make the recovery of 3D structure from 2D image motion simpler and better posed. However, relatively little consideration has been given to the issues raised when one asks how should the vision system start to fixate and track, and when should it stop.

In our work, we have been exploring the role of scene and image motion in this cycle of distraction by, and attraction to, objects in the scene—a cycle fundamental to any surveillance and many navigation tasks. Such reactive vision requires considerably more from the visual apparatus, not only in terms of raw mechanical performance of the steerable cameras, but also in terms of the bandwidth of the visual sensors and plant controllers.

In this paper, we describe the steps involved in the development of such a system. Bandwidth and accuracy considerations for reactive control of gaze rule out mounting the cameras either on industrial manipulators or the use of commercial surveillance apparatus, and there is a need to build a specialised mechanism. We start with a task specification, described in the next section, and use this to motivate the design specification and implementation, described in Sections 3 and 4. We discuss the system architecture as it pertains to control and vision in Sections 5 and 6, respectively. The performance of the system is assessed by mechanical and visuo-mechanical tests in Section 7. For an account of experimentation into active vision using the system, the reader is referred to [6, 7].

2. TASK SPECIFICATION

To focus the design specification we shall consider a surveillance system able to act in the following scenario. The head system is gazing idly in a room where there is no activity. A person now enters the room. The system is attracted by the moving person and shifts its gaze to initiate tracking of the person as (s)he crosses the room. While the system is tracking, someone else enters the room. The system must be capable of rapidly redirecting its attention to the new target. It must be able to maintain the target in sharp focus and, if so desired, zoom in on the target. Finally, the system must be able to respond to threatening behaviour in order to protect the cameras.


Given the scenario above, we can quantify the task specification for the complete gaze control system as follows:

(i) Nominal surveillance distance: The nominal distance to the target under surveillance is \( \sim 5m \), the range being from 2m to \( \infty \).

(ii) Target tracking: The target could reach a maximum velocity of \( 8ms^{-1} \) (e.g., a person sprinting across the area under surveillance).

(iii) Target capture: The system should be able to capture the target travelling at maximum velocity at a minimum distance of 2m from the surveillance head.
(iv) **Target detail:** A magnification of 10:1 should be attainable and the target should be maintained in sharp focus.

This task specification is distinct from the design specification in that it addresses the task, i.e. **what** the system is required to do, rather than the implementation, i.e. **how** the system performs the task.

### 3. DESIGN SPECIFICATION & IMPLEMENTATION: DYNAMICS

In order to address the issues raised by the typical surveillance scenario above we must supplement the active vision paradigm with the further requirement that the system be **reactive**, in that it must be able to redirect attention **rapidly** in response to “interesting” events. To do so requires relevant apparatus: a high performance locally controlled mechanical device to steer the cameras, high bandwidth visual feedback, and a gaze controller to integrate the numerous feedback loops and minimise the effects of processing delays. This section focuses on the design and implementation of the steerable camera mount.

#### 3.1. Dynamic Specification

Given the task specification we can determine the required minimum dynamic specification for the mount. To do so requires that we specify the maximum delay permissible between a fast moving target entering the surveillance area and the gaze system locking onto that target. This delay will be a combination of a controller/mechanical delay and a processor delay in the sensor feedback — the former being largely determined by the drives used, and the latter being dependent on the visual processing.

We originally estimated that a visual motion detector algorithm, running at frame rate (25Hz in our case), might have a latency of up to 150\(\text{ms}\)^1. As the motion detector is used to initiate saccades (i.e. fast redirection of attention based on a fed-forward trajectory determined from the motion detector), it is essential that the positional and velocity estimates closely match those of the target. For this reason the motion detector process may take up to three frames to yield a close estimate of the target position and velocity resulting in an additional maximum latency of 120\(\text{ms}\) — i.e. the target may be within the field of view for a maximum of 120\(\text{ms}\) before three images are grabbed (minimum 80\(\text{ms}\)).

The response to a demand for rapid redirection of gaze (as in a saccade) will lie principally in the performance of the vergence and elevation axes. The pan axis need not have as high a performance — indeed there are strong arguments for decoupling the pan and vergence axes by limiting the bandwidth of the pan axis. The specifications can be determined from the following:

(i) **Resolution & Repeatability:** these must be preserved to sub-pixel accuracy at full zoom.

(ii) **Response:** the vergence and elevation must be able to detect and track and stabilise on an object with a maximum velocity of 8\(\text{ms}^{-1}\) at a minimum distance of 2\(\text{m}\) from the camera, under the assumptions of a nominal field of view of 40° and a maximum measurement delay of 270\(\text{ms}\). The time constraint is that the target is centred before the range limits are violated.

---

^1Subsequent implementation of the algorithm have indicated that the maximum latency is 115\(\text{ms}\).
Table 1: Design Specification for a head/eye camera platform.

Using a nominal range limit of 60° on the vergence axis and calculating the trajectory, the worse case time constraint requires the vergence/elevation axes to be capable of saccading through 60° within 200ms, while the pan axis must be capable of attaining a velocity of 60°s⁻¹ within the same time period, and with a maximum pan velocity of 150°s⁻¹. Maximum payload for the vergence axis is to be 5kg, giving a range of load inertias (for nominal dimensions given below) from 0.02kgm² (at zero offset) to 0.22kgm² (at an offset of 200mm). The inertia for the pan axis will be considerably larger, and depend on the inertias of the elevation and vergence actuators, sensors, and links. The design specifications derived from the data given is presented in Table 1.

The elevation and vergence inertias have been increased to allow for servo actuator and link structure inertias. On the assumption that all targets encountered will be less reactive in the vertical plane, the elevation acceleration specifications have been reduced accordingly. A further allowance that may be made is that the collaborative interaction of pan axis with the vergence axis will allow us to lower this design specification slightly.

3.2. Motors & Transmission.

The above task specification, and resulting design specification, highlight the key performance requirement: that of high accelerations and decelerations. A further important requirement is that the tracking capability of the system be smooth and the system be vibration free. It is interesting to note that stepper motors, used in several recent designs [8, 9], although easy to interface, suffer from low accelerations, poor velocity tracking performance and poor friction characteristics, making such drives less than ideal for a responsive gaze control system. Other designs have vibrational problems [10] either due to the mechanism itself or the platform the head is sited on.

The high accelerations and smooth tracking requirements argue for DC direct or geared drives. The principal benefits of using direct drive are zero backlash and negligible friction. These are outweighed, however, by the susceptibility of direct drive (DD) trains to disturbance torques, lower resolution and, when limited by size and weight, lower accelerations. The differences between geared and direct drive are detailed below.

3.2.1. Accelerations vs Speed. While DD motors have very high output speed capability, they are limited by the acceleration when loaded. Conversely, a geared drive will be limited in speed but will have higher acceleration. This distinction becomes crucial when the axis is limited in operating range, as is the case for a gaze control system. Fig. 1 illustrates the different performances for geared and direct drive transmissions and the times taken to execute saccades up to 200°. The graphs show that for limited travel the geared drive has a considerably faster response. The figures for these graphs were taken from experimental data on two drives that were kinematically similar and were carrying the same inertial load.
Figure 1: (a) Maximum velocities and (b) times to execute saccades for geared (—) and direct drive (−−) transmissions. The dotted line indicates the DD performance where the range of motion is limited to ±60°.

3.2.2. Load torques. In DD transmissions the load inertia will dominate over the rotor inertia so that each axis will need to be statically balanced whenever the loads (e.g., cameras) are changed. Furthermore, very careful consideration must be given to the wiring design to minimise load torques due to the wiring loom. The effect of changes in inertia can be seen using the simplified torque equation (i.e., ignoring friction):

\[
\ddot{\theta} = \frac{N_{\text{gear}} \tau_{\text{motor}}}{N_{\text{gear}}^2 (I_{\text{motor}} + I_{\text{gear}}) + I_{\text{load}}}
\]

where \(\ddot{\theta}\) is the acceleration, \(N_{\text{gear}}\) is the gear ratio, \(\tau_{\text{motor}}\) is the torque generated at the motor, and \(I_{\text{motor, gear, load}}\) are the relevant inertias. For a DD motor \(N_{\text{gear}} = 1\), \(I_{\text{gear}} = 0\) and \(I_{\text{motor}} \ll I_{\text{load}}\), so that the acceleration can be closely approximated by \(\tau_{\text{motor}} / I_{\text{load}}\), which will vary as the load varies. On the other hand, for a geared drive the inertias of the motor and gearbox will usually be greater than the load inertia and the accelerations can be approximated by \(\tau_{\text{motor}} / N_{\text{gear}} (I_{\text{motor}} + I_{\text{gear}})\), a constant. The robustness of a geared drive to changing loads has been verified and will be presented in Section 7.

3.2.3. Dynamic Response. Even with gravity compensated devices, where the load on each joint is reduced to a pure rotational inertia, the cross-coupling between joints during a dynamic trajectory introduces Coriolis forces which are inversely proportional to the gear ratio. Thus for a multi-link device which uses direct drive transmissions, precise modelling is necessary to be able to compensate for these Coriolis forces, especially during gross motions such as a saccade. For geared multi-link systems, the compensation required will be small and can be accommodated in the error recovery of the servo controller.

3.2.4. Resolution. For the same encoder or resolver, a geared drive has the advantage of increasing the resolution of the link by a factor equal to the gear ratio, but the disadvantage of requiring some form of additional calibration sensor. For DD, very high resolution can be obtained by using laser encoders or precision resolvers, but at greater expense.

3.2.5. Backlash. Backlash must be kept to within the repeatability specification, i.e. sub-pixel. DD motors have zero backlash while geared drives require precision gear-boxes or complicated anti-backlash mechanisms, again increasing the cost.

3.2.6. Compliance. Similarly compliance, not present in DD, can be reduced in the more
expensive geared servos.

3.2.7. Friction. As with stepper motors, geared drives suffer from friction and stiction, both of which are negligible in DD transmissions. Although stiction introduces a phase lead at low frequencies [11], friction increases the deceleration capabilities and can, therefore, help in the dynamic braking of the joint.

3.3. Implementation.

To enable each axis to achieve the higher accelerations required, especially when using higher inertial loads due to the motorised lenses, and to minimise the effects of load inertias and Coriolis forces, we have chosen a geared transmission, as detailed in Fig. 2.

The transmission for each axis is comprised of a DC motor with a Harmonic Drive gearbox. The gear-boxes incorporate flexispine harmonic gears admitting minimal backlash (less than 27 seconds of arc), though friction is noted to be quite high. A gear ratio of 50:1 was chosen to maintain the maximum required acceleration and velocity. The output of the gearbox is coupled to the driven link via a torsionally rigid flexible coupling — the flexibility allows for slight misalignment between the actuator and link shafts.

At maximum zoom, the field of view is reduced to 5°. To maintain sub-pixel accuracy the positional resolution of the vergence and elevation axes must be better than $5^\circ/512\text{pixels} \equiv 0.01^\circ/\text{pixel}$. Consequently, the motor shafts are instrumented with high precision (5000 line) incremental optical encoders offering a maximum motor positional resolution of 0.018° (using
full quadrature counters) which translates to an output or link resolution of 0.00036°. This resolution is within the repeatability of the drive 0.0075°, which, in turn, is well within the sub-pixel accuracy requirement. The output shaft of the gearbox also carries a small magnet to allow precise calibration of the axis on start-up via bounce-free Hall sensor chips.

The theoretical acceleration of the unloaded servo-actuators, instrumented with the precision encoders is 9500°s⁻², with a maximum velocity of 600°s⁻¹. More details of the actual performance of the full 4-axis platform will be presented in Section 5. Finally, each link is coupled to the previous link via pre-loaded precision bearings which minimise any lateral load torques on the actuator shaft.

4. KINEMATICS: SPECIFICATION & IMPLEMENTATION

Two practical kinematic designs are available, as sketched in Fig. 3 (from [12]):

(i) Helmholtz (or common elevation): in which both cameras tilt up and down and verge independently about axes perpendicular to the elevation plane; and

(ii) Fick (or gun-turret): where each camera verges around an axis which remains vertical, and elevates about an independent horizontal axis.

The imaging conditions are different in each case, but we have been unable to find any overriding benefits for employing one configuration over the other. The extra degree of freedom in the form of the pan axis increases the field of view and enables symmetry to be maintained in stereo viewing in steady state.

To preserve the accuracy of the higher acuity 3D tasks, such as accurate depth reconstruction, the limited resolution of the cameras requires that the baseline between the cameras be increased as the distance of the object increases. On the other hand a fixed baseline reduces the complexity of the vision algorithms considerably. Thus a fixed baseline is specified. For surveillance distances of ~ 5m, a baseline of ~ 500mm is appropriate.

Some active vision algorithms assume that the effect of head motion on vision is that of pure camera rotation, i.e. that all axes of rotation pass through the camera optical centres. For any stereo head (where both cameras are separated from the pan axis) this assumption is invalid. However, with a baseline of 500mm, our calculations indicate that the visual error due to the pan axis offset will be un-measurable in the image (ie less than a half of a pixel) for distances (≥ 2m). Thus, for objects at distances ≥ 1.5m, to eliminate the effects of axis offset in the
Figure 4: Modularity of design and alternative configurations: Helmholtz on the left, and Fick on the right.

...elevation and vergence axes, we limit the allowable offset between an axis of rotation and its corresponding optical axis to \( \leq 375\,mm \).

\subsection{Implementation.}

One of the principal aims of the design has been versatility without sacrifice to performance or weight. The mount is therefore designed to be modular in that each of the vergence axes and elevation axes drive trains are identical, the azimuth or pan axis is simply an enlarged version of the design shown in Fig. 2. The addition of an extra degree of freedom is simple to implement, as shown in Fig. 4.

The versatility of the design allows the system to be configured in the standard common-elevation (or Helmholtz) model or reconfigured with ease into the independent “gun-turret” (or Fick) model, Fig. 4. Both models offer advantages and disadvantages: while the common-elevation model allows both elevation axes to be mechanically coupled thus reducing the degrees
of freedom to four, the gun-turret model has the advantage that each camera can “maintain
the vertical”, having zero rotation or cyclotorsion about the optic axis. In both configurations,
one axis is coincident with one of the optical axes (vergence axes for the Helmholtz config-
uration and elevation axes for the Fick configuration), the other having a small offset. The
configuration implemented is the common-elevation model, where the mechanical coupling is
achieved at the motor shafts via a torsionally rigid coupling. The baseline is 550mm and the
offset between the elevation axis and the camera/lens attachment bracket is 50mm, which is
within the kinematic design specification. Figure 5 shows a CAD view of the final design, as
well as a cutout sectional view. This design has been implemented at Oxford as Yorick\textsuperscript{2}, as
shown in Fig. 6.


The camera/lens specification comprises an externally synchronised 512 × 512 CCD camera,
with electronic shutter and a zoom lens with the following functionality: auto iris, with manual
override, and a maximum aperture of f1.6; motorised focus; and motorised zoom giving a field of
view from \(\sim 75^\circ\) to \(\sim 5^\circ\). The computer vision community as a whole are currently restricted
by the lack of light weight lenses with motorised zoom, focus and iris (without resorting to
military specifications and their associated cost). Commercially available camera/lens combi-
nations which incorporate these extra three degrees of freedom tend to be reasonably large (up
to \(\sim 200\)mm \(\times 60\)mm \(\times 60\)mm) and heavy (some weighing up to 4kg), thus requiring more
powerful drives, with the inherent disadvantages of size, drive train weight and, not least, cost\textsuperscript{3}.

5. SYSTEM ARCHITECTURE & CONTROL

The essential features of a gaze control system are sketched in much simplified form in Fig. 7.
The key feature of such a system is the communications between the controller and the visual
sensors. These communications, we emphasize, are intimately related: in addition to visual
events providing the primary error signals for gaze control, many vision algorithms can be
improved through direct communication with gaze control to determine the current status
of the mount. Processing delays in machine vision, if unknown, will seriously degrade the
performance, so the issue of time-stamping is crucial to the success of the system. At a higher
level, visual processes drive the attention redirection mechanisms through the controller, which
must in turn mediate switching of visual competences to match the current behaviour.

The centrepiece of system architecture is a configurable network of general purpose proces-
sors (transputers) which have on-chip hardware for supporting communication protocols. The
host development systems for control (an Intel 386 machine) and vision (a Sparc-station) are
independent but the homogeneity of the transputer network facilitates communication between
gaze control and vision.

5.1. The Effect of Delays.

One of the principal potential causes of instability and poor performance is rooted in the
latencies of the system under control — sensor delays from image capture and vision processing,
and actuator delays in the controller and mechanics [9, 13, 14, 15, 16]. Since a time delay merely
shifts a signal in time and has no effect on its magnitude, it is seen that the delay introduces a

\textsuperscript{2}Hamlet, Act V, Scene I.

\textsuperscript{3}Such light weight lenses are, however, readily available in the related “still” photography market.
Figure 5: CAD drawings showing a cut out view of the assembly.

Figure 6: Yorick.
phase shift that grows increasingly negative in proportion to the frequency [17]. Larger delays will introduce larger phase shifts, thus reducing the phase margin and pushing the system toward instability. One common method of reducing the phase margin and pushing the system (ie. a control delay) is to use Smith predictors. The Smith predictor, however, relies not only on precise knowledge of the delay but also requires an exact model of the plant. Furthermore, we cannot assume, as in [9], that the delay is principally a control delay, and not a sensor delay. In fact, quite the contrary: the major delay in the system is in the image capture and vision processing. Delays ranging between 40ms and 150ms are common for the simple 2D processes, but might extend to several hundreds of milliseconds, if not seconds, for more sophisticated 3D processes. Any predictive filtering introduces further delays in the bootstrapping phase. For example, a coarse motion algorithm with a processing delay of up to \( \sim 150ms \) might take 3 frames in the bootstrapping phase to provide an accurate cue to initiate a saccade, yielding a total delay of 270ms.

Although the sensor delays are much greater than the actuator delays, both have a distinct effect on the stability of the system. To achieve an optimum response both delays must be known. If any delay is unknown, the system will have to be detuned to provide a stable, though very sluggish, response to visual cues, based on some estimate of the maximum delay. This strategy would be completely unacceptable for a reactive vision sensor. Therefore, it is imperative that the gaze control system knows the precise delay involved in the capture and appropriate processing of each image and subsequent control delay of the mount. To that end we have adopted a system architecture which explicitly allows (a) the precise determination of the sensor latencies and (b) the minimisation of control latencies. The architecture is shown in Fig. 8.

5.2. Local Servo Control.

The principal cause of actuator delays in many systems lies not in the mechanical response of the actuator itself, but in the local controller and the interface to that controller. The interface of a higher gaze controller (which interprets the vision results) to a commercial robot controller, such as VAL-II, will incur delays that can be long and varied due to the separate subsystems involved, effectively giving rise to an asynchronous servicing of the demand. By
implementing the controller and vision processing on transputers, which have hardware support for inter-processor communications, these delays are minimised, or eliminated [18].

The control architecture is divided into two subdivisions: the low level servo controller and a higher level gaze controller. To overcome the problems of clock skews between different independent processors the sample rate (500Hz), and therefore system clock, is defined by the servo controller. The servo carries out all synchronous controls, time-stamping, ring buffers and trajectory limiting, as well as (asynchronously) the inverse and forward kinematics. At the highest priority the gaze controller generates a synchronous demand for the servo and uses the remaining bandwidth for the higher level asynchronous constructs: pursuit, saccade, predictive filtering, focus, zoom, attentional shifts, and so on.

5.2.1. Controller/Mount Interface. Standard 12-bit DACs interface the transputer servo controller directly to the motor amplifiers. The encoders are similarly interfaced via a custom designed board which accommodates the encoder counters, index pulses and Hall sensor pulses for all four axes as well as generating a velocity estimate from the encoder pulse width. The velocity is given by

$$\dot{\theta}_i = \left\{ \frac{2\pi}{N_{li}N_{gi}} \right\} \frac{Clock}{COUNT_i}$$

where $N_{li}$ is the number of (quadrature) encoder lines per motor revolution, $N_{gi}$ is the gear ratio, $Clock$ is the onboard clock and $COUNT_i$ is the clock count for the duration of the encoder pulse width. Linear amplifiers are used for the elevation and vergence axes to minimise noise in the interface rack, while pan axis uses a stand-alone PWM amplifier.
5.2.2. Servo Controller Latency. The resulting latency from reading the latest positions and velocities to applying the appropriate current to the motors for all four axes is $280\mu s$ — the axes delays therefore vary between $\sim 100\mu s$ for the left vergence axis to $\sim 280\mu s$ for the pan axis. These servo delays are, however, almost constant, deviating from the stated figures by at most $20\mu s$ to any service request for odometry from vision. As the local controller and mechanical delays are constant, the controller can be tuned to give the required response.

5.2.3. Servo Controller. The servo control implemented is based on a “square root” control (SRC) scheme. The control law is very similar to that of a PID controller where the positional error is replaced by the square root of the positional error, with appropriate modifications $[15]$, see Appendix A. The general effect of introducing this nonlinearity into the control scheme is to generate equivalently infinite positional gains as the zero positional error tends to zero, while limiting saturation of the actuator at high positional errors $[20, 15, 19]$, see also $[21]$. Figure 9 shows the response of the vergence axis to a varied trajectory, where the SRC is augmented with integral action with varying success. The bottom graph, Fig. 9(d) illustrates the response of the vergence axis to sub-pixel demands, where the amplitude of the step is equivalent to 1 pixel with a standard field of view lens.

6. SYSTEM ARCHITECTURE & VISION

The vision processes can be categorised into two types: coarse, peripheral processes which are designed to provide early triggers for saccades, and higher acuity, quasi-foveal processes mainly used for tracking. The vision system is implemented on a large hybrid parallel processor, consisting of pipelined processors for early image-wide operations and with a fixed delay of the order of $25ms$, followed by a configurable network of transputers for 2D and 3D vision processes, with varying delays. Currently, the former are used for operations such as image capture, convolution and frame-store whilst the latter operates on either sub-sampled versions of the image (e.g coarse motion detector) or on small image windows for higher acuity operations such as corner detection and tracking. Each image is tracked from time of image capture to receipt of the results at the gaze controller as described below.

6.1. Time-stamping.

The low-level servo controller, as part of its $500Hz$ control loop, maintains a ring buffer of mount status data, such as position and velocity and control mode (e.g. saccade or smooth pursuit) at the time of image capture. Clearly, the availability of this precise odometric and state information from the mount and low-level controller can greatly simplify many of the vision tasks to be performed (see, for example $[6]$).

The fixed (and precisely known) pipeline delay is then used to determine the position in the controller’s ring buffer corresponding to the field capture time (relative to the controller’s clock). This capture time, along with the relevant mount parameters is propagated back through all vision pipelines, having the dual advantage that:

(i) vision algorithms can make use of mount status information; and
(ii) the precise delay involved for any one motion process can be computed by the gaze controller upon receipt of vision results (current system clock time minus time at image capture).
Figure 9: Simple trajectory following using the Square Root Controller a) without integral action, b) with simple integral action, c) with selective integral action, and d) illustrating the sub-pixel control of the head.
<table>
<thead>
<tr>
<th>Description</th>
<th>Vergence</th>
<th>Elevation</th>
<th>Azimuth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axis Range</td>
<td>360°</td>
<td>360°</td>
<td>360°</td>
</tr>
<tr>
<td>Max Acceleration</td>
<td>6000° s⁻²</td>
<td>5000° s⁻²</td>
<td>3000° s⁻²</td>
</tr>
<tr>
<td>Max Deceleration</td>
<td>10000° s⁻²</td>
<td>9000° s⁻²</td>
<td>5000° s⁻²</td>
</tr>
<tr>
<td>Max Slew Rate</td>
<td>400° s⁻¹</td>
<td>400° s⁻¹</td>
<td>240° s⁻¹</td>
</tr>
<tr>
<td>90° Slew Time</td>
<td>0.28s</td>
<td>0.29s</td>
<td>0.76s</td>
</tr>
<tr>
<td>360° Slew Time</td>
<td>0.95s</td>
<td>0.97s</td>
<td>1.69s</td>
</tr>
<tr>
<td>Time to Max Speed</td>
<td>0.073s</td>
<td>0.08s</td>
<td>0.114s</td>
</tr>
<tr>
<td>Angle Resolution</td>
<td>0.00036°</td>
<td>0.00036°</td>
<td>0.00018°</td>
</tr>
<tr>
<td>Repeatability</td>
<td>0.0075°</td>
<td>0.0075°</td>
<td>0.0025°</td>
</tr>
<tr>
<td>Min Velocity</td>
<td>0.027° s⁻¹</td>
<td>0.027° s⁻¹</td>
<td>0.014° s⁻¹</td>
</tr>
</tbody>
</table>

Table 2: Performance figures for Yorick.

It is important to note that the delay will vary within each motion process depending on the data in the image — a corner tracking process will have a longer delay as more corners within the image are tracked. Therefore the delays must be computed for each and every set of results that are received by the gaze controller.

Requests to the controller from vision for mount status information are serviced at high priority meaning that the time required for the information to propagate back to the heads of each vision pipeline is negligible (our timing tests suggest the order of 50μs). We remark that this method of time-stamping images is significantly more robust than that used in the Marvin system [22, 23], wherein odometry is encoded as an analogue signal at the start of the image data.

To date, we have implemented six different behavioural reflexes on the system: coarse motion detection [6], corner tracking [24], looming (alarm detection), opto-kinetic response and smooth pursuit [7], and a bright spot tracker [15]. The last, although of limited visual interest, provides a convenient way to test the system, to debug the communications between the different sections and to demonstrate the performance of Yorick.

### 7. PERFORMANCE

The vergence, elevation and pan results in Table 2 were obtained from experiments carried out on the unloaded mount, using the square root local controller. The design specification given in Table 1 is seen to be matched by the actual performance in all but two requirements— the vergence and elevation maximum velocities fall short by 33% and 11% respectively. This is solely because the current limit on the amplifiers used for the elevation and vergence axes falls short of the peak current capacity of the motors used. Open loop tests have yielded maximum accelerations for the vergence axis of the order of 8000° s⁻² and top speeds of 650° s⁻¹, both figures tying in well with the theoretical limits.

When the figures given in Table 2 are translated into pixels and frames rates the performance of the head becomes readily apparent. Using cameras with a field of view of 40°, and a frame rate of 25Hz, the vergence axis can, for example,

(i) accelerate from zero to maximum speed (205 pixels/frame) within two frames
(ii) decelerate from maximum sped to zero in one frame;
(iii) step 38 pixels (~ 3°) within each frame;
(iv) has a resolution in excess of 1/250th of a pixel;
(v) has a repeatability in excess of 1/20th of a pixel; and
(vi) can measure velocity down to 0.014 pixels per frame.

We recall that the system was designed to be able to carry cameras with both wide angle and high zoom lenses. With a wide angle of view the high acceleration capability becomes more important. Conversely, at the highest specified zoom, where the field of view may be reduced to 5°, the system can maintain sub-pixel accuracy and sub-pixel repeatability.

7.1. Effects of varying loads.

As required by the design specification, the performance of the system should be robust under varying loads. Figure 10 shows the step response of the vergence and elevation axes with varying loads weighing up to 2.5kg, giving a maximum change in load inertia of 0.0077kgm\(^2\) for the vergence axis and 0.025kgm\(^2\) for the elevation axis. For the common elevation implementation, one of the optical axes is coincident with the vergence axis and therefore the change in inertia is small, giving a negligible change in the step response (the inertia of the motor, gearbox and encoder hub (at the output) is 0.075kgm\(^2\)). On the other hand, the camera platform is offset from the elevation axis by 50mm so the effect of a larger change in inertia can be readily seen. The response is seen to improve slightly as the load inertia increases — this is due to the increased inertia balancing the inertia of the transmission and structure of the vergence link. Note that for the step responses shown, the tuning is chosen to give a slightly overdamped response to facilitate smooth switching from a saccade to tracking behaviours.

For the vergence axes, the highest inertia of 0.0077kgm\(^2\), when reflected back to the input, is \(3 \times 10^{-6}kgm^2\), which is about 60% that of the motor/gearbox inertia. However, the precision encoder has an inertia of \(30 \times 10^{-6}kgm^2\), which therefore swamps any disturbance at the link.

A Bode plot of acceleration vs current supplied to the motor for the open loop vergence axis amplifier/motor pair is given for varying loads in Fig. 11, where the acceleration was measured using an accelerometer attached to the link and the current measured at the motor using a
current probe. The resonant frequency at 83Hz for the unloaded case is seen to reduce to 23Hz for the maximum load tested. This resonance is due to the compliance in the gearbox. The resonant frequency is still beyond the bandwidth of the vision processes (Nyquist frequency is 12.5Hz), though care must be taken to ensure that the controller does not excite these modes. The phase lead exhibited at low frequencies is due to the stiction of the gearbox [11].

7.2. Visual Response.

To demonstrate closing the loop around the vision processes we present some results from a simple corner tracking algorithm: a corner’s position and velocity, determined using a modified version of the algorithm suggested by [25], is used as the mount demand. When the corner being tracked disappears it is replaced by a nearby one with similar velocity.

Figure 12 shows the response of the vergence and elevation axes to demands from the basic corner tracking algorithm, where we adopt a simple control strategy in which position and velocity demands can potentially compete. Position demands are filtered by a low-pass (1Hz) first-order Butterworth filter, while velocity demands are filtered at the Nyquist frequency for vision (ie. 12.5Hz), meaning that the system is much more responsive to changes in velocity than in position. In addition the velocity is used to provide the system with an elementary predictive mechanism.

It was noted, by matching image sequences to the tracking results, that image directions appeared to move in the opposite direction to the corner velocities. This is the effect of system delays caused by the inherent processing delay and compounded by the necessary filtering. The desired fixation point was also seen to switch from corner to corner between frames — this can be clearly seen in 12. To begin with, the position demands are stable as a single corner is consistently tracked, resulting in smooth, but delayed control. However at approximately time 1.8 seconds, the position demands begin to fluctuate considerably as the algorithm switches corners. The response remains stable only through use of the filtering described above. Details of this, and a more sophisticated corner tracker which eliminates this problem are given in [24].
Figure 12: Vergence and elevation response from the corner tracker using interpolated vision error signals.

6. CONCLUSIONS

We have presented the process behind the design and development of a high performance stereo platform for use in a reactive vision sensor surveillance. The resulting implementation and testing of the design, as Yorick, has demonstrated the success of the mechatronic approach adopted — ie the design process involved the integration of mechanical and control engineering, systems architecture and machine vision research, and resulted in a real time gaze control system that is not limited by the poor performance of the camera platform, or the communications between that platform and the gaze controller and vision sensors.

The physical mount has been designed with the aim of versatility without sacrifice to performance or weight. The basic hardware has been tested with a variety of controllers, implemented within the transputer network, enabling the close relationship between vision and control to be fully exploited. The mount and controller statistics presented herein demonstrate that high performance can be maintained even under considerable load, so that the basic system may be used with different camera/lens configurations.

A direct result of this design process is the speed with which new algorithms can be implemented: in the six months since the real time gaze control system has been installed, we have implemented six different behaviours, all working concurrently at video rate.

Acknowledgements The design and implementation of Yorick has been funded by SERC Project GR/G30003, while the research into visual processes and the servo and gaze controllers was funded by ESPRIT Project 5390: Real Time Control of Gaze. We are grateful for contributions from other members of the Robotics Research Group, University of Oxford, and from discussions with our European collaborators, especially Donald Weir and Nigel Gent at the GEC-Marconi Research Centre.
References

APPENDIX A: The Square Root Controller

The servo controller implemented is based on a Square Root Controller (SRC) with added integral control. The motor torques in a SRC are generated from the following control law:

$$\tau = K_p \sqrt{\theta_{err}} + K_v \dot{\theta}_{err}$$

(1)

$$\theta_{err} = \theta_{dem} - \theta, \quad \dot{\theta}_{err} = \dot{\theta}_{dem} - \dot{\theta}$$

(2)

where $\theta$ is the joint angle, $\theta_{dem}$ is the demanded joint angle, $\tau$ is the torque applied to the motor, $K_p$ and $K_v$ are the proportional and derivative gains respectively. The response to this type of controller is very good in the steady state with errors being typically reduced to the resolution of the encoder. It is not, however, without its drawbacks. A major problem occurs as the positional error tends to zero, when limit cycles are seen to arise. This is due to the increasing “effective gain” which tends to infinity in the limit as the errors tend toward zero, i.e. the transfer function of the controller (neglecting velocity) becomes

$$H = \frac{\tau}{\theta_{error}} = \frac{K_p}{\sqrt{\theta_{error}}} \rightarrow \infty, \quad \theta \rightarrow \theta_{dem}$$

(3)

In practice the effective gain is limited by the resolution of encoder. Nevertheless, the controller will tend to “chatter” as the error tends to zero.

A secondary problem arises from the dynamics induced into the damping ratio. The effective positional gain increases as the error reduces. Thus, if the controller is tuned to achieve critical damping, at $\theta_{err} = 1$ say, then the controller will be effectively underdamped when $\theta_{err} \leq 1$ and overdamped when $\theta_{err} \geq 1$. For a reactive visual sensor this seems to be precisely the opposite type of response to have. Ideally the sensor should be underdamped at large errors, allowing a quick response to a saccade demand, and slightly overdamped at small errors, allowing the sensor to begin to track smoothly objects of interest towards the end of a saccadic movement.

In order to retain the advantages of the SRC, we have introduced some modifications into the control law:

$$\tau = K_p \phi + K_v \varphi + \tau_{int}$$

(4)

$$\tau_{int}(t) = \tau_{int}(t-1) + K_i (\phi(t) + \phi(t-1))$$

(5)

$$\phi = \sqrt{\theta_{err}}, \quad \varphi = \sqrt{\theta_{err}}$$

(6)

where $\phi$ and $\varphi$ are linearised in the neighbourhood of $\theta_{error} \simeq 0$ and $\dot{\theta}_{error} \simeq 0$ respectively; $\tau_{int}$ is the integral torque and $K_i$ the integral gain. This linearisation of the errors in the neighbourhood of zero has the effect of eliminating the limit cycling at low errors but also reduces the repeatability capability (in the absence of integral action). By applying the square root of the velocity error we can minimise the variation in the damping ratio. The response of this controller (without integral action) is shown in Fig 9(a). It is seen that, in order to track a moving object, some integral action is required.

The addition of integral action, as expected, enables the controller to integrate out the error in dynamic as well as static trajectories (see Fig 9(b)). However the accumulation of the integral torque $\tau_{int}$, even with anti-integral windup as in Fig 9(b), degrades the overall performance of the controller.
Fig 9(c) shows the system response to the same demanded trajectory when a selective integral square root controller was used. The modified integral torque is given by the equation

$$\tau_{int}(t) = Q\{\tau_{int}(t - 1) + P(\phi(t) + \phi(t - 1))\} \quad (7)$$

where $P$ and $Q$ are nonlinear weighting functions which are dependent on the states of the system. The weight $P$ is zero outside a specified positional error and increases smoothly and continuously to a maximum at zero positional error (anti-windup). The function $Q$ is piecewise continuous and, in effect, exponentially reduces the integral torque to zero when either the demanded acceleration or demanded velocity are outside specified minima/maxima. Even using the same gains the response (Fig 9(c)) shows a marked improvement as the trajectory changes dynamically.