Reactions to Peripheral Image Motion using a Head/Eye Platform*

David W Murray, Philip F McLauchlan, Ian D Reid and Paul M Sharkey
Department of Engineering Science, University of Oxford,
Parks Road, Oxford OX1 3PJ, UK

Abstract

In this paper we demonstrate four real-time reactive responses to movement in everyday scenes using an active head/eye platform.

We first describe the design and realization of a high bandwidth four degree-of-freedom head/eye platform and visual feedback loop for the exploration of motion processing within active vision. The vision system divides processing into two scales and two broad functions. At a coarse, quasi-peripheral scale, detection and segmentation of new motion occurs across the whole image, and at fine scale tracking of already detected motion takes place within a foveal region. We detail several simple coarse scale motion sensors which run concurrently at 25Hz with latencies around 100ms. We demonstrate the use of these to drive the following real-time responses: (i) head/eye saccades to moving regions of interest; (ii) a panic response to looming motion; (iii) an opto-kinetic response to continuous motion across the image and (iv) smooth pursuit of a moving target using motion alone.

1 Introduction

The embedding of visual feedback in sensing-perception-action loops that enable the control of what in the scene is looked at and how it is looked at, promises to address one of the principal difficulties of the data-driven 3D reconstructionist paradigm, namely the need to build and maintain an omniscient dynamic representation of the surrounding environment.

Of the benefits of such active vision, most explored are those arising from (i) making known movements and (ii) fixating. It has long been known that the recovery of structure from known motion is inherently simpler and better conditioned than that when the camera motion has to be recovered as well. Moreover, it appears that a range of shape-from-X recovery tasks that are ill-posed when camera motion is unknown become well-posed when it is known [1]. Turning to fixation, one benefit is that exocentric coordinate frames can be established at points in the scene [2], allowing the exploration of local structure [10]. A second lies in the elimination of motion blur, of which there are several demonstrations using active heads (eg [6, 19, 20]).

In all of these studies however, once making a known motion, or once fixating, the camera is set to continue on its fixed trajectory or to carry on fixating, and the need for gaze control is reduced considerably. A new set of problems is raised by asking how the system starts attending to, or fixating upon, an object and how, some time later, it moves onto some new object or area of interest. One aspect of this problem of where to look next has been explored by Rimey and Brown [24, 25] who move the gaze direction successively to areas of the scene which provide some maximally discriminatory information for the task in hand. Rimey and Brown’s work is performed in a static environment, and the task is high level. It is obvious though that at a lower level, many decisions about where to look next can be (and in biological systems are) driven by motion in the viewed scene, with qualitative and quantitative analysis of the projected motion in the image providing cues for the segmentation of moving regions, for allocating visual attention to them, and for pursuing or tracking them.

Our contribution here is to show how straightforward but real-time high-bandwidth visual motion processing, when coupled in the feedback loop to a controller and fast mechanical head/eye platform, can elicit such motion responses — responses, or “gaze tactics”, which can then be built up into a gaze control strategy. The responses we demonstrate here are all driven from coarse scale, quasi-peripheral visual motion and include

- the initiation of motion saccades;
- the firing of panic reactions to threatening looming movement in the scene;
- an analogy to the primate opto-kinetic response; and
- the smooth pursuit of an object using motion alone.

A sister paper [22] describes the use of foveal vision for tracking.

A key motivation for our approach is the belief that rapid responses to motion events at the level of the 2D image provide an essential clue for the more deliberate head/eye movements for 3D information recovery that characterize a good deal of the work undertaken in active vision.

2 Some mechatronics

The essential features of a gaze control system are the visual feedback loop, the controller and a head/eye platform as controlled plant. Without doubt, the severest challenge in behavioural gaze control lies within the gaze controller itself. How should appropriate demands be generated to achieve the current visual task, what visual feedback should be selected, and how can cooperation be obtained between the several sensing-action loops? To begin to explore these issues, we have established several independent high-bandwidth sensing-action loops based on motion understanding, and constructed a high performance head/eye platform.

---

*This work is supported by grants from the UK Science and Engineering Research Council (Grant GR/G30003), and from the EC Esprit Programme (Project 5390).
2.1 The visual feedback loop

Applied to everyday scenes, an active vision system will encounter a wide range of motion magnitudes and characteristics. When tracking successfully, successive images will have displacements near zero, whereas unexpected motion may give rise to much larger displacements. There is of course no absolute upper limit, though for surveillance and navigation applications, angular velocities of \(\sim 3^\circ\) (say 30 pixels) per frame are typical. Again, when tracking, interest is focused at the image centre, whereas distracting unexpected motion is more likely to occur at the image periphery. A single process is unlikely to be able to deal with such a range of motions and motion traits. Instead we use multiple simple processes, some coarse, fast and robust, others more refined and stately, each utilizing different representations of image motion. Our architecture supports two sets of concurrent processes at distinct scales, one set at coarse, quasi-peripheral scale over the entire but sub-sampled image and the other set at fine, quasi-foveal scale in a central sub-image. As well as simplifying the specification of each motion knowledge source, this division limits and balances the data throughput between the sets.

Dealing with change and the unexpected requires rapid response. As well as high process sample rates, the vision processes should have minimal latency, principally because the stable control of systems with delayed feedback requires low gain, giving sluggish response. It is of course possible, indeed necessary [4, 5], to use prediction on the vision output to compensate for delay, but the larger the delay, the larger the uncertainty in the prediction, again requiring conservative gain settings to ensure stability. To achieve high rate and low latency, we have adopted a balance of pipelined and spatial MIMD parallelism, associating short wide-diameter pipelines with each motion process. In the sketch of the overall architecture (Figure 1) we show the division of the visual feedback loop into foveal and peripheral sections, and within each show a couple of pipes. Apart from image capture and initial smoothing, all the vision processing discussed in this paper runs concurrently on nine 8Mip 32-bit T805 Transputers. These devices are equipped with 4 bi-directional inter-processor links which are fully integrated into the model of concurrency, facilitating the construction of communication protocols between the several vision processes and, importantly, between vision and control, the latter also being implemented on Transputers.

2.2 The controller

As is apparent from Figure 1, the controller in our system is divided into two parts: the high level gaze controller which knows something about vision and behaviour, and the low level servo-controller which knows about head kinematics, joint angles, motors and encoders. Part of the high level gaze controller operates asynchronously at a rate determined by the vision processing (typically 25Hz) and selects and predicts visual output to drive gaze constructs such as pursuit, saccade, and so on. In this paper the selection stage is manual, but more recently [21] we have considered the combination of feedback loops and behaviours. The other part of the high level controller, from the interpolation stage onwards, runs synchronously at 500Hz outputting a gaze direction and velocity in head coordinates to the servo-controller. The servo in turn performs all synchronous controls, as well as the forward and inverse kinematics, trajectory limiting, receiving feedback from encoders on the motor shafts. The servo-controller also has an important role as system clock. The need to combine prompt head data with delayed vision results for prediction makes timing an important issue, the more so as motion sensors may have different rates and will almost certainly have different latencies.

As part of its 500Hz control loop, the servo maintains a ring buffer of mount status data, such as position and velocity and control mode (saccade, smooth pursuit, etc) at the time of image capture [23, 26], data which can be requested by the vision processes and prediction stage.

![Figure 1: The overall architecture of the visuo-control loop developed for our work. The vision system provides parallel feedback loops, grouped into peripheral and foveal channels. The different delays in the different processes require timed head encoder data to be stored in a ring buffer, so that they can be used in prediction.](image)

The primate visual system is driven not only from visual feedback, but also from proprioceptive information, as evidenced by our own ability to perform controlled eye movements with our eyes shut. In our system, the servo-controller effects this by obtaining feedback from encoders on the motor shafts. Via the forward kinematics, these measure an absolute gaze direction, as opposed to a gaze direction relative to the visual scene. This allows the head to move as a pointing device without visual feedback, at much higher gains and speeds. This is essential during saccadic fast motions, where images are severely blurred and vision is effectively useless for feedback. The servo runs a square-root controller (SRC) [13] with added integral control, and is described in [27].

2.3 The head/eye platform: Yorick

We have argued elsewhere [17] that to redirect gaze quickly and accurately there is a need for specialized apparatus. It is perhaps too early in the evolution of active vision for a consensus design to emerge, but our work on reactions to motion has made clear to us that as well as high speed, mechanical stiffness, precision, and simplicity, the head platform requires high acceleration.

The head mechanism, Yorick\(^1\), has five powered axes, each with

---

1. After the famous skull in Hamlet: *This same skull, sir, was Yorick’s*
the same modular design, configured as a common elevation platform with the two elevation axes mechanically linked. (A two axis, monocular version is also in use.) One design aspect which has proved itself is the use of DC motors with negligible backlash gearboxes (from Harmonic Drive) in the drive trains. Geared drives maintain high acceleration, good tracking ability at low velocities, even under large changes of load (eg, a change of cameras).

![Diagram of drive axis](image)

**Figure 2:** The modular drive axis (a); five such axes configured as a common-elevation platform with four degrees of freedom (b); and the head/eye platform “Yorick” as built (c).

In Table 1 we show axis performance specifications which were obtained from experiments on the platform.

<table>
<thead>
<tr>
<th>Description</th>
<th>Vergence</th>
<th>Elevation</th>
<th>Pan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axis Range</td>
<td>360°</td>
<td>360°</td>
<td>360°</td>
</tr>
<tr>
<td>90° Slew Time</td>
<td>0.28 s</td>
<td>0.29 s</td>
<td>0.76 s</td>
</tr>
<tr>
<td>360° Slew Time</td>
<td>0.95 s</td>
<td>0.97 s</td>
<td>1.69 s</td>
</tr>
<tr>
<td>Max Slew Rate</td>
<td>400°/s</td>
<td>400°/s</td>
<td>300°/s</td>
</tr>
<tr>
<td>Max Acceleration</td>
<td>6000°/s²</td>
<td>5000°/s²</td>
<td>500°/s²</td>
</tr>
<tr>
<td>Max Deceleration</td>
<td>10000°/s²</td>
<td>9000°/s²</td>
<td>800°/s²</td>
</tr>
<tr>
<td>Backlash</td>
<td>0.0075°</td>
<td>0.0075°</td>
<td>0.0025°</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</td>
<td>0.027°/s</td>
<td>0.014°/s</td>
</tr>
</tbody>
</table>

Table 1: Measured performance of the geared drive trains for the four axes: vergence ($\times 2$), elevation and pan (or neck).

---

**skull . . . a fellow of infinite jest, of most excellent fancy.**

---

### 3 Peripheral motion processes

To obtain the required performance from vision on finite hardware requires compromise, and one must expect outputs to contain not only statistical but also gross error. What is important is that each process acts as a motion knowledge source, able to advise when image conditions are appropriate or inappropriate for its operation.

The division between the quasi-peripheral and -foveal processes is not merely one of scale but also of functionality. Peripheral processes are there to alert to motion which may be of interest or be threatening — motion which might then be attended to by redirecting gaze. They are not required to be highly accurate, though they must give some degree of quantitative information to the controller.

#### 3.1 Detection and segmentation

Although gradient-based optical flow was found in earlier structure from motion studies to be too error-prone for detailed quantitative recovery of scene structure (eg [16]), recent work on qualitative motion understanding has revived interest in its use [29, 18, 8], and it is this approach we adopt. The initial data for all the peripheral processes are edge-normal components $\mathbf{v}$ of the optical flow field $\mathbf{r}$ derived from spatio-temporal gradients of the smoothed and sub-sampled image irradiance $E$ using the motion constraint equation [12, 14]

$$E_t + \mathbf{r} \cdot \nabla E = 0$$

whence

$$\mathbf{v} = -E_t \nabla E / |\nabla E|^2.$$

The deficiencies in this equation are well-explored [28], but by more heavily weighting motion with large $|\nabla E|$, we find that the motion derived is sufficiently good for the qualitative and quantitative interpretation we require. Indeed, observing the output of the real-time motion detector over extended periods of time, what is strikingly apparent is not the gross and statistical errors made in each frame, but the overall temporal coherence of the computed motion.

To start to segment out objects moving independently of the background we subtract the image motion arising from known motion of the camera on the head platform, $\mathbf{r}_h$. The motion, and its component, arising from the scene alone is then

$$\mathbf{r}_s = \mathbf{r} - \mathbf{r}_h \text{ and } \mathbf{v}_s = \mathbf{v} - (\mathbf{r}_h \cdot \mathbf{v})\mathbf{v},$$

where $\mathbf{v}$ is a unit vector. If the rectilinear and angular velocities of the camera with respect to the static background are $\mathbf{V}$ and $\Omega$, then the image motion due to head motion is

$$\mathbf{r}_h = \frac{1}{Z} (\mathbf{V} - \mathbf{r} (\mathbf{V} \cdot \hat{z})) - \Omega \mathbf{r} - r (\Omega \mathbf{r} \cdot \hat{z})$$

where the image position $\mathbf{r}$ and the depth of the scene $Z$ are scaled by the camera’s focal length. If the body carrying the head has motion $(\mathbf{V}_{body}, \Omega_{body})$ then $(\mathbf{V}, \Omega)$ are found from the forward kinematics as nonlinear functions of body motion, joint angles $\Theta$ and velocities $\dot{\Theta}$

$$\mathbf{V} = f(\Theta, \dot{\Theta}, \Omega_{body}) + \mathbf{V}_{body}$$

$$\Omega = g(\Theta, \dot{\Theta}) + \Omega_{body}.$$
In our work there is no body movement, and axis pivot length is so small compared with the typical depth of the scene that the rotational components of $\mathbf{r}_n$ dominate over the translational and so we neglect the translational terms. Indeed, we conclude that for many applications there is little need to have a mechanism to ensure that the optic centre of the camera is always at the rotation centre.

Image regions which correspond to background will, within a noise tolerance, have $\mathbf{v}_s \sim \mathbf{0}$, and are excluded from further consideration.

Foreground motion regions are grown by spatially grouping non-background vectors $\mathbf{v}_s$ [15], and their averaged image velocity $\langle \mathbf{r}_s \rangle$ calculated by least squares fitting to all the normal flow data in a region [7]. The motion constraint equation is rewritten as

$$\begin{bmatrix} \mathbf{H} \end{bmatrix} (\langle \mathbf{r}_s \rangle) = \mathbf{m},$$

where $\mathbf{m}$ is a vector of measurements

$$m_i = E_{hi}$$

and $[\mathbf{H}]$ has an $i$-th row of $H_i = -\nabla E_i^T$. We find a weighted classical least squares solution from

$$[\mathbf{H}^T][\mathbf{A}][\mathbf{H}](\langle \mathbf{r}_s \rangle) = [\mathbf{H}^T][\mathbf{A}]{\mathbf{m}}$$

using LU decomposition of $[\mathbf{H}^T][\mathbf{A}][\mathbf{H}]$. The diagonal weight matrix $[\mathbf{A}]$ has $A_{ii} = |\nabla E|^p / \Gamma^p$, where the power $p$ is raised to increase the weighting given to strong irradiance gradients. The role of $\Gamma$ is to allow a residual to be calculated and compared to the chi-squared distribution; moving regions with too high a residual are labelled and the velocity information for that region is ignored.

The detection and segmentation algorithms are implemented on our head/eye platform at the frame rate of 25 Hz with a latency of $\sim 75$ms using a network of five transputers [15].

Figure 3 shows 8 frames (spacing 80ms — ie, every other frame from the 25Hz results) of two people walking past each other on the street. The outline of each detected moving region is shown along with the velocity vector and the error ellipse, both magnified six-fold for clarity. At the start of the displayed sequence the people are just passing each other and are detected as one moving region, the nearer person moving to the right dominating the result (frames 1,3). As they separate the estimated error increases greatly as there are equal amounts of image data travelling in opposite directions (frames 5,7). At that point the algorithm separates the two people and the velocities are estimated with greater precision. The error ellipse is obtained directly from the residual of the least squares fit and its eccentricity, obtained from the covariance matrix, indicates the degree to which the aperture problem has been overcome, a large eccentricity indicating little constraint in the direction of the major axis.

### 3.2 Detecting looming motion

The human visual system reacts powerfully to large looming motions indicating imminent collision. We have implemented a looming detector in Yorick’s feedback loop that triggers an alarm when a large object is close to and on collision course with the camera. The patches derived in the segmentation are analyzed to determine whether any is large and exhibits substantial divergence, using an affine flow field approximation. If such a patch is found, its focus of expansion and the time-to-contact of the approaching object are determined and then used to trigger a panic response. As it uses intermediate results from the coarse motion algorithm, the detector is able to run at 25Hz using only one extra T805 transputer.

The affine approximation to the flow field due to the scene is [29, 7, 8]:

$$\mathbf{\dot{r}}_s = \mathbf{\dot{r}}_0 + [\mathbf{M}] \mathbf{r},$$

where

$$[\mathbf{M}] = \begin{pmatrix} \frac{\partial h}{\partial x} & \frac{\partial h}{\partial y} \\ \frac{\partial \phi}{\partial x} & \frac{\partial \phi}{\partial y} \end{pmatrix},$$

which can be rewritten in terms of the first-order differential invariants of the motion field, $\text{div}(\mathbf{r})$, $\text{def}(\mathbf{r})$ and $\text{curl}(\mathbf{r})$. Again
taking errors as normally distributed and unbiased we can rewrite
the motion constraint equation as

$$
\mathbf{m} = [\mathbf{H}] \mathbf{s}
$$

where we wish to solve for

$$
\mathbf{s} = (\dot{x}_0 \ y_0 \ M_{11} \ M_{12} \ M_{22})^T.
$$

The vector \( \mathbf{m} \) is still as in Equation 1, but the \( i \)-th row of \([\mathbf{H}]\) here is

$$
H_i = -(E_{xi} \ E_{yi} \ x_i E_{xi} \ y_i E_{yi} \ x_i E_{yi} \ y_i E_{yi}).
$$

Again we solve using least squares, adopting the same weighting
scheme as outlined earlier.

Using the resulting \( \mathbf{r}_0 \) and \([\mathbf{M}]\), the focus of expansion is determined from

$$
\mathbf{r}_{\text{FOE}} = -[\mathbf{M}]^{-1} \mathbf{r}_0
$$

and, in the absence of deformation of the field, the time to contact is found from

$$
t_c = \frac{2}{\text{div}(\mathbf{r})} = \frac{2}{M_{11} + M_{22}}.
$$

When there is flow field deformation, upper and lower bounds on
the time to contact are [11]:

$$
t_{c\pm} = \frac{2}{\text{div}(\mathbf{r}) \pm \text{def}(\mathbf{r})}
$$

and use of the lower bound provides a cautious trigger for evasive
action.

Figure 4 shows frames from the response of the looming detector
as the left camera is attacked with a book. Although the process
is running continuously, only when the divergence is substantial
does the detector display output. The white lines are regular samples
from the fitted affine flow field and the circle marks the position
of the focus of expansion. Also shown are graphs of the variation
of the upper bound on, lower bound on, and zero deformation
approximation to, the time to contact. Typically we find
that using the lower bound is not overly pessimistic, and does not
give rise to premature panic. In this run, the platform motion was
disabled, and one sees the time to contact continue downwards.
As two frames are required to compute motion, a time to contact
of less than 80ms will never be logged. We show an example of
this process driving evasive actions by the head in Section 4.

### 3.3 Whole image retinal slip

This process advises when the image motion across the entire
periphery can be well fitted by a single image velocity \( \mathbf{v} \),
using the least squares technique given in Section 3.2. This velocity
is used to drive an artificial opto-kinetic reflex, described in the next
section.

#### 4 Behaviours from peripheral motion

The peripheral motion sensors described above have been used to
generate a variety of responses on Yorick. We stress that as the
sensors run concurrently, each response is potentially available at
all times, though to date we have explored only their individual re-
responses, manually selecting feedback from the appropriate motion
detector.

### 4.1 Saccade initiation

The first behaviour we demonstrate is the initiation of saccades
from the coarse scale motion detection and segmentation process.
When independent foreground motion is detected, the head ro-
ates as quickly as possible to match the position and velocity of
the region. During this rapid movement, feedback from the joint
encoders is used to control the head. The head then continues to
move with that angular velocity until it reaches its end stops. Note
that no tracking is initiated, although obviously it is during the pe-
riod where the velocity is matched that tracking should be initiated
using, for example, the corner cluster tracking method described
in these proceedings [22]. This transition from saccade to pursuit
has been demonstrated in recent work in our laboratory [21].

Figure 5 shows eight stills cut from a video recording taken from
behind the head platform during the response as a person passes
in front of the head. From frames 1-4 (2-4 not shown) the mo-
tion magnitude and position of the segmented region are derived
and by frame 5 (shown) the head platform is beginning to move
to place the region of motion at the image centre and to match
its velocity. From frames 9 onwards the target is centered, and a
tracking phase could be initiated.
Figure 5: Initiation of a saccade from motion input, viewed from just behind the head platform (note camera at bottom center of each picture). The interframe spacing of these stills cut from video is 160ms and there are 3 more result frames between each shown.

4.2 Response to looming motion

The response to looming motion we have explored is one which drives the vergence and elevation axes as fast as possible to defensive positions on detection of a time to contact of less than 0.3s and a focus of expansion lying within or close to the physical image.

Figure 6(a) shows stills taken from videos of the head’s response and (b) shows a set of corresponding views through the camera during it. At frame 5 the head is just starting to move, raising the elevation axis and turning the cameras inwards, and the image (5b) becomes blurred by this motion. At frame 6, some 300ms later, the movement is complete, the elevation axis raised through 90° and the cameras rotated by 90°, giving a view (6b) of wiring along the back of the head.

A possible refinement to this response would be to exploit the fact that the affine approximation used in the looming detector is precise for a translating front-parallel plane (with possible camera rotation about its optic axis). If the camera is viewing a typical spheroidal projectile, the approximation is more closely satisfied by turning to look directly at the approaching object. Such a strategy also reduces the deformation, so improving the estimate of the time to contact, and makes the computation of the FOE, on which one decides whether there is to be a collision, better conditioned. This response would take longer to perform, and requires the detection of looming from smaller image patches. We have not yet explored whether the affine approximation will give sufficiently accurate determination of divergence in these smaller patches to make this approach feasible.

4.3 The opto-kinetic reflex

The opto-kinetic reflex is evoked in the primate visual system when large parts of the image move coherently. In a typical experiment, the subject is placed at the axis of a drum with a textured interior which is turned at a constant rate to produce a constant velocity field on the retina. The response is one of slow-phase eye movement to follow the motion, followed by a quick-phase fly-back when the eyes approach their extremes of rotation. The response is driven primarily by nulling the optical flow, rather than positional information.

In our experiments we placed the head in front of a continuous belt of scenery, as shown in Figure 7. The belt was moved in the direction of the arrow in frame 1 at a (roughly) constant speed. A
least squares fit was made for a constant flow field across the entire peripheral image (although because the scene is not at constant depth this model will never be satisfied) and the camera moved so as to null the motion using the square root controller. Frames 1-5 show the first slow sweep of the camera, followed between frames 5 and 6 by a rapid flyback, and the process repeats in frames 7–10.

In primates, flyback is in fact triggered at some point earlier than physical limits of eye rotation, for reasons which appear poorly understood. In our experiments we have simply placed an angular limit on the camera rotation, and flyback occurs to the opposite limit.

4.4 Smooth pursuit

The simplest (and over-simplified) model of the smooth pursuit response in the primate visual system regards it as a foveal version of the OKR. The eye is driven to null retinal slip on the foveal window rather than over the whole image [3, 9].

Figure 8 shows stills from a video showing the head platform moving to stabilize the image as a newspaper is moved around in front of the camera. As in the case of the OKR, stabilization is achieved without using positional data. This contrasts with the approach in the computer vision literature where tracking is almost without exception driven from positional data, albeit often filtered in some way.

5 Discussion

In this paper we have described the design and realization of a head/eye platform and visual feedback loop for the exploration of motion processing within active vision. We have gone some way to incorporating peripheral and foveal models into the vision system, although to date the division is a straightforward one between two scales and between two broad functions, detection of new motion and tracking of already detected motion.

We have given details of several real-time motion sensors. In
implementing vision algorithms our aim has been to create simple self-contained motion sensors which run concurrently and quickly, and which have well defined limits on expected functionality.

Finally we have demonstrated the real-time reactions or responses that they give rise to on the head platform. In particular we have discussed the responses driven off “peripheral”, coarse scale, motion: the initiation of motion saccades, a panic response to looming, the opto-kinetic response and smooth pursuit.

Experimentation with the motion sensors leads us to make the following remarks. First, simple processes such as those presented here will make gross “mistakes” (errors of fact rather than of precision) at an instant, but provide data which has remarkable underlying temporal consistency. This implies that it is more profitable to put effort into improving results over time by appropriate temporal integration, than into making ever more perfect snapshot processes. We have already applied some temporal integration to the output of the peripheral segmentation stage using a Kalman filter with both image position and velocity measurements as input.

Indeed, the velocity of a patch is usually known more reliably than its position because the moving object is usually extended. Figure 9 shows two frames from the start and end of a sequence where a chair is wheeled across the floor, and the tracking output of a constant acceleration Kalman filter run over the twenty intermediate frames. Such integration is of benefit in initiating saccades, where a “look before you leap” strategy, waiting a few frames to accumulate better velocity and position estimates for the target before moving, yields better hit rates.

![Figure 9](image1.png)

Figure 9: The start and end of a sequence from the detection and segmentation stages and the output from a constant velocity Kalman filter on the intermediate results. The filter uses both position and velocity data as input.

Our second remark is that, in the context of a real-time gaze control system, it is not the “perfection” of the flow vectors that matters, but the number of sensible responses that each module gives rise to. A question for the future is how to assess the pertinence of a response.

A sister paper describes some recent work in fine scale foveal motion detection and tracking [22], and our future work will explore the combination of behaviours driven from both peripheral and foveal motion knowledge sources into an increasingly sophisticated gaze control strategy. Initially we have experimented with more definite behavioural sequences, such as linking motion saccades to smooth pursuit [21]. One might regard these as skills comparable with those discussed by Rimay and Brown [24]. There are many interesting questions here alone. For instance, how should one manage the deceleration phase between saccade and pursuit, at what point should visual feedback take over from proprioceptive feedback, and so on. Our present apparatus affords the opportunity for detailed experimentation of each.

References


