10 Attentive Visual Servoing

James J. Clark and Nicola J. Ferrier

A panel of computer vision researchers recently convened and produced a document concerning the relatively new field of "Active Vision" [36]. Although there is no precise definition of what active vision is, most researchers would agree that active vision is concerned with controlling camera parameters, such as position, focal length, aperture width, and so on, in ways that serve to make vision processing more robust and more closely tied to the activities that a robotic system may be engaged in.

In this chapter of the book we will concern ourselves with two aspects of the active vision paradigm. First, we will motivate the desirability of directed, or attentive, trajectory control of camera position. Second, we will describe details of how such control can be implemented in an actual robotic system. In particular, we detail the implementation of a visual servo control system which implements attentive control of a binocular vision system through specification of gains in parallel feedback loops. The servo type of control that we propose is based on models of mammalian oculomotor control systems [10, 23, 31].

10.1 Directed and Attentive Vision

Humans and other advanced animals process visual data in a dynamic fashion, where instead of applying image analysis operations to a single "snapshot" of the environment, they operate in an purposive and integrative manner on a temporal and spatially disparate sequence of images.

One can view the dynamic nature of animate vision in two ways: on one hand, processing of image sequences is a neccessity because animals move about in a changing environment and the vision system must deal with these changes; on the other hand one can argue that sensor motion actually helps in extracting information from visual sense data. Both of these viewpoints have been put forth as motivation for active vision research. The view taken in this paper is the latter (this is not to say that the former viewpoint is without merit). In accepting this viewpoint we need to answer the important question: What does the ability to move image sensors give us in terms of solving vision tasks? There have been produced a number of replies to this question, some obvious, some

subtle, and some which are summarized below.

Occasionally a viewpoint of a scene will be such that there are accidental alignments of objects or illuminants which make interpretation of the image difficult. A gross motion of the camera can then result in a viewpoint where these accidental alignments are no longer present. Thus camera motion can be used to increase the robustness of vision algorithms by improving the quality of data being operated on. An example of this is in viewing a highly specular surface. If the camera is aligned with the specular angle relative to the light source, then the intensity of the light incident on the camera may be very high. If the camera has poor dynamic range, and becomes saturated, then details such as surface albedo changes (surface markings) will be washed out. If the camera is moved a little away from the specular direction, the camera will no longer be saturated and the surface details will be more readily distinguished. Similar examples can be found in obtaining depth maps from viewpoints where very rapid changes in depth are present (i.e. those caused by highly slanted surfaces rather than true depth discontinuities caused by one object occluding another). In this case a small motion of the camera in the proper direction will result in a viewpoint where the change in depth is smaller. Thus directed motion of cameras is useful in obtaining *generic* views of objects [17].

Aloimonos *et al* [1] showed that camera motion can be used to make a vision problem that is ill-posed in the single image case into a well posed one. This is possible due to the availability of extra constraints from the additional images obtained in the active vision process. These added constraints may be enough to convert an underdetermined problem into an overdetermined one, and hence allow a robust solution to be obtained. Examples of the application of this type of active vision include shape from shading. Aloimonos and Basu[2] show how additional information can come from the optical flow induced in the image by the motion of the camera relative to the object surface. Images of points on the surface move in directions that are functions of the surface normal. Knowing how the images of points move (the optical flow), and the lighting, along with the observed intensity, the object shape (surface normal vectors) can be uniquely determined.

Motion of the camera can induce object dependent optical flows that can be used to determine object shapes and estimate depths to points in the scene (the *structure from controlled motion* process). Much work

has been done on structure from motion algorithms, even before the current interest in active vision. What the active approach brings to the structure from motion problem is the control over the motion, thereby both simplifying the problem [2] and allowing specification of optimal motions. An example of using directed motion to aid in a visual processing task was detailed by Ferrier [17]. In her work she determines the motion needed to make the active shape determination process of Aloimonos and Basu most robust to noise in the data..

Small, controlled motions of the camera can aid in solving the very difficult problem of feature correspondence in binocular stereo vision. Geiger and Yuille [19] describe a stereopsis algorithm which relies on small controlled eye movements to simplify the binocular feature correspondence problem. This is an example of a class of active vision algorithms in which eye or camera movements are used to provide constraints that simplify the computation of visual features [7]. A related process is that of *incremental stereo* [27] where the baseline between a pair of cameras is slowly increased from a small distance. At each incremental step the correspondence problem is easily solved. Controlled eye movements can also be used to help in a calibration process, wherein geometrical information regarding the imaging system is obtained [7].

Burt [9] describes active sensing (or "smart" sensing) as the selective, task oriented gathering of information. In this form of active vision one focusses the "attention" of the visual system on a portion of the scene that is important to the task at hand. As the demands of the robotic task evolve this focus of attention may shift. Such a form of active vision could be referred to as *attentive vision* to distinguish it from the active vision in which movements are made in order to provide additional constraints for solving a given vision problem. Bajcsy [3, 4] extends the concept of active perception to include the presence of feedback. In this extension, information obtained through the visual process, both high and low level information, is used to control the data acquisition process.

An obvious application of camera motion is in exploration, where one moves the camera so as to bring into view parts of the world that were occluded or hidden from view in the previous camera position. In this case, active vision serves to help the vision task by providing new information about the scene. In general the concept of visual exploration is more subtle than one may at first think. Consider that the real reason for needing camera motion for exploration purposes is due to the spatially

limited nature of the camera. The camera only samples a finite portion of the entire scene. In order to obtain information about more of the scene, the camera must move. The need for camera motion is not limited to the strict limits on the sensitive areas of the camera. For reasons of computational efficiency (as eloquently argued in Tsotsos' complexity level analysis [38]) one may wish to concentrate the computational resources one has at their disposal on a portion of the camera data. In the situation of a *foveal* camera, where the density of photorecpetors in the camera is nonuniform, with greatest concentration of sensors in a central region (or *fovea),* this shifting of computational resources implies moving of the camera, so as to *foveate* on different parts of the scene. By acquiring a part of the scene within the fovea, a detailed analysis can be made of this region. Visual tasks such as object recognition require the movement of the eyes to closely examine areas of interest for the particular task, while allotting only a small portion of the computational resources to the part of the scene which is viewed peripherally. Experiments have shown that directing attention to a location for one task increases visual capabilities for other tasks in that region [34, 35]. This region of increased visual attention has been likened to a spotlight or variable powered lens [14, 21]. Such capabilities are desirable in an active vision system. An efficient vision process would benefit from such a "spotlight" in which to devote most of its computational power, while processing the rest of the field at low resolution.

In most of the examples of active vision given above, the required motion of the camera is *directed.* In such cases, the vision system carrying out the vision task speficies a definite trajectory for the camera. We will refer to this form of active vision as *directed vision.* The other types of active vision, such as the standard structure from motion methods, which operate with arbitrary (although possibly constrained) camera motions, will be referred to here merely as generic active vision.

10.2 Camera Motions for Directed Vision

The camera motions in a system using directed vision must come from the vision subsystem. Depending on the type of robotic system the camera is attached to, this motion can be carried out in a number of ways. The most common situations are:

- Camera rigidly fixed to a mobile platform. In this situation, the vision system will have control over the motion, typically left-right-forward-backward in a plane, of the mobile platform.

- Camera rigidly attached to a robot arm, which can itself either be fixed to a base (such as a typical industrial robot), or to a mobile platform. In this case the vision system will have control over the motion of the robot arm, and perhaps control over the base if it is mobile. If the vision does not have any control over the motion of the base then there may be information directly transmitted to the arm controller from the base controller to permit any motion of the base to be compensated for by the arm, without needing any additional visual processing.

- Camera attached to a *head,* which is attached to a (mobile or fixed) base. This arrangement has less flexibility and more restricted range of motion than where the camera is attached to a six degree of freedom arm, but will typically be faster and more precise. In this case the vision system will have control over the relative position of the camera with respect to the *neck* of the head. This approach has the advantage that the motion of the arm and base are free for manipulation activities.

  The form of the control that we have over camera position is clearly application dependent and will vary with the particular directed vision process that is needed. For gross exploration, where the camera needs to peer behind objects or look into rooms, clearly the use of a mobile base is indicated. For situations where the exploration is only needed in a limited region, such as the workspace of a fixed industrial region, mounting a camera on a dextrous robot arm may be sufficient. This type of system could used in the active object mapping approach of Ferrier [17] as that method requires that the camera be moved about an object, but does not require motion over large distances. In cases where directed exploration is not required, a head type of camera motion may be all that is required. A head system may be used, for example in the optimal active shape from shading technique described in [16] or the incremental stereo technique of Geiger and Yuille [19]. In both of these examples the small precise motions available in a head camera system are all that are required.

  An important class of directed vision processes, and ones that are appropriate for head camera systems, are those involving what is commonly

termed *attentive* vision. By attentive vision we mean visual processes in which the computational resources available to the vision system are focussed on different parts of the scene. This focussing can operate either by moving the camera so that different parts of the scene become visible, or by changing which parts of the image are being processed at any given time. This mode of attention forms the basis for Ullman's visual routine paradigm [39], in which sequences of elementary image analysis operations are performed to obtain properties of, and relations between objects, in a scene. Focus of attention may also refer to the selection of a given set of image processing operations that are to be used to extract information from the scene. For example, a given visual task may require that corners of objects be detected, while another visual task may require that the color of objects be determined. In each of these two cases different features would be attended on.

As we are concerned here with active vision, where camera motions are paramount, we will only talk about the type of attentive vision in which shifts in the focus of attention are created through camera motions. Furthermore, since most of the attentive vision algorithms are those that are suitable for head camera systems we will for the rest of the paper concern ourselves only with head camera systems, although much of what we say will be applicable to arm and base mounted camera systems as well.

10.3 Saliency Based Feedback Control of Camera Motion

In all of the proposed directed and attentive active vision algorithms there are two common tasks to be performed. One is to figure out where to direct the camera gaze next (the *next look* problem [36]), and the other is to then carry out the motion that will let one look there.

In directed vision applications there are two different forms that the camera motion can take. The first is a *saccadic* motion which brings the camera to view on a particular part of the scene. An exploratory algorithm will typically generate a sequence of such camera motions. The second form of camera motion is a *pursuit* motion where the desired motion is a non-constant smooth trajectory. Examples of algorithms which generated this type of motion include those which track moving objects, the optimal shape from shading technique of Ferrier and Clark

[16] as well as their active object mapping algorithm [17].

One can think of an approach for deciding what camera motion to carry out that bases the decision on a measure of the saliency of a given point in the camera's positional configuration space. That is, based on the visual data coming from the camera, and depending on the particular directed vision algorithm being performed, a saliency value is computed for each possible position (or perhaps each increment in position) that the camera can acheive. The camera is then moved to the configuration of maximum saliency. At this point a region of interest (ROI) processor may perform more complicated visual tasks. A pursuit motion or other trajectory can be obtained by constantly changing the saliency map so as to shift the point of maximum saliency over time in the required direction.

Assuming that one adopts such a saliency based camera motion control scheme, one must answer the question of how to determine what is the saliency measure to be used, and how is it to be computed? In the context of human visual search, or exploration, Treisman and Gelade [37] identified a "preattentive" stage wherein certain features, primitives, are detected in parallel across the visual field. Possible primitives include colour, line ends (terminators), spatial frequency, motion, line orientation, binocular disparity, and texture (see [5, 8, 21, 18, 26, 37, 40]). These features could then be combined to produce a saliency map by forming a weighted combination of the feature values. Depending on the precise weights, the point of maximum saliency will appear at different points. Changing these weights corresponds to shifting the point of maximum saliency, and hence, in our view, changing the focus of attention.

One can describe the control of a mechanical system through a differential equation relating the effect of control inputs to the state of the system as follows:

$$\dot{x}(t) = f(x(t)) + G(x(t))v(t) \ ; \ y(t) = h(x(t)) \tag{10.1}$$

where $x(t)$ is an n-dimensional state vector, $v(t)$ is a m-dimensional vector of controls and $y(t)$ is a p-dimensional vector of sensor signals (which depends on the system state x). The system state usually includes the positions of the various mechanical degrees of freedom of the structure. The function $G(\cdot)$ relates the effect of the control vector $v(t)$ on the system state. The control vector can be independent of the sensor variables $y(t)$ in which case we have *open loop* control, or it can depend on the

sensor variables in which case we have *closed loop* control (assuming, of course, that the sensor variables are functions of the system state).

To make the open loop/closed loop distinction more explicit one can write the control vector as the sum of an open loop component and a closed loop component as follows:

$$v(t) = u(t) + k(y(t)) \tag{10.2}$$

The term $u(t)$ represents a vector of open loop control inputs, or setpoints, that we wish the system to follow. The function $k(\cdot)$ operates on the sensor variables $y(t)$ to provide the state feedback required for closed loop control.

The above formalism captures both the physical nature of the system (through the G and f functions) and the activities the system is to undertake (through the u, k, and y functions). One can absorb the definition of the $y(t)$ functions into the k function by assuming that all possible observations are available and that the k *selects* the observations that are used in any given control scheme. The k function is of critical importance in our saliency based scheme. Let us assume that k is a linear operator, i.e.

$$\vec{v}(t) = \vec{u}(t) + \mathbf{k}(t)\vec{y}(t) \tag{10.3}$$

where $\mathbf{k}(t)$ is a time varying matrix of feedback gains. Let us further interpret the vector $\vec{y}(t)$ as a feature vector, derived from the camera image. It is clear that by altering the elements of $\mathbf{k}(t)$ we alter the relative effect that the various features have on the control signal $\vec{v}(t)$ and hence on the position of the camera.

In the remainder of the chapter we present details of of an attentive vision system that we have implemented that is based on this servo model of attention. This system is a dual level system. The first, or inner, level performs automatic vergence and pursuit operations based on set points and mode controls supplied by the outer level. The outer level sends setpoints and feedback gains to the inner level which are themselves based on a set of setpoints and feedback gains provided by the user as input to the outer level. In this case the outer level gains k's describe what visual routines, or modes, are to be applied to the binocular visual input (the $y(t)$'s) to generate the control signals (the $v(t)$'s). Changes in attention are implemented by supplying the outer level motion control component with a new set of feedback gains. Visual routines which involve many

shifts in attention are implemented by sending the controller a mode containing a sequence of feedback gains and setpoints.

10.4 The Harvard Head Oculomotor Control System

In this section we describe the physical configuration of our robotic "head" and describe the implementation of the low level oculomotor control system for our attentive binocular vision system. This control system is based on models of mammalian oculomotor control systems.

The mechanical structure of our binocular image acquisition system is shown in figure 11.1. This mechanism can be attached to a mobile platform or it may be rigidly fixed to a worktable overlooking the workspace of a robot for assembly or inspection tasks. The "head", shown in figure 11.2, has seven degrees of freedom that must be controlled. Three of these degrees of freedom are associated with the orientation of the cameras, while the other four have to do with the state of the cameras' aperture and lens focus. The three mechanical degrees of freedom are: 1) Pan, which is a rotation of the inter-camera baseline about a vertical axis, 2) Tilt, which is a rotation of the inter-camera baseline about a horizontal axis, and 3) Vergence, which is an antisymmetric rotation of each camera about a vertical axis. With these three degrees of freedom one can theoretically place the intersection of the optical axes of the two cameras (what we will refer to as the fixation point) anywhere in the three dimensional volume about the head. In practice, the volume of accessible fixation points will be restricted due to the limited range of motions of the degrees of freedom.

The distance to the surface of exact focus can be controlled with the electronic focus on the lens. This distance ranges from a near distance of about 30 cm to essentially an infinite distance away. The focus control is an integral part of any attentive vision system as it allows us to focus on the point of fixation. With no focus control, the features that we are fixating on may be out of focus. The ability to control lens focus also allows us to obtain depth information monocularly through focusing [25], or through defocus measurements [22, 29]. Our system also allows control over the lens aperture, which affects the amount of light received by the image sensor, and the depth of focus (not to be confused with the depth of the surface of exact focus). It is important to be able to adjust the aperture to maintain sufficient light levels for the image sensor. The

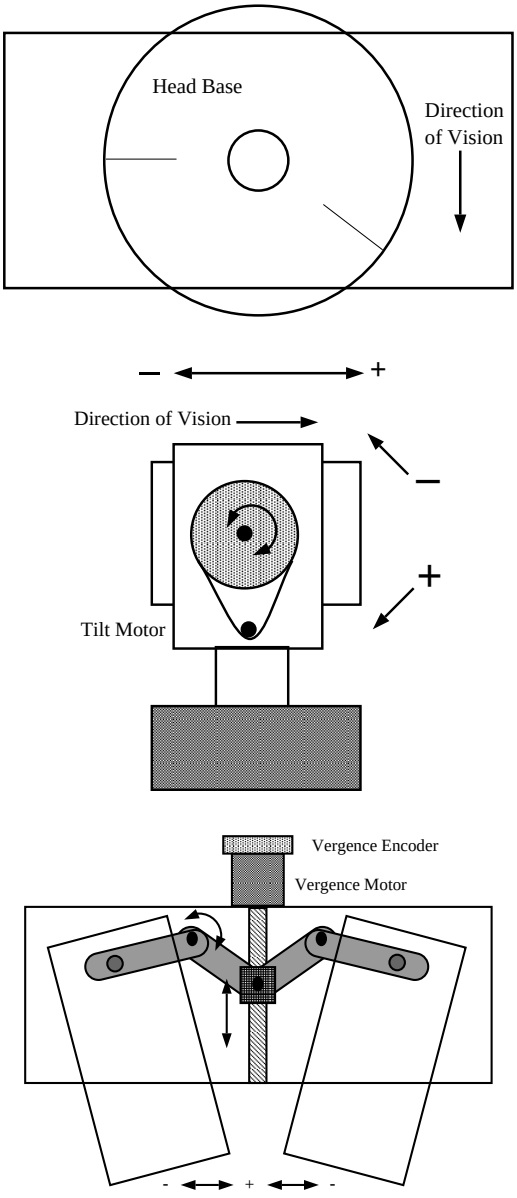


Figure 10.1
A schematic view of the Harvard head showing the sign conventions for the pan, tilt, and vergence angles.

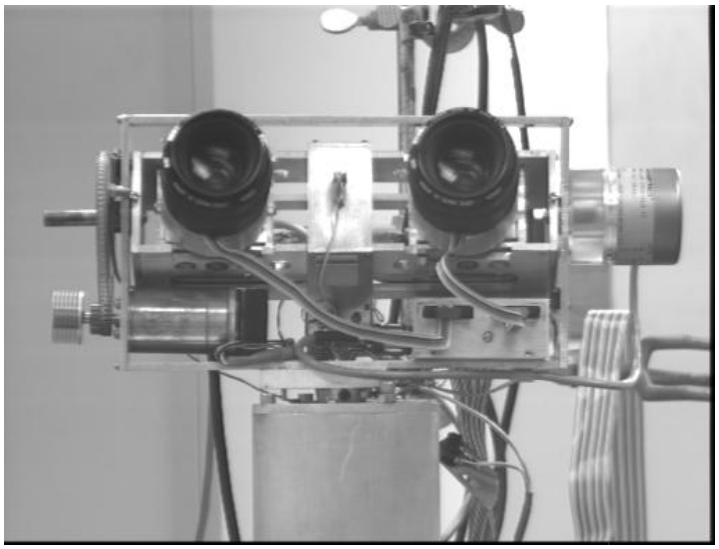


Figure 10.2
A frontal view of the Harvard head.

aperture control in our system is automatic, and responds to changing light levels, and is not dependent on any attentive inputs. DC motors are used to drive the pan, tilt, and vergence axes. The pan axis is driven directly, while the tilt axis is belt driven, mainly due to space considerations. The vergence motor drives a lead screw, which then causes the camera rotations through a kinematic chain. The relationship between the vergence motor rotation (or the lead screw displacement) and the camera vergence angle is approximately linear (within 1 percent over the range of travel) which makes the programming of the vergence control simple. The focus motion is generated via a motor encased in the lens housing. Control signals to this motor are generated by an integrated circuit also located in the lens housing. A digital data stream, suitably encoded, must be sent to the focus motor driver I.C., to command a change in focus. The manufacturer of the lens, Canon, would not release details on the specifications of the required command data streams, so we determined the proper data sequences ourselves. These details are available from the authors, subject to certain disclosure conditions.

One can partition the control of the pan, tilt, and vergence axes of the head mechanism into three descriptive regimes. These are, *saccades, pursuit,* and *vergence.* Taken together, these three modes of operation allow control over shift in attention, and maintenance of attention. A saccade is a rapid motion of the pan and tilt axes which causes a coupled motion of the optical axes of the two cameras, resulting in a change in the direction of gaze of the cameras. In a saccade, both cameras move in the same direction. This motion is not enough to allow independent control of the gaze direction of each camera. To obtain this one uses a vergence movement. A vergence movement is a coupled motion of the two cameras wherein the the two cameras rotate in opposite directions. Taken together, the saccadic and vergence systems allow the fixation point of the binocular camera system to be arbitrarily controlled. Once the saccadic and vergence systems have fixated the cameras on a feature in the scene, the pursuit system is then used to track the feature. The pursuit system adjusts the velocity of the pan and tilt axes so as to minimize the retinal velocity (the velocity as measured in the camera images) of the fixated feature. This will keep the feature fixated as long as it does not move in depth. If it moves in depth the vergence system will adjust the vergence angle (the relative angle between the two cameras) to maintain fixation.

The human oculomotor system is very complex and it is not yet fully understood. It contains many interacting functional modules, such as [23] the Frontal Cortex (for making plans and intentions), the Occipital Cortex (for visual reflexes and smooth pursuit tracking movements), the Pontine Reticular Formation (for both saccadic and pursuit movements), the Cerebellum (for coordinate transformations), the Superior Colliculus (for relating visual input to oculomotor commands), and the Vestibular System (for allowing the eyes to compensate for body motions). The control system that we have described in this chapter takes on the functionality of many of these modules, and we do not claim our system as a model for any particular part of the human oculomotor control system. We have, however, used some models of the human oculumotor system in developing our system.

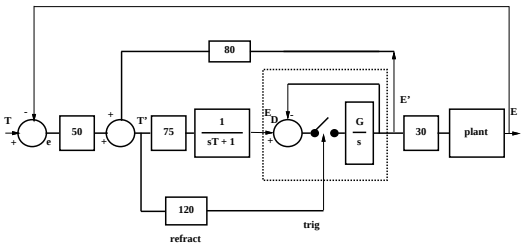
In humans, the physiological evidence indicates that saccades are controlled with a sampled data system, while pursuit motions are continuously controlled [31, 33]. The latency, or reaction time of the human saccadic system has been determined to be about 200 milliseconds [31],

although it has been observed that anticipatory behaviour can reduce this latency time [11]. This latency is the time it takes from the moment of change in retinal position of an attended feature to the moment that a motor command is given to generate the saccade. Presumably the bulk of this time is taken up in processing the retinal image to determine the position of the feature. During this period the oculomotor system is insensitive to further changes in the retinal position of the feature, and the saccade that is generated is that appropriate to the retinal position of the feature as it was 200 milliseconds prior to the generation of the saccade. If the feature moves during this refractory period the saccade will result in a position error. From this observation came the sampled data model of the oculomotor control system, originally proposed by Young and Stark [43].

Young and Stark treated the pursuit system as a sampled data system as well. Upon further psychophysical examination (e.g. see [32]) this assumption turned out to be incorrect, and the pursuit system is now thought to use a continuous time data system, or at least a sampled data system in which the sampling rate is much higher than the sampling rate for the saccadic system [32]. It has been observed [10] that pursuit movements are not always smooth, but will include saccadic components if the visual feature being pursued has a large retinal velocity. Presumably these saccades are neccessary if the pursuit system can not keep up with the moving object. In this case a cumulative position error builds up, and when this error reaches a certain threshold a saccade is generated in order to reduce the position error.

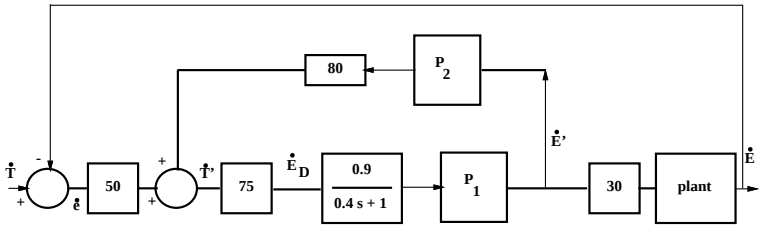
The control scheme that we use to control the pan, tilt, and vergence degrees of freedom of our head system is based on the model of human oculomotor control described by Robinson in [31]. This model postulates separate subsystems for pursuit and saccadic motion. These subsystems are depicted in figure 11.3 (adapted from [31]).

There are two interesting features of Robinson's model. The first is that the sampled data nature of the saccadic system. The desired retinal position, E_D is sampled (with a pulse sampler), and held by a first order hold (an integrator). The output of this sample/hold is then used as a setpoint to the plant (in this case the local motor controller). The actuator will then try to move the camera to the desired position. During the period between sampling pulses, the output of the sample/hold is being held constant, and hence the desired eye position is being held



Robinson's Oculomotor Control System Model

-- Saccadic Motions



Robinson's Oculomotor Control System Model

-- Pursuit Motions

Figure 10.3
Robinson's model for the human oculomotor control system. TOP: Saccades, BOTTOM: Pursuit.

constant, even though the image of the feature to be attended to may be moving. A sample/hold does not appear to be present in the pursuit system.

The second feature of Robinson's model to be noted is that there is internal positive feedback in the control loop. This positive feedback is neccesary in the case of the pursuit system (figure 11.3b) to prevent oscillations due to delays in the negative feedback loop. The negative feedback is provided by the vision system which, in the case of the pursuit system, detects the velocity of a feature, computes the retinal velocity error (which is equal to the retinal velocity since the desired retinal velocity is zero for tracking purposes), and causes the eye to move in a manner to reduce this error. However, these computations can not be done instantaneously, so there is a delay between the time at which an visual observation is made and the time at which the control command based on this observation is available. To eliminate the oscillations that can occur with this feedback, a compensatory internal positive feedback is inserted into the loop. This is done by adding a delayed "efference copy" of the current eye velocity to the computed retinal velocity error. The delay is such that the efference copy that is added to the velocity error is that measured at the same time that the visual observation (that the retinal velocty error is based on) is made. The sum of the retinal velocity error and the delayed efference copy gives a new desired eye velocity which is input to the plant (eye muscles or motor driver). The effect of this positive feedback path is to essentially eliminate the negative visual feedback. The saccadic system is modeled in the same way, except that position control is being done instead of velocity control. In the saccadic system, however, the internal positive feedback is not really needed to ensure stability, as stability is gained through the use of the sample/hold. Nontheless, the available evidence indicates that the human saccadic system does use internal positive feedback to compensate for delays.

Note that the internal positive feedback scheme implies that the saccadic system directs the eye to move to an absolute position, in head coordinates, rather than to move by a certain displacement in a given direction. The issue of whether saccadic control of eye movements is head coordinate based or retinotopic coordinate based has been long a subject of discussion among neurophysiologists. The current evidence, according to Robinson [31] and others, suggests that head based coordinates

are used.

Details on a model for the vergence system are sketchy, but Robinson [31] indicates that the vergence system is continuous (no sample/hold is used) and uses internal positive feedback (although this is by no means certain). This is similar to the pursuit system save that position control is being done instead of velocity control and that the vergence system responds more slowly than the pursuit system.

Based on Robinson model as described above we have implemented the control scheme that is depicted schematically in figure 11.4 for the Harvard head. The pan, tilt, and vergence motors are driven by a pulse

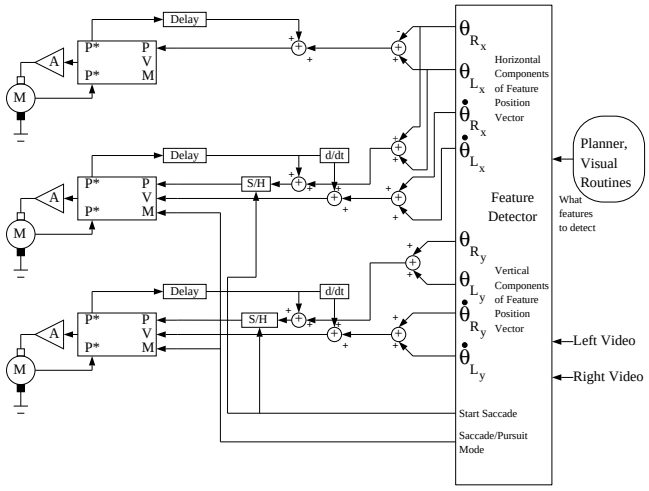


Figure 10.4
The control system used for the Harvard head.

width modulated MOSFET amplifier. The input to this amplifier is derived from the output of a Dynamation motor controller board [13]. The Dynamation board is indicated in figure 11.4 by the box taking in the shaft encoder position from the motor and which outputs a drive signal to the motor amplifier. The Dynamation board takes set point inputs over a VME bus connection to a SUN computer. These setpoints can either be position setpoints (in the case of vergence or a saccade) or velocity setpoints (in the case of pursuit). The Dynamation can output to the VME bus (and then on to the SUN computer) an efference copy

of the current motor position. This efference copy is delayed, in the SUN computer, by a time equal to the time taken to perform visual feature localization, and added to the current position errors, determined by the visual feature localization process. The Dynamation board does not have a tachometer, so that an velocity efference copy is not available. Thus we generate one by differentiating the position efference copy. The sampling rate of the Dynamation board is very high (more than 1000 samples per second), however, so that this estimate of velocity should be accurate.

The feature detection and localization is performed in a special purpose image processing system, manufactured by Datacube [12]. This system can do image processing operations such as 8x8 convolution, histogramming, and logical neighborhood operations on a 512x512 pixel image at video rates (30 frames per second). Thus the latency per operation is 33 milliseconds. Most feature detection operations require more that one frame time however. In our initial experiments we implemented a feature detector that could detect black blobs or white blobs, in about 3 frame times. Therefore the latency of our feature detector was about 100 milliseconds. The Datacube system, after it detected the presence of a feature, would output the position and velocity of the feature over the VME bus to the SUN workstation. The SUN workstation then computes the quantities $\theta_{R_x} + \theta_{L_x}$, $\dot{\theta}_{R_x} + \dot{\theta}_{L_x}$, $\theta_{R_y} + \theta_{L_y}$, $\dot{\theta}_{R_y} + \dot{\theta}_{L_y}$, and $\theta_{R_x} - \theta_{L_x}$, where θ_{R_x} is the x component of the retinal disparity in the right camera, θ_{R_y} is the y component of the retinal disparity in the right camera, θ_{L_x} is the x component of the retinal disparity in the left camera, θ_{L_y} is the y component of the retinal disparity in the left camera, and $\dot{\theta}$ indicates a retinal velocity. The difference in the left and right x components of the retinal position is added to the delayed position efference copy of the vergence motor. Thus this difference will be driven to zero. The sum of the left and right retinal position errors in both the x and y directions are added to the delayed position efference copies of the pan and tilt motors respectively. This will, during a saccade, drive these sums to zero. Combined with the driving of the difference of the x retinal position errors to zero by the vergence, the result will be that the x and y retinal position errors in both cameras will be driven to zero, as desired. A saccade trigger signal (that opens up the sample/hold) is generated by the feature detection system when the retinal position error is greater than threshold value. During the saccade, visual process-

ing is turned off to prevent saccades being generated while the saccadic motion is being performed.

During pursuit the sum over the two cameras in each of the x and y retinal velocity errors will be driven to zero. If the system has the correct vergence, then the x and y component of the retinal velocity error will be driven to zero in each eye, and not just the sum of the errors in the two eyes.

We have performed simple blob tracking experiments which show that the system operates as desired, in that the vergence and saccadic modes result in fixation of the feature as we move it about in space.

10.5 Modal Control of Attention

The inner level control loop described in the previous section is controlled by an outer loop which implements attentional shifts in camera positions.

The first stage in our visual attention model acquires the images and extracts "primitives" in parallel across the visual field. The results from this stage are a set of feature maps $y_i(x, y, t)$ which indicate the presence or absence of a feature at each location in the image. Simple feature maps may indicate the presence of a specific color or line orientation. Complex feature maps may perform texture and figure-ground segmentation or more complex feature maps may implement inhibition from neighboring regions to compute which regions are different from their surroundings.

The next stage of the model combines the results from the feature maps. The output from the feature maps are "amplified" with different "gains", $k_i(t)$ for each map y_i and then these amplified values are summed to form the saliency map, $S(x, y, t)$. The value of the map at each location is a numeric indicator of how "salient" is the information at that location. Hence finding the location with the maximum value will give the most salient location with respect to the given amplifier gains, $k_i(t)$. As the notation indicates, these gains may vary over time, thus changing the location of the most salient feature. If more than one location shares the same maximum value, one location must be chosen (it does not make sense to attend to a location in the middle of two salient features, one or the other location must be picked). Figure 11.5 shows this attention model.

It can be seen that this model incorporates many of the psychophysical

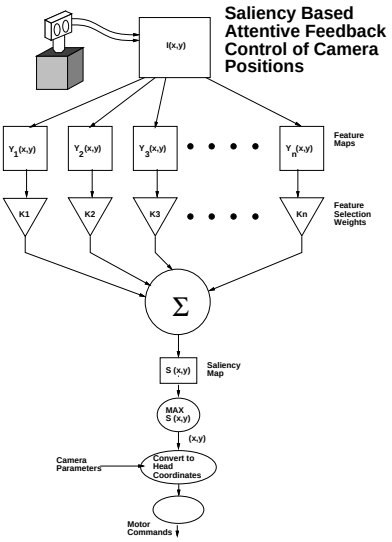


Figure 10.5
Saliency based attentive feedback control of camera position.

results observed earlier. Adjusting the gains of a particular feature map will direct attentional resources to occurences of that feature. A decaying gain function, $k(t)$, will decrease the saliency of a location over time and hence another location will become more salient and attention will change to a new location. For example, consider the detection of a red T in a field of green L's. Suppose attention is first be directed at the red T. As the gain in the feedback path corresponding to the color red decreases and the gain corresponding to the color green increases, the focus of attention will change locale, to attend to the nearest green L. Another psychophysical result which is captured in our model is that higher cognitive levels can actively select which features to attend to by adjusting $k_i(t)$. Human attention can be consciously applied to a visual task so humans must be able to consciously select the more salient features.

Koch and Ullman [24] describe the Winner-Take-All (WTA) network which will locate the most conspicuous location (one whose properties differs most from the properties of its neighbors). The locations which differ significantly from their neighbors are singled out and a numeric

value representing the "conspicuousness" is assigned. The results from each primitive detector are combined into a global saliency map which combines the value from each feature map and assigns a global measure of conspicuity. The WTA network finds the maximum value of "conspicuity" and locates that maximum. Attention can be allocated to the position which gave the highest value for further processing.

It can be seen that the WTA scheme uses the same models of attention. The values assigned in the global saliency map of Koch and Ullman corresponds to the saliency map of this model when using an appropriate set of gains. The WTA scheme is an *implementation* which deals with the problem of finding the maximum of the saliency map and localizing it. The notion of winner-take-all is appropriate since only one location can be attended to at one time. Koch and Ullman actually suggest the idea of a higher cognitive process adjusting the "conspicuousness" of a feature to selectively inhibit or attend on a specific feature, which corresponds to changing $k(t)$ in this model.

We implement our modal attention scheme with two nested feedback loops. The gains of the inner feedback loop which is concerned with setpoint control of the head positioning motors remains constant, as the load on the head motors remain roughly constant. One need only determine the position feedback gains k once, such that the step response of the motor to the inner level setpoints is critically damped. These gains are set in the Dynamation controller board, which handles the inner level control loop. The sensory input to the inner level is the motor shaft position, measured with the shaft encoders. The velocity of the motor shafts are not measured directly but are computed from the position measurements through differentiation as described in the previous section. The inner control loop is switched between position control and velocity control by the outer control level. This is done, in effect by sending a (u, k, T) triple (following Brockett [6], we refer to this triple as a *mode* , where T is a time interval, u is a setpoint or trajectory during this time interval, and k are the feedback gains to be used in this interval) in which the k's decide which measurement (position or velocity) will be used to control the motor. The setpoints u that are input to the inner level control loop also come from the outer control loop in these (u, k, T) triples.

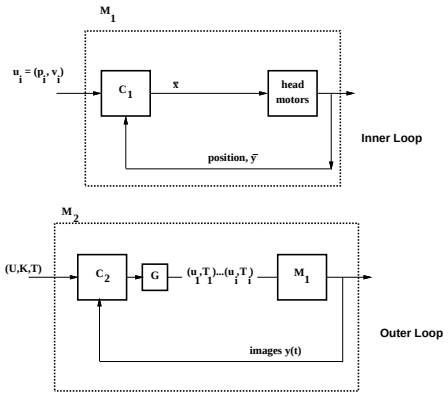
The k's in the (u, k, T) motion control system definitions concerned with the outer, visual, feedback loop will change due to changes in the

focus of attention. The feedback selection process at this level is much more complicated than the inner level feedback selection in which only direct position or velocity feedback was being selected for. In the outer level, one still selects for position or velocity feedback but, in addition, one must select the feature(s) to be used to detect the scene element whose position or velocity is fed back. This feature selection is performed by adjusting the weight we apply to a given feature in the control feedback loop.

The outer control level consumes *modes* which allocate attention to specific features and produces different modes for the inner loop. The output modes consist of position and velocity setpoints and a time interval in which to apply these setpoints. The modes consumed by this outer level are again of the form (u, k, T) where u is the desired postion (always 0 for foveation – to center target on visual field), k is a vector which represents which features to detect (the amplifier gains) and T is the time period in which the mode is to be applied.

In the language given earlier, $y(t)$ is the feedback vector. In this case, $y(x, y, t)$ is a pair of images (left and right "eyes"). Referring to the model given earlier, $K(t) = (k_1(t), k_2(t), ..., k_n(t))$ is a vector containing the "weights" to be applied to the results from the primitive operations (feature maps). With these gains, the saliency map can be computed and the maximum found. The location of the maximum must then undergo a coordinate transform in order to obtain the setpoints in head coordinates. This transformation will depend on the camera parameters and the particular configuration of the "head" and hence can be absorbed in the $G(\cdot)$ term in equation (1). The idea that alteration of the gains of visual feedback paths result in shifts in attention, (or vice versa) has some support from physiological studies [20, 28, 30, 41, 42] which indicate that the responses of neurons involved in visual perception are modulated by changes in the focus of attention.

Figure 11.6 shows the lowest two stages of the modal control. A mode, (u, k, T), which was generated at a higher level, is "fed" into the intermediate level (denoted M2). Over a time period, $0 \leq t \leq T$ the weights associated with the feature maps will be $K(t) = (k_1(t), k_2(t), ..., k_n(t))$. At each instant of time, t, a location (x, y) will be output as the "most salient feature" of the image. These positions are output to the inner loop (denoted M1) where they generate positional errors used to drive the head motors.



The Two Control Loops

Figure 10.6
The two lowest stages of the modal control system.

10.6 Summary

In this chapter we have provided a brief explanation of the desirability of active vision. We gave a number of typical applications of active vision. We described a control system for a binocular camera mechanism which allows shifts in focus of attention to be made in a natural, device independent manner. Shifts in focus of attention is accomplished via altering of feedback gains applied to the visual feedback paths in the position and velocity control loops of the binocular camera system. By altering these gains we can perform a feature selection operation, by which the *saliency,* in the sense of Koch and Ullman [24], of a given feature is enhanced, while the saliency of other features are reduced.

The control system that we have described in this system is a two level one. The first, or inner, level performs the direct control over the position and velocity of the motors attached to the cameras. This level is based on models of the human oculomotor control system. The outer level controls the focus of attention, in that it determines what features are going to be used in determining where to look next.

Acknowledgments

The mechanical components of the Harvard Head system were designed and constructed by J. Page, W. Labossier, and M. Cohn, with input from R. Brockett, J. Clark and E. Rak. M. Cohn reverse engineered the Canon electronic focus controls (details can be found in [15]). The electronics for the motor drivers and associated systems was put together by J. Page. Software for the motion control systems was written by N. Ferrier, and P. Newman. Software for the visual processing modules was written by N. Ferrier with some assistance from M. Lee and E. Rak. Ideas and enthusiasm concerning the development of the head and its motion control were supplied in great abundance by R. Brockett. The authors would like to thank J. Daugman for bringing to our attention some of the psychophysiological work concerning attention.

This research was supported in part by the Office of Naval Research under grant N0014-84-K0504, the NSF University of Maryland/Harvard University Systems Research Center funded through grant CDR-88-03012, and by the Harvard-MIT-Brown Center for Intelligent Control Systems.

References

[1] John (Yiannis) Aloimonos and Amit Bandyopadhyay. Active vision. In *Proceedings of the IEEE 1st International Conference on Computer Vision*, pages 35–54, London, December 1987. IEEE.

[2] John (Yiannis) Aloimonos and Anup Basu. Combining information in low-level vision. Technical report, University of Maryland, 1988.

[3] Ruzena Bajcsy. Active perception vs. passive perception. In *Proceedings 3rd IEEE Workshop on Computer Vision*, pages 55–59, Bellaire, April 1985.

[4] Ruzena Bajcsy. Perception with feedback. In *Proceedings of the Darpa Image Understanding Workshop*, Cambridge, MA, April 1988. DARPA.

[5] Jacob Beck and Bruce Ambler. The effects of concentrated and distributed attention on peripheral acuity. *Perception and Psychophysics*, 14:225–230, 1973.

[6] Roger W. Brockett. On the computer control of movement. In *Proceedings of the 1988 IEEE Robotics and Automation Conference*, Philadelphia, 1988.

[7] Christopher Brown. Progress in image understanding at the University of Rochester. In *Proceedings of the DARPA Image Understanding Workshop*, Cambridge, MA, April 1988. DARPA.

[8] David Burr and J. Ross. Visual processing of motion. *Trends in Neuroscience*, 9, 1986.

[9] Peter Burt. Algorithms and architectures for smart sensing. In *Proceedings of the Darpa Image Understanding Workshop*, Cambridge, MA, April 1988. DARPA.

[10] Han Collewijn and Ernst Tamminga. Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds. *Journal of Physiology*, 351:217–250, 1984.

[11] P.J. Dallos and R.W. Jones. Learning behavior of the eye fixation control system. *IRE Transactions on Automatic Control*, 8:218–227, July 1963.

[12] Datacube, Inc., Peabody, MA. *Maxvideo System Documentation.*

[13] Dynamation, Inc., Mountain View, CA. *Dynamation Motion Controller Board Documentation.*

[14] Charles W. Eriksen and James St. James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception and Psychophysics*, 40(4):225–240, 1986.

[15] Nicola Ferrier. The Harvard binocular robotic head. Technical Report 91-9, Harvard Robotics Laboratory, September 1990.

[16] Nicola Ferrier and James Clark. Optimal motions for active sensing. In *International Symposium on Robotics and Manufacturing*, Burnaby, BC, Canada, July 1990.

[17] Nicola Joy Ferrier. *Trajectory Control of Active Vision Systems.* PhD thesis, Harvard University, 1992.

[18] J.P. Frisby and J.E.W. Mayhew. Spatial frequency tuned channels: implications for structure and function from psychophysical and compuatational studies of stereopsis. *Phil. Trans. R. Soc. Lond. B*, 290:95–116, 1980.

[19] Davi Geiger and Alan Yuille. Stereopsis and eye movement. In *Proceedings of the First IEEE Conference on Computer Vision*, pages 306–314, London, January 1987. IEEE.

[20] P.E. Haenny, J.H.R. Maunsell, and P.H. Schiller. State dependent activity in monkey visual cortex: visual and non-visual factors in V4. preprint, 1985.

[21] Anya Hurlbert and Tomaso Poggio. Do computers need attention? *Nature*, 321(12), 1986.

[22] Ten Lee Hwang, James J. Clark, and Alan Yuille. A depth recovery algorithm using defocus information. Technical Report 89-2, Harvard Robotics Laboratory, 1989.

[23] E.R. Kandel and J.H. Schwartz. *Principles of Neural Science*, chapter 34. Elsevier/North-Holland, New York, 1981.

[24] Christof Koch and Shimon Ullman. Selecting one among the many: A simple network implementing shifts in selective visual attention. Technical Report 770, MIT AI Laboratory, Jan 1984.

[25] E. Krotkov. Focussing. *Internation Journal of Computer Vision*, 1(3), 1987.

[26] David Marr and Shimon Ullman. Directional selectivity and its use in early visual processing. *Proc. R. Soc. Lond. B*, 211:151–180, 1981.

[27] Larry Matthies, Takeo Kanade, and Richard Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[28] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[29] Alex Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, 1987.

[30] S.E. Petersen, D. L. Robinson, and W. Keys. Pulvinar nuclei of the behaving rhesus monkey: visual responses and their modulation. *Journal of Neurophysiology*, 54:867–886, 1985.

[31] D. A. Robinson. The oculomotor motor control system: A review. *Proceedings of the IEEE*, 56(6):1032–1049, June 1968.

[32] D.A. Robinson. The mechanics of human smooth pursuit eye movement. *Journal of Physiology*, 180:569–591, 1965.

[33] D.A. Robinson. Why visuomotor system don't like negative feedback and how they avoid it. In M. Arbib and A. Hanson, editors, *Vision, Brain and Cooperative Computation*, pages 89–107. MIT Press, Cambridge, MA, 1987.

[34] D. Sagi and Bela Julesz. Enhanced detection in the aperture of focal attention during simple discrimination tasks. *Nature*, 321(12):693–695, June 1986.

[35] G.L. Shulman, R.W. Remington, and J.P. McLean. Moving attention through visual space. *Journal of Experimental Psychology: HP & P*, 5(3):522–526, 1979.

[36] M. J. Swain and M. Stricker. Promising directions in active vision. Technical report, University of Chicago, November 1991.

[37] Anne M. Treisman. Strategies and models of selective attention. *Psychological Review*, 76:282–299, 1969.

[38] John K Tsotsos. A 'complexity level' analysis of vision. In *Proceedings of the IEEE 1st International Conference on Computer Vision*, pages 346–355, London, December 1987. IEEE.

[39] Shimon Ullman. Visual routines. Technical report, MIT AI Laboratory, June 1983.

[40] H.R. Wilson. A four mechanism model for threshold spatial vision. *Vision Research*, 19:19–32, 1979.

[41] R.H. Wurtz, M.E. Goldberg, and D.L. Robinson. Behavioral modulation of visual responses in the monkey: stimulus selection for attention and movement. *Progress in Psychobiology and Physiological Psychology*, 9:43–83, 1980.

[42] R.H. Wurtz, B.J. Richmond, and W.T. Newsome. Modulation of cortical visual processing by attention, perception, and movement. In *Dynamic Aspects of Neocortical Function*, pages 195–217. Wiley & Sons, New York, 1984.

[43] L.R. Young and L. Stark. Variable Feedback Experiments Testing a Sampled Data Model for Eye Tracking Movements. *IRE Transactions on Human Factors in Engineering*, 4:38–51, September 1963.