Modal Control of an Attentive Vision System *

James J. Clark
Nicola J. Ferrier

Division of Applied Sciences
Harvard University
Cambridge, MA

Abstract

A vision system for use in a mobile robot system, or in a fixed multi-tasking industrial robot requires attentive control. Attentive control refers to the process by which the direction of gaze of the visual sensors are determined, along with the determination of what processing is required to be applied to the sensed images based on the goals of the robot and the tasks it is performing.

This paper describes the implementation of a motion control system which allows the attentive control of a binocular vision system. Attentive inputs to the system specify the type of visual feedback that the oculo-motor control system will use. The MDL language developed by Brockett [7] is used to communicate between the attentive planner and the motion controller.

1 Introduction: Active and Attentive Vision

An emerging theme in current vision research is that of "Active Vision" [1,8,4,10], in which the vision process is dynamic rather than static. Instead of applying image analysis operations to a single "snapshot" of the environment, the active vision process in its most general implementation applies image analysis operations in an purposive and integrative manner on a temporal and spatially disparate sequence of images.

There have been put forth a number of definitions for active vision. Active vision in the sense used by Aloimonos et al [1] is 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. 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 [8]. Controlled eye movements can also be used to help in a calibration process, wherein geometrical information regarding the imaging system is obtained [8].

In each of the above types of active vision one must be able to control the position of the cameras so as to obtain independent visual constraints as required to make a given visual processing operation well posed, or to simplify a given visual processing operation. There are, however, situations in which one would like to change our field of view for reasons other than the ones given above. For example, Burt [10] 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 [4,5] 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.

Changing of the focus of attention can refer to changes in the spatial region in the scene upon which our visual system is concentrating (or "foveating" at a high resolution). Much of the information in the field of view is not needed to perform many tasks. This is witnessed in the human eye where much of the field of view, the periphery, is viewed at a very low resolution. Only the fovea, a small percentage of the total field of view, contains high resolution detail. By acquiring a part of the scene within the fovea, a detailed analysis can be made of this region. Visual tasks require the movement of the eyes to closely examine areas of interest for the particular task, paying little attention to the rest of the scene which is viewed peripherally. Since we want to devote computational power only to regions containing salient information, these regions must be identified and then foveated. As the environment changes we want attention to move so that we are constantly acquiring the regions with salient information in the fovea. (There is evidence that attention can be directed to a location other than the point of fixation, such a mechanism will only be useful for simple tasks as the low resolution in the periphery will not be adequate

*This research was supported in part by the University of Maryland/Harvard System Research Center and by the Harvard-MIT-Brown Center for Intelligent Control Systems.

for complex visual tasks [3]). Foveation can be accomplished by mechanically adjusting the direction of view of the image sensor or it can be accomplished by moving a processing window about in an internal representation of the image (e.g. see the multiresolution foveator of Burt [10]). This mode of attention forms the basis for Ullman's visual routine paradigm [40], 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 colour of objects be determined. In each of these two cases different features would be attended on.

In all of the proposed active and attentive vision paradigms there are two common tasks to be performed. One is to figure out where to look next, and the other is to then carry out the motion that will let one look there. The question arises as to how to determine what is the salient information in a scene. Treisman [39] 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 [39,6,21,41,18,27,9]). However, detection of conjunctions and more difficult recognition tasks are believed to require serial processing. Experiments show that a conspicuous feature such as a red T amongst a field of green P's will quickly "pop out". This is not true of a red T in a field of red and green P's since the detection of the red T requires detection of a conjunction of features. Also, subjects can selectively attend on a specific feature or "suppress" a feature, such as when subjects "ignore" features performing over a long period of time. The attentional mechanism can thus be dynamically changed and consciously programmed.

Experiments have shown that directing attention to a location for one task increases visual capabilities for other tasks in that region [38,37]. This region of increased visual attention has been likened to a spotlight or variable powered lens [21,17]. 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.

We propose a model of attention that captures the above aspects of attention. The most salient feature is found and centered on the field of view. At this point a region of interest (ROI) processor may perform more complicated visual tasks. We describe how our system can be applied to implementing changes in visual attention, and to controlling the motions of a binocular image acquisition system to change gaze directions based on these changes in focus of attention. The control technique used in our system is based on MDL [7] which is a high level motion description language that allows device (actuator) dependent control of the robot's oculomotor system. This system uses selective visual feedback, based on a high level description of the sequence of foci of attention or "modes" of activity, to control the position and velocity of the cameras in a binocular image acquisition system. The primary purpose of this system is to allow the specification of attentive behaviour of an oculomotor system at a high level, independent of the type of actuator used in the system.

2 The MDL Motion Description Language

The role of any motion control system is to allow the motion of a mechanical system to be specified to a desired level of precision.

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:

$$\dot{x}(t) = f(x(t)) + G(x(t))v(t) \ ; \ y(t) = h(x(t)) \qquad (1)$$

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:

$$v(t) = u(t) + k(y(t)) \qquad (2)$$

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 can be thought of as a generalized compliance. For example consider the case where x is a position of some kind, and y is a sensed force. Then k converts forces to changes in position. Thus the system acts as a spring with compliance (1/stiffness) k. If force components are sensed in different directions and different positional degrees of freedom are controlled by these sensed forces, then k is a matrix which relates the effect of a force in a given direction to a change in position in some other direction. This system is then a generalized spring system. If k is diagonal, then its elements determine the compliance of the system in different directions. Such a system could be more stiff in one direction than in another. In general, one may not have force sensors, or use position control. In such a case the k's will not represent compliances, but will still relate the effect of the individual sensory inputs on the control of the system. This is an important point as it shows that, by controlling the k's one can select different types of feedback mechanism. A simple example is that of hybrid position and force control used in some robotic manipulators [33]. In hybrid control both force and position is sensed and motor torque is controlled. The level of motor torque affects both position and applied force of the robot manipulator (through the motor dynamics and manipulator kinematics which are modelled with the f and G terms in the above differential equation). If the k values are such that the position sensor information has a much greater effect on the motor torque than the force sensor information, then the manipulator will act as a very stiff spring and the manipulator will track position (the position setpoint component of u) very well, but the desired force will not be followed as closely. In the opposite situation, the system will act as a very loose spring and force will be controlled accurately but position will not. This example shows how one can select between two types of feedback using the k functions. In section 4 we will extend this idea to the control of visual attention, wherein we change the values of the k's that select for different visual sensing operations in order to attend on a given scene element.

Based on the above control scheme Brockett [7] proposed an MDL (for a Motion Description Language) device which would accept the open loop controls u and the feedback processing functions k and produce the correct actuator signals which would force the state vector $x(t)$ to be a solution of the equation:

$$\dot{x} = f(x(t)) + G(x(t))(u(t) + k(y(t))) \qquad (3)$$

In most complex robotic activities such those encountered in industrial assembly tasks or in mobile robot navigation, different actions must be performed at different times. Thus we will want a control system that allows for the changing of the user definable parameters of the control system in order to allow the carrying out of the various desired operations. As we have seen, the important user definable parameters are the setpoints u, and the feedback selection functions k. The MDL device of Brockett consumes (u, k, T) triples which specify the adaptive nature of the control. Each (u, k, T) triple describes the type of control law that is to be used over the epoch T. Thus given a string of triples such as $(u_1, k_1, T_1), (u_2, k_2, T_2), \ldots (u_n, k_n, T_n)$ an MDL device would execute a motion which follows the state $x()$ given by:

$$\dot{x} = f(x) + G(x)(u_1 + k_1(y)) \ ; \ 0 \leq t \leq T_1$$
$$\dot{x} = f(x) + G(x)(u_2 + k_2(y)) \ ; \ T_1 \leq t \leq T_1 + T_2$$
$$\vdots$$
$$\dot{x} = f(x) + G(x)(u_n + k_n(y)) \ ; \ \sum_{i=1}^{n-1} T_i \leq t \leq \sum_{i=1}^{n} T_i$$

Thus, in order to produce a given complex motion, one would supply a string of (u, k, T) triples to the MDL controller. Brockett [7] refers to these strings as *modes.* One could store a number of modes, each of which corresponds to a certain complex motion, in a table where they would be available for accessing when required. These modes could be hardwired, or they could be learned through some optimization process (training and practice). A control system using motions defined as modes, that are input to an MDL controller can be called a modal control system. Note that the modes are described at a high level, and hence the modal definition of a complex motion is "device independent". Only the MDL interpreter, which converts the (u, k, T) strings into actuator signals, need be designed for each mechanical system.

In the remainder of the paper we present an MDL based implementation of an attentive vision system. This system will control the motion of a pair of cameras in such a way as to facilitate the execution of varying robotic tasks. The system that we are proposing 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 (u, k, T) triples to the inner level based on a set of (u, k, T) triples provided by the user as input to the outer level. In this case the outer level 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 (u, k, T) triple. Visual routines which involve many shifts in attention are implemented by sending the controller a mode containing a string of (u, k, T) triples. The implementation of the outer level component is described in detail in section 4.

3 Oculomotor Control Systems

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 physical motion required to adjust the positions of cameras attached to a robot can be obtained in many ways depending on the mechanical structure of the robot. For example, if the robot is mobile and can move with three degrees of freedom (translation in x, and y and rotation about the z axis) in a plane, then the direction of view of a camera, fixed to the robot, can also be controlled with these three degrees of freedom. In general, however, it is more convenient to decouple the attitude of the camera(s) from the attitude of the body of the robot. This allows the camera to look in a given direction independently of the direction in which the robot is pointed. Furthermore, the time constants of a system that positions the camera alone will be, in general, much smaller than that of a system that positions the robot. So, by controlling the camera orientation independently of the orientation of the robot one, obtains an increase in flexibility and speed, over the case in which the camera orientation is rigidly coupled to that of the robot.

The mechanical structure of our binocular image acquisition system is shown in figure 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 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 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, or optic flow, (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.

In humans, the physiological evidence indicates that saccades are controlled with a sampled data system, while pursuit motions are continuously controlled [35,36]. The latency, or reaction time of the human saccadic system has been determined to be about 200 milliseconds [35], although it has been observed that anticipatory behaviour can reduce this latency time [12]. 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 time 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 [45].

Young and Stark treated the pursuit system as a sampled data system as well. Upon further psychophysical examination (e.g. see [34]) 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 [34]. It has been observed [11] 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 necessary 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.

Vergence motions are the motions by which the direction of gaze of two spatially disparate eyes or cameras are brought to intersect at a given point in space. Coordination of the movement of the two cameras is of obvious importance in this regard. Ditchburn [14] has suggested based on his experiments that saccades are generated in both eyes at the same time and that the decision to make a saccade is based on information from both eyes. However, the magnitudes of the saccades can be different in the two eyes, and these magnitudes are determined wholly on the information from the individual eyes. Enright [16] presents the results of experiments which indicate that relatively large vergence movements are superimposed on saccades (if required) followed by slow vergence motions after the saccade. This is in opposition to the long held view of Yarbus [44], Alpern [2] and others, who postulated that vergence motions were slow and symmetrical, and were superimposed on balanced saccades (saccades of equal magnitude for each eye). In our system, however, we cannot independently control the magnitude of the saccades of the two eyes, as they are, by the physical nature of our mechanism, balanced. Likewise, our vergence rates are constant so that the model we will be using to control vergence will be that of Yarbus, and not of the currently held models of human vergence control.

There is evidence that the control of focus in humans is linked to the vergence mechanism. When vergence changes, the depth to the plane containing the fixation point also changes. In order to keep the fixation point in focus, the focus must change as the vergence changes. The focus control cannot be completely slaved to vergence, however. One often wants to control the focus independently of vergence in order to obtain monocular depth information (via focusing or defocus information [22,25,29]). In addition, precise control over the focus allows one to bring fixated features into exact focus, when the focussing due to slaving of the vergence results in only near exact focus conditions.

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 [35]. This model postulates separate subsystems for pursuit and saccadic motion. These subsystems are depicted in figure 3 (adapted from [35]). 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 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 necesary in the case of the pursuit system (figure 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 velocity 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. Recent evidence, according to Robinson [35] and others, suggests that head based coordinates are used.

Details on a model for the vergence system are sketchy, but Robinson [35] 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 4 for the Harvard head. The pan, tilt, and vergence motors are driven by a pulse width modulated MOSFET amplifier. The input to this amplifier is derived from the output of a Dynamation motor controller board [15]. The Dynamation board is indicated in figure 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 [13]. 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 processing 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.

4 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 colour 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. However, there is psychophysical evidence that humans will, under certain conditions, attend to a location in the middle of two salient features). Figure 5 shows this attention model.

It can be seen that this model incorporates many of the psychophysical results observed earlier. Adjusting the gains of a particular feature map will direct attentional resources to occurances 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. In the example of the red T in the field of green L's, attention will first be directed at the red T. As the gain decreases, attention will change locale. Since the red T is the only different feature and thus has a high saliency value, attention will go back to the T. The brief saccades to other areas of the visual field are found when subjects fixate on one target for a long time. 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 have chosen to express this model in the paradigm of the motion control language (MDL) described earlier. This paradigm allows a description of motion control of the head/eyes based on visual feedback which is "independent" of the underlying hardware or implementation. Using the MDL will allow a mechanism to control attention as a "high level language".

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 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, in the MDL paradigm, 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 second level are again of the form (u, k, T) where u is the desired position (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 [20,28,30,42,43] which indicate that the responses of neurons involved in visual perception are modulated by changes in the focus of attention.

Figure 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.

There are advantages in using the MDL description for the control of attention. The same description can be used with simple vision routines or with more complicated algorithms depending on the available hardware. The complexity of the feature maps used will determine what tasks can be performed. A large set of feature maps with maps at many scales detecting a large group of primitives will allow for sophisticated visual processing.

5 Experiments

Two experiments have been performed to demonstrate modal control of attention.

The first experiment involved tracking "blobs", regions of a range of intensity values. The features are black or white blobs against a neutral background. We used paper objects suspended on fish line. The task was to locate either the black or white feature and follow it. The objects were placed 0.5 to 2.0 meters from the head. The head was able to fixate on an object to within 2 pixels. The vision system for simple blob detecting tasks could process on the order of 5 frames per second. Taking into account the communication time between the vision system and the head control system, an overall rate of 3 frames per second could be achieved.

The second experiment was designed to demonstrate the attentive control system on a more complex scene. The features used are the 0^{th}, 1^{st} and 2^{nd} moments of each object and the intensity value. The scene is segmented into connected components, the various features are computed and the saliency map is built (as described in previous sections). Stereo correspondence is performed using the peak saliency values. As the task is to find the most salient feature with respect to the feature gains, k_i, the most salient points are the only ones that need to be considered in computing stereo correspondence. Since only a few points will be maximal (with well chosen gains), the correspondence problem is easily solved. With this done, the disparity values are computed and used to drive the head motors as described above. Using a combination of black and white, circular and rectangular objects, the attention system can successfully locate geometric shapes at different orientations and fixate on them. Altering the feature map gains, k_i, alters the direction of gaze to fixate on the object most salient with respect to the new gains. This experiment is much slower than simply distinguishing between a black and a white object. At present, depending on the complexity of the scene, the attentive system may between 1 to 10 seconds to fixate on the most salient object. The vision system is the culprit. This implementation uses a hybrid vision system employing both Suns and the Datacube and is not yet optimized. Future work to incorporate the entire vision system on the Datacube is already underway. Given that the vision system could work arbitrarily fast, the attentive control system is successful at tracking objects of interest.

6 Summary

We have described a control system for a binocular image acquisition mechanism which allows shifts in focus of attention to be made in a natural, device independent manner. The control method is based on the modal control technique proposed by Brockett [7]. 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.

The advantages of using active and attentive vision in a mobile robot application instead of "snapshot" vision are obvious; active vision algorithms can be more robust than static algorithms, and are often computationally efficient since irrelevant information is ignored. Furthermore, there are some tasks which are naturally suited to active vision, and for which conventional vision systems find very difficult to perform. An example is object recognition. The ability to obtain multiple views, and multiple views that are intelligently selected, helps enormously in performing model based object recognition. One of the drawbacks of active vision has been the requirement that real-time image processing operations are necessary to maintain real-time operation. However, recent advances in image processing hardware, exemplified by Datacube's [13] Maxvideo system, and the Pipe system [31] produced by Aspex, have made it possible for researchers to perform dynamic image processing operations at video rates on sequences of images obtained from video cameras, so there are few practical reasons why vision systems for mobile robots should not use active vision techniques. The control system we have described in this paper will extend the abilities of active vision systems in that it provides a method by which attentive behaviour can be conveniently obtained.

7 Acknowledgements

The Harvard Head project has been a collaboration involving many people. The mechanical component of the system was designed and constructed by J. Page, W. Labossier, and M. Cohn, with input from R. Brockett, J. Clark and E. Rak. M. Cohn should be singled out for the job he did in reverse engineering the Canon electronic focus controls. 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, V. Eng, M. Cohn, 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 introducing us to some of the physiological work concerning attention.

References

[1] Aloimonos, Y., Weiss, I., Bandyopadhyay, A., "Active vision", Proceedings of the 1st IEEE Conference on Computer Vision, London, 1987, pp35-54

[2] Alpern, M. "The position of the eyes during prism vergence", A.M.A. Arch. Opthalmol., Vol. 57, pp 345-353, 1957

[3] Ambler, B. and Finklea D., "The Influence of Selective Attention in Peripheral and Foveal Vision", Perception and Psychophysics, *19(6)*, 1976, pp. 518-524.

[4] Bajcsy, R., "Active perception vs. passive perception", Proceedings 3rd IEEE Workshop on Computer Vision, Bellaire, pp 55-59, 1985

[5] Bajcsy, R., "Perception with feedback", in the Proceedings of the 1988 Darpa Image Understanding Workshop, pp 279-288

[6] Beck, J. and Ambler B., "The Effects of Concentrated and Distributed Attention on Peripheral Acuity", Perception and Psychophysics, *14*, 1973, pp. 225-230.

[7] Brockett, R.W., "On the computer control of movement", Proceedings of the 1988 IEEE Robotics and Automation Conference, Philadelphia

[8] Brown, C.M., "Progress in image understanding at the University of Rochester", in the Proceedings of the 1988 Darpa Image Understanding Workshop, pp 73-77

[9] Burr, D., and Ross, J., "Visual Processing of Motion", Trends in Neuroscience, 1986.

[10] Burt, P., "Algorithms and architectures for smart sensing", in the Proceedings of the 1988 Darpa Image Understanding Workshop, pp 139-153

[11] Collewijn, H. and Tamminga, E.P., "Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds", J. Physiology, Vol. 384, pp 217-250, 1984

[12] Dallos, P.J., and Jones, R.W., "Learning behaviour of the eye fixation control system", IEEE Transactions on Automatic Control, Vol. 8, pp 218-227, 1963

[13] Maxvideo System Documentation, Datacube Inc., Peabody, MA

[14] Ditchburn, R.W., **Eye Movements and Visual Perception** Oxford: Clarendon Press, 1973

[15] Dynamation motor controller board documentation, Dynamation Inc., Mountain View, CA

[16] Enright, J.T., "Changes in vergence mediated by saccades", J. Physiol., Vol. 350, pp 9-31, 1984

[17] Eriksen, C. W. and St. James J., "Visual Attention within and around the field of focal attention: A zoom lens model", *Perception and Psychophysics*, Vol. 40 (4), 1986, pp. 225-240.

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

[19] Geiger, D., and Yuille, A., "Stereopsis and eye movement", Proceedings of the 1st IEEE Conference on Computer Vision, London, 1987, pp306-314

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

[21] Hurlbert, A., and Poggio, T., "Do Computers Need Attention?" *Nature*, 321, 12, 1986.

[22] Hwang, T.L., Clark, J.J., and Yuille, A.L., "A depth recovery algorithm using defocus information", in preparation, 1988

[23] Hughes, H.C., and Zimba, L.D., "Spatial maps of directed visual attention," Journal of Experimental Psychology: Human Perception and Performance 11, pp 409-430, 1985

[24] Koch, C., and Ullman, S., "Selecting one among the many: A simple network implementing shifts in visual attention", MIT AI Memo No. 770, January, 1984

[25] Krotkov, E., "Focussing", Int. J. of Computer Vision, Vol. 1, No. 3, 1987

[26] Krotkov, E., Fuma, F., and Summers, J, "An agile stereo camera system for flexible image acquisition", IEEE Journal of Robotics and Automation, Vol. 4, No. 1, pp 108-113, 1988

[27] Marr, D, and Ullman, S., "Directional selectivity and its use in early visual processing", Proc. R. Soc. Lond. B, 211, 1981, pp. 151-180.

[28] Moran, J., and Desimone, R., "Selective attention gates visual processing in the extrastriate cortex," SCIENCE 229, pp 782-784, 1985

[29] Pentland, A., "A new sense for depth of field", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9, No. 4, pp 523-531, 1987

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

[31] Pipe system documentation, Aspex Inc.,

[32] Poggio, T., et al, "The MIT vision machine", in the Proceedings of the 1988 Darpa Image Understanding Workshop, pp 177-198

[33] Raibert, M.H., and Craig, J.J., "Hybrid position/force control of manipulators", in **Robot Motion: Planning and Control**, Brady, M. et al editors, MIT Press, 1982

[34] Robinson, D.A., "The mechanics of human smooth pursuit eye movement", J. Physiology (London), Vol/ 180, pp 569-591, 1965

[35] Robinson, D.A., "The oculomotor control system: A review", Proceedings of the IEEE, Vol. 56, pp 1032-1049, 1968

[36] Robinson, D.A., "Why visuomotor systems don't like negative feedback and how they avoid it", in **Vision, Brain and Cooperative Computation**, Arbib, M. and Hanson, A. eds., MIT Press, Cambridge, MA, 1987

[37] Sagi, D., and Julesz, B., "Enhanced detection in the Aperture of Focal Attention during Simple Discrimination Tasks", Nature, Vol. 321, No. 12, June, 1986, pp. 693-695.

[38] G.L. Shulman, R.W. Remington and J.P. McLean , "Moving Attention Through Visual Space", Journal of Experimental Psychology: Human Perception and Performance, *5(3)*, 1979, pp. 522-526.

[39] Treisman, A. M. and Gelade,G., "A Feature-Integration Theory of Attention", *Cognitive Psychology*, Vol. 12, 1980, pp. 97-136.

[40] Ullman, S., "Visual routines", Cognition, Vol. 18, pp 97-159, 1984

[41] Wilson, H. R., "A Four Mechanism Model for Threshold Spatial Vision", *Vision Research*, 19, pp.19-32.

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

[43] Wurtz, R.H., Richmond, B.J., and Newsome, W.T., "Modulation of cortical visual processing by attention, perception, and movement," in Dynamic Aspects of Neocortical Function (New York: Wiley and Sons), 195-217, 1984

[44] Yarbus, A.L., "Eye movements during changes of the stationary points of fixation", Biophysics Vol. 2, pp 679-683, 1957

[45] Young, L.R., and Stark, L., "Variable feedback experiments testing a sampled data model for eye tracking movements", IEEE Transactions Human Factors in Electronics, Vol. 4, pp 38-51, 1963

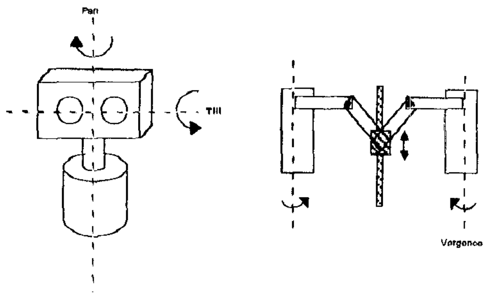

Figure 1: The mechanical structure of the Harvard head illustrating the positional degrees of freedom.

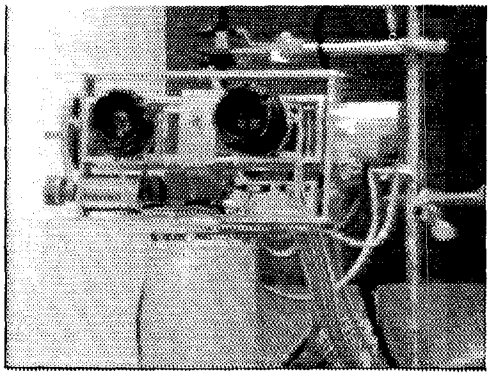

Figure 2: A photograph of the Harvard Head.

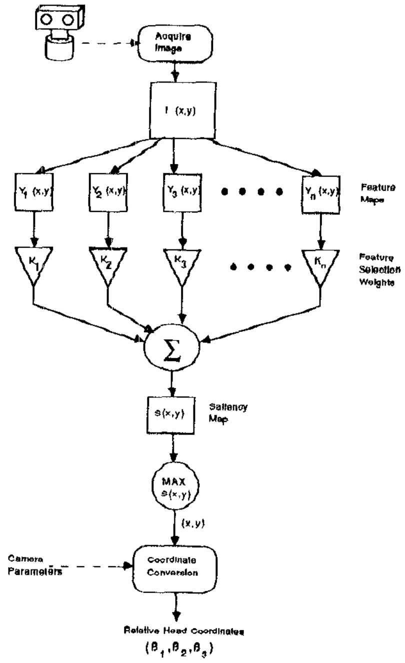

Figure 5: The feedback selection model of attention.

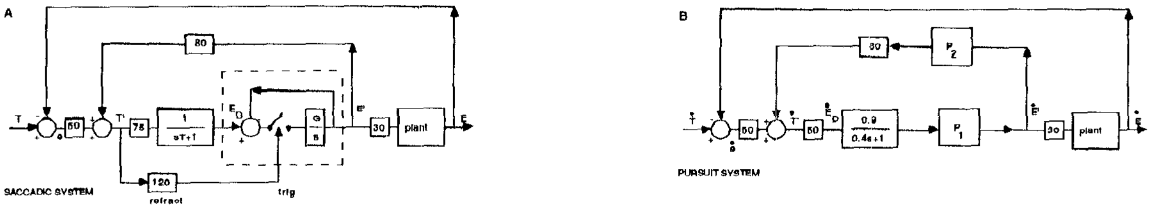

Figure 3: Robinson's model for the human oculomotor system. a)Saccadic system. b)Pursuit system.

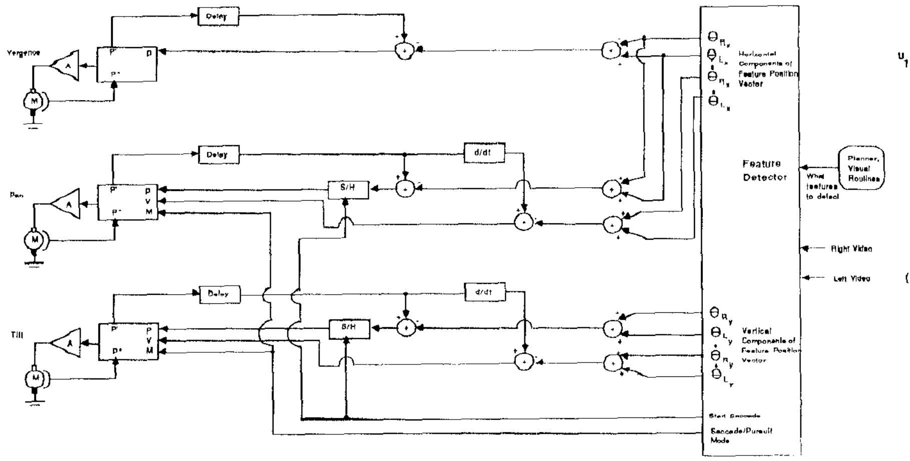

Figure 4: The control system used in the Harvard head system.

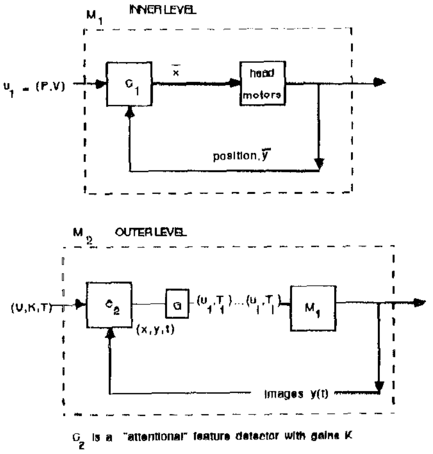

Figure 6: The two levels of the attentive control system.