NISTIR 4826

Real-time Smooth Pursuit Tracking for a Moving Binocular Robot

David Coombs

Sensory Intelligent Group

Christoper Brown

University of Rochester
Computer Science Department
Rochester, NY 14627

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Robot Systems Division
Bldg. 220 Rm. B124
Gaithersburg, MD 20899

NIST

NISTIR 4826

Real-time Smooth Pursuit Tracking for a Moving Binocular Robot

David Coombs

Sensory Intelligent Group

Christoper Brown

University of Rochester
Computer Science Department
Rochester, NY 14627

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Robot Systems Division
Bldg. 220 Rm. B124
Gaithersburg, MD 20899

April 1992

U.S. DEPARTMENT OF COMMERCE
Barbara Hackman Franklin, Secretary

TECHNOLOGY ADMINISTRATION
Robert M. White, Under Secretary for Technology

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

Real-time Smooth Pursuit Tracking for a Moving Binocular Robot

David Coombs
coombs@cme.nist.gov
Robot Systems Divsion
Manufacturing Engineering Laboratory

Christopher Brown
brown@cs.rochester.edu
University of Rochester
Computer Science Department
Rochester, NY 14627

National Institute of Standards and Technology
Technology Administration
U.S. Department of Commerce
Gaithersburg, Maryland 20899

NISTIR 4826

Abstract

This paper examines the problem of a moving robot tracking a moving object with its cameras, without requiring the ability to recognize the target to distinguish it from distracting surroundings. A novel aspect of the approach taken is the use of controlled camera movements to simplify the visual processing necessary to keep the cameras locked on the target. A gaze holding system implemented on a robot's binocular head demonstrates this approach. Even while the robot is moving, the cameras are able to track an object that rotates and moves in three dimensions.

The key observation is that visual fixation can help separate an object of interest from distracting surroundings. Camera vergence produces an horopter (surface of zero stereo disparity) in the scene. Binocular features with no disparity can be extracted with a simple filter, showing the object's location in the image. Similarly, an object that is being tracked will be imaged near the center of the field of view, so spatially-localized processing helps concentrate on the target. Instead of requiring a way to recognize the target, the system relies on active control of camera movements and binocular fixation segmentation.

This work was done largely while Coombs was affiliated with the University of Rochester. This material is based upon work supported in part by the National Science Foundation under Grants numbered IRI-8903582, CDA-8822724, and IRI-89220771, and by ONR/DARPA research contract number N000114-82-K-0193. Product endorsement disclaimer: references to specific brands, equipment, or trade names in this document are made to facilitate understanding and do not imply endorsement by the National Institute of Standards and Technology.

1 Introduction

Primate visual systems offer an existence proof for gaze holding capabilities, and models of these systems can offer hints for the design of robot visual systems. However, most models for holding gaze do not address the visual processing necessary to implement this function. For instance, it is generally assumed that full-field optical flow is the visual signal that is used to stabilize gaze under egomotion, but optical flow is homogeneous only for rotation of the eye about its optical center. Similarly, "retinal slip of the visual target" is the visual signal commonly assumed to drive smooth pursuit eye movements that follow a moving object. In this case, it is being assumed that *the target's* retinal slip has been distinguished from the retinal slip of the rest of the scene. However, the optical flow signal is complex in general, and it is not clear how to parse the optical flow to determine the retinal slip of the target object. In both cases, what is needed is a mechanism that distinguishes the retinal slip of the visual target from that of the rest of the scene. This problem has received little attention in the biological literature. However, it has been suggested that disparity filtering mechanisms may play a role in the interpretation of optical flow by gaze stabilization systems [Howard and Simpson, 1989, Miles *et al.*, 1991], and disparity filtering of a similar sort is used by the gaze holding system that is demonstrated.

The goal of *smooth pursuit* contrasts with computer vision's traditional *passive tracking* task. In passive tracking, the cameras move without regard to the goal of tracking the target object. For instance, the cameras on a mobile robot may point straight ahead like automobile headlights. The optical flow observed by the robot will result from the three dimensional structure of the scene and the robot's motion. Further, the target will move about in the cameras' images. In contrast, during active visual following, the cameras rotate to follow the target. Consequently, the target's retinal slip is minimal, and the flow of the surrounding scene is in the direction opposite the camera's rotation. In addition, the target's image is held near the center of the field of view.

In order to be as general as possible, the pursuit system should be able to follow a moving object without necessarily recognizing it first. A robot that must recognize an object to be able to follow it will only be able to use that facility in domains where every object is known to it and in which it has a means to recognize all objects. Such a robot's applicability will clearly be limited. Therefore, the system must use *precategorical* visual cues (*i.e.*, cues available prior to object recognition) in order to distinguish the visual target from distracting objects and the surrounding scene. Unfortunately, it is not clear how to extract this information from the visual signal.

Some visuomotor research has recently begun to appear in the computer vision literature, and various visual tracking and servoing systems have been reported. The systems perform saccadic tracking, smooth servoing on position or velocity, or combinations of both smooth pursuit and catch-up saccades. Similarly, approaches to visual processing include both optical flow processing and object identification and location.

The goal of this work has been to build a robot gaze holding system whose only knowledge of the target is essentially that the cameras are initially pointed at it. The situation is depicted in Figure 1(a). The gaze holding problem is to maintain fixation on a moving visual target from a moving platform. In order to do this, the errors in camera orientation must be determined, so the location of the target's image on the retina must be found. How can this be done without recognizing the target? The approach taken in this work exploits binocular cues and the fact that the cameras are actively following the target.

1.1 Control of Sensors and Simplified Sensing

As a consequence of gaze holding, the visual target is easier to pick out. Thus, it is easier to actively follow an object with moving cameras than to track its images in stereo images with static vergence and no control of camera movement [Coombs *et al.*, 1990]. For instance, during active following, motion blur de-emphasizes the background. Further, simple visual sensing techniques that are uniquely available during gaze holding can be used to segment the object being fixated, as illustrated in Figure 1(b). Foveal vision emphasizes the fixated object simply by spatially localized processing or enhanced resolution, and disparity filtering picks out features near the *horopter* (the set of points in the scene whose disparity is zero). The demonstration system locates the target by foveally filtering the features found by the zero-disparity filter (ZDF), effectively

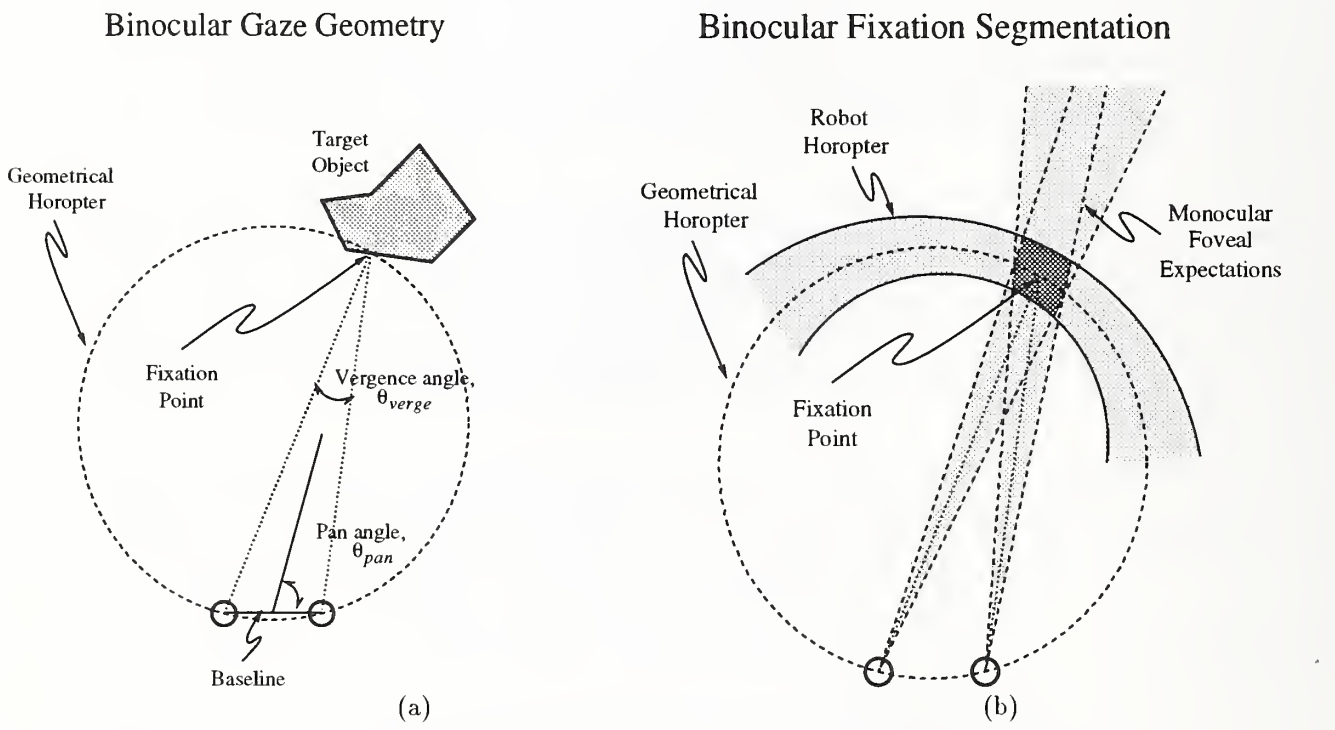


Figure 1: (a) Top view of binocular gaze geometry. The goal of gaze holding is to keep the eyes or cameras fixed on a common world point or visual target. The gaze vector, $\vec{\theta}$, consists of the gaze pan and tilt angles and vergence angle. In order to keep the world point fixated, the gaze holding system must generate gaze and vergence angles that keep the cameras directed toward the target. The result of fixating a target is that the object lies near the *horopter*, which is the set of world points whose disparity is zero. The stereo images of an object that lies near the horopter have a narrow range of disparities. (b) In this top view of binocular fixation, the lightly shaded areas are the regions of space that are highlighted by foveal and disparity filtering. The intersection of these areas is shaded darker, and it corresponds to the area around the fixation point in which objects can be segmented.

producing the intersection of the fovea and ZDF. The target's retinotopic location provides the error signals the gaze control system needs to control the gaze and vergence angles.

The gaze holding problem is comprised of pursuing the visual target with the cameras, and for binocular systems, verging the cameras on the object. We will call the direction of the cameras the gaze angle or direction. The pursuit system rotates the cameras in tandem to keep gaze directed toward the target. The vergence angle is the angle between the visual axes of the cameras. Vergence system control consists of rotating the cameras in opposite directions so the visual axes of the cameras intersect at the distance of the object of interest. The demonstration system exploits binocular cues and deliberate control of the cameras that enable precategorical segmentation of the fixation target. These ideas are embodied in the ZDF, which reveals only objects that lie at the fixation distance. The next three sections discuss the issues and related work in visual processing and motor control for robot pursuit systems, and the remaining sections describe the approach taken on the robot head.

2 The Pursuit Problem

The goal of smooth pursuit behavior is to keep the visual target's image steady and centered in the camera's image. Thus the system must be able to determine the target's retinotopic position and retinal slip.

First, consider taking a hint from primate pursuit models that might help us build a robot pursuit system. Krauzlis and Lisberger [1989] present a recent model of the smooth pursuit system. However, like most such models, their model does not address the extraction of the retinal slip of the the target. Even the treatment of [Lisberger *et al.*, 1987], which is entitled "Visual Motion Processing and Sensory-Motor Integration for Smooth Pursuit Eye Movements", goes no farther than arguing that retinal slip and retinal image acceleration must be signals to which the pursuit responds, based on analysis of the behavior of the system. The means by which the retinal slip of the target is distinguished from other optical flows experienced by the visual system is so heavily influenced by cognitive factors that no studies of biological visual systems have clearly illuminated this question. Therefore we must rely on the computer vision literature.

2.1 Parsing Optical Flow

One approach is to try to parse the optical flow to find the target's retinal location or slip. This is easy for passive tracking if the robot is stationary and the target is the only object in motion. Simply hold the cameras still and detect the moving object. This approach will not work for gaze holding, however, since the camera movements make the entire scene appear to move, and it will not work even for passive tracking if there are several moving objects or the surrounding scene appears to move because the robot is moving. It is necessary to distinguish the target or its optical flow from the rest of the scene and its flow. Consider that a robot moving in a 3D scene generates optical flow as a function of the distance of the objects as well as the robot's motion. How will the visual system isolate the target or its slip from the rest of the scene?

Optical flow parsing techniques in computer vision range from detection of moving objects to reconstruction of the scene. In a simple scene, Allen [1989] is able to use spatio-temporal motion energy to servo a camera on an object moving on a blank floor. In general, however, detecting moving objects against a more complicated background is computationally expensive [Heeger and Hager, 1988]. Nonetheless, Nelson [1991] has shown that object motions that result in flows that are inconsistent with the flow that arises from egomotion can be detected in real time. For instance, if the robot is rotating to the left, all optical flow due to the robot's egomotion is constrained to move to the right. Any optical flow inconsistent with this (*e.g.*, up and down) indicates a moving object. Thus it is possible to detect moving objects under some rather constrained camera movements, but the optical flow that arises during visual following of a moving object is quite complex, and it is not yet known how to parse the optical flow signal efficiently.

The optical flow patterns that arise from camera movement and egomotion in a three-dimensional scene can be quite complex. For instance, consider the simple case of a robot translating linearly along a smooth ground plane with the cameras pointed straight ahead. The optical flow bursts smoothly outward from the horizon in a radial pattern. Now if the observer simply fixates a point on the ground plane to the right of

the direction of travel, the instantaneous camera rotation adds a uniform flow field to the optical flow due to the observer's translation. The optical flow that results from adding these two flow fields expands from the point of fixation on the ground plane and the flow vectors swirl outward in a counterclockwise pattern. A robot moving through a scene with less uniform structure would observe distortions of this optical flow due to the varying depths of objects in the scene. In general, the optical flow signal is complex, and it is not easy to distinguish the retinal slip of the target from that of the surrounding scene.

Both Burt *et al.* [1989] and Heeger and Simoncelli [1989] describe techniques to estimate egomotion by iteratively refining the egomotion estimate from image motion. Given the complexity of the flow signal, it is hardly surprising that these techniques are computationally expensive and do not converge quickly. On the other hand, Heeger and Jepson [1990, 1991] present a parallelizable, non-iterative method for computing three-dimensional motion and depth from optical flow signals for a rigid static scene. While this work represents an important step, the method has not yet been shown to extend robustly to correct interpretation of non-rigid scenes, and it depends on the ability to extract optical flow reliably.

Without explicitly estimating egomotion, it is possible to segment images based on independent coherent motion under roughly the same conditions that allow binocular stereo segmentation (*i.e.*, that the change of the view of each rigid component of the scene can be approximated locally by an image shift). A method that might be considered "generalized difference of images" is presented in [Shvaytser, 1988]. Interestingly, the operator is not only given with a frequency domain interpretation, but also it is derived in a probabilistic formulation. Such approaches either require many fast local operations or they rely on fast implementations of Fourier transforms on many local image patches.

Woodfill and Zabih [1991] have achieved real-time motion and stereo segmentation of objects on a Connection machine. However, there is currently no real-time image capture mechanism available for such machines. Their method relies on the object, whose initial segmentation is given, not to change appearance much between frames. Image patches that belong to the target are correlated with nearby patches in stereo and motion disparity spaces to maintain the segmentation of the object. The object is not required to be rigid, but the appearance of the object must not change too drastically between sample frames. Using both stereo and motion provides redundancy and more robust segmentation, since each modality can sometimes fill in when the other lacks reliable information. Here again, the required correlation of many image patches is computationally expensive.

2.2 Matched Filters

Another approach to locating the target is to use matched filters. Several variants of this technique have been employed. Clark and Ferrier [1988] suggest using conjunctions of features in "saliency" images to locate an object of known properties. Thus, it is assumed that choosing an appropriate set of features in saliency images will suffice to make the target "pop out". However, the phenomenon of "pop-out" occurs in humans only for a limited number of features (such as color, and motion) and only a few conjunctions pop out [Triesman, 1985]. Similarly, computational attempts to dynamically conjoin arbitrary sets of features have met with limited success. Corke and Paul [1989] locate binary thresholded objects whose moments are known *a priori*. These features might be initialized by taking the conjunction of the features in an image window that is known to contain the target image. However, computing moments of thresholded grayscale images is likely to be brittle to changes in lighting, point of view, etc. Papanikolopoulos *et al.* [1991] tracks a set of correlation features on a rigid object, but it is difficult to automatically select features that will provide robust correlation results [Thorpe, 1983].

Swain and Ballard [1990] introduce a fast scheme for recognizing and locating multi-colored objects. However, since the representation is not view-invariant, several "characteristic views" of each object are used. This provides a relatively fast method for reliably locating an identified object, but it would not follow a novel object.

A general problem with matched filters is that this approach does not work if the object rotates. The difficulty is that a matched filter is view-dependent. The problem can be illustrated by considering a simple form of matched filtering. A simple strategy is to use the last subimage believed to contain the target image as a correlation template to locate the target in the next image. Unfortunately, this technique relies on a

view-dependent model of the target (*i.e.*, its appearance from the last viewpoint of the observer) so it will fail if the appearance of the target image changes considerably (*e.g.*, due to rotation of the target object). This problem can be overcome by updating the correlation template with the new image of the target. However, without some other means of locking onto the target, this procedure is prone to drift off the target onto the background.

2.3 Fixation Segmentation

Considering that gaze is to be actively held on the target, it is possible to exploit expectations of the target's retinal location and disparity. Miles *et al.* [1991] suggest that both peripheral and foveal optical-flows contribute to gaze stabilization in monkeys. The foveal flow is attributed to gaze target, and the peripheral flows are from other depths. Howard and Simpson [1989] have found that the gain of optokinetic nystagmus (OKN) in humans is inversely proportional to binocular disparity. They argue that the activity of cells responsive to direction of motion and zero disparity selectively augments OKN, thus enabling humans to stabilize images in the plane of regard without interference from competing motion signals arising from other distances. These observations suggest mechanisms that take different approaches to extracting useful information from the heterogeneous flow field present under head translation. The differences in flow velocities are due to motion parallax caused by the presence of objects at multiple depths. The Howard and Simpson observations suggest a mechanism similar to the zero disparity segmentation described here that ignores flow signals from depths other than that of gaze fixation.

3 Related Work

A few attempts have been made to pursue moving objects based on optical flow and most of them have used simple approaches to motion segmentation. *E.g.*, Lee and Wohn [1988] use stop-and-look (which they call "static look and move (SLAM)") tracking in which the target is expected to be the only moving object. Allen [1989] uses spatio-temporal motion energy to servo a robot arm-mounted camera on a target moving in front of a blank background. Tölg [1991] simulates the traditional pursuit model [Young, 1971], with internal positive feedback and catch-up saccades. The motion-based target segmentation assumes the object is moving coherently (and therefore it must appear essentially flat). Jenkin [1991] uses stereomotion channels to estimate the target's motion. However, it is not clear how the trajectory detectors distinguish the motion of the target from the motion of the surrounding scene.

The other approaches assume the object is preselected and rely on view-dependent features and therefore suffer from the shortcomings of matched filters. Corke and Paul [1989] smoothly follow the centroid of a blob that is translating in a fronto-parallel plane, by translating the camera in a fronto-parallel plane. The moments (which are known *a priori*) are computed on a binary image obtained by thresholding the greyscale input image. This approach is likely to be brittle to changes in lighting, point of view, etc. Papanikolopoulos *et al.* [1991] smoothly tracks pre-selected features (image patches). It is not known how to select such features automatically, although this problem has been and will doubtlessly continue to be a subject of investigation [Matthies *et al.*, 1989, Thorpe, 1983]. Waxman *et al.* [1988] saccadically track a set of features whose spatial relationship is pre-selected, and the object is therefore identifiable.

Clark and Ferrier [1988] present a rare binocular tracking system that saccadically tracks the appropriate conjunction of features that locate an object whose properties are known *a priori*. Their system has demonstrated saccades and position-servoed vergence to binocularly fixate a target as it moves in three dimensions. The left and right images are processed independently, and the location of maximum "salience" (the desired combination of features determined to be relevant) in each image is taken to belong to the target object. However, neither is there a guarantee that the feature values will be invariant to point of view, nor are the maximum salience values necessarily unique. Consequently, the correspondence problem may not be trivial enough to yield to salience images.

4 Strategies for Pursuit Control

There are two natural and obvious measures of target following performance: position error and velocity mismatch. Restated, a pursuit system could attempt to center the target image, and at the opposite extreme, the system could try to stabilize the target's image on the retina (reducing slippage) by matching the target's velocity without regard to the retinal position of the target's image.

Note that the goals of image-centering and slip-minimizing can conflict since smooth camera movements can only improve one of these measures at the expense of the other. The target's image is repositioned by slipping it across the sensor array, and as the target accelerates, the image is stabilized by sacrificing the position slippage that has already occurred to prevent slip since the target would have to be slipped back to its desired position.

One approach to this problem is give precedence to one of these goals. Thus one simple pursuit system could attempt to keep the target image centered without regard to the image slip required. Another simple system could try to minimize target slip. Both of these behaviors can be elicited from monkeys [Lisberger, 1990]. However, it is generally believed that the primate pursuit behavior consists of a combination of both smooth servo control that matches velocity to minimize slip and catch-up saccades that recenter the target image when it deviates too far from the fovea. A more realistic model includes a small position-error response as well as the velocity-matching response in the smooth component of the system.

This is a clever solution to the dilemma of how to minimize both velocity and position error simultaneously. As previously noted smooth movements alone cannot achieve both goals, and certainly saccadic movements cannot reduce motion blur since they don't match velocity. The catch-up saccades perform the important function of correcting accumulated position error while introducing minimal target slip, and target slip is minimized by the smooth component that matches the target velocity with the camera velocity.

Unfortunately, implementing saccades in a binocular system requires careful attention, especially to the vergence angle. The various complications involved in coordinating saccades have led us to employ the simple strategy of using only smooth camera movements in the demonstration system.

One of the most challenging problems faced by visuomotor control systems is coping with the delays in the system. Delays can cause a system to be unstable. For instance, if a feedback system can respond more quickly than it can measure the error, it may respond to old error signals and over-react. Consider driving a car on a slippery road. The driver turns the wheel, but the car does not immediately change its course, so the driver turns the wheel further and consequently over-steers the car. Since cameras are quickly maneuverable and visual processing is slow, visuomotor systems face a similar control problem due especially to the delays of visual processing. There are a couple of obvious ways to prevent overshooting responses. One approach is to model the delays present in the system and anticipate the response of the system with an internal model of the the visuomotor system. Some simple predictive control is described here, and some simulation investigations of predictive gaze control are described in [McDonald *et al.*, 1983, Brown, 1990a, Brown and Coombs, 1991, Coombs, 1992]. Another, simpler approach prevents overshoots simply by reducing the responsiveness of the system enough to make it stable. A combination of these methods is used in the demonstration system.

5 Pursuit on the Robot

A demonstration pursuit system runs in real-time on a robot head. The system is diagrammed in Figure 2(a). The system has two components, vergence and pursuit, that require different visual processing. However, both components use foveally-processed visual signals. The vergence and pursuit systems perform complementary functions. The pursuit system keeps the foveas centered on the fixated object, and the vergence system keeps the cameras converged on the target. The vergence system, previously reported in [Olson and Coombs, 1991], uses an estimate of the binocular disparity between the foveal images to measure the vergence error. However, the disparity estimator provides only *what* disparity is present in the images; it does not inform the system *where* in the images the disparity arises. Conversely the ZDF does not measure disparity, but rather locates portions of the images that have zero stereo disparity. In this way the vergence

system minimizes disparity, but it cannot ensure that the target is foveated. Similarly, the pursuit system foveates on the target, but it requires that the cameras be properly verged on the object in order to locate it. The vergence and pursuit systems generate camera vergence and pan and tilt velocity commands, but the motors are not configured with those degrees of freedom mechanically. However, they are linearly related and the motor controller performs the conversion from camera coordinates to motor coordinates.

The use of binocular cues and control of the camera motions enables a simple signal processing and servo system to achieve gaze holding precategorically. *I.e.*, the system is able to hold gaze on the fixation target without the ability to recognize the object, and the target is distinguished simply because it is fixated by the robot's gaze.

As a consequence of gaze holding, the visual target is easier to pick out. Thus, it is easier to actively follow an object with moving cameras than to track its images in stereo images with static vergence and no control of camera movement. For instance, during active following, motion blur de-emphasizes the background. Further, simple visual sensing techniques that are uniquely available during gaze holding can be used to segment the object being fixated, as illustrated in Figure 1. Foveal vision emphasizes the fixated object simply by spatially localized processing or increased resolution, and disparity filtering picks out features near the *horopter* (the surface in the scene whose disparity is zero). The demonstration system locates the target by foveally filtering the features found by the ZDF, effectively producing the intersection of the fovea and ZDF. The target's retinotopic location provides the error signals the gaze control system needs to control the gaze and vergence angles.

Taking advantage of the vergence system, the pursuit system locates the target's retinal azimuth and elevation by the location of features with zero stereo disparity. The vergence system controls the vergence angle of the cameras to minimize the disparity of the foveated object. Thus the vergence system relies on the pursuit system to keep the target foveated. This symbiotic cooperation of the pursuit and vergence system enables precategorical visual processing to suffice to support gaze holding.

In the demonstration system, the cameras smoothly pursue the target by servoing on its position to minimize the retinal azimuth and elevation pointing errors of the cameras.

We have used both PID control and P control with $\alpha - \beta - \gamma$ prediction. Initial experiments employed PID control for simplicity and robustness, and later attempts were made to reduce the phase lag by the use of prediction to overcome the latency of visual processing. Both of these methods produce fairly smooth following behavior. PID control is a simple controller that offers some flexibility by responding to the integral and derivative of the error as well as the error signal itself. However, the $\alpha - \beta - \gamma$ filter is used to explicitly apply a simple linear model to the error signal in order to predict its future values. If the error signal can be successfully predicted, the latency of processing the error signal can be mitigated by controlling the system with predicted error signals and comparing the observed error with the predicted error once the visual signal is processed. The use of these control methods in the demonstration system is discussed in Section 7.

6 Visual Processing for Pursuit

The pursuit system is started up once the cameras have initially acquired fixation on the target with a saccade and matched its estimated velocity. Thus it is fair for the pursuit system to assume upon initiation that the target is roughly centered and the target velocity is approximately known. In addition, the pursuit system can assume that the target image's size is given, since the visual processing to acquire the target can be assumed to have coarsely segmented the target (*e.g.*, by motion).

If the target can be located in the image reliably, its retinal slip can be estimated. The converse is not necessarily true, since knowing the target's retinal slip does not directly lead to the target's retinal location. However, it may be possible to find the target's location by segmenting the image based on areas of uniform flow if the target's slip is unique enough. For a target moving fast enough, this may be true since the camera rotation to follow the target will give rise to opposite optical flow of a stationary scene, while minimizing the target's flow.

However, the demonstration does not segment the target by its motion. Instead, the target is distinguished by its stereo disparity. Although the idea is similar, the zero-*stereo*-disparity can be implemented simply and

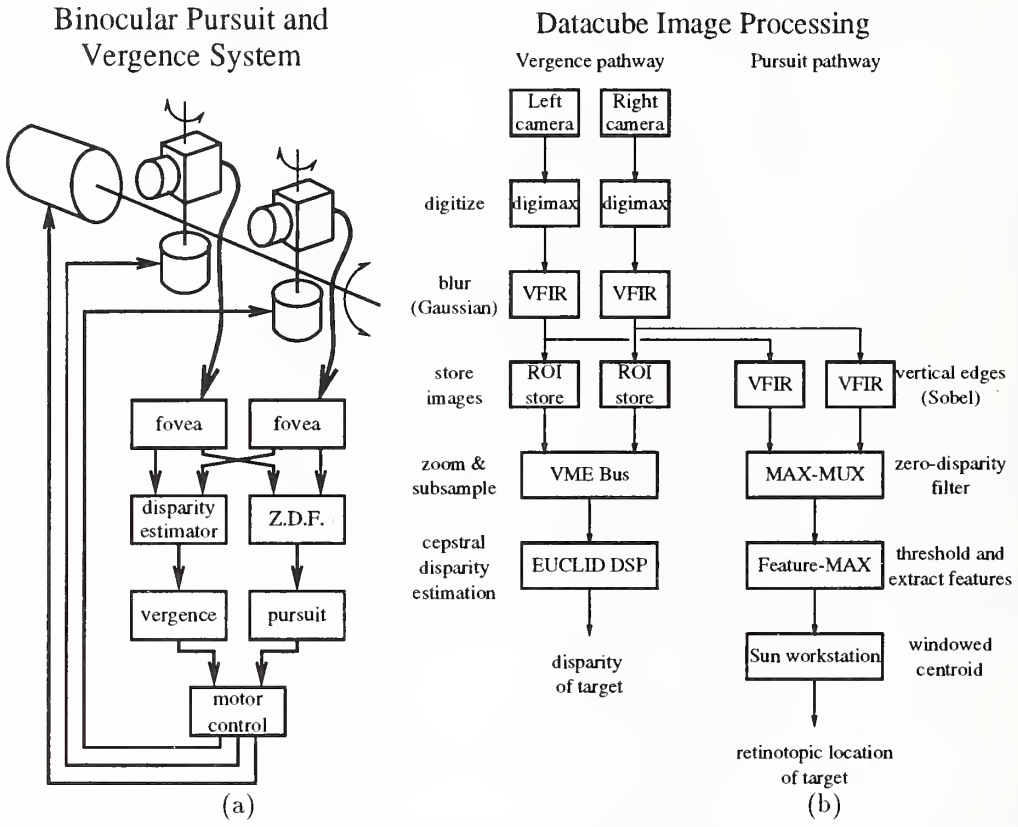


Figure 2: (a) Binocular Pursuit system: The visual processing consists of vergence and pursuit branches. The visual signal is foveally processed before use in both subsystems. The vergence system estimates the disparity of the foveal images and controls the vergence angle to minimize the disparity. The pursuit system locates the retinal image of the object that is in both the fovea and the horopter, and the cameras are panned and tilted to keep the object located centrally in the fovea. The motor controller combines the pan, tilt, and vergence velocity commands and determines by simple linear relations the required left and right pan and tilt motor velocities. (b) Datacube Image Processing: The gaze holding system does nearly all of the visual processing on a Datacube MaxVideo image processing system. There is a shared pathway in early visual processing, and two later branches for the vergence and pursuit systems. First, stereo images are digitized from synchronized cameras, and the images are blurred by convolution with a Gaussian kernel ($\sigma = 2.5$pixels). The vergence pathway begins with "zooming" and subsampling the images, and continues by using the cepstral filter [Yeshurun and Schwartz, 1989] to estimate the disparity of the "foveal" images. The pursuit pathway detects features (vertical edges) for disparity comparison, locates features with near-zero disparity and computes the centroid of these features that lie inside the "foveal" region.

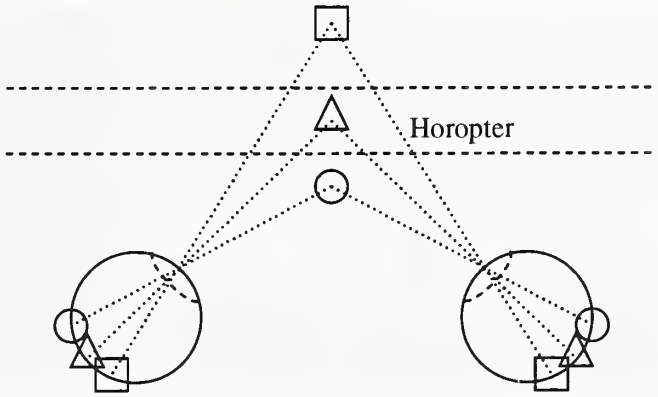


Figure 3: The *horopter* is operationally defined to be the region of space that contains objects that have no stereo disparity. It is a thin shell located at the fixation distance (the distance at which the cameras are verged). This figure illustrates the principle. The images of the triangle project to the same location in both retinae, whereas the images of the square, which lies beyond the horopter, have negative stereo disparity. Similarly, the circle, which is nearer than the horopter, results in stereo images with positive disparity.

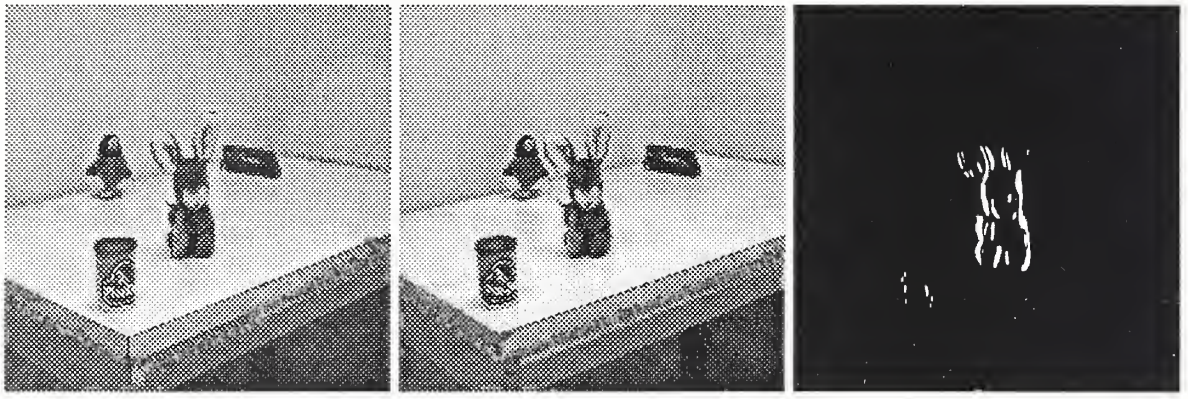
robustly even when image flows are complex and difficult to parse (assuming, of course, that the vergence system is able to keep the cameras verged on the target).

6.1 Disparity Filtering to Locate Objects

Pursuit uses vergence to isolate the target by disparity filtering. Features that have no stereo disparity can be detected in real-time using a disparity filter. When the cameras converge on an object, it projects an image onto the "retina" (CCD array) of each camera. Figure 3 depicts a scene of three objects at different depths with the cameras verged on the intermediate object. Each of the objects projects an image on each retina. However, only the middle object projects to the same locations on both retinae. The region of space that contains objects that project onto the retinae with no stereo disparity is call the *horopter*, and a simple filter can detect objects that lie in the horopter.

Disparity filtering is used to isolate the target from the background. A disparity filter can detect features that have no stereo disparity more easily than interpreting the stereo disparity of the images. A real-time nonlinear filter implements zero-disparity filtering to isolate the objects in the horopter. Figure 4 shows an example of this sort of filtering. The pursuit system relies on the vergence system to keep the disparity of the target within the range of the disparity filter. The vergence system does this by keeping the horopter on the target object by changing the vergence angle of the cameras to follow the target. With the target in the horopter, the disparity filter provides the retinal location of the target. On the assumption that gaze will normally be directed toward objects of interest, it may be appropriate for binocular agents to ignore features at large disparities. That is, disparity may be used to filter objects that are not currently of interest out of the scene.

The zero-disparity filter is a nonlinear filter that suppresses features that have non-zero stereo disparity. The features it uses are vertical edges, since they are identifiable features that can give useful information about horizontal disparity. (Clearly, horizontal edges provide no helpful information about horizontal disparity, since long horizontal edges can match over much of their length even with substantial disparity. Only their *ends* can be compared to find horizontal disparity.) The first step is to construct a vertical edge image of each image in the stereo pair. Then these images are compared in corresponding locations. If an edge is present in both images, then a feature appears in the resulting zero-disparity image. Of course the edges must be of like phase. Thus the filter detects features that have no stereo disparity.



(a) (b) (c)

Figure 4: Disparity Filtering of the scene shown in stereo images (a) and (b). Image (c) was produced by a real-time zero-disparity filter. The stereo images were first processed with vertical Sobel edge operators, and the stereo edge images were combined by a pixel-wise multiplicative 'AND' operator to produce the zero-disparity image. The effect of the filter is to suppress edges that have non-zero disparity, leaving an edge image that is dominated by objects in the horopter.

6.2 Datacube Image Processing

The visual processing for the gaze holding system is implemented in real time almost entirely on a Datacube MaxVideo^{TM} image processing system (Figure 2(b)). The processing begins by digitizing a stereo pair of images from the robot head's cameras. The cameras are synchronized so the images are taken simultaneously.

The binocular gaze holding system includes the vergence system as well as the pathway that supports pursuit. The stereo images are convolved with anti-aliasing filters (*e.g.*, Gaussian, $\sigma = 2.5$ pixels for 8-fold reduction in resolution) before being stored in frame buffer memory.[1] The EUCLID digital signal processing microcomputer included in the MaxVideo image processing system estimates disparity using the cepstral filter [Yeshurun and Schwartz, 1989]. The EUCLID computer is based on the ADSP-2100 digital signal processor [Analog Devices, 1987], which is optimized for operations such as convolution, finite impulse response filtering and Fast Fourier Transforms. The EUCLID board subsamples the stored images and performs cepstral filtering on the windowed, subsampled images. EUCLID locates the peaks in the disparity image and reports the disparity to the Sun^{TM} host. Thus the vergence error is measured.

Like vergence, visual processing for pursuit begins with the images blurred by convolution with a Gaussian kernel ($\sigma = 2.5$ pixels). This reduces the amount of aliased matches by removing some of the high frequency features from the images. Edge operators are convolved with the blurred images to produce a stereo pair of vertical edge images. These images are then disparity-filtered. The disparity filter is implemented as a nonlinear function in a lookup table on a multiplexer board. The function compares the edge energy pixel-wise to determine whether there is a zero-disparity feature, which is indicated by matching edge features with no stereo disparity. The FeatureMax board records the locations of all the features in the zero-disparity image. The Sun host then computes the centroid of the features in the central window of the image. This provides the pursuit system with the retinal location (and therefore the pursuit error) of the target.

[1] For only slight reductions in resolution, blurring is not necessary and is therefore omitted. Ideally, the amount of blurring is proportional to the resolution reduction, but due to a shortage of convolution boards, vergence either shares the pursuit blurring or omits blurring entirely when used together with the pursuit system.

7 Gaze Holding Control

The goal of the binocular pursuit system is to generate smooth camera movements that correct both gaze angle and vergence errors. The binocular pursuit control loop consists of three stages: digitization, error estimation (both gaze angle and vergence), and error correction (of both gaze and vergence angles). The timing of the system is sketched in Figure 5. Digitization is done under control of the Sun host using the MaxVideo digitizers, convolvers and frame stores (one each per camera). In addition, the zero-disparity image is thresholded by the FeatureMax and the list of above-threshold features is stored in its memory. It takes between one and two RS-170 frame times (33 to 67 milliseconds), depending on how much time remains in the current video frame when the command to acquire the next frame is issued. The Sun is free to do other things during digitization. Once the images are available in the frame store, the Sun signals EUCLID to extract the images from the frame buffers and estimate the disparity. This process takes approximately 59 milliseconds, after which EUCLID places the disparity estimate in a known location in shared memory and issues an interrupt to signal completion. The Sun converts the pixel disparity to angular coordinates by multiplying it by an empirically determined constant. While EUCLID is estimating the disparity, the centroid of the Sun computes the centroid of the zero-disparity features. The length of time required depends on the number of zero-disparity features. Once the gaze and vergence errors are estimated, the Sun applies the control law to issue the appropriate velocity commands to the camera motors. The Sun issues the motor commands *after* initiating the next digitization in order to allow digitization to proceed concurrently with motor control. This causes a slight delay in issuing the motor commands, but permits a higher overall sampling rate. The loop commonly takes 150 ms (four and one-half frame times) to complete. Thus, the system achieves a servo rate of 7.5 Hz.

7.1 The Controller

The pursuit system uses a PID controller for each of pan, vergence and tilt of the robot's gaze, as shown in Figure 6. The vergence system's error is found in the binocular disparity of the foveal images. The error signals for the pursuit controllers are the horizontal and vertical displacements of the target from the center of the fovea.

The gaze parameters, $\vec{\theta}$, map fairly directly, though not identically onto the mechanical degrees of freedom, $\vec{\phi}$, of the robot head. There is a single tilt motor, so

$$\theta_{tilt} = \phi_{tilt}.$$

The situation for pan and verge angles is sketched in "top-view" in Figure 7. The pan and verge angles are related to the left and right pan camera angles by

$$\begin{aligned} \theta_{pan} &= \frac{1}{2}(\phi_{right} + \phi_{left}) \\ \theta_{verge} &= \phi_{right} - \phi_{left}. \end{aligned}$$

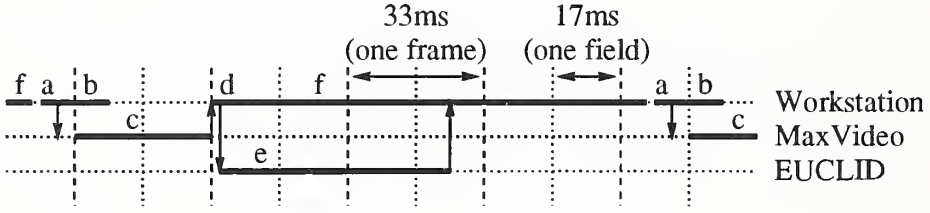
These equations relate the static angles. However, the motor controller must convert the pan and verge velocity commands to left and right pan motor velocities. Differentiating with respect to time, we obtain the motor velocities:

$$\begin{aligned} \dot{\phi}_r &= \dot{\theta}_p + \frac{1}{2}\dot{\theta}_v \\ \dot{\phi}_l &= \dot{\phi}_r - \dot{\theta}_v \\ &= \dot{\theta}_p - \frac{1}{2}\dot{\theta}_v. \end{aligned}$$

As one might expect, the pan velocity is transmitted to both camera pans, and the vergence is split evenly between them.

Each of these three gaze control systems operates independently, with no explicit cooperation. This simplifies the control laws. However, the gaze controls must be integrated to generate motor commands.

Pursuit and Vergence Timing Chart



a --- Set up frame buffers to capture images
Set up feature extractor to collect ZDF pixels

b --- Read motor angles
(Update gaze-vector display overlay)
Estimate retinotopic target location
Issue motor commands
(Update crosshair display overlays)
(Issue stimulus motor commands)

c --- Frame buffers capture images
Feature Extrator collects ZDF pixels

d --- Fork cepstral disparity estimator on EUCLID

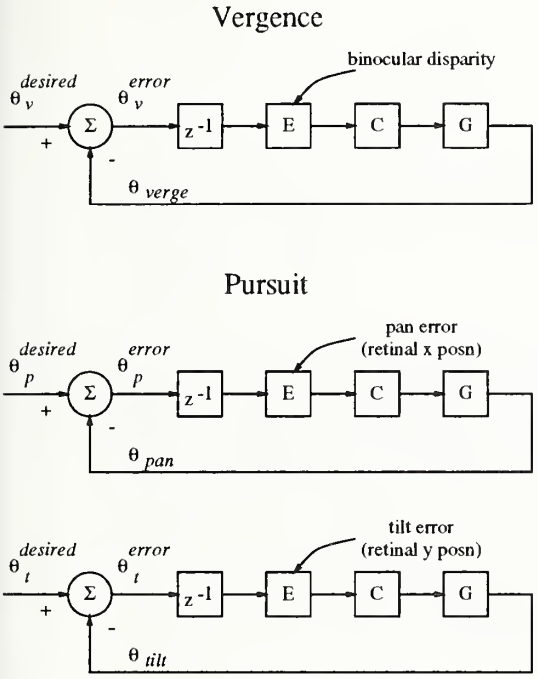
e --- EUCLID grabs subsampled images and estimates disparity

f --- Collect ZDF pixels and calculate centroid

NB: times are approximate, for illustrative purposes.

Figure 5: Binocular Pursuit Loop Timing Diagram.

Gaze Holding Controls



PID Controlling an Integrator in a Negative Feedback Loop

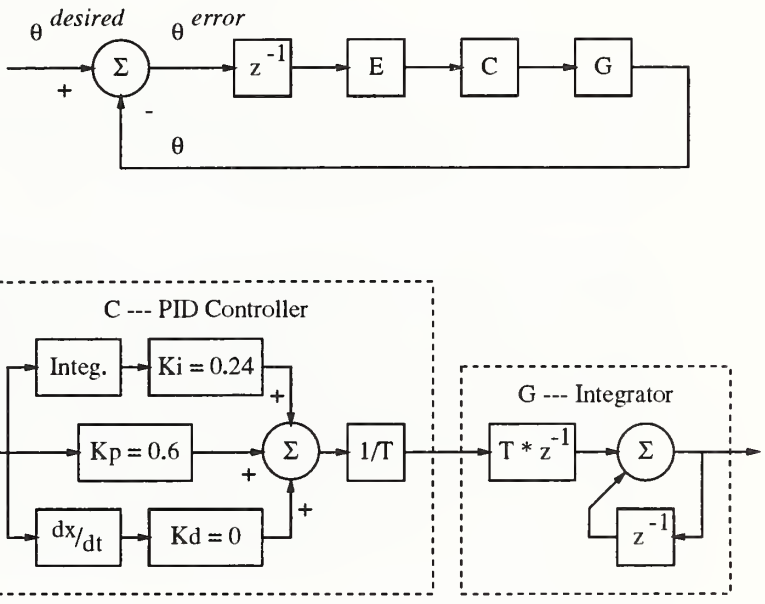


Figure 6: Gaze Holding Control Systems: The vergence and pursuit controls are driven by three independent feedback controllers, one each for pan, tilt, and vergence (a). The vergence system's error is found in the binocular disparity of the foveal images. The error signals for the pursuit controllers are the horizontal and vertical displacements of the target from the center of the fovea. Each of the controllers is a PI controller, and each degree of freedom of the gaze control system can be modeled approximately by the system shown in (b).

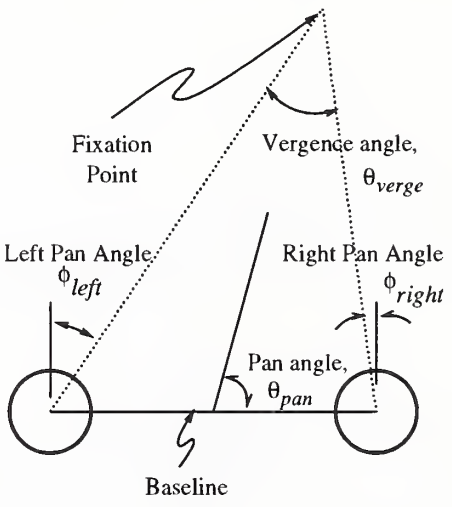


Figure 7: The relation between gaze angles $\vec{\theta}$ and motor angles $\vec{\phi}$ is sketched in this view of the "gaze plane" defined by the camera and the fixation point.

These gaze controls do not interfere with one another for this head's motor configuration, but other controls could. For instance, a vestibular gaze-stabilizing system might interfere with the pursuit system if the robot were holding gaze on an object that was moving. In this situation, the VOR would attempt to stabilize the cameras on a stationary object, but holding gaze on the moving object would require different camera movement. Because different gaze controls must share common degrees of freedom, they should be explicitly coordinated to achieve optimum performance [Brown, 1990b, Coombs and Brown, 1991].

One of the problems faced by visuomotor control systems is that the error signal is affected by every control response since the sensor is being moved. This can result in disruption of the visual signal. Similarly, in a dynamic scene, object motion may also perturb the signal. For example, as the view of the fixated object changes, the zero-disparity signal shifts and evolves, and sometimes it even drops out completely for brief periods of time.

A simple remedy for signal dropout is something like visual persistence, and this can help keep the cameras moving when the signal disappears temporarily. However, this mechanism is based on a very simple model of target behavior, and it is likely to introduce sharp changes in the error signal. Another simple and common model of target behavior is the $\alpha - \beta - \gamma$ filter, and this filter smoothes and interpolates the error signal. The $\alpha - \beta - \gamma$ filter assumes the target signal has constant acceleration, so this estimation must be done in a coordinate system in which the signal is relatively stable. For gaze-holding target coordinates, this means the head-centered reference frame is much better for estimating the target position than any visual coordinate system. The visual coordinate system is based on retinal image coordinates, and the retina is constantly being moved by the gaze control system. In order to estimate the head-centric target position, the retinotopic location of the target must be combined with the camera's angle within the head.

Using the $\alpha - \beta$ predictor in the loop to predict the delayed signal can lead to more accurate tracking. Figure 8(a) shows the camera movement and tracking error as the camera tracks the image of a dark object in approximate harmonic motion with a period of (360/70) seconds. The object is rotating in a plane and thus its distance from the camera varies and its velocity is not purely sinusoidal. The error is measured as the off-axis angle the centroid of the object's image. There is approximately a 100 ms delay in the system. The small phase difference between the sinusoidal waveforms of the target and camera motion induces a surprisingly large error. In Figure 8(a) an $\alpha - \beta$ filter is used ($\lambda = 1$) with no predictive advance, so the

tracking signal is smoothed somewhat. In Figure 8(b) the filter extrapolates the signal 50 ms into the future. The result is livelier tracking (in fact more advance destabilizes tracking) and reduced errors.

7.2 Gaze Holding Performance

Figure 9 shows a stereo robot's-eye view of a typical stimulus setup and the measured camera pan and convergence angles[2] and visual error signals for a target object moving through a field of distractors. The bunny was fixed to the end of the rotating stick and thus inscribed a circle, rotating while moving laterally and in distance. These measurements were recorded from a run with the stimulus rotating at 0.1 Hz, and the pan angle trace reveals rotational camera velocities as high as $13\,\mathrm{deg}/s$, with the cameras lagging a bit behind the apparent target velocity, as indicated by the non-zero observed retinal error. The pan and vergence traces are similar since they use similar controllers. Note that they are not entirely smooth because performance is sometimes perturbed by the changes in the scene as the target moves and the surrounding scene changes due to object motion, camera movements, and egomotion.

7.3 Ablation Experiments

In order to illustrate the function of each component of the gaze holding system, selected components were removed from the system, and the resulting behaviors are compared with the behavior of the complete system. Traces of camera pan angle can be seen in Figure 9(e). The first trace show the uninjured system's pan angle in the typical bunny-following scenario. The second trace illustrates the effect of loss of foveal processing in vergence and pursuit. The vergence system verges the cameras on any object that captures the disparity estimator's attention, resulting in vergence bouncing around the scene as the pan angles and bunny position change. Similarly, the pursuit system attempts to center gaze on all objects that lie in the horopter. The third trace shows the behavior that results from fixing the vergence angle. Early in this run (not shown), the system wandered until it locked onto an object that shared the horopter with the target's initial location. The fourth trace shows the result of eliminating the ZDF and using instead the edge energy of one image to drive pursuit. (The large step inputs that resulted make the system unstable, so foveal reduction was also eliminated for this experiment. Including extra-foveal edges in the "target" centroid calculation dilutes the effect of features entering and leaving the "foveal" area that caused the instability.) Clearly the centroid of the edge energy was influenced by the target's motion, but it was also significantly anchored by the surrounding stationary scene. Obviously, each piece of the system contributes to the performance, and it is the combination of the simple components that allows each part to be simple.

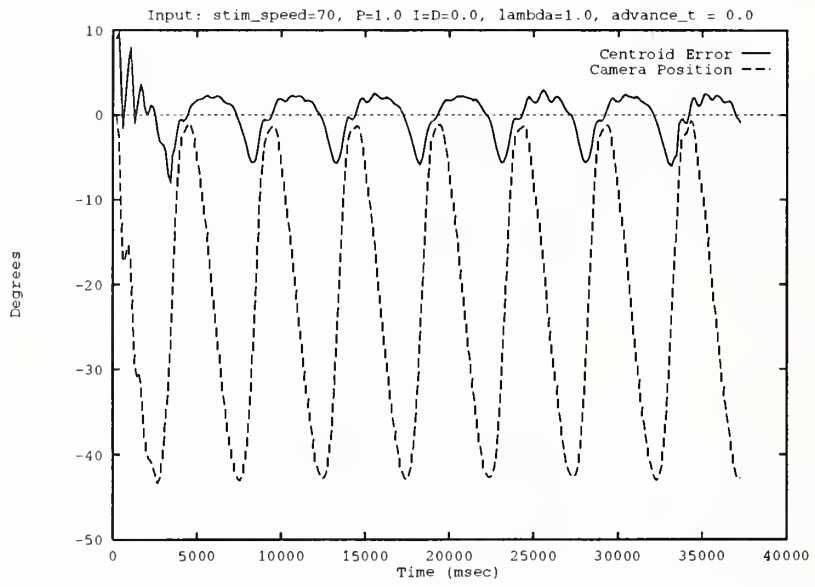
8 Conclusion

Eye movements are pervasive in the animal kingdom, and they have recently begun to play a prominent role in computer vision as well. Robots, like animals, inhabit a world of moving and stationary objects, and robots and animals themselves move about. Consequently, the ability to hold gaze on an object is crucial to seeing it clearly. Binocular foveal vision requires that the robot hold its foveae simultaneously on the visual target. In addition, motion blur degrades spatial resolution if the image is not prevented from slipping across the retina.

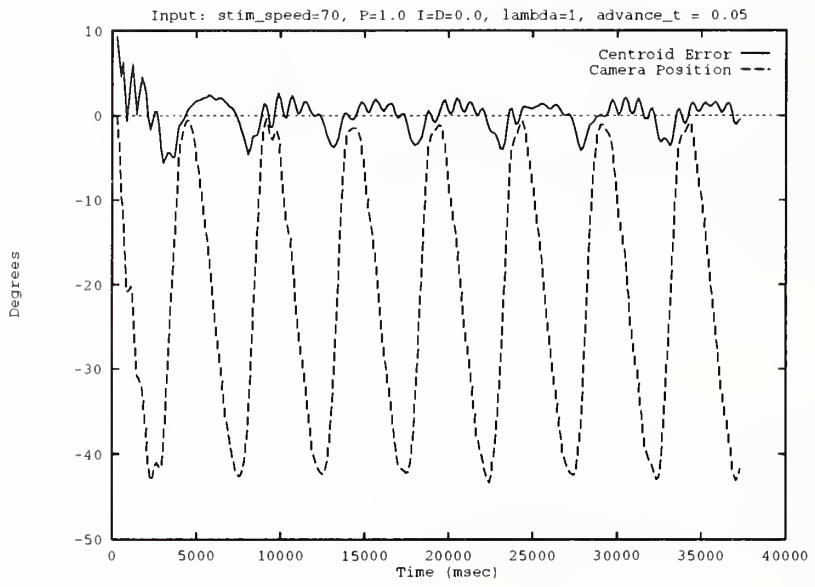
This paper examines the problem of holding a robot's gaze on an object while both are moving using only *precategorical* visual processing (*i.e.*, without requiring the ability to recognize the target). The approach is based on the premise that the control of camera movements should be considered an integral part of visual perception. By exploiting constraints that can be maintained by active control of camera movement, simplified visual processing is sufficient to hold the robot's gaze.

A system running in real-time on a moving robot demonstrates the idea, holding the binocular gaze of the robot on an object that moves through a cluttered scene. The vergence and pursuit components of

[2]The tilt trace is omitted for brevity.



(a)



(b)

Figure 8: Camera motion and error when tracking an object in approximate harmonic motion. (a) Delay of approximately 0.1s induces small phase lag but large tracking errors. (b) Camera motion and error using $\alpha - \beta$ predictor to advance the signal and overcome a delay of approximately 0.1s. Here the advance in the filter is 0.050s.

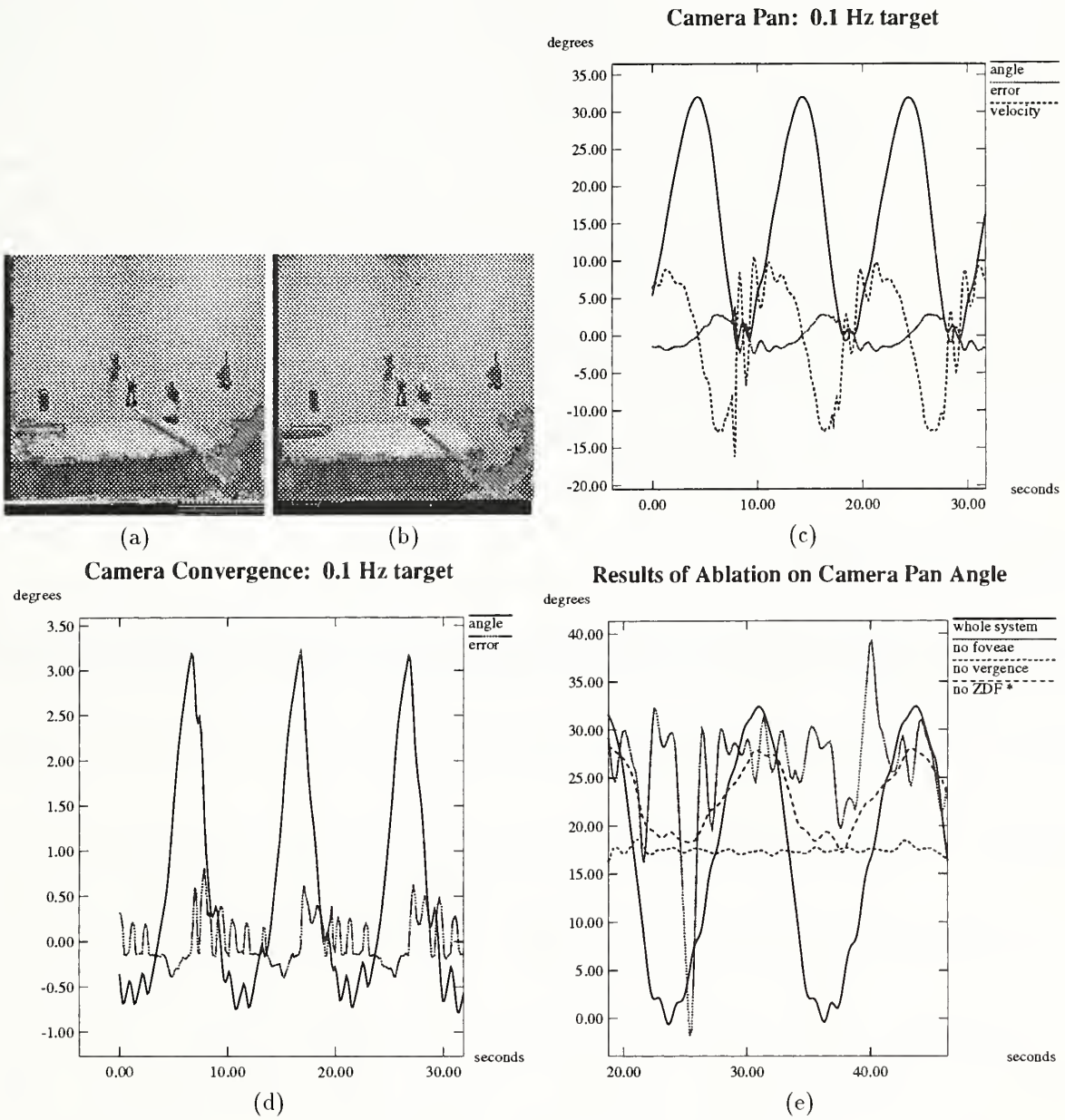


Figure 9: Gaze holding camera traces. Measured traces of the pan (c) and vergence changes (d) show the performance of the gaze holding system in following a target moving in 3-D through a field of distractors. A robot's-eye stereo view of a typical stimulus setup is shown in (a,b). (e) Ablation Camera Traces: The first trace shows the behavior of the unimpaired system for comparison. The second trace illustrates the loss of focus on the target object that results from the removal of foveal processing (or peripheral suppression). The third trace demonstrates the system's inability to follow the target if the vergence angle is cemented. The final trace shows how the system is distracted by objects at all distances when zero-disparity filtering is eliminated.

the system cooperate to simplify the visual processing required, as illustrated by Figure 1. The vergence system controls the vergence angle between the cameras to minimize the stereo disparity of the foveated target. A fast correlation-based technique estimates the most prominent disparity in foveal stereo images. The pursuit system controls the pan and tilt angles of the cameras to center them on the foveated object that has no stereo disparity. A simple zero-disparity filter locates features that have no stereo disparity. Thus the vergence system maintains zero disparity of the target for pursuit, and pursuit keeps the target foveated for vergence. The system is able to maintain these invariants in the retinal images by its active control of the camera angles.

It is important to note that it is easier to detect the tracking signals for active visual following than for tracking an object in passive stereo-motion image sequences. First, motion blur emphasizes the signal of target over the background. In passive visual following, the target's image slips across the retina and may thus be degraded by motion blur. During active pursuit, however, the eyes move to follow the target and stabilize the retinal image. Thus the image of the surrounding scene rather than the target moves across the retina and suffers from motion blur. The result is that image of the target is emphasized over the image of the background. Second, maintaining vergence isolates the target by disparity filtering. Holding vergence on the target enables the object to be isolated by simple zero-disparity filtering that detects objects at the fixation distance. Thus maintaining vergence on the target makes it possible to locate the target for pursuit control with simple precategorical visual processing. Third, active visual following also enables localized visual processing. The target's retinal location is roughly known because the pursuit system is keeping it near the center of view. This permits spatially localized visual processing.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation under Grants numbered IRI-8903582, CDA-8822724, and IRI-89220771, and by ONR/DARPA research contract number N000114-82-K-0193.

References

[Allen, 1989] Peter Allen. Real-time motion tracking using spatio-temporal filters. In *Proc. of the DARPA Image Understanding Workshop*, 1989.

[Analog Devices, 1987] Analog Devices, Inc., Norwood, Massachussetts. *DSP Products Databook*, 1987.

[Brown and Coombs, 1991] Christopher Brown and David Coombs. Notes on control with delay. Technical Report 387, University of Rochester, Computer Science Department, Rochester, New York 14627 USA, August 1991.

[Brown, 1990a] Christopher Brown. Prediction and cooperation in gaze control. *Biological Cybernetics*, May 1990.

[Brown, 1990b] Christopher Brown. Prediction and cooperation in gaze control. *Biological Cybernetics*, May 1990.

[Burt *et al.*, 1989] P. Burt, J. Bergen, R. Hingorani, R. Kolczinski, W. Lee, A. Leung, J. Lubin, and H. Shvaytser. Object tracking with a moving camera: An application of dynamic motion analysis. In *Proceedings of the Workshop on Visual Motion*, 1989.

[Clark and Ferrier, 1988] James Clark and Nicola Ferrier. Modal control of an attentive vision system. In *Proc. of ICCV'88, the Second International Conference on Computer Vision, (Tampa, FL, December 5–8, 1988)*, 1988.

[Coombs and Brown, 1991] David Coombs and Christopher Brown. Cooperative gaze holding in binocular vision. *IEEE Control Systems*, June 1991.

[Coombs *et al.*, 1990] David Coombs, Thomas Olson, and Christopher Brown. Gaze control and segmentation. In *Proc. of the AAAI-90 Workshop on Qualitative Vision*, Boston, MA, July 1990. AAAI.

[Coombs, 1992] David Coombs. Real-time gaze holding in binocular robot vision. Technical Report 415, University of Rochester, Department of Computer Science, Rochester, New York 14627 USA, June 1992. Ph.D. Thesis.

[Corke and Paul, 1989] Peter Corke and Richard Paul. Video-rate visual servoing for robots. Technical Report MS-CIS-89-18, Department of Computer and Information Science, University of Pennsylvania, GRASP Lab, Philadelpha, PA 19104, February 1989.

[Heeger and Hager, 1988] David Heeger and Gregory Hager. Egomotion and the stabilized world. In *Proc. of ICCV'88, the Second International Conference on Computer Vision, (Tampa, FL, December 5–8, 1988)*, pages 435–440, December 1988.

[Heeger and Jepson, 1990] David Heeger and Allan Jepson. Simple method for computing 3D motion and depth. In *Proc. of ICCV'90, the Third International Conference on Computer Vision, (Osaka, Japan, December 4-7, 1990)*, 1990.

[Heeger and Jepson, 1991] David Heeger and Allan Jepson. Subspace methods for recovering rigid motion I: Algorithm and implementation. *International Journal of Computer Vision*, 1991. in press.

[Heeger and Simoncelli, 1989] David Heeger and Eero Simoncelli. Sequential motion analysis. In *Symposium on Robot Navigation*, pages 24–28, Stanford, CA, March 1989. AAAI.

[Howard and Simpson, 1989] Ian Howard and W. Simpson. Human optokinetic nystagmus is linked to the stereoscopic system. *Experimental Brain Research*, 1989.

[Jenkin, 1991] Michael Jenkin. Using stereomotion to track binocular targets. In *Proc. of CVPR'91, the IEEE Conference on Computer Vision and Pattern Recognition (Maui, Hawaii, June 3–6, 1991)*, 1991.

[Krauzlis and Lisberger, 1989] R. Krauzlis and S. Lisberger. A control systems model of smooth pursuit eye movements with realistic emergent properties. *Neural Computation*, 1989.

[Lee and Wohn, 1988] Sang Wook Lee and K. Wohn. Tracking moving objects by a mobile camera. Technical Report MS-CIS-8-97, University of Pennsylvania, Computer and Information Science Department, November 1988.

[Lisberger *et al.*, 1987] S. Lisberger, E. Morris, and L. Tychsen. Visual motion processing and sensory-motor integration for smooth pursuit eye movements. *Annual Review of Neuroscience*, 10:97–129, 1987.

[Lisberger, 1990] S. Lisberger, 1990. personal communication.

[Matthies *et al.*, 1989] L. Matthies, T. Kanade, and R. Szeliski. Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision*, 3:209–236, 1989.

[McDonald *et al.*, 1983] J. McDonald, A. Bahill, and M Friedman. An adaptive control model for human head and eye movements while walking. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):167–174, April 1983.

[Miles *et al.*, 1991] F. Miles, U. Schwarz, and C. Busettini. The parsing of optic flow by the primate oculomotor system. In A. Gorea, editor, *Representations of Vision: Trends and Tacit Assumptions in Vision Research*, pages 185–199. Cambridge University Press, Cambridge, 1991.

[Nelson, 1991] Randal Nelson. Qualitative detection of motion by a moving observer. In *Proc. of CVPR'91, the IEEE Conference on Computer Vision and Pattern Recognition (Maui, Hawaii, June 3–6, 1991)*, pages 173–178, 1991.

[Olson and Coombs, 1991] Thomas Olson and David Coombs. Real-time vergence control for binocular robots. *International Journal of Computer Vision*, 7(1):67–89, November 1991.

[Papanikolopoulos *et al.*, 1991] N. Papanikolopoulos, P. Khosla, and T. Kanade. Vision and control techniques for robotic visual tracking. In *International Conference on Robotics and Automation*. IEEE, 1991.

[Shvaytser, 1988] Haim Shvaytser. Detecting motion in out-of-register pictures. In *Proc. of CVPR'88, the IEEE Conference on Computer Vision and Pattern Recognition (Ann Arbor, MI, June 5–9, 1988)*, pages 696–701, 1988.

[Swain and Ballard, 1990] Micheal Swain and Dana Ballard. Indexing *via* color histograms. In *Proc. of ICCV'90, the Third International Conference on Computer Vision, (Osaka, Japan, December 4-7, 1990)*, pages 390–393, 1990.

[Thorpe, 1983] Charles Thorpe. An analysis of interest operators for fido. Technical Report CMU-RI-TR-83-19, Carnegie-Mellon University, December 1983.

[Tölg, 1991] Sebastian Tölg. A biological motivated system to track moving objects by active camera control. In O. Simula, editor, *Proceedings of the International Conference on Artificial Neural Networks*, Espoo, Finland, June 1991. ICANN-91, Elsevier.

[Triesman, 1985] Anne Triesman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31:156–177, 1985.

[Waxman *et al.*, 1988] A. Waxman, W. Wong, R. Goldenberg, S. Bayle, and A. Baloch. Robotic eye-head-neck motions and visual navigation reflex learning using adaptive linear neurons. *Neural Networks Supplement: Abstracts of 1st INNS Meeting*, 1:365, 1988.

[Woodfill and Zabih, 1991] John Woodfill and Ramin Zabih. Real-time motion and stereo tracking. In *Proc. of the National Conference on Artificial Intelligence*. AAAI, July 1991.

[Yeshurun and Schwartz, 1989] Yehezkel Yeshurun and Eric Schwartz. Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), July 1989.

[Young, 1971] L. Young. Pursuit eye tracking movements. In P. Bach y Rita, C. Collins, and J. Hyde, editors, *Control of Eye Movements*. Academic Press, 1971.

NIST-114A
(REV. 3-90)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NISTIR 4826

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
APRIL 1992

4. TITLE AND SUBTITLE

Real-time Smooth Pursuit Tracking for a Moving Binocular Robot

5. AUTHOR(S)

David Coombs and Christopher Brown

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
GAITHERSBURG, MD 20899

7. CONTRACT/GRANT NUMBER

8. TYPE OF REPORT AND PERIOD COVERED

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

10. SUPPLEMENTARY NOTES

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

Abstract

This paper examines the problem of a moving robot tracking a moving object with its cameras, without requiring the ability to recognize the target to distinguish it from distracting surroundings. A novel aspect of the approach taken is the use of controlled camera movements to simplify the visual processing necessary to keep the cameras locked on the target. A gaze holding system implemented on a robot's binocular head demonstrates this approach. Even while the robot is moving, the cameras are able to track an object that rotates and moves in three dimensions.

The key observation is that visual fixation can help separate an object of interest from distracting surroundings. Camera vergence produces an horopter (surface of zero stereo disparity) in the scene. Binocular features with no disparity can be extracted with a simple filter, showing the object's location in the image. Similarly, an object that is being tracked will be imaged near the center of the field of view, so spatially-localized processing helps concentrate on the target. Instead of requiring a way to recognize the target, the system relies on active control of camera movements and binocular fixation segmentation.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

Pursuit Camera Movements; Active Vision; Binocular; Mobile Robots

13. AVAILABILITY

X	UNLIMITED
	FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).
	ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.
X	ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
25

15. PRICE
A02

ELECTRONIC FORM