Localized Binocular Attention and Real-time Smooth Pursuit in Moving Robots

David Coombs
National Institute of Standards and Technology *
Building 220, Room B-124
Gaithersburg, MD 20899
coombs@cme.nist.gov

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

The importance of eye movements to biological visual systems is obvious from their ubiquity. Even insects exhibit eye movements, although they are accomplished by head movements [Land, 1975]. In contrast, controlled camera movements have played a small role in computer vision research. However, there is growing interest in the role of camera movements in robotic visual perception, and some lessons for computer vision systems may be learned from studying biological vision systems. Both have limits on resolution and field of view, and they must therefore direct their visual sensors toward areas of the environment that are of interest. Also, both animal and robot visual systems exist to provide visual perception of the dynamic environments in which their owners operate.

There is a growing trend in computer vision to consider the visual system in the context of the behavior of a robot interacting with a dynamic environment. The *active vision* approach [Krotkov, 1989, Aloimonos *et al.*, 1988, Bajcsy, 1988] observes that constraints derived from known camera motion can replace other assumptions (*e.g.*, smoothness) that had previously been employed to solve mathematically ill-posed problems. *Animate vision* [Ballard, 1991] considers that visual perception is one of several behaviors employed by a creature in order to achieve its goals in a dynamic environment. Interest in animate vision has been sparked at least partly by the recent availability of real-time image processing equipment, which has enabled researchers to consider visual perception as a viable sensory input to a robot interacting with a dynamic world. Thus vision has begun to be considered as a means for gathering information that is relevant to the task in which the robot is engaged.

One of the principal tenets of animate vision is that sensing and motor behavior interact closely with one another [Ballard, 1991]. In this highly synergistic relationship, sensors provide perception to inform the creature's behavior, and motor actions make the creature an animate observer, using its sensors to maximum advantage. A corollary of this idea is that the sensor and the motors that move it must be considered together to arrive at the description of the perception system. One way of viewing this is to consider each creature as a sensory-motor system, consisting of perception, control, and effectors.

This work 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 the Rochester 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 a 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. Thus, it is demonstrated that deliberate control of the cameras can simplify the sensory processing that is required.

The central idea is that localizing attention in 3D space makes simple precategorical visual processing sufficient to hold gaze. The precategorical nature of the visual sensing means the algorithms are simple and do not require delicate tuning.

*Robot Systems Division, Manufacturing Engineering Laboratory, Technology Administration, U.S. Department of Commerce.

Precategorical Gaze Holding

Holding gaze is a fundamental capability of biological visual systems. 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.

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. 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.*, 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.

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. 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 Gaze Geometry

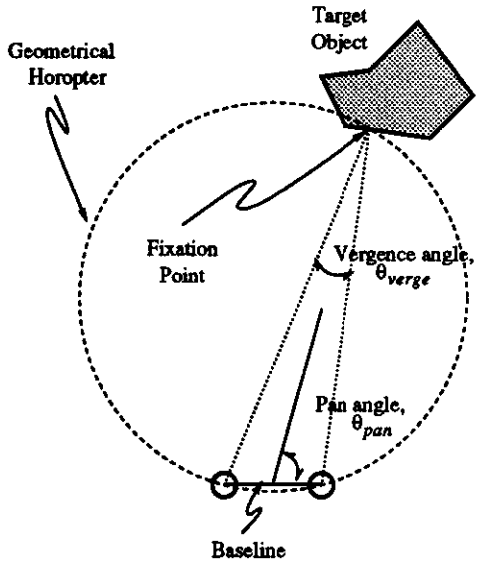


Figure 1: Top view of binocular gaze geometry. The goal of gaze holding is to keep the eyes or cameras fixed on a common world point or visual target. The gaze vector, $\vec{\theta}$, consists of the gaze pan and tilt angles and vergence angle. In order to keep the world point fixated, the gaze holding 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 binocular disparity is zero. The stereo images of an object that lies near the horopter have a narrow range of disparities.

binocular cues and the fact that the cameras are actively following the target.

Control of Sensors and Simplified Sensing

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. The Vergence control system rotates the cameras in opposite directions so the visual axes of the cameras intersect at the distance of the target object.

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 it in stereo

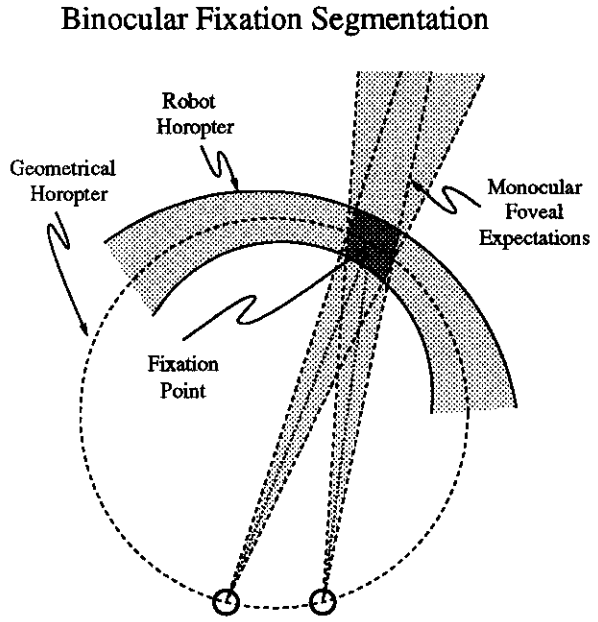


Figure 2: In this top view of binocular fixation, the lightly shaded areas are the regions of space that are highlighted by foveal and disparity filtering. The intersection of these areas is shaded darker, and it corresponds to the area around the fixation point in which an object can be segmented by this technique.

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 2. 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 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. Thus, the demonstration system exploits binocular cues and deliberate control of the cameras that enable precategorical segmentation of the fixation target.

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

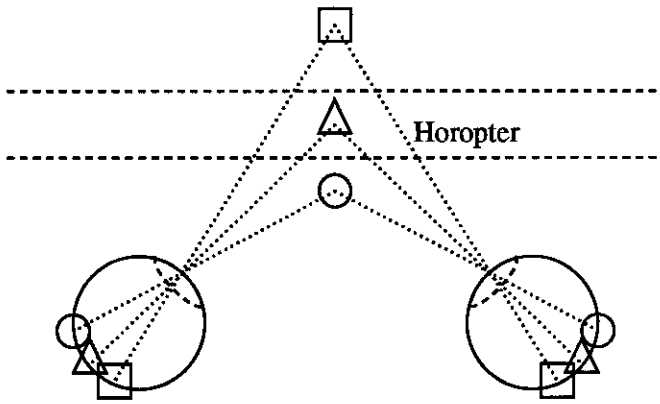
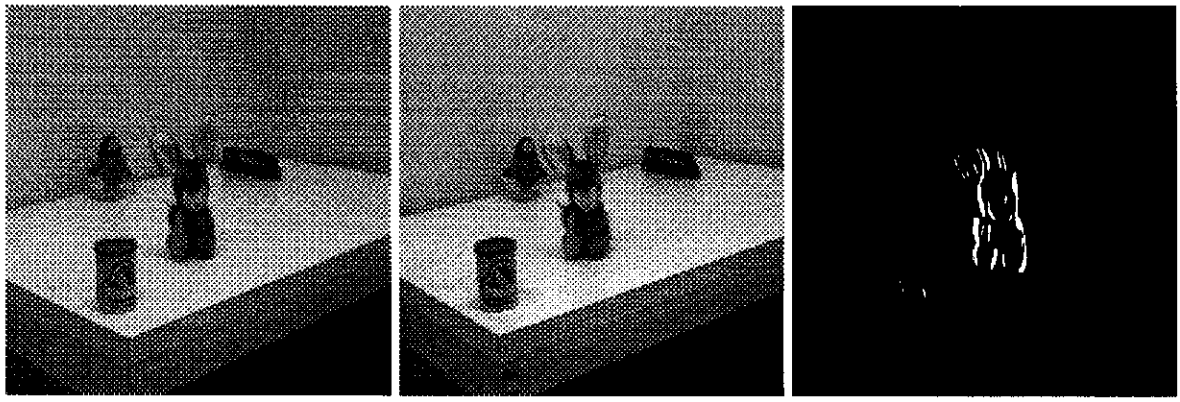


Figure 3: The *horopter* is operationally defined to be the region of space that contains objects that have no stereo disparity. It is a thin shell located at the fixation distance (the distance at which the cameras are verged). This figure illustrates the principle. The images of the triangle project 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.

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 called the *horopter* [Reading, 1983], and a simple filter can detect objects that lie in it.

Disparity filtering is used to ignore the background and foreground, leaving only the objects that have no binocular disparity. A disparity filter can detect features that have no stereo disparity (or any single fixed 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. In the demonstration gaze holding system, 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 in and out). With the target in the horopter, the disparity filter provides the retinal location of the target [Coombs and Brown, 1991, Coombs *et al.*, 1990].

The zero-disparity filter that is used by the demonstration system is a nonlinear filter that suppresses fea-



(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.

tures that have non-zero stereo disparity. (We have also experimented with correlation-based methods for implementing ZDFs [von Kaenel *et al.*, 1991]. These filters are less susceptible to aliasing, but the feature-based approach will illustrate the point.) 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. Thus the filter detects features that have no stereo disparity.

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 work examines the problem of holding a robot's gaze on an object while both are moving using only *pre-categorical* 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 the system cooperate to simplify the visual processing required, as illustrated by Figure 2. 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 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

[Aloimonos *et al.*, 1988] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, January 1988.

[Bajcsy, 1988] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 76:996–1005, 1988.

[Ballard, 1991] Dana Ballard. Animate vision. *Artificial Intelligence*, 48:57–86, 1991.

[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.

[Krotkov, 1989] Eric Paul Krotkov. *Active computer vision by cooperative focus and stereo.* Springer-Verlag, 1989.

[Land, 1975] Michael Land. Similarities in the visual behavior of arthropods and men. In Michael Gazzaniga and Colin Blakemore, editors, *Handbook of Psychobiology*, pages 49–72. Academic Press, 1975.

[Reading, 1983] R. Reading. *Binocular Vision: Foundations and Applications.* Butterworth, Boston, 1983.

[von Kaenel *et al.*, 1991] Peter von Kaenel, Christopher Brown, and David Coombs. Detecting regions of zero disparity in binocular images. Technical Report 388, University of Rochester, Computer Science Department, Rochester, New York 14627 USA, August 1991.