Reference Frames for Animate Vision

Dana H. Ballard*
Computer Science Department
University of Rochester
Rochester, New York, 14627 U.S.A.

Abstract

Animate vision systems have gaze control mechanisms that can actively position the camera coordinate system in respose to physical stimuli. Compared to passive systems, animate systems show that visual computation can be vastly less expensive when considered in the larger context of behavior.

1. What is Vision For?

We are accustomed to thinking of the task of vision as being the construction of a detailed representation of the physical world. Furthermore, this constructive process is regarded as being independent of larger tasks. However, a new paradigm that we term *animate vision* argues that vision is more readily understood in the context of the visual *behaviors* that the system is engaged in, and that these behaviors may not require elaborate categorical representations of the 3-D world. Brooks has argued against internal representations in a larger context [1986], and others have demonstrated the importance of integrating vision with behavior [Bajcsy and Allen 1984; Chen and Kak 1989] as well as demonstrating the advantages of knowing camera motions [Aloimonos et al. 1987]. The main purpose of this paper is to summarize the computational advantages of the animate vision paradigm, emphasizing problems of spatial localization.

The study of visual behaviors concerns the movements of the imaging system in the process of solving complex tasks. In this regard it is most instructive to examine data from eye movements in the human visual system.

This research was supported by NSF Grant No. DCR-8602958 and NIH Grant No R01 NS22407-01.

The human eye is distinguished from current, electronic cameras by virtue of having much better resolution near the optical axis. It has a high-resolution fovea where over a 1° range the resolution is better by an order of magnitude than that in the periphery. One feature of this design is the simultaneous representation of a large field of view and high acuity in the fovea. With the small fovea at a premium in a large visual field, it is not surprising that the human visual system has special fast mechanisms (saccades) for moving the fovea to different spatial targets. The first systematic study of saccadic eye movements in the context of behavior-was done by Yarbus [1967]. A selection of his data are shown in Figure 1. Subjects were given specific tasks pertaining to a familiar picture. The figure shows the traces for three minutes of viewing as a subject attempts to solve different tasks: (a) give the ages of the people; and (b) remember the position of the people and the objects in the room. The second is most remarkable, since it is so similar to the task of so many computer vision programs: we conjecture that since the eye movement traces show a specialized signature for this task, it is *not* done routinely. Instead, the overall impression of these traces is that the visual system is used to subserve problem-solving behaviors and such behaviors may or may not require an accurate model of the world in the traditional sense of remembering positions of people and objects in a room.

2* The Animate Vision Paradigm

We term the collection of different mechanisms for keeping the fovea over a given spatial target *gaze control.* The single most distinguishing feature of an animate vision system is high speed gaze control mechanisms. An important factor in gaze control is image stabilization. As animals, we move in relativelv fixed environments, but we also have to deal with other moving objects, animate and inanimate. Although we must function in the presence of different kinds of motion, our visual

Figure 1 (after [Yarbus 1967]) Reproduction from I.E. Repin's picture "An Unexpected Visitor" (a) and two records of eye movements. The subject examined the reproduction with both eyes for three minutes each time. Before the recording sessions, the subject was asked to: (b) give the ages of the people; and (c) remember the position of the people and objects in the room.

system works best when the imaged part of the world does not move. The spatial target at which the two optical axes intersect is termed the *point of fixation.* Binocular systems work to stabalize the images in the neighborhood of the point of fixation but cannot achieve complete stabilization owing to the 3-D nature of the world.

Gaze control mechanisms fundamentally change computational models of vision. Without them the visual system must work in isolation, with the burden of solving difficult problems with many degrees of freedom. With them a new paradigm emerges in which the visual calculations are embedded in a sensory-motor repertoire that reduces degrees of freedom and has the following computational advantages.

1) An animate vision system can move the cameras in order to get closer to objects, change focus, and in general use visual search [Pentland 1985; Krotkov 1988J. Often this visual search is more effective and less costly than algorithmic search on a single image, which may not even have the desired object in its field of view [Nelson and Aloimonos 1988].

2) Animate vision can make programmed camera movements. These provide additional constraints on the imaging process [Aloimonos et al. 1988]. In turn this facilitates the computational process dramatically: properties that are difficult to compute with a fixed camera system are much more easily computed with a moving camera system. One of the first demonstrations of this advantage was Bandopadhay's computation of rigid body motion parameters [Bandopadhay 1987].

3) Gaze control systems can be used to focus attention or segment areas of interest in the image precategorically. That is, one can isolate candidate visual features without first associating them with models using the degrees of freedom of the gaze control mechanisms. For example, one can use the blurring introduced by self motion while fixating to isolate the region around the point of fixation [Brown et al. 1988]. Similarly, one can use regions of near zero disparity produced by a binocular vergence system

4) The ability to control the camera's gaze, particularly the ability to fixate targets in the world while in motion, allows a robot to choose external coordinate frames that are attached to points in the world (see Figure 2). Behaviors based on fixation point relative coordinates allow visual computations to be done with less precision.

5) The fixation point reference frame allows visuo motor control strategies that servo relative to thai- frame. These are much simpler than strategies that use ego-centric coordinates.

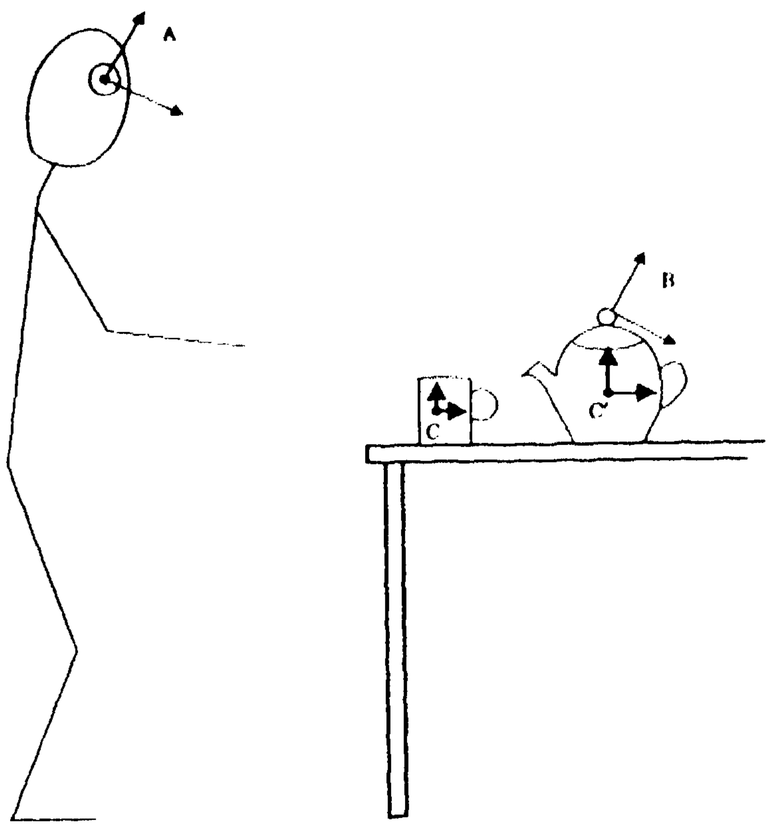

Figure 2 Much previous work *in* computational vision has assumed that the vision system is passive and computations are performed in a viewer-centered frame (A). Instead, biological and psychophysical data argue for a world-centered frame (B) This frame is selected by the observer to suit information-gathering goals and is centered at the fixation point. The task of the observer is to relate information in the fixation point frame to object-centered frames (C).

6) Gaze control leads naturally to the use of object centered coordinate systems as the basis for spatial memory. Object-centered coordinates have a great advantage over ego-centric coordinates in that they are invariant with respect to observer motion.

3. Gaze Control

We argue that the ability to control gaze can greatly simplify the computations of early vision, but what of the complexity of gaze control itself? If that should turn out to be prohibitively difficult it would negate the value of this paradigm. Fortunately, all our experimental work to date argues that this will not be the case [Ballard 1989|. Figure 3 shows our animate vision system. Currently we use a "dominant eye" control protocol whereby the dominant camera controls the system pitch and its own yaw coordinate using a simple correlation tracking scheme |Brown et al. 1988]. The non-dominant camera uses a novel vergence correction algorithm fOlson and Potter 1988] based on the cepstral filter [Yeshurun and Schwartz 1987J to correct its own yaw error. Brown [1989] has recently shown how these and other components can work together synergetically These components run in real time. At the moment there are many differences with a reasonable human model, but the performance is sufficiently good to allow us to explore vision while fixating in real time. Details may be found in [Brown et al. 1988].

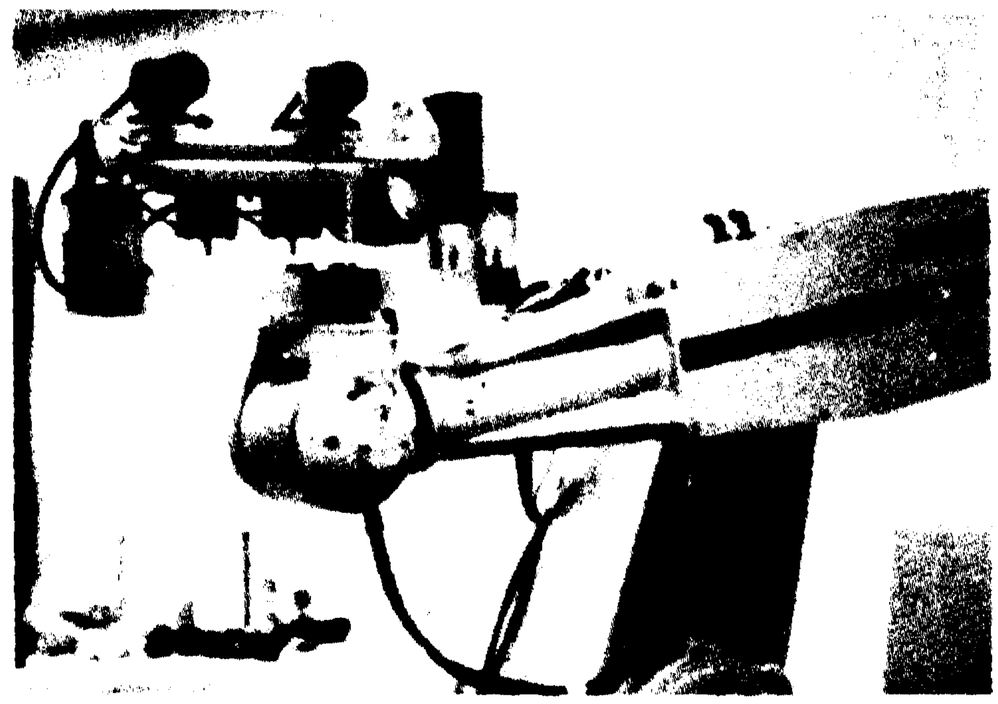

Figure 3 The University of Rochester's animate vision system. The "robot head" has three motors and two CCD high-resolution television cameras providing input to a MaxVideo image processing system. One motor controls pitch of the two-eye platform, and separate motors control each camera's yaw. The motors have a resolution of 2,500 positions per revolution and a maximum speed of 400% second. The robot arm has a workspace with a two-meter radius, and a top speed of about one meter/second.

The importance of vergence in gaze control is dramatically demonstrated in [Olson and Potter 1988]. Without vergence, very large disparities on the order of half the image dimension can be obtained These pose difficulties for algorithms that use stereo to build depth maps. With vergence, the disparities for the objects of interest can be kept small. In fact, most models of human stereopsis posit a fusional system that brings the disparities within the range of a detailed correspondence process IErkelens and Collewijn 1985; Marr 19821

4. The Fixation Frame

Early vision builds retinotopically-indexed maps of important environmental features such as depth, color, and velocity. Despite extensive work in this area over the past decade, the construction of such maps with computational models has proven to be very difficult. A primary reason for this may have been the assumption of a passive vision system. In an animate vision system, the degrees of freedom of the cameras are under the control of the animal. Aloimonos [et al. 19881 show in a general way how such assumptions can stabilize the computation of those features but their analysis misses the following vital point. *A passive vision system is more or less constrained to use the coordinate system dictated by the camera optics. In contrast, an active system that can fixate an environmental point can use an object-centered frame of reference centered at that point.* The calculations of early vision are greatly simplified given this ability. Note that this is a very different assertion than that of Marr [1982], who emphasized that the calculations were in viewer-centered coordinates. We assert that the calculations are more correctly represented as being in world-centered coordinates. As shown in Figure 2, the world-centered frame is viewer-oriented, but not viewer-centered.

To illustrate the advantages of using the fixation frame, we developed a computational model of motion parallax. Motion parallax, or kinetic depth, is the sensation of depth obtained by moving the head while fixating an environmental point in a static scene. Objects in front of the fixation point appear to move in the opposite direction to the motion while objects behind the fixation point move in the same direction. The apparent velocity is proportional to the distance from the fixation point [Cutting 1982]. Under these conditions it is easy to compute scaled depth (depth / fixation depth), which is a monotonic function of spatial and temporal derivatives of the image intensity function and has a zero value at the fixation point. By implementing this strategy on our robot we verified that a depth estimate can be obtained in real time over a 400X400 pixel image without iteration [Ballard and Ozcandarli 1988]. This result shows that the early vision computations of animate vision are decidedly simpler than fixed camera vision, as first noted by [Aloimonos et al. 1988]. Table 1 compares the two paradigms.

Animate Vision	Fixed Camera Vision
well-posed	usually underconstrained (the "ill-posed problem")
fixation frame	camera frame
behavioral state supplies necessary constraints	smoothness conditions required to stabilize the computations
computations are local in image space; can be solved in place in constant time	iteration over entire image is required

Table 1 A comparison of early vision computations done by animate vision and fixed camera systems.

5. The Importance of Behavior

When fixating a stationary point, the optical flow map can be interpreted as a depth map, but when pursuing a moving target, this interpretation is no longer valid. This motivates the dissection of visual activity into dominating behaviors. A compelling example of behavior in an animal system comes from Maunsell and Van Essen's work [1986] on the macaque monkey. The macaque contains a very distinct retinotopic cortical map that is sensitive to motion. Regular electrode sampling across this map showed that the visual area where the hands would be in hand-eye coordination was over-represented with respect to other areas (Figure 4). It may be that the visuo-motor system is best thought of as a very large amount of distinct special-purpose algorithms where the results of a computation can only be interpreted if the behavioral state is known. Ramachandran [1987] has raised a similar point, arguing from psychophysical grounds that the visual system may best be thought of as many different algorithms that exploit different cues, but that do not always work and may not be simultaneously satisfiable. Brooks [1986] has also noted this point, using the term "sensor fission" to emphasize that different sensors may be used in

different tasks. Recent work by Pentland on the shape from shading problem has also shown very simplified solutions for dominant special cases that depend on the behavioral milieu [1988], and there have long been special case solutions to the motion problem that depend on behaviors.

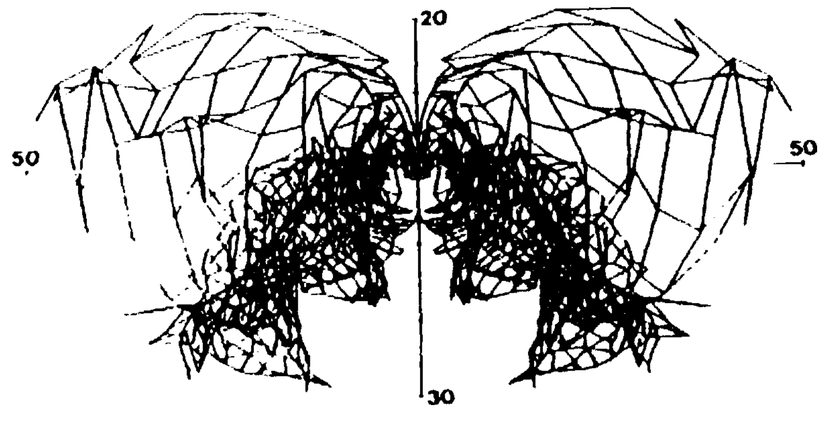

Figure 4 Maunsell and Van Essen's plot of the visual field in a macaque monkey's representation of optic flow shows that the portion of the visual field where the hands would be in a hand eye coordination task is more densely represented than other areas [1986]. Numbers mark degrees. Data from one hemifield is reflected about the midline.

6. Relative Vision

Many psychophysical tasks suggest that the way the image is interpreted depends on occlusion cues such as shown in Figure 5. It is not easy to make such judgements in an arbitrary position, as would be required by a viewer-centered hypothesis. The kinetic depth result suggests that the notion of a fixation point may be implicit behind the analysis even though we might not be aware of it. Our perceptual system is structured to make acccurate judgements relative to an object-centered frame at the fixation depth. Simplistically, imagine that one keeps two maps: one for structures that are judged to be in front of or at the fixation depth, and one for structures that are behind the fixation depth. The different interpolation rules can be fixed for each map.

The relative system has the virtue of requiring much less mathematical precision than the computations done in absolute coordinates. This is because the foveas provide the best precision only at the fixation point and an animate vision system can

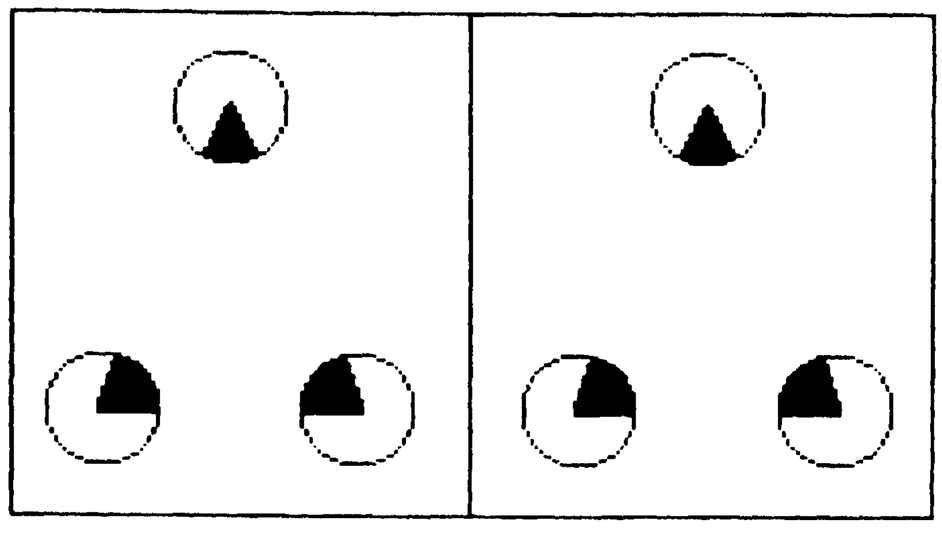

Figure 5 Ken Nakayama's illusion of subjective contours using stereo (not to scale). When fused, if the relative disparities are such that the triangle is in front of the circles, subjective contours are seen; if behind, then they are not.

control the location of its fixation point. In contrast, a fixed resolution system would have to be at least ten times larger, with even greater increases in computational costs, which scale by at least a low order polynomial factor [Tsotsos 1987]. In visually guided reaching, an arm out of the plane of fixation can be guided in depth to a target at the fixation plane by using only relative disparities of the manipulator as seen by the visual system. This scheme has the virtue of using the natural output of the stereo system which is in terms of fixation-relative coordinates.

7, Spatial Memory

Animate vision systems have the fundamental problem of representing space. One extremist solution is to have very high-resolution maps of the spatial environment and update these maps when something is changed. But for a variety of reasons, such a solution is not practical for animate systems. The foremost of these is the errors in the measurement system itself, which are a function of the relative positions of the robot and target object. Another reason is that such maps are very expensive in terms of size, since only a small portion of the material is relevant to tasks that require it to be identified.

We have argued that animate vision allows the perception of properties of the world to be related to a coordinate frame that is attached to the world by using the abilities to fixate or pursue. However, this coordinate frame is only valid for the duration of the

camera fixation; some additional structure is necessary for spatial memory. Our model of spatial memory represents features with respect to object-centered frames. An elegant way of relating this coordinate frame to object-centered frames (OCFs) posits an explicit representation of *transformations* between OCFs and the current view. If one assumes that the model and view have primitive parts, for example, line segments, matches between these parts determine particular values of the transformation that relates the stored model to the current view [Ballard 1981; Hinton 19811.

Figure 2 can be used to summarize the central ideas. The *current view* represents similar features *but with respect to a frame that is centered on the current fixation point* (as opposed to the camera frame used by passive systems). For example, if the fixation point is the object-centered frame origin, the transformation will only differ by a rotation, having a translation value of zero. This leads naturally to models of spatial memory that store relationships between object-centered frames. In a computational theory of active vision, eye movements have an integral role in the storing and retrieval of spatial information in the following ways: (1) the view transform T_{bc} contains the information necessary to foveate a visible object that has been recognized; and (2) stored relationships between objects, T_{cc}', can be used to transfer gaze from one object to another. In contrast, ego-centric or camera-centered systems attempt to maintain the transformations T_{ac} and T_{ac}', which is more computationally intensive.

A fovea is an elegant solution to the problem of simultaneously having high spatial resolution and a wide field of view given a fixed amount of imaging hardware. The price paid is that the high resolution fovea must be centered on the visual target. Thus small objects in a cluttered periphery can be effectively invisible. We think this difficulty can be minimized by having a stored model data base whereby small objects are linked to larger objects. To illustrate this proposal, we have built a two-dimensional eye movement simulator. Figure 6 shows the results from a test simulation. The problem is to locate a cup that is initially invisible in the periphery. Knowing that the cup is on the table, we first locate the table via a Hough transform technique and then use the pose information to center the gaze. In this instance, once the gaze is centered on the table, the cup is within the high resolution fovea and can be found by using the same Hough transform technique, but now with the cup as the stored model. Here again, application of a system with a high precision fovea avoids having to make fine-grained measurements over the full field of view.

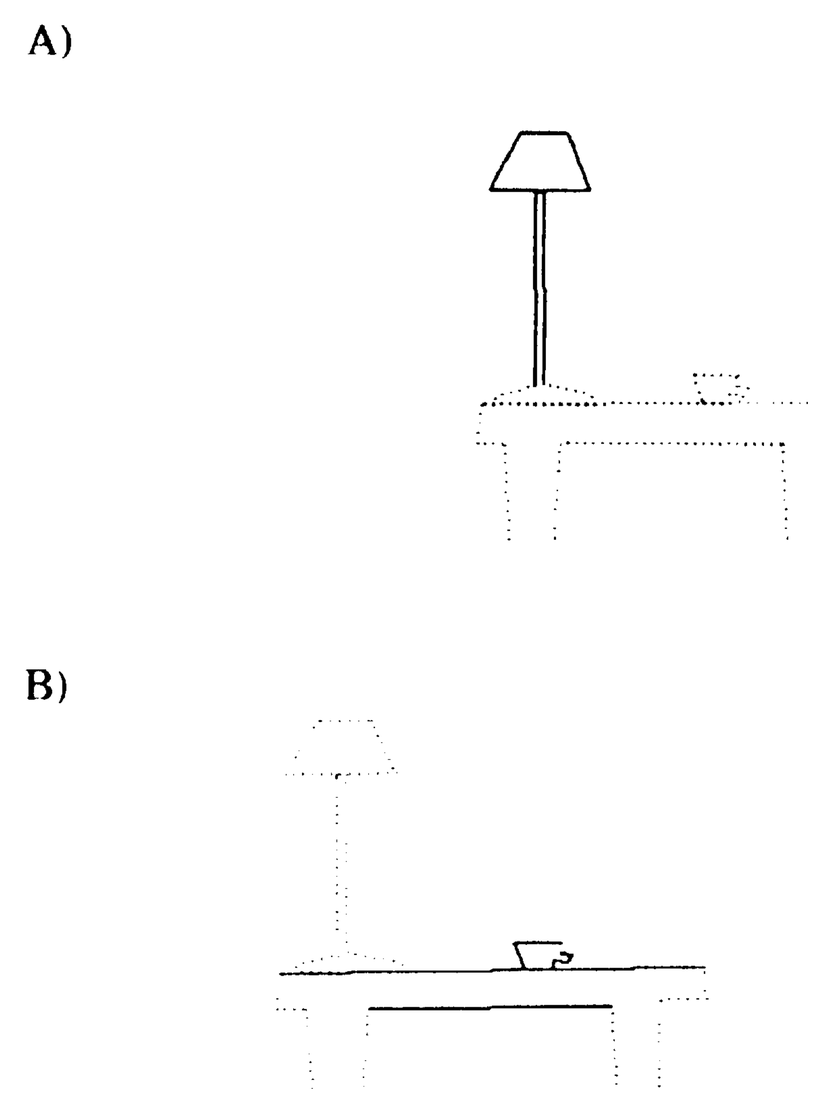

Figure 6 A foveal vision system is an elegant solution to the problem of high spatial resolution and a wide field of view. The price paid is that small objects on the periphery are hard to see. However, known relationships with large objects can help. In (A), the cup cannot be easily seen, but in searching for the cup, one can first look for the table (B), which in this case brings the cup near the fovea, where it can be found.

8. Conclusions

An animate vision system with the ability to control its gaze can make the execution of behaviors involving vision much simpler. Gaze control confers several advantages in the use of vision in behavioral tasks, and these have been summarized in Section 2.

Animate systems that rapidly change their coupling with the real world place a great premium on maintaining elaborate representations of the world. However, it may be the case that memorizing such representations is unnecessary, since they can be rapidly and *incrementally* computed on demand.

The study of animate vision is in its infancy, but we can already project that this paradigm will greatly extend the capabilities of all kinds of computer vision systems, but particularly those of mobile vision platforms. We believe such platforms will form the mainstay of future vision applications.

Acknowledgements

I would like to thank the University of Rochester team of researchers, Christopher Brown, Tim Becker, David Coombs, Roger Gans, Nat Martin, Tom Olson, Randal Nelson, Robert Potter, Ray Rimey, Dave Tilley, Steve Whitehead, Lambert Wixson, and Brian Yamauchi, all of whom have greatly helped refine the ideas herein. Peggy Meeker is responsible for the pleasing format of this manuscript.

References

Aloimonos, J., A. Bandopadhay, and I. Weiss. Active vision. In *Proc, 1st Intl. Conf. on Computer Vision,* pp. 35-54, June 1987. Also in *Int'l. J. Computer Vision 1*, 4, pp. 333-356, 1988. Bajcsy, R. and P. Allen. Sensing strategies. *U.S.-France Robotics Workshop,* U. Pennsylvania, Philadelphia, Nov. 1984. Ballard, D.H. Behavioral constraints on computer vision. *Image and Vision Computing 7,* 1, Feb. 1989. Ballard, D.H. Generalizing the Hough transform to arbitrary shapes. In *Proc, Intl. Conf. on Computer Vision and Pattern Recognition,* 1981 Ballard, D.H. and A. Ozcandarli. Eye fixation and early vision: Kinetic depth. *Proc, 2nd IEEE Intl. Conf. on Computer Vision,* December 1988. Bandopadhay, A. A computational study of rigid motion. Ph.D. thesis, Computer Science Dept, U. Rochester, 1987. Brooks, R.A. Achieving artificial intelligence through building robots. TR 899, Massachusetts Inst, of Technology, 1986. Brown, CM. Gaze controls with interactions and delays. TR 278, Computer Science Dept., U Rochester, March 1989. Brown, CM. (Ed), with D.H. Ballard, T.G. Becker, R.F Gans, N.G. Martin, T.J. Olson, R.D. Potter, R.D. Rimey, D.G. Tilley, and S.D. Whitehead. The Rochester robot. TR 257, Computer Science Dept., U. Rochester, August 1988. Chen, C.H. and A.C. Kak. A robot vision system for recognizing 3-d objects in low-order polynomial time. To appear, *IEEE Trans. SMC* (Special Issue on Computer Vision), 1989. Cutting, J.E. Motion parallax and visual flow: How to determine direction of locomotion. *4th Meeting, Int'l. Soc for Ecological Psychology,* Hartford, CT, 1982. Erkelens, CJ. and H. Collewijn. Eye movements and stereopsis during dichoptic viewing of moving random-dot stereograms. *Vision Research 25,* pp. 1689-1700, 1985. Hinton, G.E. Shape representation in parallel systems. In *Proc, 7th Intl. Joint Conf on Artificial Intelligence,* August 1981. Krotkov, E. Focusing. *Intl. J. Computer Vision 1,* 3, pp. 223-238, 1988. Marr, D.C *Vision.* W. H. Freeman and Co., 1982 Maunsell, J.M. and D. Van Essen. The topographic organization of the middle temporal visual area in the macaque monkey: Representational biases and the relationship to callosal connections and mylo-architectonic boundaries. *J. Comparative Neurology 266,* pp. 535-555, 1986. Nelson, R. and J. Aloimonos. Obstacle avoidance: Towards qualitative vision. In *Proc, 2nd Intl. Conf on Computer Vision,* December 1988. Olson, T.J. and R.D. Potter. Real-time vergence control TR 264, Computer Science Dept., U. Rochester, November 1988. Pentland, A. A new sense of depth of field. In *Proc, Intl. Joint Conf. on Artificial Intelligence,* pp 988-994, August 1985. Pentland, A. Shape from shading: A theory of human perception. In *Proc, Intl. Conf. on Computer Vision,* Tarpon Springs, FL, December 1988. Ramachandram, V.S. Interactions between motion, depth, color and form: the utilitarian theory of perception. In *Proc, Conf on Visual Coding and Efficiency,* September 1987. Tsotsos, J. A complexity level analysis of vision. In *Proc, IJCCV,* London, June 1987. Ullman, S. Visual routines. In S. Pinker (Ed). *Visual Cognition.* Cambridge, MA: Bradford Books, pp. 97-160, 1984. Yarbus, A.L. *Eye Movements and Vision.* Plenum Press, 1967. Yeshurun, Y. and E.L. Schwartz. Cepstral filtering on a columnar image architecture: A fast algorithm for binocular stereo segmentation. Robotics Res. TR 286, Courant Inst., New York U, March 1987.