Proceedings of the 1996 IEEE International
Conference on Control Applications
Dearborn, MI • September 15-18, 1996

Control of a Camera for Active Vision: Foveated Vision, Smooth Tracking and Saccade

Héctor Rotstein
Dept. of Electrical Engineering
hector@ee.technion.ac.il

Ehud Rivlin
Dept. of Computer Science
ehudr@cs.technion.ac.il

Technion – Israel Institute of Technology
Haifa 32000 – Israel

Abstract

In an active vision system, the information available for feedback has to be computed from images acquired on real-time. This image processing task can be seen as a degree of freedom that the designer has at hand to maximize tracking performance; for instance one can define a region of the image where heavy processing is performed: the fovea. Based on control considerations, this paper shows how to compute the optimal size of this fovea, and presents a two-tracking mechanism which is necessary if foveated vision is to be implemented on a real environment.

1. Introduction

"Active Vision" refers to the ability to move an image acquisition system in a controlled manner, in order to facilitate or allow certain machine vision tasks [1]. Active vision systems usually consist on one or more cameras mounted in such a way that their orientation and imaging parameters (focus, zoom, aperture) can be adjusted in real-time. The first active vision systems were constructed in the early eighties and were relatively slow and limited in scope [2]; recent advances in the technology of cameras and motors allow for systems with high dynamic performance, in some respects comparable with the human oculomotor system [3]. This has created the need for highly efficient dedicated image processing tools and for control systems capable of exploiting the potential characteristics of the mechanisms.

The gaze control of robot heads is usually modeled after the human visual system. It consists of a number of low level control loops which interact and –hopefully– cooperate to direct the attention of the system to a desired location. Gaze control can be divided into two primary categories [1]: gaze stabilization or fixation and gaze change. In this paper we will only consider the former, which is more closely related with classic control problems; the latter category usually involves high level planning tasks. See [1] for a discussion on potential advantages obtained by using gaze control.

Human Visual System

Two attributes of the human visual system are of interest for this paper: non uniform resolution and eye movement. The former refers to the fact that the human eye has a relatively small region with a high concentration of visual sensors in the center, the *fovea* and a more or less exponential decay of the concentration of receptors towards the periphery.

The fact that the fovea is relatively small, implies that it should be possible to reorient it in order to place the objects of interest in the region of high resolution, hence the need of ocular motion. Eye movement has been the subject of intensive studies and there is a large amount of sometimes conflicting data available; the interested reader is referred to the review paper [4] for an account of the activity with a systems flavor. For our purposes, we will consider the two main mechanisms responsible for tracking. The first one is called *saccade*, and consists of rapid shift in the eye position. It is characterized by a fast velocity and a relatively large time delay, presumably caused by processing time of the retinal information. The second mechanism is called *smooth pursuit*, which is slower than saccade and involves a smaller time delay. It is usually modeled as a continuous time system and its purpose is to keep the target within the fovea.

2. Setup and Modeling Considerations

For the purpose of addressing the basic problems of foveal vision and tracking mechanism, it suffices to consider a configuration with a single camera and one degree of freedom illustrated in Fig. 2. The camera is mounted on a motor and the angle θ that forms the

0-7803-2975-9/96/$5.00 © 1996 IEEE

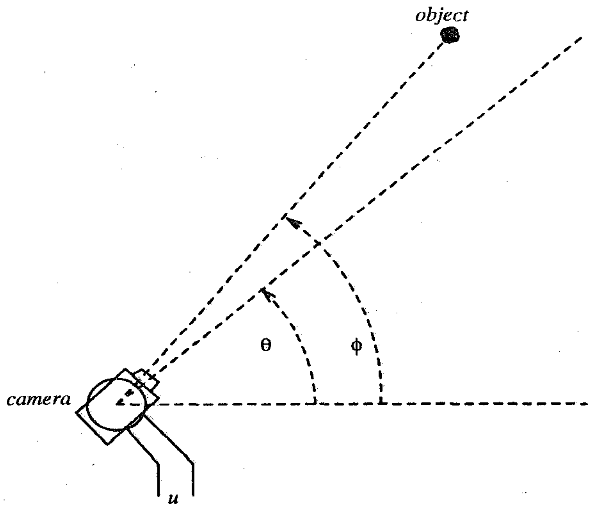

Figure 1: A simplified setup.

optical axis with the horizontal is the only degree of freedom. The image of the object is acquired by the camera connected to a vision card, which entails a sampling process, at a typical rate of at most 30 Hz. Each image should be processed in order to obtain the angle $\phi - \theta$. Processing time depends of the amount of data present, i.e., the "size" of the image, and the sophistication of the image processing algorithm. For control purposes, the image processing stage can be modeled, together with other effects like control law computation and communication times, as a pure *time delay* τ proportional to the size of the image (plus some overhead). If τ is larger than the sampling period T, the sequence of images has to be down-sampled by, say, q unless parallel processing units are employed. In the simplest possible case, q will be equal to the smallest integer larger than τ/T, but it can be made smaller subject to hardware availability. For ease of exposition, the former case is considered in the sequel. Assuming that the hardware and the image processing algorithms are fixed, the sub-sampling rate will only be a function of the size of the image x; the notation q^x will be used to stress this fact.

The feedback block-diagram, including the motor and the load, is shown in Fig. 2. The block S_T is a sampler with sampling period T, which is followed by a down-sampler with down-sampling rate q; the continuous-time signal is then sampled with sampling period qT. Note that the position of the object and the angle θ are continuous time functions, but the acquired image and the input to the controller are discrete time signals with different sampling rates whenever $q > 1$. Neither the continuous-time error $e(t) = \phi(t) - \theta(t)$ nor the one resulting from the "fast" sampling $e(kT) = \phi(kT) - \theta(kT)$ can be measured if $q > 1$; only $\hat{e}(k) = \phi(kqT) - \theta(kqT)$ is available for control. The block H_{qT} represents a hold function (typically a zero-order hold) which translates the discrete time output of the controller into a continuous-time signal. Finally, the system dynamics are lumped into the plant P_{in}, and a feedback controller C_{in} is included in order to obtain good position regulation and desensitazion of the electro-mechanical system from plant variations, possible neglected nonlinearities and disturbances.

In this paper, continuous time signals will be denoted as, e.g., $w(t)$, $\theta(t)$ and sometimes the dependence on t will be dropped when no confusion can arise. Discrete time signals will be denoted by, e.g., $\hat{e}(k)$.

3. Is Non-Uniform Resolution Beneficial?

Most cameras available commercially have uniform resolution, raising the question of whether it is beneficial to implement a fovea. A foveal window should be easy to implement in hardware or a combination of hardware and software, by keeping high resolution on a specified region of the image and reducing the resolution, e.g., by filtering and down-sampling, on the rest. The existence of a region of high resolution reduces computational times, which leads to faster sampling-rates and smaller time-delays and suggest the potential of a better performances. At the same time, for tracking purposes the image of a target should remain inside the region of high resolution, and this specifications becomes tighter if this region is smaller. This describes the basic tradeoff involved in deciding the potential benefits of implementing multi-resolution sensing. The purpose of this section it to formulate this tradeoff in a systematic manner, which will allow the computation of the size of the fovea in some optimal sense.

Consider the feedback configuration illustrated in Fig. 3. A reference model M has been included which generates the position $\phi(t)$ of the object as a function of the external signal $w(t)$. Inclusion of M does not necessarily imply an *a priori* knowledge on the behavior of the target since, for instance, M could be a single or double integrator which corresponds to assuming that ϕ is generated by the velocity or acceleration of the target which should then be characterized in some useful sense. It is worth stressing that this does not imply that $w(t)$ is available for feedback: the control system is driven by the positional error alone, since this is the only quantity that can be measured. The signal $w(t)$ is introduced as an artifice for designing the controller C.

When $w(t)$ denotes the acceleration of the target, a feasible controller should drive $e(t)$ asymptotically to 0 whenever $w(t) \equiv 0$, i.e., zero asymptotic error for constant velocity. This is a desirable characteristic also observed in the human visual system, cannot be achieved

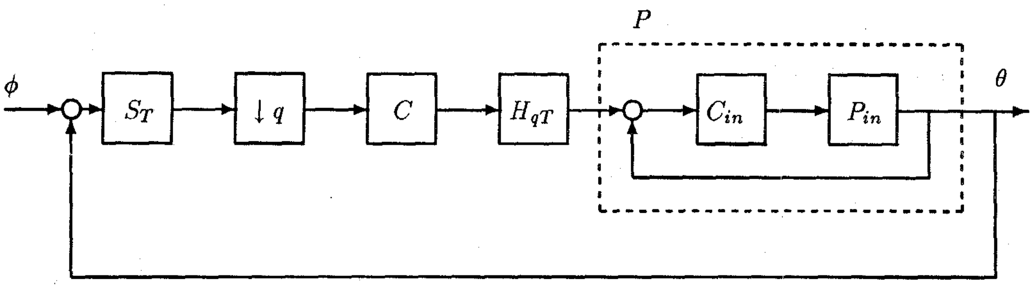

Figure 2: Closed-loop system.

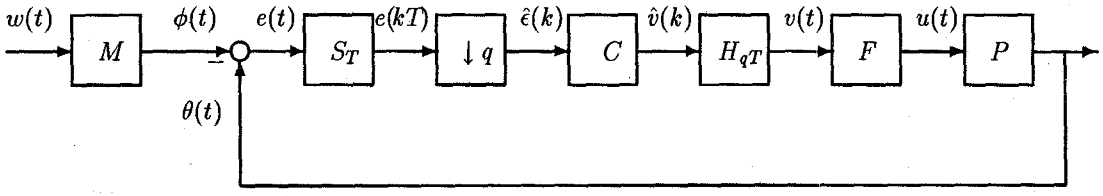

Figure 3: The feedback configuration with reference model

by using the discrete time controller C alone, but requires a pure integrator or, in general, a filter $F(s)$, between the output of the controller and the input of the plant.

The signal w is assumed to be an integrable function belonging to a set $\mathcal{W}(\alpha)$ parameterized by a positive real number α. Examples are the sets

$$\mathcal{W}(\alpha)_\infty \doteq \{w \text{ s.t. } |w(t)| \leq \alpha \ \forall \, t \geq 0\} \qquad (1)$$

$$\mathcal{W}(\alpha)_2 \doteq \left\{w \text{ s.t. } \int_0^\infty |w(t)|^2 \leq \alpha^2\right\}. \qquad (2)$$

$\mathcal{W}(\alpha)_\infty$ and $\mathcal{W}(\alpha)_2$ correspond to signals which are uniformly bounded for each time t and signals with bounded energy respectively. More generally, $\mathcal{W}(\alpha)$ should satisfy a monotone inclusion property as a function of α: $\mathcal{W}(\alpha_1) \subset \mathcal{W}(\alpha_2)$ if $\alpha_1 < \alpha_2$. The set $\mathcal{W}(\alpha)$ is a degree of freedom available for design, and its choice is dictated by the class of movements that the system is expected to track.

The final ingredient is the *half* size of the fovea, denoted x and measured in the same units as θ and ϕ. If $e(t) = \theta(t) - \phi(t)$ denotes the difference between the position of the camera and the target at time t, the control objective is to design a discrete-time controller C such that

1. The closed-loop system is stable, and
2. $|e(t)| \leq x$ for each $t \geq 0$, whenever $w \in \mathcal{W}(\alpha)$.

Notice that the specification in 2. is made in terms of the continuous time error $e(t)$ and not the sampled one $\epsilon(k)$ which is available to the controller. The reason for this is that concentrating in $\epsilon(k)$ may result in the target not remaining within the fovea during inter-sample time, which may be undesirable for image processing purposes; moreover, it may lead to oscillatory responses which should be avoided since the velocity of the object with respect to the camera must be relatively small to prevent image blurring.

The existence of a controller that satisfies the above criterion will in general depend of α, since $e(t)$ cannot be guaranteed to be small for arbitrarily "large" signal. It is then natural to consider the optimization problem:

Problem 1 (Maximum Size of Input) *Given x, find the largest α^x for which there exists a controller C^x that guarantees $|e(t)| \leq x$ for any $w(t) \in \mathcal{W}(\alpha^x)$.*

In mathematical terms, Problem 1 may be written as:

$$\alpha^x = \sup\{\alpha \ : \ \inf_{C \in \mathcal{C}} \sup_{w \in \mathcal{W}(\alpha)} |e(t)| \leq x, \ \ t \geq 0\} \qquad (3)$$

where $C \in \mathcal{C}$ is used to denote that the controller is stabilizing. This problem is closely related to optimal control problems with an induced-norm criterion. To see this, let $T_{ew}(C)$ denote the transfer function between w and e for a given controller C. For C stabilizing, $T_{ew}(C)$ is stable and it is possible to define the system norm:

$$\|T_{ew}(C)\|_{\infty,i} \doteq \sup_{w \in \mathcal{L}_i} \frac{\|e(t)\|_\infty}{\|w(t)\|_i}$$

The relevance of this norm is that, given an input w, it is possible to bound the norm of the output as:

$$\|e(t)\|_{\infty} \leq \|T_{ew}(C)\|_{\infty,i}\|w\|_i$$

and the bound is tight in the sense that there always exist an input w such that it holds as an equality. The controller C can be chosen optimally as a solution to the problem

$$\gamma_i^x \quad = \quad \inf_{C \in \mathcal{C}} \|T_{ew}(C)\|_{\infty,i}$$

A solution C^x to this problem is *min-max* optimal in the sense that it guarantees that the norm of the output will remain smaller than $\gamma_i^x\|w\|_i$ for a given input $w \in \mathcal{W}(\alpha)_i$. It follows that $\|e(t)\|_{\infty} \leq x$ if $\alpha \leq \alpha^x \doteq x/\gamma_i^x$; otherwise, there exists $w \in \mathcal{W}(\alpha)_i$ such that the constraint on the norm of $e(t)$ is violated. From a computational point of view, C^x as above may be found by using control theory techniques: in the case $i = 2$, C^x is known as generalized $\mathcal{H}_2$ controller while for $i = \infty$ it is called an $\mathcal{L}_1$ controller.

The performance γ^x depends of x via the sub-sampling rate q^x, and hence will be piece-wise constant: only variations of x large enough to change the integer q will affect it. Given that x is positive and bounded above by the physical size of the camera, the computation of only a finite number of values for γ^x is required. Notice that α^x will be small both for $x \approx 0$ and, usually, also for large values of x. The maximum of α^x will then be achieved for some finite value of x:

$$\alpha^* \doteq \max_x \alpha^x.$$

In principle, the maximum can be achieved for more than one value of x, so let x^* denote the largest such value less than the half size X of the camera. Non-uniform resolution is beneficial whenever $0 < x^* < X$, since then a controller may be designed such that w belongs to the largest possible set $\mathcal{W}(\alpha^*)$ such that $|e(t)| \leq x$. The associated controller $C^s = C^{x^*}$, which for the cases considered above is linear and time-invariant, will be referred to as a smooth tracking controller. C^s guarantees that the target will remain inside the fovea for the worst case $w \in \mathcal{W}(\alpha^*)$, although w can potentially not be in $\mathcal{W}(\alpha^*)$ and still the objective $|e| < x^*$ be satisfied.

4. Smooth-Pursuit and Saccade

The purpose of this section is to develop a control strategy for the case when the target moves out of the fovea or a fixation shift is specified by a higher level controller, which are characterized by $|e(t_v)| > x$ for some time t_v. The objective is not only to center the target in the fovea at some time $t_s > t_v$ but also to guarantee that the smooth controller will be able to perform satisfactory if the assumption on $w(t)$ is satisfied for $t \geq t_s$. Since performance is very poor in the interval $[t_v, t_s]$, a natural objective is to make this interval as short as possible.

4.1. Switching Between Controllers

The purpose of this section is to show how to perform the switching between smooth pursuit, saccade and vice-versa, and at the same time to provide the necessary tools for a systematic formulation of the saccadic action. A complete discussion is lengthy and rather technical [5] and hence only the main idea will be outlined. Let S denote the closed-loop transfer function from w to e for a controller design as in the previous section, i.e., $S = T_{ew}(C^s)$. Since this is a sampled-data system, it is linear but periodically time-varying. Given a state-space representation as in [6], an initial state $\hat{x}_0$ at time k_0 and some (integrable) function $w(t)$, let $\mathcal{F}_S(k, \hat{x}_0, k_0, w)$ denote the linear function mapping x_0 into the state trajectory $\hat{x}_S(k)$:

$$\hat{x}_S(k) = \mathcal{F}_S(k, \hat{x}_0, k_0, w),$$

where

$$\hat{x}_S(k) = \begin{bmatrix} \hat{x}_M \\ \hat{x}_C \\ \hat{x}_P \end{bmatrix}$$

where $\hat{x}_M$, $\hat{x}_C$, $\hat{x}_P$ represent the vector of states for the reference model, controller and plant respectively. Consider the *Reachable Set* $\mathcal{R}_S$ of S, defined as the set of all states that can be reached from 0 in a finite number of samples by using inputs $w \in \mathcal{W}(\alpha^*)$:

$$\mathcal{R}_S \doteq \{\hat{x}_S = \mathcal{F}_S(k_f, 0, 0, w) \text{ for } k_f \geq 0, \;\; w \in \mathcal{W}(\alpha^*)\}.$$

Given $\phi(t_s)$ at some future time $t_s = qk_s$, the objective of saccadic control is to synthesizes a control action that would allow to switch the smooth controller back into the loop at time t_s. A moment of reflection shows that it is not enough to guarantee that $|e(t)|$ will remain smaller than x for $t > t_s$ since, for instance, there may be a large velocity mismatch between the camera and the target at t_s. The problem is then to find an appropriate "target set" for the saccade; a similar problem was addressed in [7] for the discrete-time case.

Definition 1 (Target Set) *Given an internal state of the reference model $\hat{x}_M^o$, the state $\hat{x}_P$ belongs to the* target set $\mathcal{O}(\hat{x}_M^o)$ *if there exists k_s and $w \in \mathcal{W}(\alpha^*)$ such that*

$$\begin{bmatrix} I & 0 & 0 \end{bmatrix} \mathcal{F}_S(k_s, 0, 0, w) = \hat{x}_M^o$$
$$\begin{bmatrix} 0 & I & 0 \end{bmatrix} \mathcal{F}_S(k_s, 0, 0, w) = \hat{x}_P.$$

The set $\mathcal{O}(\tau, x_M^o)$ contains the states of the plant which can be reached by signals $w \in \mathcal{W}(\alpha^*)$ in a finite number of sample intervals if the internal state of the reference model is constrained to be equal to the one at τ, $\hat{x}_M(\tau)$. The important observation [8] is that if u^{sac} is now computed so that $\hat{x}_P(\tau) \in \mathcal{O}(\hat{x}_C(\tau))$, then the smooth controller can be switched back into the loop at time $q\tau$ by initializing its internal mode to $\hat{x}_C(\tau) = [0\ 0\ I]\mathcal{F}_S(k_s, 0, 0, w^v)$ where $w^v \in \mathcal{W}(\alpha)$ is such that

$$\begin{bmatrix} x_M(\tau) \\ \hat{x}_P(\tau) \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \mathcal{F}_S(k_s, 0, 0, w^v).$$

It follows from the reasoning in [7] (see also [8] for a more detailed treatment) that $|e(t)| < x$ for $t \geq \tau$ if the future disturbances $|w(t)| \leq \alpha^*$. The reason is that the closed-loop system will behave for $t > \tau$ as if the past input to the system would have been w^v (a similar interpretation can be made for the case of normed bounded signals).

4.2. Saccadic Control

The discussion in the previous section provides the framework for the systematic treatment of saccadic control. Following the approach in [7], four different stages are considered.

Switch On

Suppose that the constraint on $|e(t)|$ is violated at time t_v so that the smooth controller can no longer guarantee good performance or even continue its normal operation. A saccadic action is then trigger, which requires relatively lengthy computations. Meanwhile, the camera should somehow be operated in a way that will possibly facilitate the future correction. In the absence of additional information about the variations of the position of the target, then one could select a fictitious signal w^{tv} in such a way that the error criterion remains constant from $t_v = k_v h$ and up to the instant where the saccadic control is employed.

Modeling

In order to reduce the error signal bellow x at some future instant τ, it is necessary to predict the values of the signal $\phi(t)$ for $t \geq \tau$, based on measurements which are usually costly to obtain and potentially contaminated by noise. The success of the saccadic control action may depend of the accuracy of these predictions. As an example, suppose that the target changes its position to some stationary point lying outside the foveal window (this is a standard experiment when evaluating human saccades [4]); then the modeling problem reduces to determine the new position, which can presumably be done accurately.

The computation of models for prediction under different sets of assumptions is considered in detail in [9], which contains an array of different algorithms. The specific algorithm should be selected depending on the standing assumptions for $\phi(t)$ and the noise which is possibly corrupting the measurements. This selection is important since it will determine the time lag required to have a prediction of future position and how accurate that prediction will be. A popular choice in the active vision field is to select an $\alpha - \beta$ or $\alpha - \beta - \gamma$ filters, which have the advantage of their simplicity. Coefficients of this filters are usually selected by using the steady-state solution of a Kalman filtering problem [9]. However, much better predictions can be made if a priori knowledge of the variations of $\phi(t)$ are available and exploited, for instance, if the objective is not to track a moving target but to do a gaze shift.

Saccade

Once the model is available at time, say, t_p, it is possible to compute $\hat{x}_M(t_s)$ for some future time instant and hence the time-varying target set $\mathcal{O}(\hat{x}_M(k_s))$. The problem is now to generate the control signal $u^{sac}(t)$ that drives the plant from $\hat{x}_P(t_p)$ to $\mathcal{O}(\hat{x}_M(t_s))$. A natural objective is to do this in the shortest possible time, not only because of the tracking objective but also since the future prediction of $\hat{x}_M(t_s)$ potentially deteriorates with time. It is implicitly assume that the internal state of the plant is measurable for feedback; this can be achieved at least approximately if the internal control loop discussed before is designed so that P can be accurately approximated by a second order system, for which both position and velocity are measured.

The computation of the saccadic control appears to be challenging; it can be approximated by using fast-sampling, i.e., replacing the continuous-time virtual input $w(t)$ by a piece-wise constant function:

$$w(t) = \hat{w}(k) \quad k\hat{h} \leq t < (k+1)\hat{h}$$

where $\hat{h} << T$. This reduces the problem to a discrete-time multi-rate one. The advantage is that in that case linear programming based algorithms exist for solving these problems, and they allow the inclusion of additional constraints, like bounds on the tolerable control actions. Notice that the constraint on the target set is a linear one, and so can be incorporated with minor modification into the formulation.

Switch Off

Linear optimal controllers such as an $\mathcal{L}_1$-optimal, assume that the initial state for the plant to be controlled is zero. If the initial state is non-zero and unknown, then the controller can no longer guarantee the desired performance and should be replaced by a usually more complicated one (e.g., non linear, time-varying). As claimed above, if the initial state is known, then the same controller can be used if it is properly initialized, since it amounts to finding a fictitious but legal disturbance that would drive the state of the plant to the actual non-zero initial state, when the plant is interconnected with the optimal controller. Then, it is possible to "read-out" the state of the controller and initialize the actual configuration so that the optimal performance can be guaranteed.

5. Conclusions

In this paper some of the fundamental problems regarding the control of an active vision system have been addressed. It has been shown that the benefit of implementing foveal vision can be formulated as an optimization problem, since a trade-off appears between having a small window which would yield small computational delays but tighter control objectives or relaxing the control objectives but obtaining more challenging dynamics. Following the current approach, the size of the fovea is chosen as the one giving best tracking capabilities, as measured by the size of the signals which the system is guaranteed to track. It was also shown that foveal vision is tightly related with smooth pursuit, since the solution to the former provides a controller which makes the latter meaningful.

Since in a realistic environment the smooth controller will eventually fail to keep the target inside the fovea, a two-mode controller has been introduced. The second mode is a saccadic controller which replaces the smooth one whenever the tracking error becomes large. The computation of saccadic control includes modeling the evolution of the target, generating a signal which will drive the system while computations are performed, and then position the camera in such a way that the smooth pursuit controller can be switched-back into the loop and perform according to specifications. It was established that, in the light of recent developments in optimal control theory, all this requirements can be formulated in a systematic manner.

Several assumptions have been introduced in this paper in order to facilitate the exposition, and adjustments are required in order to implement the different stages in a practical system. Computation of the optimal size of the fovea and the smooth controller is straightforward. On the other hand, saccadic control involves intensive on-line computations, and hence should be implemented carefully to obtain satisfactory results. Computations can be speed-up by using the a priori knowledge on the type of target the system is expected to track. In this respect, notice that the modeling stage has been de-emphasized although it is critical for achieving good performance, and hence constitutes the degree of freedom that the designer has at hand for tailoring saccades to specific applications.

As stated in the introduction, a simple setup was considered in order to highlight the most important problems involved in tracking by an active vision system. The extension to the more interesting case of two degrees of freedom tracking is relatively simple, and requires rewriting the theory in this paper in terms of a system with two inputs and two outputs. Since for existing systems the two channels are essentially decoupled, the extension is straightforward. The problem becomes more challenging if a system with two cameras is considered, since then the two cameras must be operated in a coordinate manner in order to provide stereo vision and improved performance.

References

[1] Michael J. Swain and Markus A. Stricker. Promising directions in active vision. *International Journal of Computer Vision*, 11(2):109–126, 1993.

[2] Nicola J. Ferrier and James J. Clark. The Harvard binocular head. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(1):9–31, 1993.

[3] John Fiala, Ronald Lumia, Karen Roberts, and Albert Wavering. TRICICLOPS: A tool for studying active vision. *International Journal of Computer Vision*, 12(2/3):231–250, 1994.

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

[5] Héctor Rotstein and Ehud Rivlin. Control of a camera for active vision: Foveal vision, smooth tracking and saccade. Submitted, 1995.

[6] Bassam Bamieh, J. Boyd Pearson, Bruce Francis, and Allen Tannenbaum. A lifting technique for linear periodic systems with applications to sampled-data control. *Systems & Control Letters*, 17:79–88, 1991.

[7] Héctor Rotstein, Ehud Rivlin, and Yeoshua Zeevi. Two-mode control: An oculomotor-based approach to tracking systems. In preparation, 1995.

[8] Hector Rotstein and Leonid Mirkin. On static feedback for the $\mathcal{L}_1$ and other optimal control problems. In preparation, 1995.

[9] Yaakov Bar-Shalom and Thomas E. Fortmann. *Tracking and Data Association*. Mathematis in Science and Engineering. Academic Press, 1988.