NISTIR 5513

On Visual Pursuit Systems

John C. Fiala
Boston University
Department of Cognitive and
Neural Systems
Boston, MA 02215

and

Albert J. Wavering
Intelligent Controls
Intelligent Systems Division

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

NIST

NISTIR 5513

On Visual Pursuit Systems

John C. Fiala
Boston University
Department of Cognitive and
Neural Systems
Boston, MA 02215

and

Albert J. Wavering
Intelligent Controls
Intelligent Systems Division

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

October 1994

U.S. DEPARTMENT OF COMMERCE
Ronald H. Brown, Secretary

TECHNOLOGY ADMINISTRATION
Mary L. Good, Under Secretary for Technology

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
Arati Prabhakar, Director

On Visual Pursuit Systems

John C. Fiala
Boston University
Department of Cognitive and Neural Systems
Boston, MA 02215
fiala@cns.bu.edu

Albert J. Wavering
National Institute of Standards and Technology
Intelligent Systems Division
Gaithersburg, MD 20899
wavering@cme.nist.gov

Abstract

Visual pursuit systems are reviewed using a generalized form with the principal delay in the visual feedback pathway. The performance of several different control methods are compared through analysis of the tracking error. A high-performance active vision system, called TRICLOPS, is used to obtain experimental data on the tracking performance of the various methods, some of which had previously only been tested through simulation. Image processing delay is shown to be the primary limiting factor in performing high-speed tracking. Prediction is necessary to achieve high performance tracking, and the effectiveness of prediction depends on several factors, including the coordinate system used for motion modeling and the noise content of the target motion signal. It is demonstrated that, using relatively simple tracking algorithms which incorporate prediction with an advanced robotic device like TRICLOPS, it is now possible to outperform humans in the domain of visual pursuit of simple targets.

1. Problem Statement

A visual pursuit system in its most basic form consists of a moveable visual sensor, such as an eye or camera, capable of rotating in two degrees of freedom to keep a moving target in the field of view. In the discussion which follows, a camera is used as the visual sensor, but direct analogy with biological pursuit systems is maintained. Assuming a pinhole camera model and that the camera rotation is about the optical center, the visual kinematics for one of the degrees of freedom is as shown in Figure 1. The position of the target object on the image is given by

$$i = f\tan\rho \tag{1}$$

Visual pursuit aims to keep the image of the object in a fixed position, usually the center of the image, even though the object is moving. This can be accomplished by rotating the camera based on sensed visual information. To center

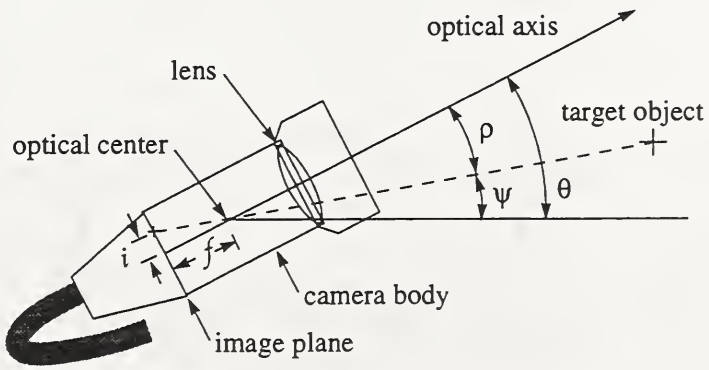

Figure 1. Simplified visual kinematics.

the object image, i.e. place the image of the object at $i = 0$, the axis of the camera should be rotated by an amount

$$\rho = \tan^{-1}\left(\frac{i}{f}\right) \tag{2}$$

The absolute angle of the camera axis (or joint) rotation will be given by θ. By defining a zero reference for θ, the target object position can also be expressed as an absolute angle o with respect to this zero reference. The imaging angle is thus

$$\rho = \theta - \psi \tag{3}$$

With a narrow field-of-view camera used for high-resolution imaging the imaging angle ρ will be small (as is the case for the small foveal region of high resolution in human vision), therefore

$$\tan\rho \approx \rho \tag{4}$$

A generic block diagram for visual pursuit is depicted in Figure 2. The real image position of the object is given by i. The desired object image position i_d is assumed to always be zero; therefore, pursuit may be thought of as a regulator problem. Since the desired image position is zero, i also represents the error during tracking. The tracking performance can be quantitatively characterized by this tracking error signal i.

The amount of tracking error in the pursuit system is determined by G, the controller and plant, and H, the visual processing delay, as well as by the object motion.

$$\text{tracking error} = i = \left[\frac{fGH}{1+fGH} - 1\right] f\psi = [fHT - 1] f\psi \tag{5}$$

where $T = \dfrac{G}{1+fGH}$

To further quantify the effects of G and H on tracking performance, it is useful to consider the system response to sinusoidal object motions. For this, a frequency domain analysis can be performed. Let the closed-loop transfer function $T(j\omega)$ have a particular amplitude a and phase ϕ at the frequency ω.

$$T(j\omega) = ae^{-j\phi} \tag{6}$$

Likewise, let $H(j\omega)$ include only the visual processing delay.

$$H(j\omega) = e^{-j\delta} \tag{7}$$

The magnitude of the tracking error in the frequency domain can be written

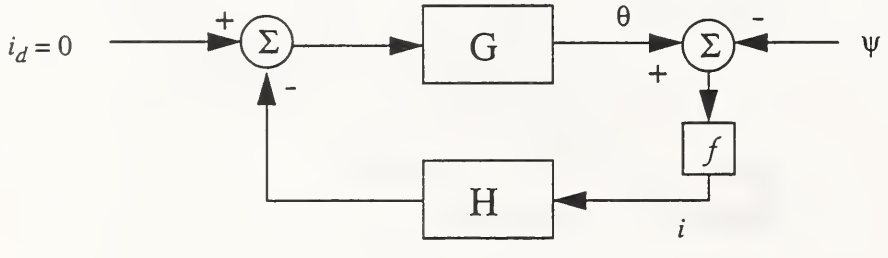

Figure 2. Generic block diagram for visual pursuit.

$$|i(j\omega)| = \sqrt{1 + a^2f^2 - 2af\cos(\phi + \delta)}\ \ f|\psi(j\omega)| \tag{8}$$

The amplitude a of $T(j\omega)$ is in the interval $[0,1/f]$, such that when the input i_d is reproduced exactly at the output θ, the transfer function includes the scale factor $1/f$ which relates image position to joint position. In the ideal case, $a = 1/f$, $\phi = 0$, $\delta = 0$, and there is no tracking error as a result. The tracking error may be normalized by dividing both sides of equation (8) by $f|\psi(j\omega)|$. Figure 3 shows the normalized tracking error as a function of the product af and the total phase lag, which consists of the closed-loop transfer function phase plus the visual processing delay ($\phi+\delta$). This plot indicates that the relationship between the magnitude of the tracking error and the closed-loop transfer function gain is dependent on the total phase. If the total phase is small, the tracking error decreases as the closed-loop transfer function gain increases. For larger amounts of phase lag, the tracking error can increase as the closed-loop gain increases, and in fact the tracking error can be larger than the object motion itself. Increasing the closed-loop transfer function phase or the visual processing delay always increases the tracking error.

A quantitative measure of tracking performance, *tracking error bandwidth*, can be defined as the frequency of object motion at which the tracking error exceeds, say, 20% of object motion amplitude, i.e. when

$$\frac{|i(j\omega)|}{f|\psi(j\omega)|} = \sqrt{1 + a^2f^2 - 2af\cos(\phi + \delta)} = 0.20 \tag{9}$$

The goal of the visual pursuit system is to minimize the tracking error magnitude and maximize the tracking error bandwidth. Several approaches to this problem are considered in the following. Experimental results for a number of different tracking algorithms are presented in Section 4.

2. Tracking Algorithms

The image error (i_d - i) can be used to construct a direct proportional-derivative (PD) servo for pursuit. The form of this system is depicted in Figure 4. Here the actuator/camera plant is modeled as a mass M, the visual delay time is D, and K_p and K_v are proportional and derivative gains, respectively. A circumflex is used to differentiate parameter

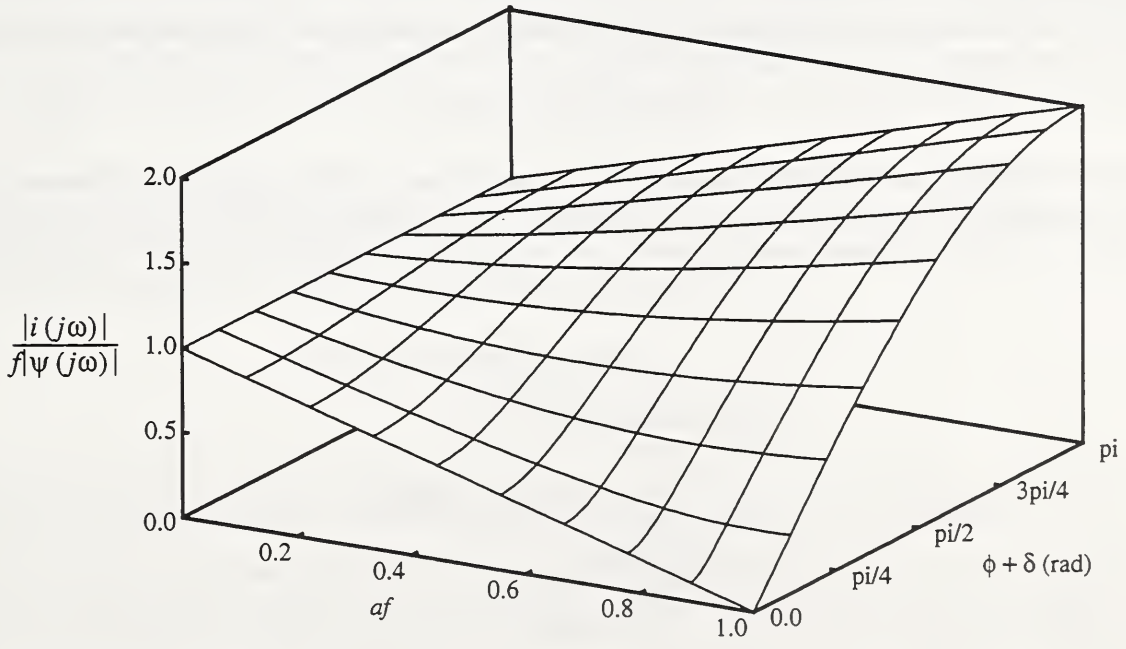

Figure 3. Dependence of normalized tracking error on closed-loop transfer function gain and total phase.

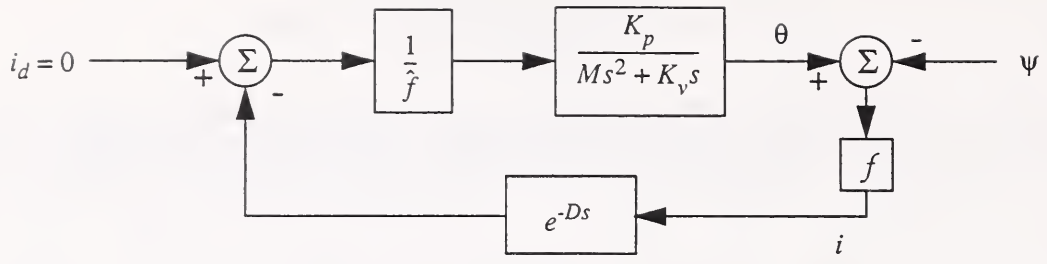

Figure 4. Image-based servoing system.

values which are estimated or measured for use in control equations from the true values which exist in the physical system. However, in most of the equations to follow, it will be assumed that these are essentially equivalent. The values of H and T for the system of Figure 4 are

$$H(s) = e^{-Ds} \tag{10}$$

$$T(s) = \frac{K_P/f}{Ms^2 + K_v s + e^{-Ds}K_p} \tag{11}$$

It can be seen that the visual processing delay D in the servo loop limits the size of K_p for which the system is stable, which results in a small amplitude of T and a corresponding large tracking error.

Since controlling a device in a coordinate system other than actuator coordinates often involves delays which limit the servo performance, robot manipulator controllers have traditionally used an inverse kinematic approach. In this scheme, the servo control is performed by a high-bandwidth joint controller as shown in Figure 5. The commanded motion, usually formulated in a Cartesian coordinate system, is transformed into joint position commands by an inverse kinematics function. This *kinematic* approach allows the controller to move in the Cartesian system with the maximum possible tracking performance.

For the visual pursuit system, the (simplified) transformation from image to absolute joint position is given by

$$\theta_d(t) = \frac{i}{f} + \theta(t - D) \tag{12}$$

where $\theta_d(t)$ = the desired joint position at time t. The absolute joint position of the target object is the image position of the object, transformed to joint angle units and offset from where the camera optical axis was when the image was

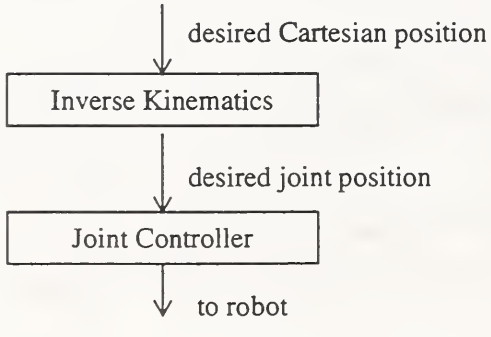

Figure 5. Traditional inverse kinematic controller.

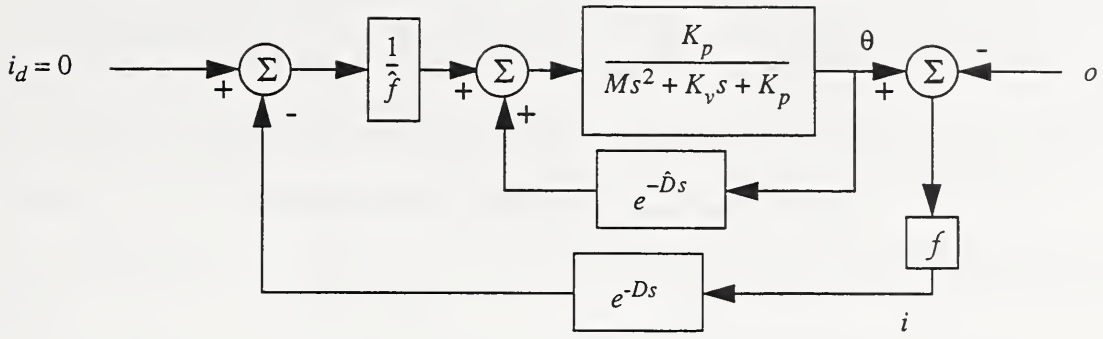

Figure 6. Kinematic approach to pursuit.

taken.

Using this kinematic approach for the pursuit controller, Figure 6 is obtained, which has

$$H(s) = e^{-Ds} \tag{13}$$

$$T(s) = \frac{K_p/f}{Ms^2 + K_v s + K_p} \tag{14}$$

Since the delay is eliminated from T, a much higher value of K_p can be used and the resulting controller produces less tracking error than the previous one. The image processing delay now affects the tracking error, but not the stability of the system. The caveat to this statement is that although the *magnitude* of the image processing delay does not affect the stability of the system, the *accuracy* of the estimate of the delay does. If the delay estimate is accurate, then the measurement of the target position is minimally affected by motions of the camera, and it does not matter if the camera is pointing directly at the target or not, as long as the target is visible to the camera. In this case, the system is effectively open-loop with respect to the visual information, and the high-bandwidth joint servos track the stream of goal positions as if the target position estimates were coming from an external measurement system. However, any errors in the delay estimate applied to the joint positions will result in coupling between the joint motions and the predicted position. A feedback loop is created in which the feedback loop gain is equal to the delay error times the velocity (Brown and Coombs 1991). If the gain of this loop is negative (caused by underestimating the delay), then the effect is primarily increased tracking error. However, if the visual delay is overestimated, a positive feedback loop is created which can cause instability.

The system of Figure 6 has the same form as that obtained by reasoning along different lines (Brown 1990, Clark and Ferrier 1988). Brown (1990) has previously described this type of controller as based on Smith's principle (Smith 1957). In Smith's scheme, the minimum-phase process for the system of Figure 4 would be as shown in Figure 7. This is just the high-bandwidth joint controller and plant, given in Figure 6 by

$$\frac{K_p}{Ms^2 + K_v s + K_p} \tag{15}$$

By adding to this minimum-phase controller positive and negative feedback loops that each contain the visual processing delay, the so-called *Smith predictor* controller is obtained, which is identical to that of Figure 6.

The control scheme proposed by Robinson (1988) for the human visual pursuit system uses image velocity signals rather than the image position signals that have been used here. In the form of our generic pursuit system of Figure 2, which has only a lumped visual processing delay, Robinson's scheme would be as shown in Figure 8. Here, C(s) is the control law which could be simply a velocity gain or perhaps something more complicated (Robinson 1988).

The sensed image quantity in Figure 8 is the derivative of the image error discussed previously. The derivative of

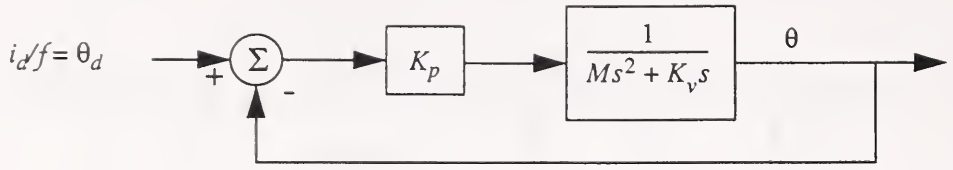

Figure 7. Minimum-phase system for Smith controller.

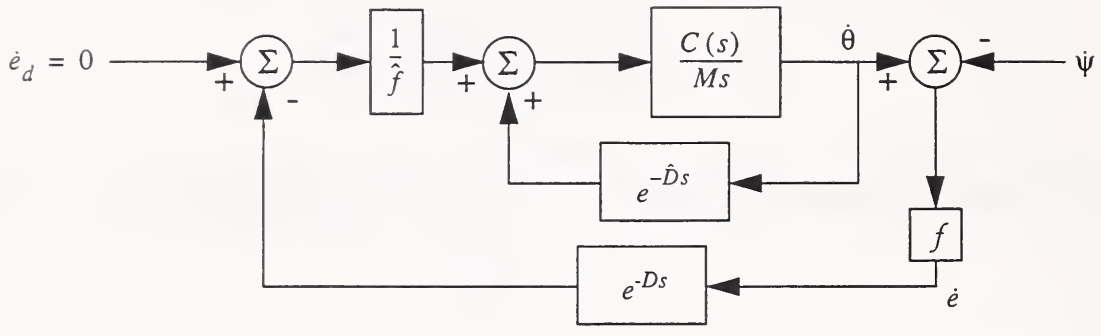

Figure 8. Pursuit system based on velocity signals.

image error can be expressed in terms of object motion.

$$\frac{di}{dt} = (fHT - 1)f\frac{d\psi}{dt} \tag{16}$$

where

$$H(s) = e^{-Ds}$$

$$T(s) = \frac{C(s)/f}{Ms}$$

Assuming that the initial image error is zero, this implies

$$i = (fHT - 1)f\psi \tag{17}$$

such that the expressions for tracking error magnitude derived in the previous section apply to the system of Figure 8 as well as that of Figure 6.

Note that, for the systems of Figures 6 and 8, $a \approx 1/f$ and $\phi \approx 0$ for frequencies of object motion well below the bandwidth of T(s). The tracking error magnitude is essentially

$$|i(j\omega)| = \sqrt{2}\sqrt{1 - \cos(\delta)}\ f|\psi(j\omega)| \tag{18}$$

with

$$\delta = D\omega$$

Thus, the tracking error bandwidth is inversely proportional to the visual processing delay.

$$BW \approx \frac{0.2}{2\pi D} Hz \tag{19}$$

The human pursuit system has visual processing delays of about 0.1-0.2 s (Yarbus 1967). This would give a tracking error bandwidth of 0.16-0.32 Hz. By direct observation, however, it is clear that the human pursuit system is ca-

pable of accurately tracking object motions up to 1-1.5 Hz (Stark, Vossius, and Young 1962). Also, the human visual pursuit system is capable of tracking continuously at speeds of 30°/s (Young 1967), and at much greater speeds for discrete intervals (Yarbus 1967). Using the systems of Figures 6 and 8, in which the visual delay is still very much a factor in tracking, would result in large tracking errors that would put the target off the fovea. For example, a visual delay of 0.13 s (Robinson 1988) at a constant speed of 30°/s, would result in a tracking error of 3.9°. Since the fovea is only about 3° of the visual field (Levine 1985), this error would prevent successful tracking in the sense that the image of the target would not be maintained within the fovea.

In order to successfully track objects the way humans do, a pursuit system must perform much better than those described in this section. It is essential to incorporate compensation for the substantial visual processing delay of the pursuit system.

3. Pursuit with Prediction

To overcome the limitations of the pursuit systems described in the previous section, an element of prediction can be introduced into the controller. The existence of a predictive capability for human visual pursuit has long been established (Stark, Vossius, and Young 1962, Young 1971). For the pursuit systems described here, the inverse kinematics from image to absolute joint position provide the object position with respect to a stationary reference. This object position signal can be used to model and predict object motion.

Figure 9 shows the pursuit system with prediction. The motion model and prediction are used to estimate the position of the object D seconds ahead of the feedback signal. If this estimate is as desired, the output of the predictive model is the present object position. Ideally, the motion model and prediction provide a forward time shift that can be modeled as e^{Ds}. Thus, the system can be represented by

$$H(s) = e^{-Ds} \tag{20}$$

$$T(s) = \frac{e^{Ds}(K_p/f)}{Ms^2 + K_v s + K_p} \tag{21}$$

The resulting tracking error is equivalent to that produced by

$$H(s) = 1 \tag{22}$$

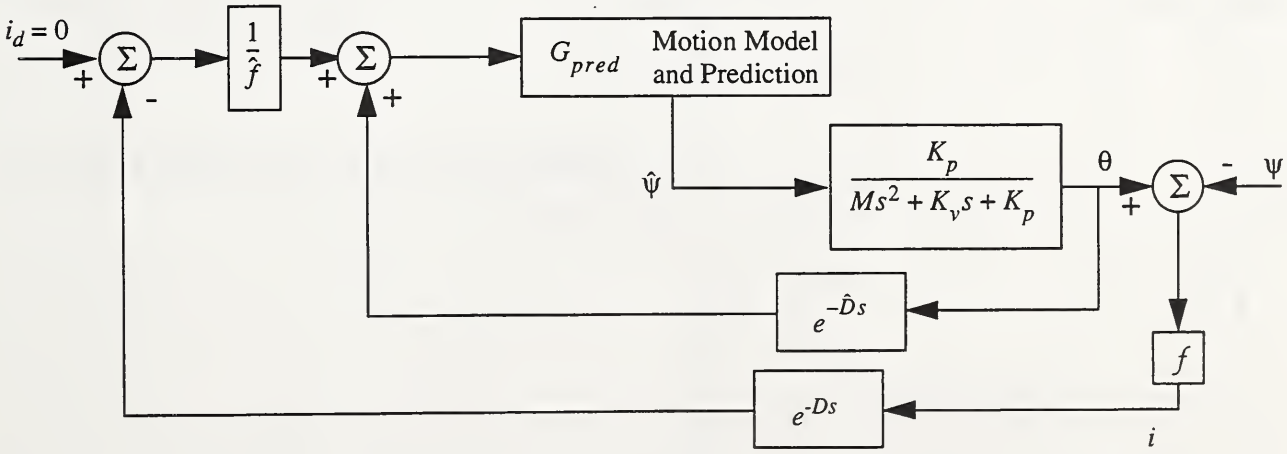

Figure 9. Pursuit system with prediction.

$$T(s) = \frac{K_p/f}{Ms^2 + K_v s + K_p} \tag{23}$$

In the ideal case, then, the prediction eliminates the effect of the visual processing delay on the tracking error. Using equation (8), the resulting tracking error is given by

$$|i(j\omega)| = \sqrt{1 + a^2 f^2 - 2af\cos(\phi)}\ f|\psi(j\omega)|. \tag{24}$$

The tracking error is determined solely by the performance of the high-bandwidth joint controller. This is essentially identical to the tracking approach advocated by Sharkey and Murray (1993).

Of course, a control element which provides the desired e^{Ds} functionality perfectly would be physically unrealizable. There are predictive filters, however, which can be used to provide a reasonable approximation of perfect prediction over a limited range of input frequency. These filters are based on the assumption that past target motions can be modeled and extrapolated to provide an estimate of target motion into the near future. Two kinds of these filters are the α–β–γ filter (Bar-Shalom 1988) and the polynomial Least Mean Squares Fit (LMSF) filter (Ng and LaTourette 1983). The α–β–γ filter is a constant-coefficient Kalman filter which estimates the target position, velocity, and acceleration, and uses this information to extrapolate the future motion of the target. Similarly, the LMSF filter involves performing a least-mean-square fit of a polynomial function of time to a sequence of previous target positions, and extrapolating to obtain an estimate of future position. The frequency domain characteristics of these filters with different parameter values are presented and discussed in (Wavering and Lumia 1993). As an example of the frequency characteristics of predictive filters, the magnitude and effective prediction for a third-order LMSF filter which uses a window of 12 previous data samples are shown in Figure 10 and Figure 11 for the case of 2.7 sample periods of prediction. This filter has a transfer function of

$$\begin{aligned} G_{pred} = {} & 2.77 + 0.540e^{-\Delta ts} - 0.729e^{-2\Delta ts} - 1.23e^{-3\Delta ts} - 1.16e^{-4\Delta ts} - 0.714e^{-5\Delta ts} - 0.087e^{-6\Delta ts} + \\ & 0.525e^{-7\Delta ts} + 0.927e^{-8\Delta ts} + 0.923e^{-9\Delta ts} + 0.318e^{-10\Delta ts} - 1.09e^{-11\Delta ts} \end{aligned} \tag{25}$$

where Δt = the sample time. The horizontal axes in Figure 10 and Figure 11 are given in terms of the normalized fre-

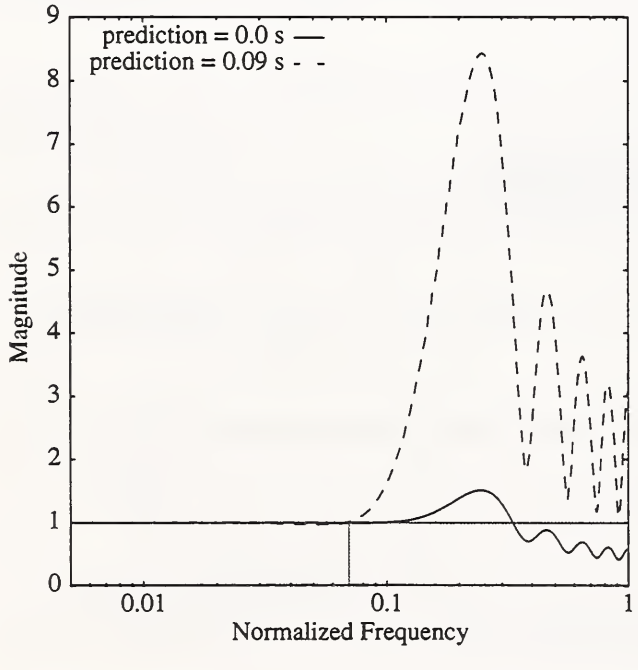

Figure 10. Magnitude response of cubic LMSF filter.

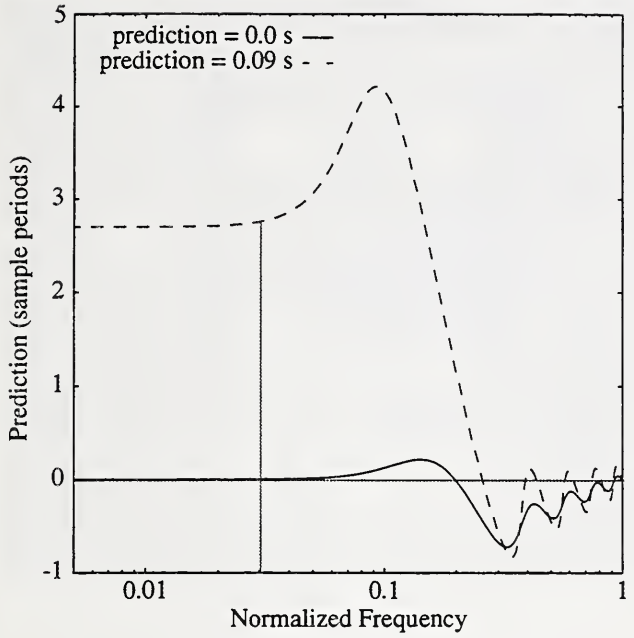

Figure 11. Prediction characteristics of cubic LMSF filter.

quency $f/f_{Nyquist}$, where $f_{Nyquist} = f_{sample}/2 = 15$ Hz for a 30 Hz sample rate. These figures show that the frequency content of the target motion signal should be below about 1 Hz (normalized frequency = 0.07) to avoid amplitude distortion, and below about 0.45 Hz (normalized frequency = 0.03) for an accurate amount of prediction. In practice these ranges are extended somewhat, since some inaccuracies can be tolerated before the target is lost from the field of view.

Wavering and Lumia (1993) also discuss an approach to prediction which uses autocorrelation of the target motion signal to identify and predict periodic target motions. This type of motion modeling and predictive control for pursuit systems has been proposed by Bahill and McDonald (1981) (see also Harvey and Bahill 1985). In their approach, termed *Target-Selective Adaptive Control* (TSAC), estimation of an object motion model is made from the feedback signals. The model motion is then used to provide a predictive input to drive the camera to the correct position. Figure 12 depicts a version of the Bahill and McDonald pursuit system with the delay isolated in the visual feedback pathway (and without the adaptive gain adjustment mechanisms discussed in Bahill and McDonald (1981) and Harvey and Bahill (1985)).

The system of Figure 12 is not in form of the general model of Figure 2. The expression for the tracking error i

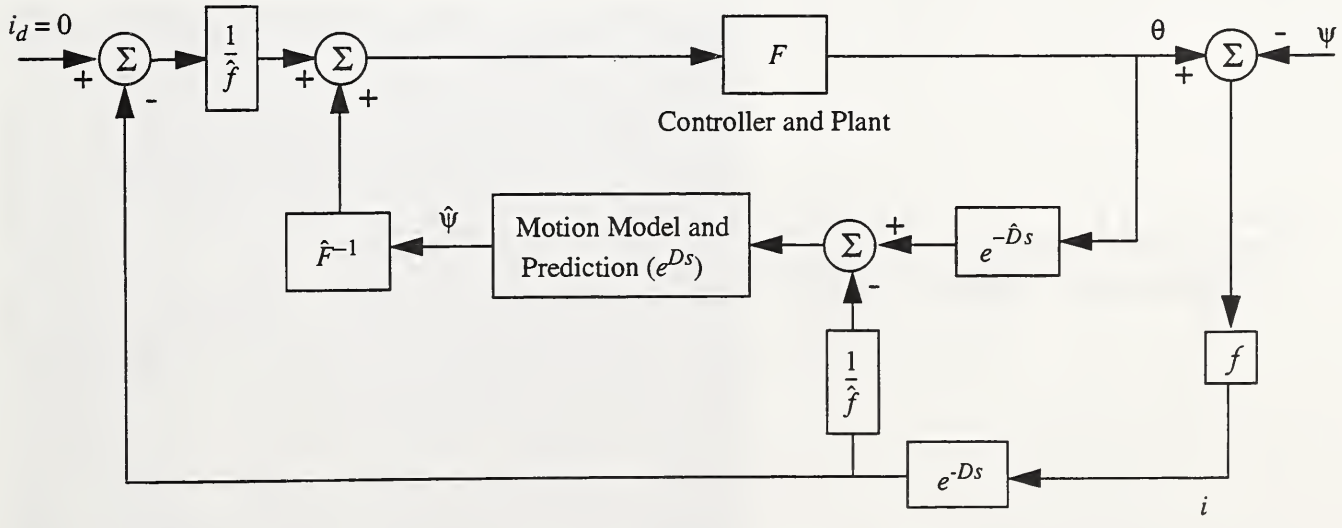

Figure 12. Target-Selective Adaptive Control pursuit system model.

can be obtained as

$$\frac{\left(-e^{-Ds}\right)i}{f}\left[F + e^{Ds}\right] = 0 \tag{26}$$

The tracking error is zero if and only if the inverse model $\hat{F}^{-1}$ and prediction are completely correct. In practice, this is not possible to achieve, and some tracking error results.

Note that the outer loop of the system is essentially an image-based servo, which includes the visual processing delay, as in Figure 4. This loop could be implemented by

$$F = \frac{K_p}{Ms^2 + K_v s} \tag{27}$$

such that the combination of prediction and $\hat{F}^{-1}$ is a feedforward signal of predicted object velocity and acceleration, scaled by the gain K_p. Obviously, it is even more difficult to predict future accelerations and velocities than it is to predict future object position, since velocities and accelerations are more sensitive to noise. As sinusoidal object motion increases in frequency, the prediction will be less accurate and the tracking error will increase (as shown in Figure 10 and Figure 11).

Consider that feedforward predicted velocity and acceleration could be used in combination with predicted position as in Figure 13. Analyzing the tracking error for this system yields a similar result to that for Figure 12.

$$i\left[Ms^2 + K_v s + K_p\right] = 0 \tag{28}$$

Thus, the tracking error i should be zero provided the prediction and model are completely accurate.

4. Experimental Results

To compare the performance of the different pursuit systems, with and without prediction, experiments were conducted using the NIST robot head, TRICLOPS (Fiala et al. 1993), shown in Figure 14. The TRICLOPS device points a stereo pair of high-resolution, narrow field-of-view (f = 15 mm) cameras. There is an independently-controlled vertical (with respect to the camera) *vergence* axis for each of the cameras. A common tilt axis rotates both vergence axes

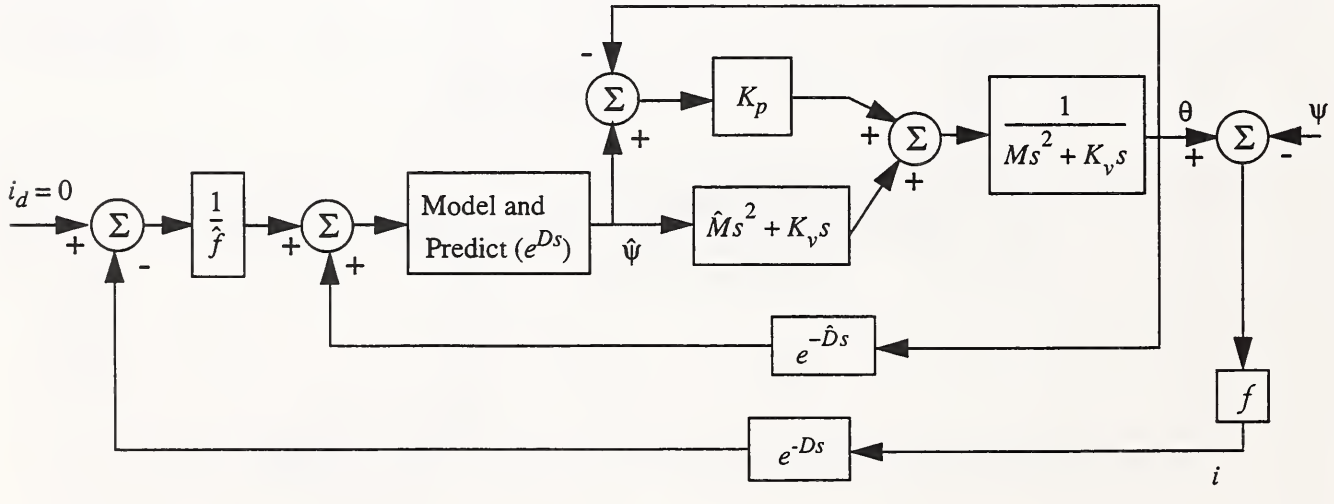

Figure 13. Pursuit system model with velocity and acceleration feedforward.

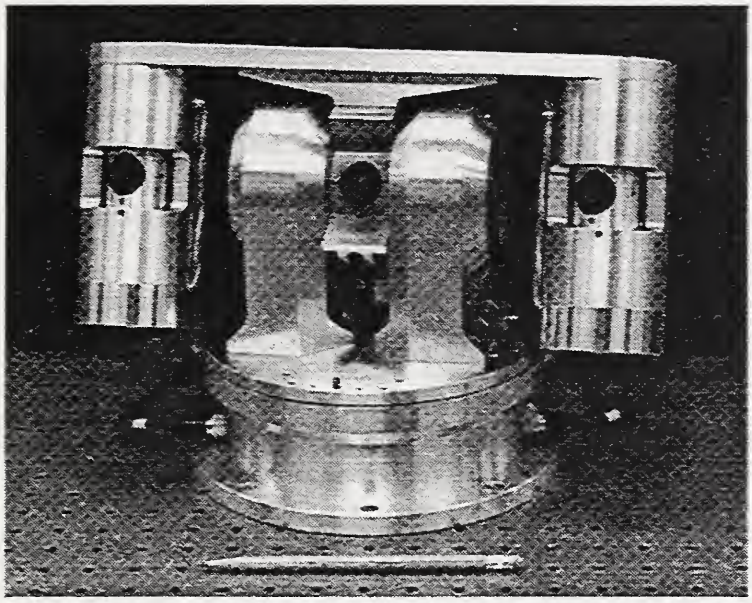

Figure 14. TRICLOPS active vision system.

up and down, and a fourth axis, the neck, pans the entire head. TRICLOPS is capable of high dynamic performance. A hierarchical multiprocessor control system based on the NASA/NBS Standard Reference Model for Telerobot Control System Architecture (NASREM) (Albus et. al. 1987) has been implemented for use with TRICLOPS. Further details concerning the control system and performance specifications of TRICLOPS may be found in Fiala et al. (1993). The joint position controller (the minimum phase controller of Figure 7) has a sample rate of 2 kHz, and easily realizes position amplitude bandwidths of 50 Hz on the vergence axes, 12.5 Hz on the tilt axis, and 4 Hz on the neck pan axis.

In the experiments, a target (small ball) is rotated about an axis parallel to the neck axis located 1.09 m from the neck axis in front of TRICLOPS, as shown in Figure 15. The circle described by the ball rotation is 0.88 m in diameter, which produces object motion encompassing about 44 degrees of the visual field on each vergence camera. Since each vergence camera has a field-of-view of only about 22 degrees, the vergence axes must track the object motion in order to keep it in the image. The goal of the tracking task is to keep the object in the center of the camera image. The movement of the centroid of the object in the image is the tracking error. In the experiments described here, only the two vergence axes and the tilt axis are used in tracking (the base rotation is stationary).

Image processing for the experiments is performed by a NASREM Level 1 perception system using a Parallel Image Processing Engine (PIPE) (Fiala et al. 1993). The pursuit system controller receives visual feedback in the form

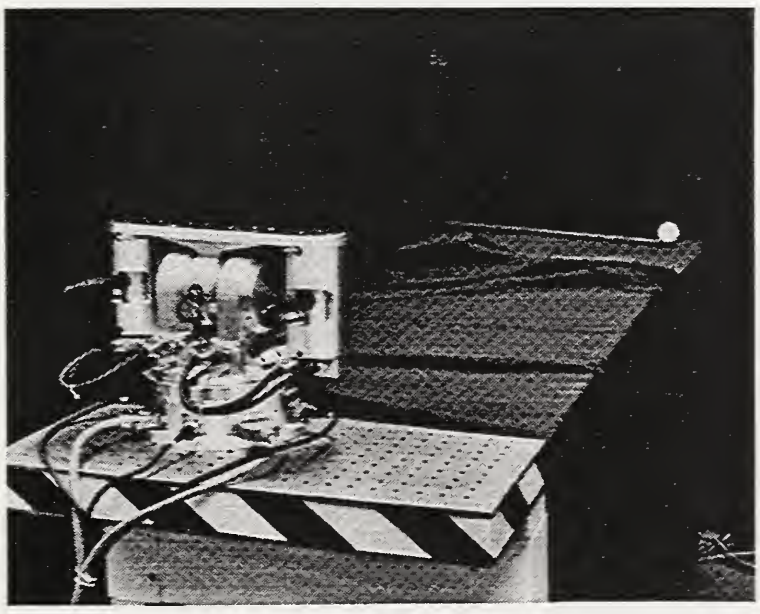

Figure 15. Experimental set-up for circular tracking.

of the centroids of the object in the left and right vergence cameras 30 times a second. This sampling rate of 30 Hz is determined by the processing of NTSC signals (60 Hz) for odd fields only. The latency associated with the centroid data is about 60 ms.With respect to the pursuit system models given above, this implies D = 60 ms. An interpolator is used to smooth the 30 Hz target position updates into 2 kHz servo command updates. Adding 30 ms to cover most of the servo interpolation time gives a total prediction time of about 90 ms (2.7 cycles). A third order, 12-sample LMSF filter is used for prediction. The frequency domain characteristics of this predictive filter were given previously in Figure 10 and Figure 11. The predictive filter produces predicted target position estimates corresponding to the visual feedback data at a rate of 30 Hz.

Our observations (both in simulations and on the real system) have indicated that overestimation of the image processing delay will result in instability if the full amount of prediction is used. However, if the prediction is decreased and the delay error is small, stability can often still be achieved. The necessity of reducing prediction to maintain stability has also been noted by other authors (Brown, Coombs, and Soong 1992). Since accurate estimation of the visual delay is so critical to the stability of the system, we have implemented a mechanism by which time stamps are used to measure the delay for each sample, rather than relying on an average empirically-determined or calculated value. This measured delay is then used to retrieve the corresponding delayed joint values (to a resolution of 0.0005 s) from a queue of previous joint positions maintained in the world model. Using this mechanism for delay compensation, the system does in fact remain stable even if more than the required amount of prediction is applied.

Figure 16 shows the tracking errors for the pursuit systems of Figure 6, without prediction, and Figure 9, with prediction. The plots are of the horizontal position of the centroid in the right vergence camera over time on two different runs. The image size is 256 pixels. The target starts at rest and then accelerates to a constant speed of approximately 1.14 rad/s (0.182 Hz). The peaks in Figure 16 occur as the tracking speed of the cameras increases when the target passes close to TRICLOPS. In this experiment, prediction is seen to reduce the tracking error by about a factor of 8. The cost of the reduced tracking error is that noise is increased somewhat by the predictive filter. Analysis of the commanded and feedback values of the joint positions indicates that about half of the remaining errors in the tracking with prediction may be attributed to servo following error.

One of the factors which affects the prediction accuracy is the coordinate system used for modeling the 3D motion. Motion modeling and prediction can be done in either joint space or Cartesian space, or with respect to another coordinate system. The primary requirement is that there be a valid trend in the position data which can be used to provide a reasonable extrapolation. So, for example, difficulties are encountered if prediction is attempted using image coordinates alone. After all, if the system is tracking well, then target will remain centered in the field of view, and the image coordinates will convey no information about how the target and/or observer is moving. The trend of the target motion through space must be captured. Although there are many choices for coordinate systems to use, it must be kept in mind that nonlinear transformations from one coordinate system to another affect the frequency content of a

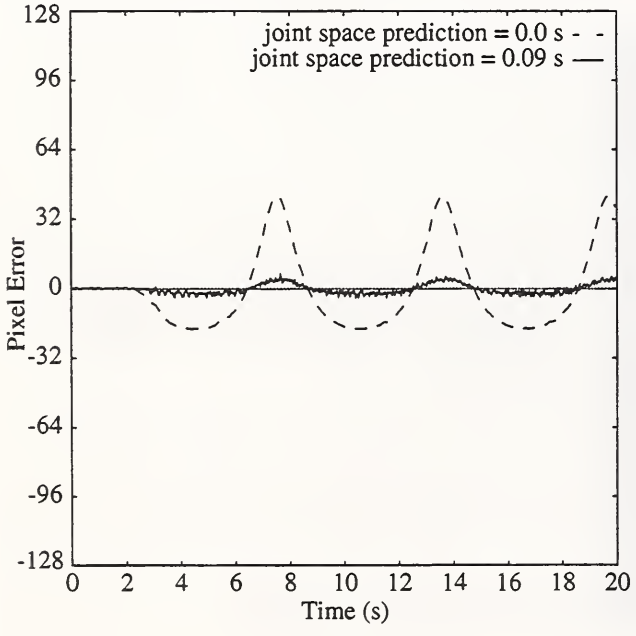

Figure 16. Tracking error for system with and without joint space prediction (target speed = 1.14 rad/s).

motion signal. Because the predictive filter accuracy decreases as the frequency of the input increases (Figure 10 and Figure 11), it is desirable to model the motion in the coordinate system that results in the fewest high-frequency components, if possible.

For the plots shown in Figure 16, only information from the right camera is used to point the camera—it is a monocular tracking algorithm. In this case, the prediction is performed on the time sequence of commanded joint angles (which should ideally result in the camera being pointed directly at the target) to estimate where the target will be. Since the target motion is not a pure sinusoid in joint space (it is sinusoidal in Cartesian space), the frequency spectrum of the joint motion signals contains an extra peak at a frequency higher than the primary peak at the motion frequency. When the target motion is only 1.14 rad/s, this extra peak is well within the region of accurate response of the cubic LMSF predictive filter. For faster target motions, however, the extra peak in the frequency spectrum moves into the region of the filter response where overshoot occurs.

This additional high frequency content causes increased prediction errors, because of the frequency characteristics of the predictive filter. For example, the frequency content of the joint motion required to track the target at a speed of 4.46 rad/s (0.71 Hz) is shown in Figure 17. At this target speed, the additional peak in the frequency spectrum occurs at about 1.45 Hz. The magnitude response of the predictive filter at this frequency is about 1.5. This predictive overshoot results in greatly increased visual tracking errors for the target speed of 4.46 rad/s. The tracking error result for this speed using joint space prediction and Cartesian space prediction are shown in Figure 18. To do the Cartesian space prediction, a triangulation algorithm is used (with two cameras) to estimate the 3D target position with respect to a fixed Cartesian coordinate system. The triangulation tracking algorithm with Cartesian space motion modeling and prediction can track the target motion along this path for much higher speeds—up to 6.9 rad/s. For the target speed of 4.46 rad/s and Cartesian space prediction, the peak rotational velocity of the camera is about 2.7 rad/s (155°/s)—significantly faster than the human capability. The peak error in this case is only about 1.3°.

Using Cartesian space prediction, there is only one peak in the frequency spectrum of the target motion signal, which is within the frequency range for accurate prediction by the filter. Therefore, as shown in Figure 18, the tracking errors are much smaller using prediction in Cartesian space (for this particular target trajectory). The same predictive filter with the same parameters was used for both the Cartesian space and the joint space motion modeling. Furthermore, if the triangulation tracking is used with joint space prediction, the results are virtually identical to those obtained with a single camera and joint space prediction (the solid line in Figure 18). Of course, if the target is moving sinusoidally in both Cartesian and joint space, then the two predictions should do equally well. And if the motion is sinusoidal in joint space, but not in Cartesian space, then the joint space prediction would perform better. In general, it is more likely that targets will be moving harmonically in Cartesian space than in the joint space of the observing device, so Cartesian space modeling may be more appropriate in most circumstances.

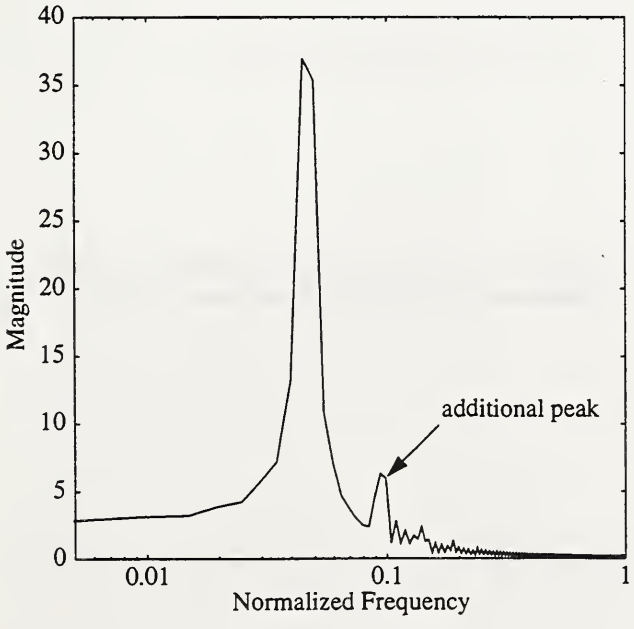

Figure 17. Frequency spectrum of joint motion required to track 4.46 rad/s circular target motion.

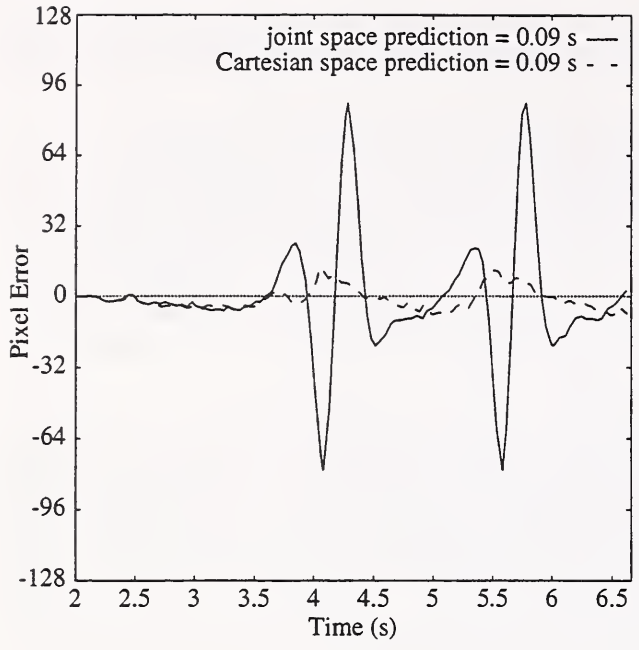

Figure 18. Tracking error for system with joint space prediction and with Cartesian space prediction (target motion = 4.46 rad/s).

3D Cartesian space motion modeling and prediction can also be performed with a single camera, provided that some method is available for estimating the range of the target. One possibility for doing this with the simple spherical target is to estimate the range of the target using

$$r = f\sqrt{A_r / A_i} \tag{29}$$

where r = range to target (mm), f = focal length (mm), A_r = actual area of target (mm^2), and A_i = area of target on CCD array (mm^2). Unfortunately, the 3D position estimate obtained using this technique is much noisier than that which is obtained using triangulation. Therefore, prediction using this estimate is not as effective. This is illustrated in Figure 19, which shows the tracking error for a slow target motion. For motions as fast as 4.46 rad/s, the increased noise causes the target to be easily lost from the field of view. So, with a single camera, it is better to use joint space prediction using the high-quality centroid position information for tracking than it is to use a single-camera 3D tracking approach. Of course, if a higher-fidelity approach to monocular range estimation (as good as triangulation) is identi-

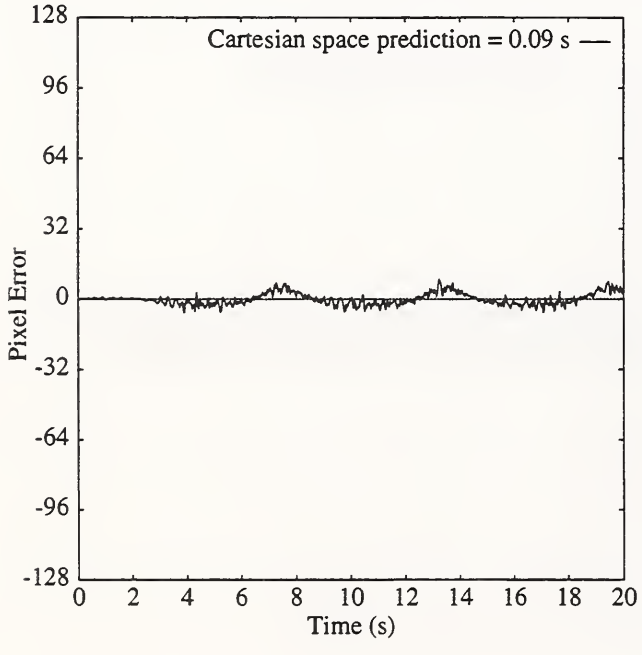

Figure 19. Tracking error for system with single-camera tracking with Cartesian space prediction.

fied, the single-camera 3D approach will improve accordingly.

Figure 20 shows the tracking performance of the TSAC-based system of Figure 12, with F as given previously. Target motion is again 1.14 rad/s. Errors in the simplified plant model affect the cancellation required for zero error, and the system is more oscillatory than before. As the speed of the object is increased the tracking error increases rapidly, due to increased errors in prediction. At an object speed of 3.9 rad/s, the tracking error becomes greater than the field-of-view of the camera. If a position error term is added to the system, the oscillations go away, and the tracking error is as shown in Figure 21 (slightly improved over the standard system of Figure 9).

Figure 22 shows the tracking error for the pursuit system with velocity and acceleration feedforward terms (Figure 13), at an object speed of 1.14 rad/s. Again, the tracking error is non-zero due to errors in modeling and prediction. In addition, the tracking error is almost as great as for the system without the feedforward part. The prediction of object position dominates the controller. This system is capable of maintaining the object in the field-of-view for speeds up to 5.2 rad/s.

The magnitude ratio of the tracking error as a function of frequency is shown in Figure 23. These plots were cre-

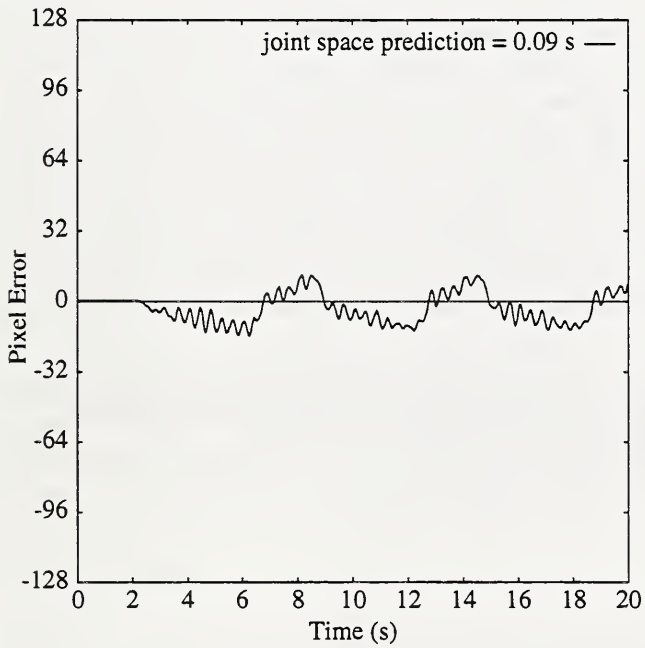

Figure 20. Tracking error for Target-Selective Adaptive Control system (system of Figure 12).

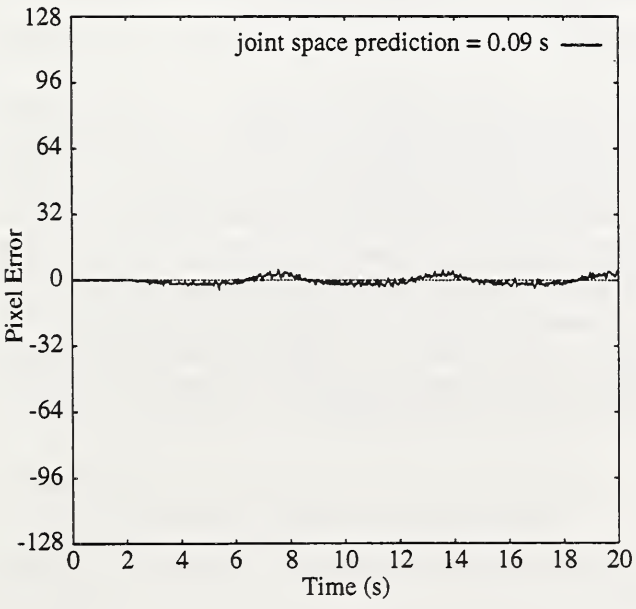

Figure 21. Tracking error for Target-Selective Adaptive Control system with position error term.

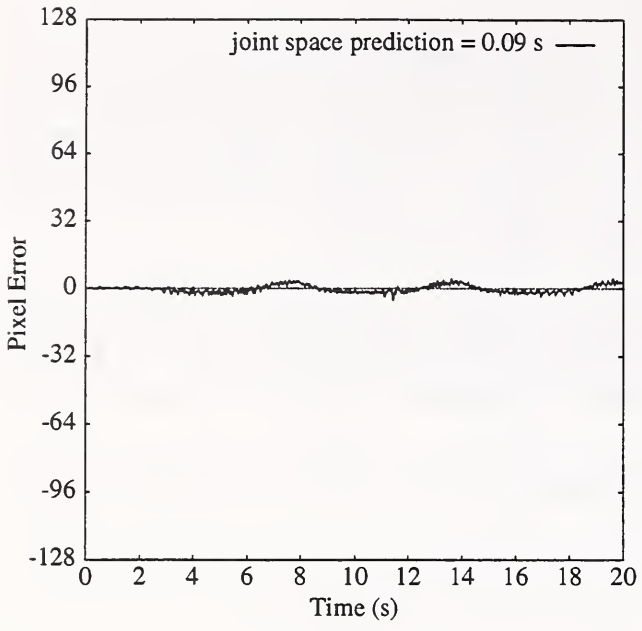

Figure 22. Tracking error with velocity and acceleration feedforward (system of Figure 13).

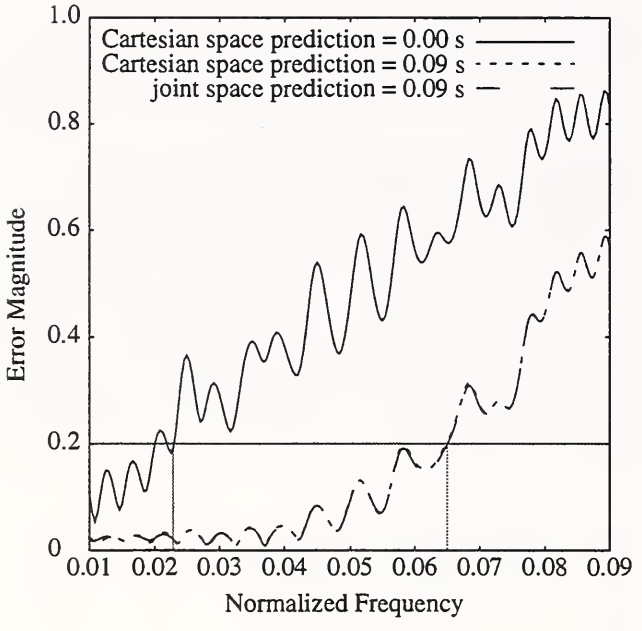

Figure 23. Experimental tracking error bandwidth results.

ated by commanding sinusoidal motion of the neck pan axis (+/- 5 deg) while tracking a stationary target at (-0.3302, 0.0, 0.7112 m). The frequency of the pan axis motion was increased linearly from 0 to 1.5 Hz in 20 s for each plot. Information about the base position is not used in the visual tracking algorithms. Image error information was recorded from the right camera. Here it is seen that the bandwidth (as defined previously) of the system without prediction agrees closely with the analytically-determined value of 0.34 Hz (normalized frequency = 0.023). It is also clear that the addition of prediction increases the tracking bandwidth by nearly a factor of three. Since the sinusoidal reference motion is very small, the target "motion" relative to the rotating camera platform is very nearly sinusoidal in both Cartesian space and joint space. This is born out by the fact that the bandwidth curves for Cartesian space prediction and joint space prediction in Figure 23 are virtually indistinguishable. The bandwidth plots for the tracking schemes of Figure 12 and Figure 13 are also similar to the plots for the other systems with prediction.

5. Conclusions

Several different approaches to visual pursuit tracking have been analyzed and investigated through experimentation with an active vision system. It is seen that the image processing delay is a primary factor in determining the limits of performance in terms of being able to track a rapidly-moving target. Therefore, the level of tracking performance is established by the effectiveness of the technique used to compensate for this delay. One of the factors which affects the accuracy of the prediction is the coordinate system in which the modeling and prediction of target motion is performed. Due to the frequency response characteristics of predictive filters, it is desirable to perform modeling and prediction in the coordinate system which results in keeping the frequency content of the motion signals confined to the lowest frequencies possible. For example, target motions which are harmonic in some coordinate system can be best predicted if that coordinate system is used for motion modeling. Conversion to other coordinate systems may introduce higher frequency components which are beyond the effective range of the predictive filter. Since target motion is more likely to be harmonic in Cartesian space than in a more specialized coordinate system (for example, joint coordinates), Cartesian coordinates are typically a good choice. Another factor which has a great effect on the performance of predictive filters is the amount of noise in the target motion signal. Again, this is due to the frequency response characteristics of predictive filters, which tend to significantly amplify high-frequency noise.

Since prediction is the limiting factor, the decision to use Cartesian space or joint space motion modeling, and whether to use single camera or multiple camera position determination techniques, is best made by examining the effects of these choices on the prediction. For example, if Cartesian space motion modeling and prediction is deemed appropriate (based on the expected target motions), and a two-camera triangulation algorithm provides less noisy Cartesian position estimates than a monocular position determination algorithm, then the two-camera system will provide the best performance. Beyond the effectiveness of the prediction, anything that can be done to make the trajectory following performance of the sensor pointing system as close to ideal as possible will improve the visual tracking performance. This is primarily achieved by high-bandwidth local position servos, but it is seen that such measures as incorporating velocity and acceleration feedforward terms provide some marginal improvement.

It has been experimentally demonstrated that, when the frequency content of the target motion signal is within the accurate frequency range of the predictive filter, very accurate tracking ($< 1.5°$ peak error) can be achieved for camera velocities up to 2.7 rad/s, whereas the maximum human eye velocity during smooth pursuit is about 0.52 rad/s. Using simple tracking algorithms with an appropriate predictive filter and an advanced robotic device like TRICLOPS, it is therefore now possible to outperform humans in the domain of visual pursuit of simple targets.

6. References

Albus, J. S., McCain, H. G., Lumia, R. 1987. NASA/NBS Standard Reference Model for Telerobot Control System Architecture (NASREM). National Inst. Standards and Technol., Gaithersburg, MD, Technical Note 1235.

Bahill, T. A., McDonald, J. D. 1981. Adaptive control model for saccadic and smooth pursuit eye movements. In *Progress in Oculomotor Research*. Fuch, A. F., Becker, W., eds. New York: Elsevier.

Bar-Shalom, Y., Fortmann, T. E. 1988. *Tracking and Data Association*. Boston: Academic Press.

Brown, C., Coombs, D., and Soong, J. 1992. Real-time smooth pursuit tracking. In *Active Vision*. Blake, A. and Yuille, A., eds. Cambridge, MA: MIT Press, pp. 123-136.

Brown, C., Coombs, D. J. 1991. Notes on control with delay. Technical Report 387. Rochester, NY: University of Rochester, Computer Science Department.

Brown, C. 1990. Gaze control with interactions and delays. *IEEE Trans. Sys., Man, & Cybern.* 20(1): 61-70.

Clark, J. J., Ferrier, N. J. 1988. Modal control of an attentive vision system. *IEEE Intl. Conf. on Computer Vision*, Tampa, FL, pp. 514-523.

Fiala, J. C., Lumia, R., Roberts, K. J., Wavering, A. J. 1993. TRICLOPS: A tool for studying active vision. *International Journal of Computer Vision* (to appear).

Harvey, D. R., Bahill, A. T. 1985. Development and sensitivity analysis of adaptive predictor for human eye movement

model. *Trans. Society for Computer Simulations* 2(4):275-292.

Levine, M. D. 1985. *Vision in Man and Machine.* New York: McGraw-Hill.

Ng, L. C. and LaTourette, R. A. 1983. Equivalent bandwidth of a general class of polynomial smoothers. *J. Acoust. Soc. Am.* 74(3):814-826.

Robinson, D. A. 1988. Why visuomotor systems don't like negative feedback and how they avoid it. In *Vision, Brain, and Cooperative Computation*, Arbib, M. A., Hanson, A. R., eds. Cambridge, MA: MIT Press, pp. 88-107.

Sharkey, P. M. and Murray, D. W. 1993. Coping with delays for real-time gaze control (The fall and rise of the Smith "Predictor"). *SPIE Sensor Fusion VI*, SPIE Vol. 2059, Boston, pp. 292-304.

Smith, O. J. M. 1957. Closer control of loops with dead time. *Chem. Eng. Prog.* 53(5):217-219.

Stark, L., Vossius, G., Young, L. R. 1962. Predictive control of eye tracking movements. *IRE Trans.* HFE-3: pp. 52-57.

Wavering, A. J., Lumia, R. 1993. Predictive visual tracking. *Proc. Intelligent Robots and Computer Vision XII: Active Vision and 3D Methods*, SPIE Vol. 2056, Boston, pp. 86-97.

Yarbus, A. L. 1967. *Eye Movements and Vision.* New York: Plnum Press.

Young, L. R. 1971. Pursuit eye tracking movements. In *Control of Eye Movements*, Bach-y-Rita, P., Collins, C. C., Hyde, J. E., eds. Boston: Academic Press, pp. 429-443.