Proc. Second Asian Conference on Computer Vision, ACCV'95, Singapore, December 5-8, 1995, and in *Recent Developments in Computer Vision*, S. Z. Li, D. P. Mital, E. K. Teoh, and H. Wang, eds., Springer-Verlag Lecture Notes in Computer Science 1035, 1996.

High-Performance Tracking with TRICLOPS

Albert J. Wavering[1], Henry Schneiderman[1,2], John C. Fiala[3]

[1] National Institute of Standards and Technology, Intelligent Systems Division
Gaithersburg, MD 20899
wavering@cme.nist.gov
[2] Carnegie Mellon University, Robotics Institute
Pittsburgh, PA 15213
[3] Boston University, Department of Cognitive and Neural Systems
Boston, MA 02215

Abstract. This article reviews the development of experimental visual tracking algorithms which have been implemented using TRICLOPS (The Real-time, Intelligently-ControLled Optical Positioning System), a high-performance active vision system designed and built at the National Institute of Standards and Technology (NIST). These algorithms range from triangulation using simple targets for very high-speed tracking, to model-based techniques which allow robust tracking of polyhedral targets amidst cluttered backgrounds. In all cases, effective prediction of the target motion to compensate for image processing delays is critical to achieving responsive tracking.

1 Introduction

The goal of visual tracking is to keep a moving target centered in the field of view of one or more cameras. This capability is useful for such applications as robotic assembly, surveillance, and autonomous driving/navigation. In order to support visual tracking and other active vision research at NIST, a trinocular robot head, called TRICLOPS, was designed and built [1]. This head has been used for extensive experimentation with a variety of tracking algorithms.

In this paper, we will first summarize the design and performance characteristics of the TRICLOPS active vision system. This is followed by a discussion of tracking algorithms which have been developed to perform high-speed tracking of simple visual targets. Finally, a brief synopsis of work performed using TRICLOPS for model-based target tracking in a complex visual environment is presented.

2 TRICLOPS Active Vision System

One of the primary design goals for TRICLOPS was to build a system with very high dynamic performance—on a par with the human oculomotor system. This emphasis on dynamic performance, along with other requirements such as high accuracy and durability, lead to the development of TRICLOPS as the first direct-drive robot head (Figure 1). The frameless DC motors and resolvers mounted directly to each of the motion

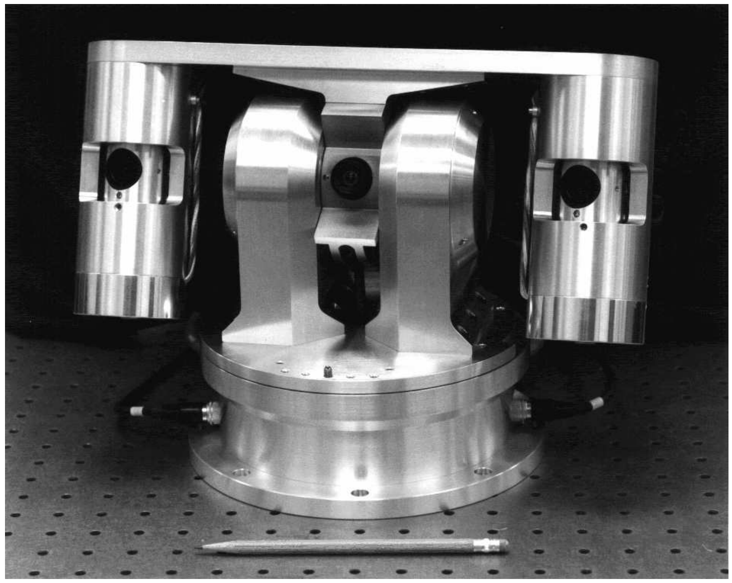

Fig. 1. TRICLOPS active vision system.

axes deliver large accelerations with low friction. Transmission backlash and compliance, which cause errors and oscillations, are effectively eliminated.

TRICLOPS has four mechanical degrees of freedom, as shown in Figure 2. The kinematic arrangement of the four axes is as follows: 1) pan (or base rotation) about a vertical axis through the center of the base, 2) tilt about a horizontal line that intersects the base rotation axis, and 3) left and right vergence axes which intersect and are perpendicular to the tilt axis.

Visual sensing for TRICLOPS is provided by three "micro-miniature" video cameras; one in each of two independently-controlled vergence units, and a third center camera which shares the base pan and tilt motions of the vergence cameras. Equipped with a wide angle (3-4 mm focal length) lens, this color camera may be used to identify general areas of interest in the wide field of view. The longer focal length lenses (15-24 mm) of the vergence cameras allow a more detailed look at specific objects or features. The combination of the center and vergence cameras thus provides the capability to view a scene at multiple resolutions simultaneously, providing a (much simplified) version of the multiresolution foveal-peripheral vision found in many biological visual systems. TRICLOPS is currently configured with a 4 mm center lens and 15 mm vergence lenses.

The performance specifications of TRICLOPS are summarized in Table 1. Further details regarding the features, design, and construction of TRICLOPS may be found in [1].

A multiprocessing controller based on the NASREM Architecture [2] has been constructed to control TRICLOPS and to perform sensory processing and world modeling operations on camera image data. The NASREM Architecture is a hierarchical

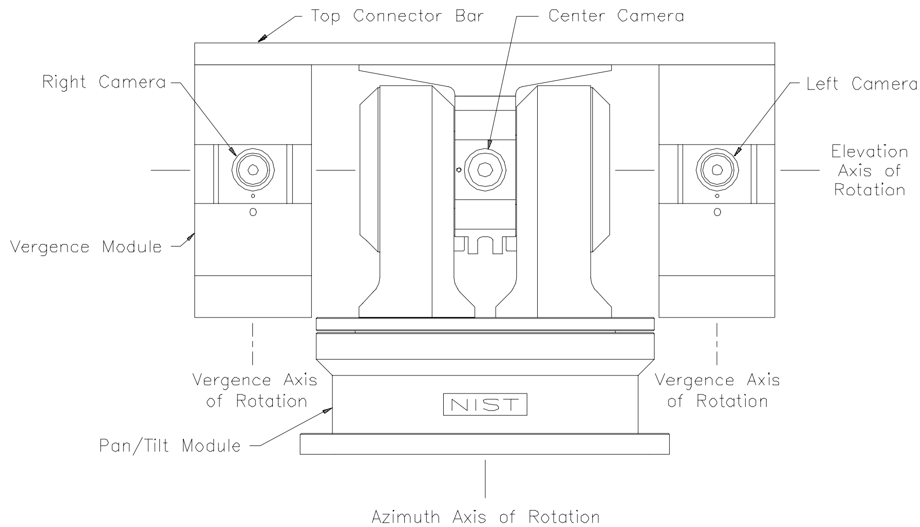

Fig. 2. TRICLOPS degrees of freedom.

Table 1: TRICLOPS Specifications.

Range of motion: Pan Tilt Vergence	 +/-1.68 rad (+/-96.3 deg) 0.48 rad down, 1.14 rad up (+27.5 deg, -65.3 deg) +/-0.77 rad (+/-44 deg)
Peak acceleration: Pan Tilt Vergence	 70 rad/s^2 (4010 deg/s^2) 320 rad/s^2 (18,300 deg/s^2) 1100 rad/s^2 (63,000 deg/s^2)
Peak velocity: Pan Tilt Vergence	 11.5 rad/s (660 deg/s) 17.5 rad/s (1000 deg/s) 32 rad/s (1830 deg/s)
Slew time: Pan (3.14 rad move) Tilt (1.55 rad move) Vergence (1.5 rad move)	 0.523 s 0.167 s 0.091 s

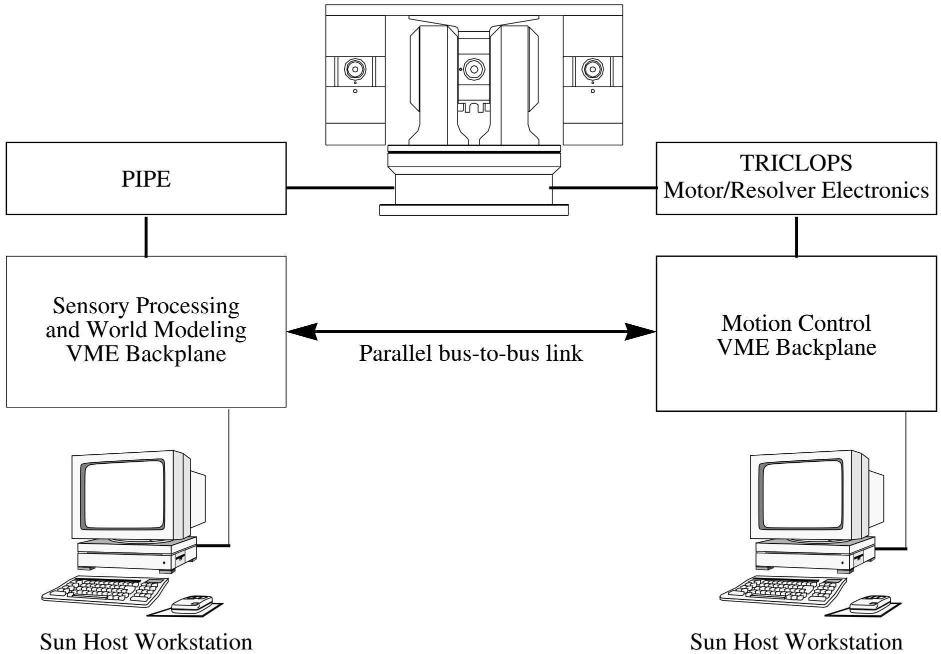

Fig. 3. TRICLOPS control and image processing hardware.

control system architecture which divides the sensing and control problem into discrete levels. At the lowest level are the servo control loops for the device actuators; at higher levels, groups of sensors and actuators are coordinated as equipment subsystems. The NASREM hierarchical control system provides a modular, easily-modifiable infrastructure for implementing and testing active vision algorithms.

The processes and operation of the TRICLOPS motion control and image processing systems are described in [1]; a brief summary is presented here. The full potential of an active vision system cannot be realized without real-time image processing capability. In our lab, this capability is provided by a Pipelined Image Processing Engine (PIPE)[1] for low-level image processing, along with a multiple-cpu VME system for higher-level sensory processing and world modeling (Figure 3). The motion control portion of the TRICLOPS control system architecture consists of a Primitive Level for producing coordinated motions of all axes, and a Servo Level for servoing the axes to commanded joint positions. The Primitive and Servo processes are implemented on five processor boards in a separate motion control VME system, also shown in Figure 3. Taking advantage of the most recent advances in high-speed microprocessors

1. Commercial equipment and materials are identified in this paper in order to adequately specify the experimental procedure. Such Identification does not imply recommendation or endorsement by NIST, nor does it imply that the materials or equipment identified are necessarily the best for the purpose.

and digital signal processing (DSP) hardware would result in a substantial reduction in the number of processor boards required for the system.

3 Fast Tracking of Simple Targets

One of the most basic behaviors desired of an active vision system is to track an object, using visual information to direct the gaze and keep the object in the field of view of the cameras. In this section, we discuss tracking algorithms which emphasize maximizing the tracking performance in terms of speed (bandwidth). Further details regarding the algorithms and experimental results presented in this section may be found in [1], [3], and [4].

To perform high-speed tracking, image processing delays must be minimized as much as possible. Consequently, the image processing used for these tracking experiments consists of computing the centroid of a thresholded intensity image of a single object which presents a high-contrast image to both the left and right vergence cameras. For these experiments, only the two vergence cameras are used. To create a simple visual scene, cellophane film is used to block out visible light, along with an incandescently-illuminated spherical target. The infrared emission of the incandescent source passes through the filter, while the ambient fluorescent lighting is filtered out, resulting in a nearly ideal image.

For this high-speed tracking, 3-D world position-based modeling and prediction of the target motion is used to generate goal fixation points. A block diagram of the tracking control system is presented in Figure 4. New centroid data are obtained from the world model every 1/30 s. Delayed joint positions which correspond in time to the centroid data are also read in at this time. The amount of delay is equivalent to the image acquisition and processing latency (about 0.084 s in this case), plus the effective delay resulting from interpolation between target position updates (0.033 s). The centroid information is corrected for lens distortion, and the corrected centroids are used along with the delayed joint positions to compute the position of the object with respect to the TRICLOPS world reference frame using a triangulation algorithm.

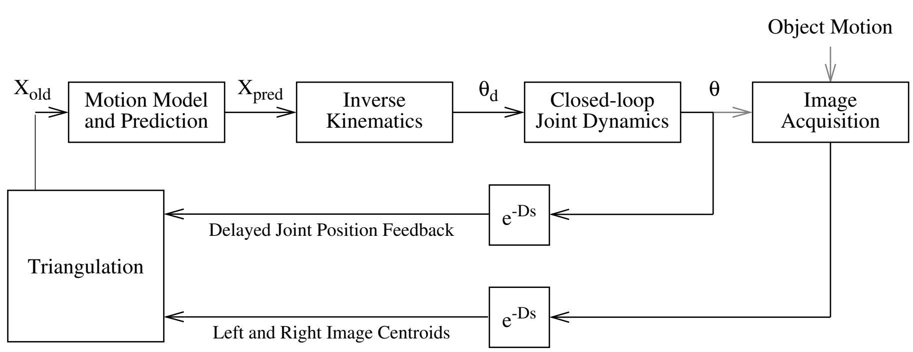

Fig. 4. Block diagram of tracking algorithm.

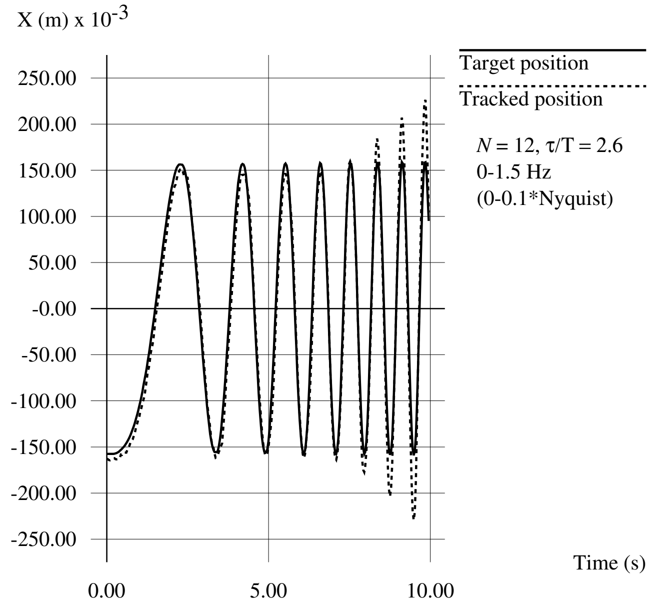

Fig. 5. Tracking performance with cubic LMSF filter.

A number of different types of predictive filters have been used for motion modeling and prediction, including constant coefficient Kalman (α–β–γ) filters and polynomial least-mean-square-fit (LMSF) filters. The equivalent transfer function forms of these filters are derived, and their frequency response characteristics are discussed, in [3].

The tracking performance of TRICLOPS using the system of Figure 4 with a 12-sample cubic LMSF predictive filter is shown in Figure 5. To obtain these tracking results, the redundancy of TRICLOPS (4 degrees of freedom for a 3 degree-of-freedom pointing task) was used to create apparent motion from a stationary target. A fixed target was visually tracked using the tilt and vergence axes while a sinusoidal motion of the TRICLOPS base rotation axis was performed. The frequency of the base motion was increased linearly with time, and the magnitude of the commanded base motion was +/-0.175 rad (+/-10 deg). Figure 5 shows the actual horizontal position of the target and the corresponding tracked position (with respect to the moving base coordinate system).

The tracking performance exhibited in Figure 5 is determined almost entirely by the frequency response of the predictive filter. This filter allows tracking without overshoot up to about 1 Hz. Beyond this frequency, the rise in the magnitude response of the predictive filter results in tracking errors. The 40% overshoot in the tracking response at 1.5 Hz is attributable to the fact that the magnitude ratio of the filter is 1.4 at that frequency. The system is clearly having difficulty tracking at 1.5 Hz, and the target is lost from the field of view of the cameras at higher frequencies.

A method of tracking higher frequency predictable motions without overshoot or

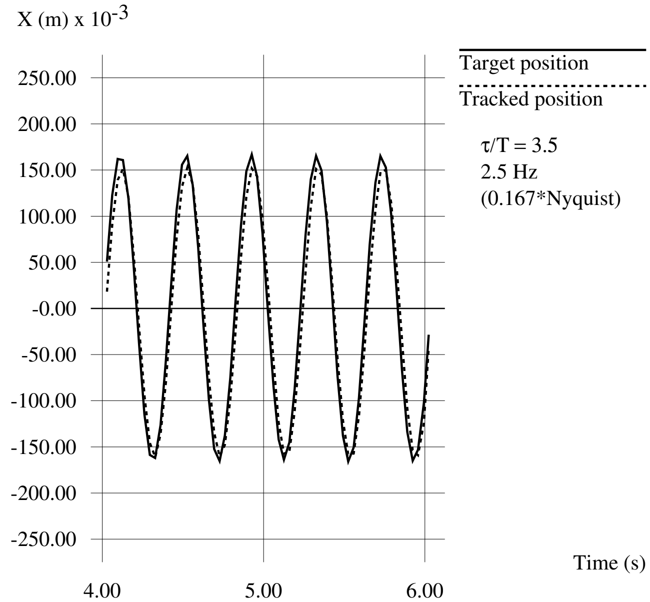

Fig. 6. Tracking performance of correlation prediction.

latency based on autocorrelation of the target signal is discussed in [3]. This method has been used to track sinusoidal target motions of 2.5 Hz and more. When the frequency of the target motion is held constant, the tracking errors with this approach are very small, even at high frequencies. An example of the tracking performance using this approach is shown in Figure 6. The correlation prediction technique makes no assumptions of the structure of the target motion signal—it should be effective for arbitrary periodic motions. However, the frequency content and amplitude of the target motion must change relatively slowly over time.

We have also combined a conventional predictive filter with the periodic motion prediction, such that both random motions with low frequency content and higher-frequency predictable motions can be tracked successfully with automatic switching between the two techniques. This has worked reasonably well, although transitioning between the two techniques successfully requires gradual changes in the target motion trajectory.

Numerous other algorithms for visual tracking have also been implemented using TRICLOPS, and are discussed in [4]. A quantitative measure of tracking performance, the tracking error bandwidth, is defined in [4] as the frequency of sinusoidal object motion at which the tracking error exceeds 20% of the object motion amplitude:

$$\frac{|i(j\omega)|}{f|\psi(j\omega)|} = 0.20. \tag{1}$$

where f = focal length, $i(j\omega)$ = image coordinates of target, $\psi(j\omega)$ = target position angle.

Using the above definition, an experimental tracking error bandwidth of approximately 1.1 Hz was obtained with a predictive filter. Without prediction, the tracking error bandwidth is about 0.34 Hz.

4 Model-Based Tracking

Of course, the tracking of simple targets in simple scenes, while useful to demonstrate the maximum performance (in terms of speed) which can be achieved, has limited application potential. A more useful capability is to be able to track an object of interest in a natural visual environment. At NIST, work toward developing such robust tracking has focused on model-based tracking of known target objects. Here, established geometric relationships between object features are used to improve the robustness of tracking. The basic elements of this approach are outlined below; additional information may be found in [5]. The approach described below assumes a rigid 3-D object with known geometry and appearance is being used as a target, and that it moves with 3 translational degrees of freedom.

The successive stages of computation for this tracking approach are:

1. Edge extraction - Extract the position and orientation for all edge points.

2. Data association - Determine likely groupings of edge points to each model feature.

3. Feature measurement - Use the grouped edge points to determine location of each feature.

4. Feature aggregation - Determine the overall location of the target by fitting the target model to the conglomerate of computed feature locations. The variance in each feature measurement is taken into account in this step to obtain the best linear unbiased estimate (BLUE) of target location.

5. Motion model update - Use the computed object location to update parameters (e.g. velocity, acceleration) that describe how the target moves as a function of time. This combination of the target location estimates over time weights each observation by the inverse of its covariance.

6a. Prediction of target location in next image - Extrapolate the motion model to predict target location in next image. Use predicted target location to determine corresponding predicted feature locations.

6b. Prediction of target location at next servo cycle - Predict the target location such that the motions of the head will keep the image of object centered in the field of view.

This strategy has been used to demonstrate 3-D target tracking using TRICLOPS 5., and also for vehicle following 6.. In the tests using TRICLOPS, the system was able to track and servo on the target as it moved at velocities of up to 1 m/s at a distance of 0.85 m. The equivalent angular velocity for this tracking speed is about 1.2 rad/s (69 deg/s).

The algorithm has demonstrated promising results in tracking the target in the presence of partial occlusion. The algorithm also was not confused by any of the background scenery it encountered. An example of a challenging scene during tracking is shown in Figure 7, with the corresponding edge image shown in Figure 8. The algorithm is also fairly robust to deviations from the assumed fixed orientation of the tar-

Fig. 7. Target with bookshelf in background.

Fig. 8. Edge image of target with bookshelf in background

get. The algorithm begins to fail when there are errors greater than 15 degrees in orientation. The algorithm will also fail if the target is moved with very sudden motions.

Progress has been made to extend this algorithm to tracking objects moving with a full 6 degrees of freedom of motion. Currently, however, the 6 degree of freedom version is not as fast, or as robust, as the 3-D version.

5 Conclusion

There are many aspects to defining the performance of visual tracking devices and algorithms. The performance is perhaps best defined by the restrictions which must be placed on the target object(s) and visual environment in order to achieve successful tracking. Examples of such restrictions include the maximum speed and frequency content of target motion, the maximum number of targets which may be tracked, lighting and background constraints required, the amount of a priori knowledge of the target object required, and allowable variability in the types of objects which may be tracked.

We have implemented a class of visual tracking algorithms using TRICLOPS which have concentrated on optimizing the performance criteria of speed and target maneuverability (at the expense of reduced target and image complexity). Using this approach, very high tracking speeds have been obtained. We have also implemented model-based tracking algorithms which lift the constraints on image complexity and work for an entire class of prismatic objects with known geometry (with some reduction in the allowable target speed). In both cases, the primary factor affecting the speed at which targets may move and be tracked successfully is the amount of image processing delay, and the effectiveness of the prediction used to compensate for it.

References

1. Fiala, J. C., Lumia, R., Roberts, K. J., Wavering, A. J., "TRICLOPS: A Tool for Studying Active Vision," *International Journal of Computer Vision*, Vol. 12, no. 2-3, pp. 231-50, April, 1994.
2. Albus, J. S., McCain, H. G., Lumia, R, "NASA/NBS Standard Reference Model for Telerobot Control System Architecture (NASREM)," NIST Technical Note 1235, 1987.
3. Wavering, A. J., Lumia, R., "Predictive Visual Tracking," *Proc. SPIE Intelligent Robots and Computer Vision XII*, Boston, MA, September, 1993.
4. Fiala, J. C., Wavering, A. J., "On Visual Pursuit Systems," NIST Internal Report 5513, October, 1994.
5. Schneiderman, H., Wavering, A. J., Nashman, M., Lumia, R., "Real-Time Model-Based Visual Tracking," *Intelligent Robotic Systems '94*, Grenoble, France, July, 1994.
6. Schneiderman, H., Nashman, M., Wavering, A. J., Lumia, R., "Vision-based Robotic Convoy Driving," to appear in *Journal of Machine Vision and Applications*.