Presented at the 31st IEEE Conf.
on Decision and Control (CDC)
Tucson, Arizona, December 16-18, 199

MONOCULAR 3-D VISUAL TRACKING OF A MOVING TARGET BY AN EYE-IN-HAND ROBOTIC SYSTEM

*N. P. Papanikolopoulos**, *B. Nelson***, and *P. K. Khosla*****

*Department of Computer Science
University of Minnesota
200 Union St. SE
Minneapolis, MN 55455

**Department of Electrical and Computer Engineering
The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Abstract

This paper applies the framework of *controlled active vision* to the problem of monocular full 3-D robotic visual tracking (three translations and three rotations). In particular, it demonstrates full 3-D tracking of a moving target by a monocular hand-eye system. A single camera is used since we believe that the tracking motion of the monocular hand-eye system can be used to create virtual stereo images. A simple adaptive scheme is proposed and the relative distance of the target from the camera is assumed partially unknown. The number of parameters that must be estimated on-line is minimal resulting in a feasible real-time implementation of the scheme. Moreover, the strong coupling of the rotational and translational degrees of freedom is treated in a way that guarantees robust tracking of the object. The limitations of our approach are discussed and the results from the application of our scheme to the TROIKABOT system (a set of three PUMA560's manipulators) are presented.

1. Introduction

This paper deals with the problem of 3-D robotic visual tracking of targets whose motion consists of 3-D translational and rotational components. The visual tracking is accomplished through a camera mounted on the robot that computer the displacements of several features that belong to the target. These visual measurements are fed to an adaptive control algorithm that provides the inputs to a cartesian robot control scheme after each measurement Numerical issues related to the strong coupling of the rotational and translational degrees of freedom are treated in a way that guarantees tracking of the object. A single camera is used instead of a binocular system because one of our main objectives is to demonstrate that relatively unsophisticated off-the-self hardware can be used to solve the 3-D tracking problem if the proper modeling and control issues are addressed.

The major differences of our algorithms from similar research efforts [1, 2, 3, 4, 5] are the use of a single moving camera, the ability to compensate for inaccurate camera parameters and unknown depth (distance of the target with respect to the camera frame), the full 3-D tracking ability, the small number of parameters that are estimated on-line, and the integration of the characteristics of the motion detection algorithm into the mathematical model for tracking. These differences allow the use of the proposed algorithms in poorly calibrated spaces or in spaces that am difficult to calibrate, such as underwater, space, and nuclear sites. This paper extends our previous work [6, 7, 8] in controlled active vision by allowing tracking of full 3-D motion (translations and rotations) and by ducing the number of parameters that should be estimated on-line. Experimental results arc presented to show the strengths and the weaknesses of the proposed approach The experiments am performed on the TROIKABOT multi-robotic system which operates under the CHIMERA II real-time operating system. The TROIKABOT system consists of three PUMA560's. One PUMA carries the camera while another holds the target.

The organization of this paper is as follows: Section 2 describes the mathematical framework under which our problem is solved. The control, filtering, and estimation strategies are discussed in Section 3. The experimental results are presented in Section 4. Finally, in Section 5, the papa is summarized.

2. Modeling of the 3-D Robotic Visual Tracking Problem

This section describes the mathematical modeling of our problem. We assume a pinhole camera model with a frame R_c placed at the focal point of the lens. Consider a target with a feature located at a point P with coordinates (X_c, Y_c, Z_c) in R_c. Moreover, the camera moves with a translational velocity $\mathbf{T} = (T_x, T_y, T_z)^T$ and with an angular velocity $\mathbf{R} = (R_x, R_y, R_z)^T$ with respect to the camera frame R_c. Since the camera and the target are moving simultaneously, we can write (using the approach described in [9]) the following equations for one feature point (d is a delay factor ($d \in \{1, 2, \ldots\}$), f is the focal length of the camera, s_x, s_y am the dimensions (mm/pixel) of the camera's pixels, (x, y) are the image coordinates of the projection of the feature point P on the image plane, T is the sampling period, q^{-1} is the backward shift operator, and $u_o(k)$ and $v_o(k)$ are the components of the optical flow induced at the time instant k by the motion of the object):

$$\mathbf{x}_F(k+1) = \mathbf{A}_F(k)\mathbf{x}_F(k) + \mathbf{B}_F(k-d+1)\mathbf{u}(k-d+1) + \mathbf{E}_F(k)\mathbf{d}_F(k) + \mathbf{H}_F(k)\mathbf{v}_F(k) \quad (1)$$

where* $\mathbf{A}_F(k) = \mathbf{H}_F(k) = \mathbf{I}_2$, $\mathbf{E}_F(k) = T\mathbf{I}_2$, $\mathbf{x}_F(k) \in R^2$, $\mathbf{d}_F(k) \in R^2$, $\mathbf{u}(k) \in R^6$, and $\mathbf{v}_F(k) \in R^2$. The matrix $\mathbf{B}_F(k) \in R^{2\times 6}$ is:

$$\mathbf{B}_F(k) = T\begin{bmatrix} \frac{-f}{s_x Z_c(k)} & 0 & \frac{x(k)}{Z_c(k)} & \frac{x(k)y(k)s_x}{f} & \frac{-f^2 - (x(k)s_x)^2}{f s_y} & \frac{y(k)s_x}{s_y} \\ 0 & \frac{-f}{s_y Z_c(k)} & \frac{y(k)}{Z_c(k)} & \frac{f^2 + (y(k)s_y)^2}{f s_x} & \frac{-x(k)y(k)s_y}{f} & \frac{-x(k)s_y}{s_x} \end{bmatrix}$$

The vector $\mathbf{x}_F(k) = (x(k), y(k))^T$ is the state vector, $\mathbf{u}(k) = (T_x(k), T_y(k), T_z(k), R_x(k), R_y(k), R_z(k))^T$ is the control input vector, $\mathbf{d}_F(k) = (u_o(k), v_o(k))^T$ is the exogenous deterministic disturbances vector, and $\mathbf{v}_F(k) = (v_x(k), v_y(k))^T$ is the white noise vector. The measurement vector $\mathbf{y}_F(k) = (y_x(k), y_y(k))^T$ for this feature is given by:

$$\mathbf{y}_F(k) = \mathbf{C}_F\mathbf{x}_F(k) + \mathbf{w}_F(k) \quad (2)$$

where $\mathbf{w}_F(k) = (w_x(k), w_y(k))^T$ is a white noise vector ($\mathbf{w}_F(k) \sim N(0, W)$) and $\mathbf{C}_F = \mathbf{I}_2$. The measurement vector is computed using the SSD algorithm which is described in [9].

*The symbol $\mathbf{I}_n$ denotes the identity matrix of order n.

One feature point is not enough to determine the control input vector $\mathbf{u}(k)$. The reason is that the number of system outputs is less than the number of control inputs. Thus, we are obliged to consider more points in our model. In order to solve for the control input that will be sent to the manipulator, it can be shown that at least three non-collinear feature points are needed. The reason for the non-collinearity requirement is investigated in [10].

The state-space model for M ($M \geq 3$) feature points can be written as:

$$\mathbf{x}(k+1) = \mathbf{A}(k)\mathbf{x}(k) + \mathbf{B}(k-d+1)\mathbf{u}(k-d+1) + \mathbf{E}(k)\mathbf{d}(k) + \mathbf{H}(k)\mathbf{v}(k) \quad (3)$$

where $\mathbf{A}(k) = \mathbf{H}(k) = \mathbf{I}_{2M}$, $\mathbf{E}(k) = T\mathbf{I}_{2M}$, $\mathbf{x}(k) \in R^{2M}$, $\mathbf{d}(k) \in R^{2M}$, and $\mathbf{v}(k) \in R^{2M}$. The matrix $\mathbf{B}(k) \in R^{2M \times 6}$ is:

$$\mathbf{B}(k) = \begin{bmatrix} \mathbf{B}_F^{(1)}(k) \\ \cdots \\ \mathbf{B}_F^{(M)}(k) \end{bmatrix}.$$

The superscript (j) denotes each one of the feature points ($(j) \in \{(1), \ldots, (M)\}$). The vector $\mathbf{x}(k) = (x^{(1)}(k), y^{(1)}(k), \ldots, x^{(M)}(k), y^{(M)}(k))^T$ is the new state vector, and $\mathbf{v}(k) = (v_x^{(1)}(k), v_y^{(1)}(k), \ldots, v_x^{(M)}(k), v_y^{(M)}(k))^T$ is the new white noise vector. The new measurement vector $\mathbf{y}(k) = (y_x^{(1)}(k), y_y^{(1)}(k), \ldots, y_x^{(M)}(k), y_y^{(M)}(k))^T$ for M ($M \geq 3$) features is given by:

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{w}(k) \quad (4)$$

where $\mathbf{w}(k) = (w_x^{(1)}(k), w_y^{(1)}(k), \ldots, w_x^{(M)}(k), w_y^{(M)}(k))^T$ is the new white noise vector ($\mathbf{w}(k) \sim N(0, \mathbf{W})$) and $\mathbf{C} = \mathbf{I}_{2M}$.

We can combine quations (3)-(4) into a MIMO (Multi-Input Multi-Output) model:

$$(1 - 2q^{-1} + q^{-2})\mathbf{y}(k) = \mathbf{B}(k-d)\mathbf{u}(k-d) - \mathbf{B}(k-d-1)\mathbf{u}(k-d-1) + \mathbf{n}(k) \quad (5)$$

where $\mathbf{n}(k)$ is the white noise vector. The new white noise vector $\mathbf{n}(k)$ corresponds to the measurement noise, to the modeling errors, and to the noise introduced by inaccurate robot control. If we assume $\mathbf{B}(k-d) \approx \mathbf{B}(k-d-1)$, then (5) can be rewritten as a MIMO ARX (AutoRegressive with auxiliary input) model. This model consists of $2M$ MISO (Multi-Input Single-Output) ARX models. In addition, the new model's equation is:

$$(1 - 2q^{-1} + q^{-2})\mathbf{y}(k) = \mathbf{B}(k-d)\Delta\mathbf{u}(k-d) + \mathbf{n}(k) \quad (6)$$

where $\Delta\mathbf{u}(k-d)$ is defined as:

$$\Delta\mathbf{u}(k-d) = \mathbf{u}(k-d) - \mathbf{u}(k-d-1). \quad (7)$$

In the next section, we present the control and estimation techniques for the 3-D visual tracking problem.

3. Control and Estimation

The control objective is to move the manipulator in a such a way that the projections of the selected features on the image plane move to some desired positions or stay at their desired positions while the target is moving. This section examines the control strategies that realize this motion and the estimation scheme used to estimate the unknown parameters of the model. Some implementation issues are also discussed.

Adaptive control techniques can be used for visual servoing around a moving object when the depth of the object is not precisely known. Adaptive control techniques are used for the recovery of the components of the translational and rotational velocity vectors, $\mathbf{T}(k)$ and $\mathbf{R}(k)$, respectively, and are based on the estimated and not the actual values of the system's parameters. This approach is called *certainty equivalence adaptive control* [11]. A large number of algorithms can be generated depending on which parameter estimation scheme is used and which control law is chosen. The rest of this section is devoted to a detailed description of the control and estimation schemes.

3.1. Selection of an Efficient Control Law

The control objective is to track the motion of certain features of the target and place their projections on the image plane at some desired positions. The tracking of the features' projections is realized by an appropriate motion of the robot-camera system. A simple control law can be derived by the minimization of a *cost* function that includes the feature positional error, the control signal, and the change in the control signal:

$$J(k+d) = [\mathbf{y}(k+d) - \mathbf{y}^*(k+d)]^T \mathbf{Q} [\mathbf{y}(k+d) - \mathbf{y}^*(k+d)] + \mathbf{u}^T(k)\mathbf{L}\mathbf{u}(k) + \Delta\mathbf{u}^T(k)\mathbf{L}_d\Delta\mathbf{U}(k). \quad (8)$$

The vector $\mathbf{y}^*(k)$ represents the desired positions of the projections of the M ($M \geq 3$) features on the image plane. In our experiments, the vector $\mathbf{y}^*(k)$ is known a priori and is constant over time. By placing waghts on the control signal, the change in the control signal, and the error, we can choose how much emphasis the controller is to place on minimizing each of the three quantities. Including the control signal and the change in the control signal in the *cost* function described by (8) causes the control input signal to be bounded and feasible. This is in agreement with the structural and operational characteristics of the robotic system and the vision algorithm. A robotic system cannot track signals that command large changes in the features' image projections during the sampling interval T. In addition, our optical flow algorithm cannot detect displacements larger than 28 pixels per sampling interval T. The term $\Delta\mathbf{u}^T(k)\mathbf{L}_d\Delta\mathbf{u}(k)$ of the *cost* function (8) introduces an integral term in the control law. This term is desirable since our mathematical model (3) has a deterministic disturbances component. One problem of the introduction of an integral term in the control law is the possible saturation of the control inputs. In order to compensate for this problem, one should turn off the integrator whenever a saturation of the control inputs occurs.

The control law is derived from the minimization of the cost function (8) by taking the derivative of $J(k+d)$ with respect to the vector $\mathbf{u}(k)$ and combining the resulting expression with the system model equation (5). The resulting control law is:

$$\mathbf{u}(k) = -[\mathbf{B}^T(k)\mathbf{Q}\mathbf{B}(k) + \mathbf{L} + \mathbf{L}_d]^{-1}[\mathbf{B}^T(k)\mathbf{Q}\{\{(d+1)\mathbf{y}(k) - \mathbf{y}^*(k+d) - d\mathbf{y}(k-1)) - d\mathbf{B}(k-d)\mathbf{u}(k-d) + \sum_{m=1}^{m=d-1}\mathbf{B}(k-m)\mathbf{u}(k-m)\} - \mathbf{L}_d\mathbf{u}(k-1)]. \quad (9)$$

Feddema and Lee [3] proposed a similar control law for the robotic visual tracking problem. The main difference of our control law is that, instead of imposing constraints on the optical flow induced by the camera motion (image plane space), we impose constraints on the components of $\mathbf{u}(k)$ of the required camera tracking motion (camera frame space). In this way, we directly control the magnitudes of the control signal and the control signal change. This fact results in a control law that is more robust and feasible than the one proposed in [3]. The design parameters in our control law are the elements of the matrices $\mathbf{Q}$, $\mathbf{L}$, and $\mathbf{L}_d$. Often, we set $\mathbf{L}$ or $\mathbf{L}_d$ to zero. In most of the experiments, we set $\mathbf{L} = \mathbf{0}$ and $\mathbf{L}_d \neq \mathbf{0}$ in order to achieve a fast and bounded response. If the matrix $\mathbf{B}(k)$ is full rank then the matrix $[\mathbf{B}^T(k)\mathbf{Q}\mathbf{B}(k) + \mathbf{L} + \mathbf{L}_d]$ is invertible. The matrix $\mathbf{B}(k)$ is singular when the M feature points are collinear [3, 10]. An extensive study of other conditions which make $\mathbf{B}(k)$ singular can be found in [9].

By selecting $\mathbf{L}$, $\mathbf{L}_d$, and $\mathbf{Q}$, one can place more or less emphasis on the control input, the control input change, and the servoing error. There is no standard procedure for the selection of the elements of these matrices. One technique is the optimization approach [12].

If we want to include the *noise* of our model and the inaccuracy of the $\hat{B}(k)$ matrix in our control law, the control objective (8) becomes:

$$J(k+d)=E\{[y(k+d)-y^*(k+d)]^T Q[y(k+d)-y^*(k+d)]$$
$$+u^T(k) L u(k)+\Delta u^T(k) L_d \Delta u(k)|F_k\} \quad (10)$$

where the symbol $E\{X\}$ denotes the expected value of the random variable X and F_k is the sigma algebra generated by the past measurements and the past control inputs up to time k. The new control law is:

$$u(k)=-[\hat{B}^T(k) Q \hat{B}(k)+L+L_d]^{-1}\{\hat{B}^T(k) Q (\{(d+1)y(k)$$
$$-y^*(k+d)-d y(k-1)\}-d\hat{B}(k-d)u(k-a)$$
$$+\sum_{m=1}^{m=d-1}\hat{B}(k-m)u(k-m)\}-L_d u(k-1)] \quad (11)$$

where $\hat{B}(k)$ is the estimated value of the matrix $B(k)$. The matrix $\hat{B}(k)$ is dependent on the estimated values of the features' depth $\hat{Z}_s^{(j)}(k)$ ($(j) \in \{(1), \ldots, (M)\}$) and the coordinates of the features' image projections. In particular, the matrix $\hat{B}(k)$ is defined as follows:

$$\hat{B}(k)=\begin{bmatrix}\hat{B}_F^{(1)}(k)\\ \ldots\\ \hat{B}_F^{(M)}(k)\end{bmatrix}$$

where $\hat{B}_F^{(j)}(k)$ is:

$$T\begin{bmatrix}\frac{-f}{s_x \hat{Z}_s^{(j)}(k)} & 0 & \frac{x^{(j)}(k)}{\hat{Z}_s^{(j)}(k)} & \frac{x^{(j)}(k) y^{(j)}(k) s_y}{f} & \frac{-f^2-(x^{(j)}(k) s_x)^2}{f s_x} & \frac{y^{(j)}(k) s_y}{s_x}\\ 0 & \frac{-f}{s_y \hat{Z}_s^{(j)}(k)} & \frac{y^{(j)}(k)}{\hat{Z}_s^{(j)}(k)} & \frac{f^2+(y^{(j)}(k) s_y)^2}{f s_y} & \frac{-x^{(j)}(k) y^{(j)}(k) s_x}{f} & \frac{-x^{(j)}(k) s_x}{s_y}\end{bmatrix}.$$

In the experiments, the delay factor d is 2, so the control law (11) becomes:

$$u(k)=-[\hat{B}^T(k) Q B(k)+L+L_d]^{-1}[\hat{B}^T(k) Q\{(3y(k)-y^*(k+2)$$
$$-2y(k-1)\}-2\hat{B}(k-2)u(k-2)+\hat{B}(k-1)u(k-1)\}-L_d u(k-1)]. \quad (12)$$

3.2. Estimation of the Depth Related Parameters

The estimation of the depth ($Z_s^{(j)}(k)$) related parameters *can* be done in multiple ways. In this section, we present some of these algorithms. If the inverse of $(s_x Z_s^{(j)}(k)/f)$ is defined as $\zeta_s^{(j)}(k)$, then, equations (1) and (2) *can* be rewritten as:

$$y_F^{(j)}(k)=2y_F^{(j)}(k-1)-y_F^{(j)}(k-2)+\zeta_s^{(j)}(k-d)B_{F_t}^{(j)}(k-d)$$
$$\Delta T(k-d)+B_{F_r}^{(j)}(k-d)\Delta R(k-d)+n_F^{(j)}(k) \quad (13)$$

where the vector $n_F^{(j)}(k)$ is a gaussian noise vector with zero mean and covariance $N^{(j)}(k)$ ($n_F^{(j)}(k) \sim N(0, N^{(j)}(k))$), and $B_{F_t}^{(j)}(k)$, $B_{F_r}^{(j)}(k)$ are given by:

$$B_{F_t}^{(j)}(k)=T\begin{bmatrix}-1 & 0 & \frac{x^{(j)}(k) s_x}{f}\\ 0 & \frac{-s_x}{s_y} & \frac{y^{(j)}(k) s_x}{f}\end{bmatrix},$$

$$B_{F_r}^{(j)}(k)=T\begin{bmatrix}\frac{x^{(j)}(k) y^{(j)}(k) s_y}{f} & \frac{-f^2-(x^{(j)}(k) s_x)^2}{f s_x} & \frac{y^{(j)}(k) s_y}{s_x}\\ \frac{f^2+(y^{(j)}(k) s_y)^2}{f s_y} & \frac{-x^{(j)}(k) y^{(j)}(k) s_x}{f} & \frac{-x^{(j)}(k) s_x}{s_y}\end{bmatrix},$$

and

$$\Delta T(k)=T(k)-T(k-1), \quad \Delta R(k)=R(k)-R(k-1).$$

By defining $\Delta u_t^{(j)}(k)$ and $\Delta u_r^{(j)}(k)$ as $\Delta u_t^{(j)}(k)=B_{F_t}^{(j)}(k)\Delta T(k)$ and $\Delta u_r^{(j)}(k)=B_{F_r}^{(j)}(k)\Delta R(k)$, equation (13) is transformed into:

$$y_F^{(j)}(k)=2y_F^{(j)}(k-1)-y_F^{(j)}(k-2)+\zeta_s^{(j)}(k-d)\Delta u_t^{(j)}(k-d)$$
$$+\Delta u_r^{(j)}(k-d)+n_F^{(j)}(k). \quad (14)$$

The final transformation of equation (14) is done by using the vector $\Delta y_F^{(j)}(k)$ which is defined as:

$$\Delta y_F^{(j)}(k)=y_F^{(j)}(k)-2y_F^{(j)}(k-1)+y_F^{(j)}(k-2)-\Delta u_r^{(j)}(k-d).$$

The new form of the equation (14) is:

$$\Delta y_F^{(j)}(k)=\zeta_s^{(j)}(k-d)\Delta u_t^{(j)}(k-d)+n_F^{(j)}(k). \quad (15)$$

The vectors $\Delta y_F^{(j)}(k)$ and $\Delta u_t^{(j)}(k-d)$ are known at every time instant, while the scalar $\zeta_s^{(j)}(k)$ is continuously *estimated*. It is assumed that an initial estimate $\hat{\zeta}_s^{(j)}(0)$ of $\zeta_s^{(j)}(0)$ is given and $p^{(j)}(0)=E\{[\zeta_s^{(j)}(0)-\hat{\zeta}_s^{(j)}(0)]^2\}$ is a positive scalar p_0. The term $p^{(j)}(0)$ can be interpreted as a measure of the confidence that we have in the initial estimate $\hat{\zeta}_s^{(j)}(0)$. Accurate knowledge of the scalar $\zeta_s^{(j)}(k)$ corresponds to a small covariance *scalar* p_0. In our examples, $N^{(j)}(k)$ is a constant predefined matrix. To simplify the notation $h(k)$ is used instead of $\Delta u_t^{(j)}(k)$.

The estimation equations are [13]:

$$^-\hat{\zeta}_s^{(j)}(k)=^+\hat{\zeta}_s^{(j)}(k-1) \quad (16)$$

$$^-p^{(j)}(k)=^+p^{(j)}(k-1)+s^{(j)}(k-1) \quad (17)$$

$$^+p^{(j)}(k)=[\{^-p^{(j)}(k)\}^{-1}+h^T(k-d)\{N^{(j)}(k)\}^{-1}h(k-d)]^{-1} \quad (18)$$

$$k^T(k)=^+p^{(j)}(k)h^T(k-d)\{N^{(j)}(k)\}^{-1} \quad (19)$$

$$^+\hat{\zeta}_s^{(j)}(k)=^-\hat{\zeta}_s^{(j)}(k)+k^T(k)[\Delta y_F^{(j)}(k)-^-\hat{\zeta}_s^{(j)}(k)h(k-d)] \quad (20)$$

where $s^{(j)}(k)$ is a covariance scalar which corresponds to the white noise that characterizes the transition between the states, the superscript (−) denotes the predicted value of a variable, and the superscript (+) denotes its updated value. The depth related parameter $\zeta_s^{(j)}(k)$ is a time-varying variable since the target moves in 3D and the camera translates along its optical axis and rotates along the X and Y axis. The estimation scheme of equations (16)–(20) *can* compensate for the time-varying nature of $\zeta_s^{(j)}(k)$ because it is designed under the assumption that the estimated variable undergoes a random change. One problem is to keep the covariance scalar $p^{(j)}(k)$ finite. Solutions for this *can* be found in [11]. In addition, we have implemented other estimation techniques which deal with time-varying parameters [9]. In addition to the previous techniques, we propose the use of a more accurate form for the state update of $\zeta_s^{(j)}(k)$. This form is based on the equation (computational delays are included):

$$Z_s^{(j)}(k+1)=Z_s^{(j)}(k)+\Delta Z_{os}^{(j)}(k)+q^{-d+1}\Delta Z_{ss}^{(j)}(k) \quad (21)$$

where $\Delta Z_{ss}^{(j)}(k)$ is defined as:

$$\Delta Z_{ss}^{(j)}(k)=-\{T_z(k)+[R_x(k)y^{(j)}(k)s_y-R_y(k)x^{(j)}(k)s_x]\frac{Z_s^{(j)}(k)}{f}\}T$$

and $\Delta Z_{os}^{(j)}(k)$ is the change in depth induced by the motion of the target. It is assumed that $\Delta Z_{os}^{(j)}(k)$ does not change significantly between two time instances. The term $\Delta Z_{ss}^{(j)}(k)$ is *created* by the motion of the camera and is derived by an algebraic computation described in [9]. Equation (21) provides an approximation of the change in the feature's depth $Z_s^{(j)}(k)$ between two time instances given the feature's image coordinates and the camera motion. This equation *can* be rewritten as:

$$Z_s^{(j)}(k) \approx 2Z_s^{(j)}(k-1)-Z_s^{(j)}(k-2)+\Delta Z_{ss}^{(j)}(k-d)-\Delta Z_{ss}^{(j)}(k-d-1). \quad (22)$$

By inverting the terms of the previous equation (22), the following equation is derived:

$$\zeta_s^{(j)}(k)=\zeta_s^{(j)}(k-1)/\{2-\frac{\zeta_s^{(j)}(k-1)}{\zeta_s^{(j)}(k-2)}$$

$$+\zeta_s^{(j)}(k-1)\frac{s_x}{f}[\Delta Z_{ss}^{(j)}(k-d)-\Delta Z_{ss}^{(j)}(k-d-1)]\} \quad (23)$$

where

$$\Delta Z_{ss}^{(j)}(k)=-\{T_z(k)+[R_x(k)y^{(j)}(k)s_y-R_y(k)x^{(j)}(k)s_x]\frac{s_x}{\zeta_s^{(j)}(k)}\}T.$$

If we substitute the values of $\zeta_s^{(j)}(k)$ with their estimates, (23) will be transformed into:

$$^{-}\zeta_s^{(j)}(k)=^{+}\zeta_s^{(j)}(k-1)/\{2-\frac{^{+}\zeta_s^{(j)}(k-1)}{^{+}\zeta_s^{(j)}(k-2)}$$

$$+^{+}\zeta_s^{(j)}(k-1)\frac{s_x}{f}[^{+}\Delta Z_{ss}^{(j)}(k-d)-^{+}\Delta Z_{ss}^{(j)}(k-d-1)]\}. \quad (24)$$

The term $^{+}\Delta Z_{ss}^{(j)}(k)$ is derived from $\Delta Z_{ss}^{(j)}(k)$ by substituting $\zeta_s^{(j)}(k)$ with $^{+}\zeta_s^{(j)}(k)$. In addition, equation (17) should be modified to incorporate the new equation for the updates of states. These estimation schemes require the estimation of one parameter per feature-point and therefore, the real-time implementation of the estimation scheme is feasible. In addition, we have implemented an estimation scheme that computes two parameters per feature point. This scheme is a variation of the previous estimation scheme and separately estimates the depth related parameters $(f/(s_x Z_s^{(j)}(k)))$ and $(f/(s_y Z_s^{(j)}(k)))$ in the X and Y directions on the image plane. In theory, this formulation can estimate the depth related parameters more accurately.

The matrices $\mathbf{B}_{F_I}^{(j)}(k)$ and $\mathbf{B}_{F_r}^{(j)}(k)$ are transformed and decomposed as follows:

$$\mathbf{B}_{F_{ri}}^{(j)}(k)=T\begin{bmatrix}-1 & 0 & \frac{x^{(j)}(k)s_x}{f} & 1\end{bmatrix}$$

$$\mathbf{B}_{F_{ri}}^{(j)}(k)=T\begin{bmatrix}0 & \frac{-s_x}{s_y} & \frac{y^{(j)}(k)s_y}{f} & \end{bmatrix}$$

$$\mathbf{B}_{F_{ri}}^{(j)}(k)=T\begin{bmatrix}\frac{x^{(j)}(k)y^{(j)}(k)s_y}{f} & \frac{-f^2-(x^{(j)}(k)s_x)^2}{fs_x} & \frac{y^{(j)}(k)s_y}{s_x} & 1\end{bmatrix}$$

$$\mathbf{B}_{F_{ri}}^{(j)}(k)=T\begin{bmatrix}\frac{f^2+(y^{(j)}(k)s_y)^2}{fs_y} & \frac{-x^{(j)}(k)y^{(j)}(k)s_x}{f} & \frac{-x^{(j)}(k)s_x}{s_y} & \end{bmatrix}.$$

The subscript i denotes the X or Y direction. The estimation equations for each feature point are $(i=1,2)$:

$$^{-}\hat{\zeta}_{si}^{(j)}(k)=^{+}\hat{\zeta}_{si}^{(j)}(k-1) \quad (25)$$

$$^{-}p_i^{(j)}(k)=^{+}p_i^{(j)}(k-1)+s_i^{(j)}(k-1) \quad (26)$$

$$^{+}p_i^{(j)}(k)=[\{^{-}p_i^{(j)}(k)\}^{-1}+h_i(k-d)\{n_i^{(j)}(k)\}^{-1}h_i(k-d)]^{-1} \quad (27)$$

$$\kappa_i(k)=^{+}p_i^{(j)}(k)h_i(k-d)\{n_i^{(j)}(k)\}^{-1} \quad (28)$$

$$^{+}\hat{\zeta}_{si}^{(j)}(k)=^{-}\hat{\zeta}_{si}^{(j)}(k)+\kappa_i(k)[A y_{Fi}^{(j)}(k)-^{-}\hat{\zeta}_{si}^{(j)}(k)h_i(k-a)] \quad (29)$$

where $\Delta y_{Fi}^{(j)}(k)$ and $h_i(k)$ denote the X or Y components of the vectors $A\mathbf{y}_F^{(j)}(k)$ and $h(k)$, respectively, and $\hat{\zeta}_{si}^{(j)}(k)$ is the estimated value of either the term $(f/(s_x Z_s^{(j)}(k)))$ or the term $(f/(s_y Z_s^{(j)}(k)))$. In practice, the experimental results from the implementation of this estimation scheme prove to be comparable with the results of the first estimation scheme. Some researchers [3] propose the use of an adaptive scheme that estimates all the elements of the block matrix $\mathbf{B}(k)$ on-line. This approach is computationally expensive and not necessary.

3.3. Implementation Issues and Robot Controllers

In the experiments, we are forced to bound the input signals in order to avoid saturation of the actuators. After the computation of the translational $\mathbf{T}(k)=(T_x(k),T_y(k),T_z(k))^T$ and rotational velocity vectors $\mathbf{R}(k)=(R_x(k),R_y(k),R_z(k))^T$, we limit the *input* signals by performing several steps that are described in [9]. Thus, the vectors $\mathbf{T}(k)$ and $\mathbf{R}(k)$ are transformed to $\mathbf{T}'(k)$ and $\mathbf{R}'(k)$, respectively.

After computing the translational velocity vector $\mathbf{T}'(k)$ and the rotational velocity vector $\mathbf{R}'(k)$ with respect to the *camera* frame R_c, we transform it to the end-effector frame R_t with the use of the transformation ${}^{t}T_c$. The transformed signals are fed to the robot controller of the PUMA which acts as the tracker. We use the Unimation controllers which are interfaced to our system through multiple Ironics IV-3230 CPU boards. The Alter line is used and the desired trajectory in cartesian space is updated every 28ms. We are currently in the process of substituting the Unimation controllers with Trident boards which can be programmed in C. Finally, the whole system runs under the CHIMERA II real-time operating system [14]. The hardware configuration of the TROIKABOT system is described in [9].

The next section describes the experimental results of our algorithms on the TROIKABOT multi-robotic system.

4. Experimental Results

The algorithms have been verified by performing a number of experiments on the TROIKABOT robotic system [15]. A camera is mounted on the end-effector of one of the PUMAs which acts as the tracker. The other PUMA holds a target and moves it accordingly. The real images are 492x510 and are quantized to 256 gray levels. The camera's pixel dimensions are: s_x=0.011mm/pixel and s_y=0.013mm/pixel. The focal length of the camera is 16mm and the objects move with full 3-D motion. The initial depth of the objects' center of mass with respect to the camera frame Z_c is 290mm. The maximum permissible translational velocity of the end-effector of the hacking robot is 10cm/sec and each of the components of the end-effector's rotational velocity (roll, pitch, yaw) is not allowed to exceed 0.3rad/sec. The objective is to move the manipulator so that the image projections of certain features of the moving object move to some desired image positions or stay at their initial positions. The objects used in the servoing examples are books, pencils, or any item with distinct features. The user uses the mouse to select features of the object to be used in tracking. Then, the system evaluates on-line the quality of the features, based on the confidence measures described in [6]. The same operation can be done automatically by a computer process that runs once and needs 2 or 3 minutes, depending on the size of the interest operators which are used. The four best features are selected and used for the robotic visual servoing task. The size of windows is 8x8 while the search area is 64x64. The maximum displacement per sampling period T that can be detected is 28 pixels. The SSD algorithm has been implemented by using the pyramidal structure described in [9]. An interesting solution to the automatic detection and selection of point features has been proposed by Tomasi and Kanade [16]. We are currently investigating the potential of this approach as an alternative to our algorithms for the selection of the best feature points.

Experimental results are presented in Figures 1 through 6. The gains for the controllers are $\mathbf{Q}=0.9\mathbf{I}_8$, $\mathbf{L}=\mathbf{0}$, and $\mathbf{L}_d=diag\{0.04,0.04,1.0,5x10^5,5x10^5,5x10^5\}$. The diagonal elements of the Q, $\mathbf{L}_p$ and $\mathbf{L}_v$ can vary by a factor of between 2 and 3 and the system will continue to track successfully. The delay factor d is 2. The vector $\mathbf{y}^*(k)$ is given every instant of time k by the dation $\mathbf{y}^*(k)=\mathbf{y}(0)$. This implies that the objective of our scheme is to keep the features at their initial positions during the motion of the target.

The computation of the $[\hat{B}^T(k)Q\hat{B}(k)+L+L_d]^{-1}$ matrix is done on a Heurikon 68030 board. The technique used is the same as the one described in [10]. The total computation time (image processing and control calculations) of $T'(k)$ and $R'(k)$ is approximately 220 ms. The knowledge of the depth Z_f is assumed to be inaccurate. For all the features, $\hat{\zeta}_f^{(i)}(0)$ is initialized to 3.63 and $p^{(i)}(0)$ is 0.1.

In the example depicted in Figures 1 through 6. the performance of the control and estimation algorithms is illustrated. The target's trajectory is plotted with respect to the frame R_f which is attached to the target at the time instant $k=0$. At the same instant, the Z axis of the R_f frame is aligned with the optical axis of the camera The estimation scheme which is used estimates one parameter per feature point, thus, four parameters are estimated in total. The forgetting factor is 0.99. The measured deviations of the features from their desired positions appear noisy. The fact that the errors on the image plane are bounded guarantees that the errors are within the search range of the SSD algorithm, thus. the SSD algorithm can accurately measure the features' positions. The errors reach a maximum value when the target changes its trajectory sharply. The control and estimation algorithms compensate quickly and after 10 seconds the errors are reduced. The error in the Z direction is large. The reason is that the noisy measurements, the camera geometry, and the experimental setup make the accurate computation of the tracking motion in the Z direction (along the optical axis of the camera) difficult. Another interesting observation is that there is a small error in pitch even though there is no pitch component in the target's motion. This phenomenon occurs since then is a strong coupling between the pitch component and the X translational component of the tracking motion The same is true for the yaw component and the Y translational component of the tracking motion. In other words, the tracking system tries to track X translational or Y translational motion of the target with the rotational degrees of freedom, R_y or R_x, respectively. Numerically, this implies that the conditio? number c ($c=\sigma_{max}/\sigma_{min}$, a ratio of singular values) of the matrix $B(k)$ is large. Appropriate selection of the feature points and the relative position of the camera with respect to the target can minimize the condition number. If the relative distance of the camera (assuming the same focal length for the camera) from the target is more than 2 meters, the condition number becomes too large and tracking is impossible. In addition, full tracking is impossible when the four feature points are close to each other, or if they are very close to the piercing point.

5. Conclusions

In this papa, we examined the problem of robotic visual tracking of full 3-D motion (three translations and three rotations) by a monocular robotic tracker. A camera is mounted on the end-effector of the robotic device and provides visual information about the motion of the target. The detection of motion is based on an optical flow technique called Sum-of-Squared Differences (SSD) optical flow. This algorithm, which has been implemented in a pyramidal scheme for computational efficiency, provides the displacement vector of certain selected features of the target. Under the general guidelines of the *controlled active* vision framework which was introduced in [7], we combine these measurements with appropriate control and estimation techniques. Adaptive control techniques are introduced to compensate for uncertainties in the model, unknown depth related parameters, and computational delays. The computational burden is reduced by estimating only one or two parameters per feature point. Our algorithms do not require accurate calibration of the workspace, and thus, can be efficiently used in assembly lines in order to track moving items. In addition, these algorithms make possible autonomous satellite docking and recovery. The algorithms were extensively tested in several experiments which were performed on the TROIKABOT multi-robotic system. The real-time experiments show the feasibility and efficiency of our algorithms. In general, these algorithms show that monocular vision in conjunction with efficient motion of the vision sensor and adaptive control algorithms can be a viable alternative to standard stereo vision techniques.

Some of the areas for future research which we are currently considering include the use of more elaborate MIMO adaptive control techniques than those that have been implemented, the computational improvement of our algorithms. and the introduction of algorithms for using edges as the source of motion information. We are currently pursuing the use of "makes" for contour servoing, the application of adaptive algorithms to model-based visual tracking and servoing, and the derivation of depth maps through appropriate motion of the robot-camera system in conjunction with simple adaptive filtering techniques.

6. Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency through ARPA Order Number DAAA-21-89C-0001, and by the U.S. Army Research Office through grant Number DAAL03-91-G-0272. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

References

1. P.K. Allen, "Real-time maim tracking using spatio-temporal filters", *Pra. DARPA Image Understanding Workshop*, 1989, pp. 695-701.
2. F. Chaumette, P. Rives, and B. Espiau, "Positioning of a robot with respect to an object, tracking it and estimating its velocity by visual servoing", *Proc. of the IEEE Int. Conf. on Robotics and Automation*, April 1991, pp. 2248-2253.
3. J.T. Feddema and C.S.G. Lee, "Adaptive image feature prediction and control f a visual tracking with a hand-eye coordinated camera", *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 20, No. 5, 1990, pp. 1172-1183.
4. A.J. Kavo and N. Houshangi, "Real-time vision feedback for servoing of a robotic manipulator with self-tuning controller", *IEEE Trans. on Systems, Man and Cybernetics*, Vol. 21, No. 1, 1991, pp. 134-142.
5. ED. Dickmanns, B. Mysliwetz, a d T. Christians, "An integrated spatio-temporal approach to automatic visual guidance of autonomous vehicles", *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 20, No. 6, 1990, pp. 1273-1284.
6. N. Papanikolopoulos, P.K. Khosla, a d T. Kanade, "Vision and control techniques f a robotic visual tracking", *Proc. of the IEEE IN. Conf. on Robotics and Automation*, 1991, pp. 857-864.
7. N. Papanikolopoulos, P.K. Khosla, and T. Kanade, "Adaptive robotic visual tracking", *Proc. of the 1991 American Control Conference*, June 1991, pp. 962-%7.
8. N.P. Papanikolopoulos and P.X. Khosla, "Feature based robotic visual tracking of 3-D translational motion", *Proc. of the 30th IEEE CDC, Brighton, UK*, December 1991, pp. 1877-1882.
9. N.P. Papanikolopoulos, *Controlled active* vision, PhD dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University, August 1992
10. N.P. Papanikolopoulos and P.K. Khosla, "Robotic visual servoing around a static target: an example of controlled active vision", *Roc. of the 1992 American Control Conference*, June 1992, pp. 1489-1494.
11. G.C. Goodwin and K.S. Sin, *Adaptive filtering, prediction and control*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, Information and Systems Science Series, Vol. 1, 1984.
12. F.L. Lewis, *Optimal control*, John Wiley & Sons, New *Yak*. 1986.
13. P.S. Maybeck, *Stochastic models, estimation, and control*, Academic Press, London, 1979.

14. D.B. Stewart, D.E. Schmitz, and P.K. Khosla, "Implementing real-time robotic systems using CHIMERA II", *Proc. of 1990 IEEE Int. Conf. on Robotics and Automation*, Cincinnati, Ohio, May 1990, pp. 598-603.

15. F.E. Acker, I. Ince, and B.D. Otinger, "TROIKABOT - A multi-armed assembly robot", *Proc. of the Robots 9 Conference, Detroit, MI*, June 3-6 1985.

16. C. Tomasi and T. Kanade, "Detection and tracking of point features", Tech. report CMU-CS-91-132, Carnegie Mellon University, School of Computer Science, 1991.

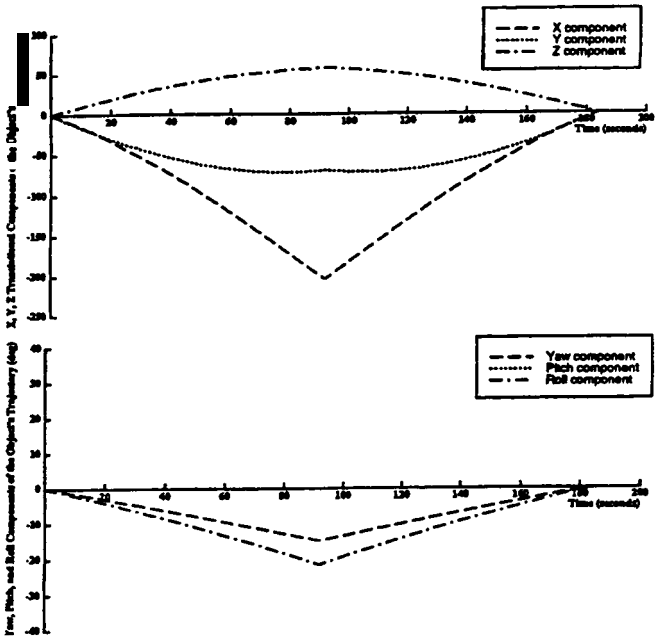

Figure 1: Translational and rotational trajectories of the moving object with respect to its initial pose (experimental).

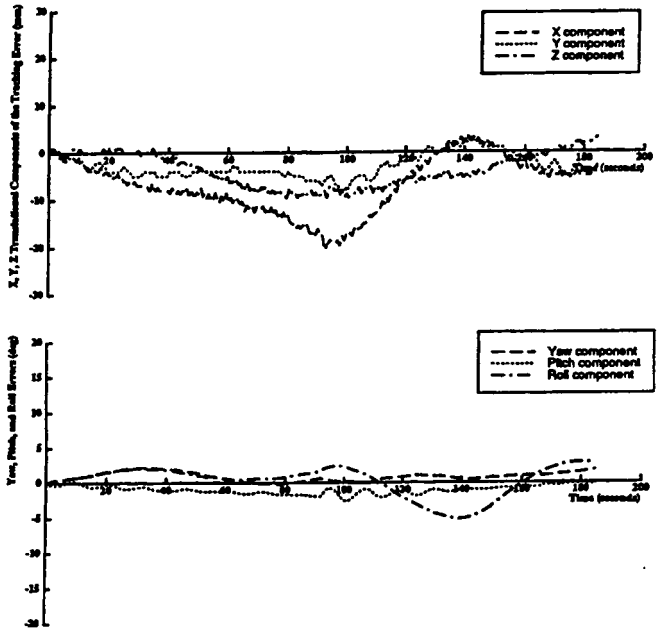

Figure 2: Translational and rotational tracking errors in the previous example (experimental).

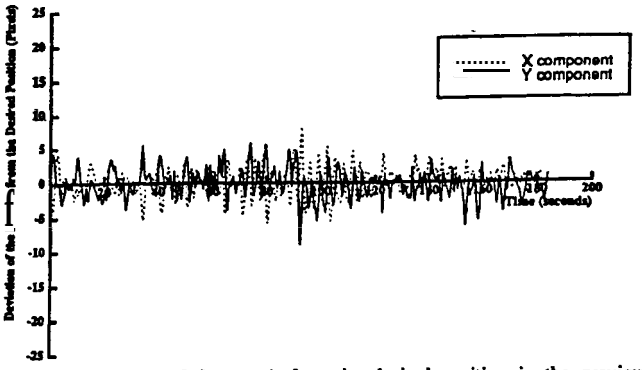

Figure 3: Deviation of feature A from its desired position in the previous example (experimental).

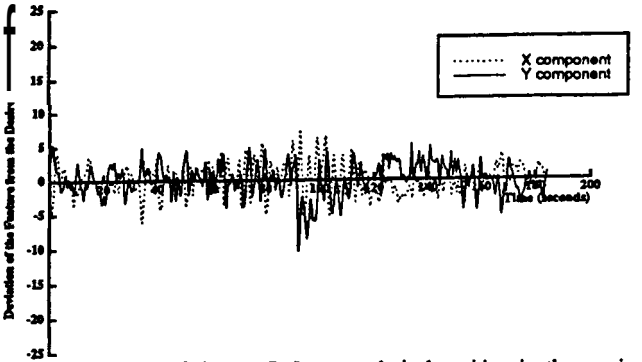

Figure 4: Deviation of feature B from its desired position in the previous example (experimental).

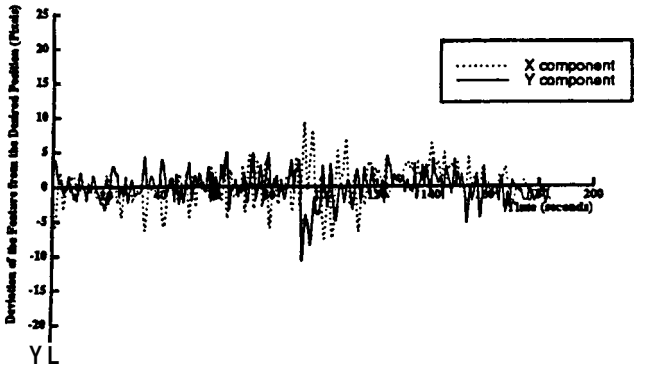

Figure 5: Deviation of feature C from its desired position in the previous example (experimental).

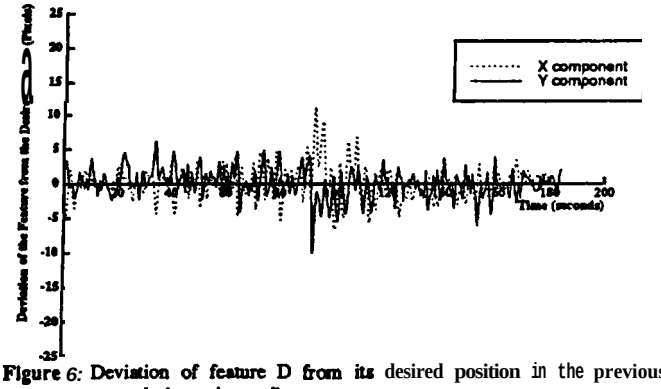

Figure 6: Deviation of feature D from its desired position in the previous example (experimental).