An Advanced On-Line Visual Tracking System†

Junghyun HWANG*, Yoshiteru OOI* and Shinji OZAWA*

A robot system equipped with an overhead CCD camera which is able to recognize, track and zoom an arbitrary 3-D object traveling at unknown velocities has been built. The problem of on-line visual tracking is formulated as a problem of combining sensor control with computer vision. For image processing, a motion and shape derivation based on eigenvalue problem in object covariance matrices is proposed to estimate an unknown moving object with low computational cost as well as noise detection. The object size, a major shape information for zoom control, is evaluated from the optimized maximum eigenvalues. In the sensor control part, an organic path planning of a pipe-line process between image processing and camera motion control is designed by using the prediction of position error and the feedback of camera motion. The performance of the proposed algorithm has been tested on Sun SS2, a controllable zoom-lens and RV-M1 robot arm, and the experimental results show that the proposed system is valid for numerous kind of moving object with 1.60 sec system sampling period.

Key Words: robot vision, adaptive sensing, visual tracking, zoom-lens control, eigenvalue problem, pipe-line process

1. Introduction

An important component of a robot vision system is the acquisition, processing, and interpretation of the available sensory information. At lower level, the sensing information is used to derive control signals to drive the robot, and at a higher level this information is used to create models of the system and the environment. The sensory information can be obtained through a variety of sensors such as position, velocity, distance, force, and vision. In this paper, the use of vision for dynamically servoing a manipulator for object tracking and zooming is proposed. Specifically, we claim that the function of zoom control is added to the visual tracking control, so we called it an advanced tracking system. The advanced tracking can be represented as more effective and accurate robot vision algorithms which allow the system to achieve the acceptable performance of the object measurement or recognition.

Several systems proposed in the literature[7]-[10] focus on estimating the object position and/or velocity from image sequence and utilizing these estimated parameters to generate control signals. Schalkoff and McVey[7] presented a model for tracking translation and rotation using 2-D affine transforms. Feddema et al.[8] studied the selection of image features based on the weighted criterion function in order to control the relative position between the end-effector of manipulator and a workpiece. Legters and Young[9] presented a mathematical model for moving object by means of a predictive Kalman filter, and Weiss et al.[12] described a hierarchical robot control structure with multiple observers. On the other hand, research in computer vision[1-6] has traditionally emphasized the paradigm of image understanding. Yachida et al.[1] proposed a system which detects and tracks moving objects from records by the linear and angular velocity on their movements, based on selecting coarse-fine mode and analyzing uncertain parts from the information of previous frames. Morikawa and Harashima[2] proposed an incremental segmentation to describe moving pictures in terms of their structural properties for video edition, video indexing, and video coding. Papanikolopoulos et al.[13] propose using the sum-of-squared differences (SSD) optical flow for the computation of the vector of discrete displacements each instant of time. Hwang
et al. proposed a tentative system which tracks and zooms a moving object by using close contour matching technique in the stable environment. The estimation part used cross-correlation in edges projections, but it needs strict assumption such as no noise on the simple environment.

When applying to a specific security zone or an unmanned factory, the visual sensing system needs large time elapse of waiting the object appearance. Moreover, a number of cameras should be required on large scale area, but it has also limitation. This paper describes an adaptive visual tracking system capable of zoom function in time varying environment. **Figure 1** shows the construction of a camera set mounted on a robot arm (eye-hand configuration). For achieving the on-line visual tracking system under real-time, an image processing of object estimation and a parallel path planning of camera motion control are designed. In image processing strategy, object features such as 2-D velocity, position, and size are first detected. The detection algorithm is based on eigenvalue problem and corresponding eigenvectors of a covariance matrix that consists of shape distribution of the moving object. However the primitive image has several problems such as much noise, multiple objects, and so on. Here a tentative algorithm to cope with noise emergence is also developed with a heuristic evaluation method which determines the optimized object shape by recursive noise detection. In control strategy, pipe-line process between image processing and camera motion control is designed to simultaneously drive two processes by the path planning by the predicted object trajectory and the feedback of sensor motions. As the result, the sensing system enables to lessen the system sampling period to nearly 60% in our experiment. The functional architecture of the sensing system is depicted in a flow chart of **Fig. 2**. To prove the efficiency of system, several experimentation is applied to either walking human or moving toy in virtual security zone.

The rest of the paper is devoted to the description of our algorithms and is organized as follows: In Section 2, we review the definition of the analysis of image by eigenvalues problem, and propose a determination of object motion and shape features. The mathematical formulation of the visual tracking problem is described in Section 3. Section 4 describes experimental results and considerations. Finally, the paper is summarized in Section 5.

**Figure 1** The manipulator of the advanced tracking system with the function of pan/tilt and zoom control

**Figure 2** Conceptual flowchart of an advanced on-line visual tracking system
2. Image Processing for Determining Object Features

One of major problems in computer vision is the processing of the large amount of data in image within a short period of time. This is of particular importance when used in the real-time tracking system. For this application, the object must be identified quickly, and its real-time location must be determined and updated when the object is moving on the field of view. In order to check whether an object has moved, scanning periodically and sampling a series of images are carried out at initial state. For achieving the tracking control, several object features such as size, position and velocity are then detected from the whole image resolution with $512 \times 432$ pixels.

From the CCD overhead camera, two gray level images are successively sampled, and two edge images, $I(t)$ and $I(t+\Delta t)$ at time $t$, are extracted by Sobel operation where $\Delta t$ is the temporal gap between the sampled images. And then, temporal differentiation of successive edge images serves to classify object edges from background and to divide into the positive and negative in intensity levels. As the result, outputs are constituted as 2 frames of binary edge images $D^+, D^-$ at time $t$, which are given as

$$D^+(x, y) = \begin{cases} 1 & \text{if } I(x, y ; t+\Delta t) - I(x, y ; t) > I_{th} \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (1)$$

$$D^-(x, y) = \begin{cases} 1 & \text{if } I(x, y ; t) - I(x, y ; t+\Delta t) > I_{th} \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (2)$$

where $(x, y)$ is a point in image space, and a threshold value $I_{th}$ serve to separate object edge from the background edge lowered by temporal differentiation. It is easy to determine $I_{th}$, because the intensity of edge is regularly high and clearly separated from background in intensity histogram of the differentiated image. Fig. 3(a), (b) and (c), (d) depict examples of original images and corresponding edge images, respectively. And (e), (f) show the results of temporal differentiation by thresholding with $I_{th}$, but by non-binaryzing. The main reason of using edge's temporal differentiation is that the essential contour of object is mostly constant against the object movement because of little overlap of thin edge. Moreover, it is useful to apply in in-door environment because the edge detection is not sensitive to the large difference of entire intensity between original images such as under the twinkling fluorescent light.

From $D^+$ and $D^-$, two means of object distribution, $(m_x^2, m_y^2)$ and $(\bar{m}_x, \bar{m}_y)$, can be measured, and their distance divided by temporal gap $\Delta t$ acts as the 2-D object velocity. To avoid duplication, the object

(a) original image at $t$  \hspace{1cm} (b) original image at $t+\Delta t$

(c) edge image $I(t)$  \hspace{1cm} (d) edge image $I(t+\Delta t)$

(e) differentiated image $D^-$  \hspace{1cm} (f) differentiated image $D^+$

(g) estimated object from initial searching space $Q^{k=0}$  \hspace{1cm} (h) detected object from optimal searching space $Q^{k+2}$

Fig. 3 An example determining the object features based on the noise elimination
features detection in each image $D^+$ and $D^-$ are developed excepting superscripts of $+,-$. Consider the center of position of object on $D$ described by

$$m_x = \frac{1}{M} \sum_{(x,y) \in \Omega} x$$  \hspace{1cm} (3)

$$m_y = \frac{1}{M} \sum_{(x,y) \in \Omega} y$$  \hspace{1cm} (4)

where $M$ is sum of object points in image as $M=\sum_{(x,y) \in \Omega} 1$, and

$$\Omega = \{(x,y) | D(x,y) = 1\}$$  \hspace{1cm} (5)

Based on the object binary image, solving the eigenvalue problem is used for the principal component analysis. Eigenvalue problem is known to analyze the geometric multiplicity of linear distribution by obtaining scalar eigenvalue $\lambda$ and eigenvector $x$ from a symmetric variance-covariance matrix $A$. The relation follows that

$$Ax = \lambda x.$$  \hspace{1cm} (6)

In this study, a variance-covariance matrix $A$ is formulated by the binary distribution of object in the image $D$ as

$$A = \frac{1}{M} \left[ \sum_{(x,y) \in \Omega} (x - m_x)^2, \sum_{(x,y) \in \Omega} (x - m_x)(y - m_y), \sum_{(x,y) \in \Omega} (y - m_y)^2 \right]$$  \hspace{1cm} (7)

Jacobi method is now introduced to determine eigenvalues, $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$, and corresponding eigenvectors, $x_{\text{max}}$ and $x_{\text{min}}$, from $A$. Each eigenvector is unit vector representing the object dominant axis or secondary dominant axis. The inclination $\phi$ of dominant axis can be obtained by

$$\phi = \cos^{-1}(x_{\text{max}} \cdot x_0)$$  \hspace{1cm} (8)

where $x_{\text{max}}$ is eigenvector of $\lambda_{\text{max}}$, and $x_0$ is $[1,0]^T$. Furthermore, eigenvalues represent the variance of image regarding to each dominant axes, which standard deviations, $\sqrt{\lambda_{\text{max}}}$ and $\sqrt{\lambda_{\text{min}}}$, are used to determine the object size quantitatively. Fig. 3(g) shows a primary result of object detection by an ellipsoid consisting of mean and standard deviation of object, but whose standard deviation is large due to noise influence.

Image processing for determining the accurate object features has several problems—being needless tiny objects, occlusions and noise—in real environment. In this study, an iterative noise reduction is designed to detect the accurate center position and size of a moving object. At the beginning, by the manners of equations (3)-(4)-(6), the candidates of object center $(m_x^0, m_y^0)$ and size $(\sqrt{\lambda_{\text{max}}^0}, \sqrt{\lambda_{\text{min}}^0})$ are determined under a searching area $\Omega^0$ at $k=0$ as entire image space. Here, $k(k=0,1,2,\cdots)$ means the iteration for the process determining the accurate object features. When the detected object points are widely distributed under $\Omega^k$, as shown in Fig. 3(g), the adjusted object searching area is required to detect the accurate object features. Consider the revised object searching area as the shape of ellipsoidal form at $k+1$ iteration described by

$$\Omega^{k+1} = \{(x,y) | (x' - m_x^k)^2/\lambda_{\text{max}}^k + (y' - m_y^k)^2/\lambda_{\text{min}}^k \leq 1, (x,y) \in \Omega^k\}$$  \hspace{1cm} (9)

where

$$x' = x \cos \phi_k + y \sin \phi_k,$$

$$y' = y \cos \phi_k - x \sin \phi_k.$$  

Although the ellipsoidal form of $\Omega^{k+1}$ is considered, the rectangular fillet form circumscribing the ellipsoid is actually used on account of reducing the computational cost. As a consequence, the object center and size are updated as the optimal object features under the adjusted object searching area $\Omega^{k+1}$. Let an assumption be defined that noise may amount to less than 10% of all pixels of object in the real image environment. Based on the assumption, to update object features is then iterated so far as the following condition is satisfied.

$$M_k > 0.9 \times M_0$$  \hspace{1cm} (10)

where $M^s = \sum_{(x,y) \in \Omega} 1$. Applying the noise reduction to the pair of $D^+,k$ and $D^-,k$, both object features are determined. Fig. 3(h) shows an example of the optimal result of object detection on $D^-$ at $k=2$ iteration.

3. Tracking and Zooming Control

In image processing two frames of image are successively sampled from the stopped camera sensor and differentiated. 2-D object motion and shape are then detected by the proposed noise elimination. Based on the result, the direction of optical axes and
the focal length are simultaneously determined for generating the servo signals of sensor. Working under on-line process, image processing and sensor control are repeated at every system sampling period, \( n(n=1,2,3,\ldots) \), while an object is moving in the field of view. First, the control of the horizontal and vertical angles of the sensor serves to correspond the direction of optical axes into the centroid of the moving object. Moreover, the control of focal length in zoom-lens enables the sensing system to obtain the well-posed object resolution in the field of view. Here, three external/ internal parameters of the sensor coordinates are defined as follows.

- The direction of the optical axes: Pan \( (\theta_{\text{pan}}(n)) \), Tilt \( (\theta_{\text{tilt}}(n)) \)
- The focal length: Zoom \( (f(n)) \)

In this control issues, the focal length is simultaneously varied for the well-posed object size, which is controlled in combination with the desired object size of operator and the condition of object motion. Moreover, the parallel process between image processing and sensor control is designed by prediction of object position and feedback of previous motion.

From the image processing, the obtained information includes position error \( p \) and the object velocity \( p' \) in image. Position error means the displacement vector between the centroid of projected object \((m_x, m_y)\) and a image point \((o_x, o_y)\) corresponding camera’s optical axis. In the \( n \)-th system sampling period, they can be defined by

\[
p(n) = \begin{bmatrix} m_x(n) - o_x \\ m_y(n) - o_y \end{bmatrix}
\]

\[
p'(n) = \frac{1}{\Delta t} \begin{bmatrix} m_x(n) - m_x(n-1) \\ m_y(n) - m_y(n-1) \end{bmatrix}
\]

where \( \Delta t \) is temporal gap of successive sampled images on the sampling periods \( n=1,2,\ldots \). The output of the tracking controller is the servo of pan and tilt. The camera rotation of pan \( \theta_{\text{pan}} \) and tilt \( \theta_{\text{tilt}} \) is derived from 2-D predicted object position \( q(n) = [q_x(n), q_y(n)]^T \) in the extended image plane and the current focal length \( f(n) \).

\[
\theta_{\text{pan}}(n) = \tan^{-1} \left( \frac{q_x(n)}{\varepsilon_x(n) + f(n)} \right)
\]

\[
\theta_{\text{tilt}}(n) = \tan^{-1} \left( \frac{q_y(n)}{\varepsilon_y(n) + f(n)} \right)
\]

where \( \varepsilon_x, \varepsilon_y \) are the vision offsets which were cali-
\[ q_n = p_{n-1} + 2 T_{sys, p} \times p'_{n-1} - q_{n-1} \]  

(17)

where

\[ T_{sys, p} = \max(\Delta t_{estim.}, \Delta t_{cont.}) \]  

(18)

Initial values \( p_0, p'_0, q_0 \) are given by zero. The period of image sampling in parallel process \( T_{sys, p} \) is determined by maximum of each processing period, and as a result it can be less than \( 0.5 T_{sys, s} < T_{sys, p} \). Fig. 4 (b) describes the algorithm of parallel processing. The combination of prediction and feedback enables the sensing system to track a moving object while feature detection of object is simultaneously processed. Assuming linear object motion, the system sampling period in parallel process is nearly half than serial process in this experiment.

Since the resolution of the tracked object in image should be controlled in accordance with the desired size, the focal length of zoom-lens is controlled by the obtained object size \( 2 \sqrt{\max(n)} \) and user's desired size \( a \). First of all, the relation of control signal to focal length has been calibrated with the mean of 10 times recursive experiment, and illustrated in Fig. 5. Consequently the control output for zoom-lens can be calculated with the combination of present focal length \( f(n) \), the desired object size \( a \), and differentiation of signal to focal length. The output value of zoom control signal \( \Delta i \) is calculated by

\[
\Delta i = \left( \frac{df(i)}{di} \right)^{-1} f(n) = \left( \frac{df(i)}{di} \right)^{-1} f(n) \left( a \left( \frac{1}{2\sqrt{\max(n)}} - 1 \right) \right)
\]  

(19)

where initial focal length \( f_0 \) is given by 10.0 mm, and the period of zoom control should be less than \( \Delta t_{cont.} \). However, for the zoom controller restriction, the maximum zoom control amount per system period is limited \( \Delta t_{cont.} \). Once the \( \Delta i \) has been determined, focal length of zoom-lens is controlled and updated for next path planning. Consequently, the adaptive sensing based on the control zoom-lens allows the system to effectively monitor an unexpected motion for successful recognition with well-posed image resolution.

4. Experimental Results and Considerations

Primarily developed for automatic tracking of unknown moving object, the complete sensing system consists of a Workstation Sun SS2, a general purpose image processor CORE including 8 frames of image buffer, a CCD overhead camera with 1/1000 shutter speed, a controllable zoom-lens Canon J10×REA2, and a robot arm Movemaster RV-M1, respectively. The function of the robot arm is to control the horizontal and vertical angles of the overhead camera. This hardware environment runs under the on-line process, performed with 1.60 sec sampling period, where the determination of object features and the sensor control are simultaneously processed. The sensor motion parameters, constituted with the servo of pan/tilt and a zoom-lens, are controlled from the Workstation through RS-232C serial port. The spatial size of image plane is 512×432 with 8-bit gray levels.

In order to prove the efficiency of the sensing system, a automobile polygon toy and a walking man were experimented on the sensing system in in-door environment of laboratory. First, the object of 29×21 mm, 7.1 meters departed from camera sensor, has been moving with 0.24 m/sec in speed on the irregular floor. The desired object size \( a \) is given 150 pixel. Fig. 6 (a) shows the moving object scene, projected on the CCD overhead camera in initial state, and (b) shows an example in steady state on tracking and...
zooming the object. Since the evaluation of shape confidence is well matched in zoom strategy, the size moving object was able to be appropriately corresponded into operator's desired size 150, as Fig. 7 (a). The detected object size, however, has fluctuated in high magnified state, due to high sensitivity in focal length of zoom-lens more than 30 mm as Fig. 5, and the improvement to cope with sensitivity is also required. Assuming an object has either high speed or random motions, the object is apt to be deviated from the screen. In that case, tracking the object with wide zoom range, so called zoom-out, is required for safety based on image understanding strategy. Fig. 7 (b) shows the tracking results of camera angles, pan and tilt, based on parallel path planning. In this experiment, the tracking deviations of the experiment are 1.5° of pan and 0.5° of tilt. From the result, we confirmed that most objects which angular velocity is less than 7°/sec can be tracked, which is equivalent to 45 km/h at 100 m distant from the sensor.

As the same manner, as tracking experiment is performed on human body who was commonly working around the tracking sensor in the laboratory. Fig. 8 shows input images as well as the corresponding results of tracking control. Since the room was too small to zoom in, the zoom control had little efficiency. The detected object features are marked by an ellipsoid constituted with the center position, deviation, and inclination of object. In this experiment, we confirm that the result of motion and shape has the confidential accuracy in spite of applying the flexible bodies under the complicated environment. Practically, the case of variable speed motion should be managed with the extensive image by zoom-out not to be deviated from the scene. Therefore a consideration of the variable speed into zoom strategy is required to stabilize the tracking control.

On the other hand, the depth of focus comes to be shallow (sensitive) according to the magnification of
the object by zoom-in. Therefore, the focusing level based on the distance between object and camera is fixed as an initial system parameter so that images are not blurred. However, a function of self-focusing is advantageous to apply to an object motion which varies the distance from camera. In the proposed image processing, the self-focusing should be stopped when two successive images are sampled.

Although the system sampling period can be improved by the higher hardware performance, this control scheme attempts to lessen the system sampling period by parallel process. Experimental results show that the proposed system is robust to detect the motion and feature of an object in spite of noise emergence and is valid for several kinds of a moving object in real scene. However, the current system is based on the assumption that one vision sensor performs to track only one object by sensor motion while an uncertain object is appeared on stationary scene. For selecting a desirable/dominant object out of multitude in the field of view, there are also needed some provisions such as image understanding strategy or human's instructions, etc., but no provision in the current system.

5. Conclusions

An advanced on-line visual sensing system has been presented. First, the image processing to detect the object motion and shape has been discussed, and then the advanced camera sensor control was also proposed. The real-time image processing utilizing the eigenvalues problem was able to reduce not only noise but computational cost, and to be improved for the accurate object shape with evaluating the shape confidence iteratively. Moreover, using prediction of object position error and feedback of sensor control, the organic path planning was facilitated to parallel process between image processing and sensor control. Experiments using several kinds of moving models—a human body, an automobile rectangular—were presented to support out theoretical base. From the result, we confirmed that the advanced tracking system yielded good issues even under the real scene environment. Assuming that more than one object appeared in the field of view, the sensing system cannot perform its ability, but it may need the help of indicating the desired object by operator or the recognition strategy with image understanding.

However, such an estimation method is still in a very early stage and not complete. The limited number of experiments does not prove that the current estimation algorithm is sufficient to be applied in visual tracking system working on out-door environment. The point is this system is designed such that the expensive segmentation procedures are not necessary, and as such it is well suited for the fast hard-
ware implementation. Currently, we are designing the extended application such as a airliner monitoring in airport, a industrial machinery guidance under construction work and so on.

References

Junghyun Hwang
Junghyun Hwang was born in Seoul, Korea, on 11 February 1967. He received the B.E degree from Inha University, Incheon, Korea, in 1989 and the M.E degree from the Keio University, Yokohama, Japan, in 1992. He is presently a doctoral student at Department of Electrical Engineering, Keio University, Yokohama, Japan. His research interests are in image processing, machine vision, and robot control.

Yoshiteru Ooi
Yoshiteru Ooi was born in Kagawa, Japan, on 7 March 1959. He received the B.S degree from the Gakushuin University Tokyo, Japan, in 1984 and the M.S. degree from the Rikkyo University, Tokyo, Japan, in 1986. He joined the INES Corporation, Yokohama, Japan, in 1986. He is currently a research engineer of the Advanced Research Center of INES. His research interest is in image processing.

Shinji Ozawa (Member)
Shinji Ozawa was born in Tokyo, Japan, on 31 May 1943. He received the B.E, M.E and Ph.D degrees from the Keio University, Yokohama, Japan, in 1967, 1969 and 1974, respectively. From 1970 to 1981, he was an Assistant Professor, and from 1981 to 1988 he was an Associate Professor. Since 1988 he has been a Professor of Department of Electrical Engineering, Keio University, where he is engaged in research on digital processing of picture and speech signals. During 1984 to 1985, he was Visiting Associate Professor of Center for Automation Research, University of Maryland, MD, USA. Dr. Ozawa is a member of IEEE and IPSJ.