Visual Tracking of Unknown Moving Object by Adaptive Binocular Visual Servoing

Minoru Asada, Takamaro Tanaka, and Koh Hosoda
Adaptive Machine Systems
Graduate School of Engineering
Osaka University, Suita, Osaka 565-0871, Japan
e-mail: asada@ams.eng.osaka-u,ac.jp

Abstract

Visual tracking of moving objects is one of the most fundamental capabilities for intelligent robots to accomplish the given task in unknown, dynamic environments. Visual servoing which utilizes direct feedback from image features to motion control has been used to realize such capability. However, the conventional visual servoing needs much knowledge on the system parameters such as kinematics and optic ones and/or on target objects. Adaptive visual servoing [1] was proposed to remedy such a limitation by adopting online estimation of image Jacobian. However, it is limited to track stationary objects or moving ones of which motion is known in advance. This paper proposes an extension of the adaptive visual servoing for unknown moving object tracking. The method utilizes binocular stereo vision system, but does not need the knowledge of camera parameters. Only one assumption is that the system needs stationary references in the both images by which the system can predict the motion of unknown moving objects. The experimental results are shown and a discussion is given.

1 Introduction

The capability of visual tracking is one of the most fundamental ones for robots to accomplish the given tasks in dynamic environments such as dexterous manipulation and intelligent locomotion. Especially, tracking moving objects is needed to catch (avoid) moving objects [2] (obstacles) of which motions are unknown.

Visual servoing which utilizes direct feedback from image features to motion control has been used to realize such capability (ex., [3][4]). However, conventional visual servoing needs much knowledge on the system parameters such as kinematics and optic ones and/or on target objects. Therefore, its use seems limited.

In visual tracking of moving object, it could be the cases that the motion of the target object and/or the system structure might be unknown or that the system parameters such as kinematics and optic ones might include much noise. Therefore, it seems hard to apply the conventional visual servoing methods to these cases.

Adaptive visual servoing (hereafter, AVS) was proposed to remedy such limitation by adopting online estimation of image Jacobian [1]. The method does not need any calibration for the system parameters in advance, but estimates a feasible image Jacobian that is not always true but sufficient to track the target. In order to obtain such a feasible Jacobian, the target motion is limited to very slow ones or known in advance. This makes it hard to apply more dynamic situations.

This paper proposes an extension of AVS for unknown moving object tracking. The method utilizes binocular stereo vision, but does not need the knowledge of camera parameters. Only one assumption is that the system need stationary references in the both images by which the system can predict the motion of unknown moving objects. The rest of this paper is structured as follows. The next section describes basic ideas how we extend the AVS method to tracking unknown moving objects by setting references in the both images. Next, we formalize the method, and construct an AVS system. Finally, the experimental results are shown and a discussion is given.

2 How can we extend AVS for unknown moving objects?

The basic ideas to extend the AVS system into one which can handle the problem of visual tracking of moving objects with unknown velocities are as follows:

1. The observed change of the image feature of the target can be separated into two parts: the first one is due to the camera motion and the second one due to the target motion itself. If we can

obtained the first one, we can apply the original AVS system to the task of visual tracking of unknown moving objects.

2. In order to estimate the first term due to camera motion, we use the binocular stereo vision system by which we can estimate the projection matrix of any image point that can predict the image feature in the next frame without any camera calibration parameters but 8 image points correspondence in advance and three references during the motion control.

The following sections will mention the details of the above ideas.

2.1 An AVS system

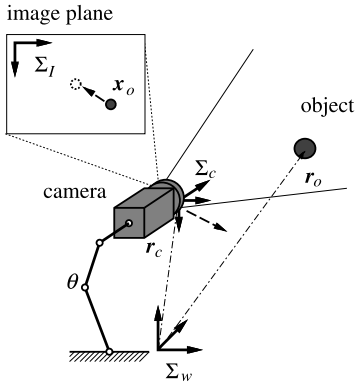


Figure 1: A hand-eye system for target tracking

First of all, we introduce a conventional AVS system shown in **Fig.1** which displays a hand-eye system with the camera in hand[1]. The task is to track target object by controlling the camera motion in such a way that observed image feature reaches the desired one. Here, we define

$\boldsymbol{x}_o$	the image feature vector of the target image
$\boldsymbol{\theta}$	the joint vector
Σ_w	the world coordinate system
$\boldsymbol{r}_c$	the position and orientation of the camera relative to Σ_w
$\boldsymbol{r}_o$	the position and orientation of the target relative to Σ_w

The following relation is obtained:

$$\boldsymbol{x}_o = \boldsymbol{x}_o(\boldsymbol{r}_c(\boldsymbol{\theta}),\ \boldsymbol{r}_o) \quad (1)$$

$$\dot{\boldsymbol{x}}_o = \boldsymbol{J}_o\ \dot{\boldsymbol{\theta}} + \boldsymbol{J}_{or_o}\ \dot{\boldsymbol{r}}_o \quad (2)$$

$$\boldsymbol{J}_o = \frac{\partial \boldsymbol{x}_o}{\partial \boldsymbol{\theta}} \quad (3)$$

$$\boldsymbol{J}_{or_o} = \frac{\partial \boldsymbol{x}_o}{\partial \boldsymbol{r}_o}, \quad (4)$$

[1]Note that generally, an AVS system does not care about the differences between system structures such as camera setting inside or outside the hand.

where $\boldsymbol{J}_o$ denotes the Jacobian which represents the relationship between the change of the image feature of the target and that of joint vector, and $\boldsymbol{J}_{or_o}$ denotes the Jacobian which represents the relationship between the change of the image feature of the target and the change of the target posture in 3-D world.

If the posture $\boldsymbol{r}_o$ and the velocity $\dot{\boldsymbol{r}}_o$ of the target are both unknown, the above equation (2) becomes nondeterministic with unknowns $\boldsymbol{J}_o$, $\boldsymbol{J}_{or_o}$, and $\dot{\boldsymbol{r}}_o$. Therefore, the original AVS systems have resolved this problem by assuming that the target is fixed to Σ_w, that is, $\dot{\boldsymbol{r}}_o = \mathbf{o}$, and estimated an image Jacobian $\boldsymbol{J}_o$ which represents the relationship between the change of the image feature and the change of the joint vector based on the following equation:

$$\dot{\boldsymbol{x}}_o = \boldsymbol{J}_o\ \dot{\boldsymbol{\theta}} \quad (5)$$

This implies that it is hard to apply the original AVS system to the situations in which $\dot{\boldsymbol{r}}_o = \mathbf{o}$ does not hold.

2.2 Virtually stationary target image feature

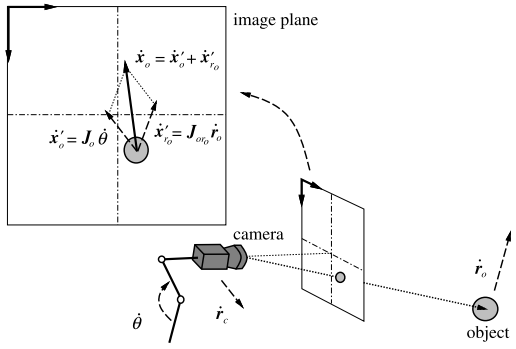


Figure 2: Image features of a moving object

Let us consider to solve the equation (2) without the assumption $\dot{\boldsymbol{r}}_o = \mathbf{o}$. The change of the target image feature in eqn.(2) can be separated into two parts as shown in **Fig.2** where the first one is the change due to the camera motion and the second one due to the target motion. If we could obtain the first one $\dot{\boldsymbol{x}}'_o$, the first term in RHS of the equation (2) can be obtained by

$$\dot{\boldsymbol{x}}'_o = \boldsymbol{J}_o\ \dot{\boldsymbol{\theta}} \quad (6)$$

We call $\boldsymbol{x}'_o$ virtually stationary target image feature (hereafter VSTIF).

If the change of VSTIF, $\dot{\boldsymbol{x}}'_o$ is known, we can estimate $\boldsymbol{J}_o$ which represents the relationship between $\dot{\boldsymbol{x}}'_o$ and $\dot{\boldsymbol{\theta}}$ as the original AVS equation (5), and further the second term $\dot{\boldsymbol{x}}'_{r_o}$ in RHS of eqn.(2).

$$\dot{\boldsymbol{x}}'_{r_o} = \boldsymbol{J}_{or_o}\ \dot{\boldsymbol{r}}_o = \dot{\boldsymbol{x}}_o - \dot{\boldsymbol{x}}'_o \quad (7)$$

$$\dot{\boldsymbol{x}}_o = \boldsymbol{J}_o\ \dot{\boldsymbol{\theta}} + \dot{\boldsymbol{x}}'_{r_o} \quad (8)$$

So, the problem is how to estimate VSTIF.

2.3 Setups of the system and assumptions

In order to estimate VSTIF of the unknown moving objects, we need

- the position of the target in the camera coordinate system, and
- the reference to the world coordinate system since the target is moving relative to the world coordinate system.

Then, we set up a binocular vision system with reference observed in the both images. Since the AVS system does not need the knowledge on any system or camera parameters, our method follows such desirable feature. Instead, we assume

1. two cameras which are fixed to each other and have the same internal camera parameters,
2. 8 points correspondence to estimate the geometrical relationship between two cameras in advance, and
3. three references in the world coordinate system that can be observed in the both images during the camera motion control.

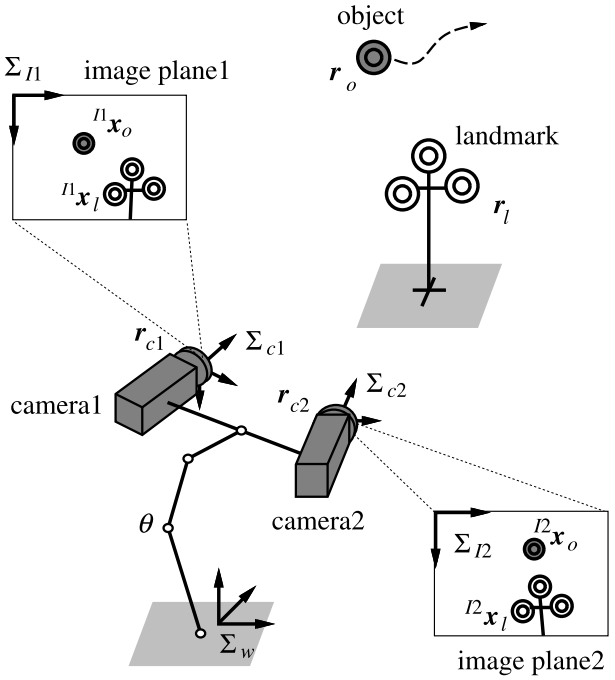


Figure 3: A system configuration

It is known in projective geometry that these assumptions are necessary not simply to estimate the geometrical relationship between two cameras but also to parallel two cameras so that the transformation matrix that describes the temporal change of any image point can be obtained from the references and the temporal changes of their disparities. Since the size of the transformation matrix is 3×3, the minimum number of the references is three. The system we used is shown in **Fig.3**.

3 Acquisition of VSTIF

Generally, we can virtually parallel two cameras that are actually not parallel but arbitrary configuration based on 8 points correspondence [5]. In the following, first we obtain the transformation matrix that describes temporal change of any image point, and then estimate another transformation matrix that describes the change of image point corresponding to any stationary point in space based on then changes of three references. Finally, we obtain VSTIF by projecting the image feature of unknown moving object by this matrix.

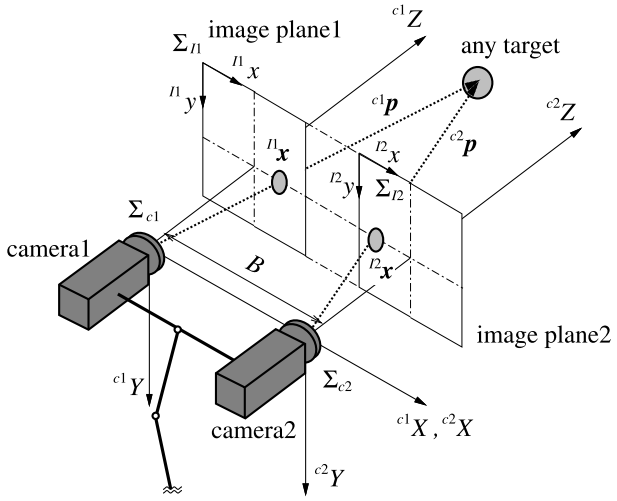


Figure 4: Model of parallel stereo camera

3.1 Temporal change transformation of any stationary point in space

Fig.4 shows a parallel stereo vision system where a projection of a stationary point to the i-th camera $ci\,(i = 1, 2)$ is denoted as $^{Ii}\boldsymbol{x} = \begin{bmatrix} ^{Ii}x & ^{Ii}y \end{bmatrix}^T$ in the image plane Σ_{Ii}, the position of the stationary point in Σ_{ci} is
$^{ci}\boldsymbol{p} = \begin{bmatrix} ^{ci}X & ^{ci}Y & ^{ci}Z \end{bmatrix}^T$. Extended vectors which include 1 at the end are $^{Ii}\widetilde{\boldsymbol{x}} \in \Re^3$, and $^{ci}\widetilde{\boldsymbol{p}} \in \Re^4$.

These vectors have the following relationship:

$$^{c1}s\ ^{I1}\widetilde{\boldsymbol{x}} = \boldsymbol{P}\ ^{c1}\widetilde{\boldsymbol{p}} \quad (9)$$

$$^{c2}s\ ^{I2}\widetilde{\boldsymbol{x}} = \boldsymbol{P}\ ^{c2}\widetilde{\boldsymbol{p}}, \quad (10)$$

where ${}^{ci}s$ denotes a scalar transformarion in terms of ${}^{ci}Z$, and $\boldsymbol{P} \in \Re^{3\times4}$ denotes a projection matrix that includes internal camera parameters such as focal length.

In parallel stereo vision,

$$ {}^{c1}\widetilde{\boldsymbol{p}} - {}^{c2}\widetilde{\boldsymbol{p}} = \begin{bmatrix} B \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{11} $$

where B denotes baseline between two optical axes.

Substituting eqns.(9)(10) into eqn. (11) leads:

$$ \boldsymbol{P}\ {}^{c1}\widetilde{\boldsymbol{p}} - \boldsymbol{P}\ {}^{c2}\widetilde{\boldsymbol{p}} = {}^{c1}s\ {}^{I1}\widetilde{\boldsymbol{x}} - {}^{c2}s\ {}^{I2}\widetilde{\boldsymbol{x}} = \begin{bmatrix} B' \\ 0 \\ 0 \end{bmatrix}, \tag{12} $$

where $[B'\ \ 0\ \ 0]^T$ is a trasnformation of $[B\ \ 0\ \ 0\ \ 0]^T$ by the projection matrix $\boldsymbol{P}$.

From eqn.(12), ${}^{c1}s = {}^{c2}s = s$ and ${}^{I1}y = {}^{I2}y$, then,

$$ s\ {}^{I1}x - s\ {}^{I2}x = B'. \tag{13} $$

Defnining a scalar variable $s\,/\,B' = \overline{s}$ leads:

$$ {}^{I1}x - {}^{I2}x = 1\,/\,\overline{s}, \tag{14} $$

where $\overline{s}$ is a invers of the disparity that can be obtained from two images. Returning to eqn.(9) with ${}^{I1}\widetilde{\boldsymbol{x}}$ and $\overline{s}$,

$$ \overline{s}\ {}^{I1}\widetilde{\boldsymbol{x}} = \frac{1}{B'}\boldsymbol{P}\ {}^{c1}\widetilde{\boldsymbol{p}} = \boldsymbol{P'}\ {}^{c1}\widetilde{\boldsymbol{p}}\,, \qquad \boldsymbol{P'} \in \Re^{3\times4}. \tag{15} $$

Next, we quantize the eqn.(15) with a sampling time T which is short enough to assume that a motion vector and its velocity can be the same.

$$ \overline{s}(k)\ {}^{I1}\widetilde{\boldsymbol{x}}(k) = \boldsymbol{P'}\ {}^{c1}\widetilde{\boldsymbol{p}}(k) \tag{16} $$

$$ \overline{s}(k+1)\ {}^{I1}\widetilde{\boldsymbol{x}}(k+1) = \boldsymbol{P'}\ {}^{c1}\widetilde{\boldsymbol{p}}(k+1). \tag{17} $$

Eqn.(16) can be

$$ {}^{c1}\widetilde{\boldsymbol{p}}(k) = \overline{s}(k)\ \boldsymbol{P'^{+}}\ {}^{I1}\widetilde{\boldsymbol{x}}(k), \tag{18} $$

where $\boldsymbol{P'^{+}}$ is a pseudo inverse matrix of $\boldsymbol{P'}$.

By the camera motion specified by a homogeneous matrix ${}^{k+1}\boldsymbol{T}_k \in \Re^{4\times4}$, the position of the stationary point ${}^{c1}\widetilde{\boldsymbol{p}}(k)$ is transformed into ${}^{c1}\widetilde{\boldsymbol{p}}(k+1)$ in the camera coordinate system.

$$ {}^{c1}\widetilde{\boldsymbol{p}}(k+1) = {}^{k+1}\boldsymbol{T}_k\ {}^{c1}\widetilde{\boldsymbol{p}}(k) \tag{19} $$

Substituting eqn.(19) into eqn.(17), we obtain:

$$ \overline{s}(k+1)\ {}^{I1}\widetilde{\boldsymbol{x}}(k+1) = \boldsymbol{P'}\ {}^{k+1}\boldsymbol{T}_k\ {}^{c1}\widetilde{\boldsymbol{p}}(k). \tag{20} $$

Deleting ${}^{c1}\widetilde{\boldsymbol{p}}(k)$ based on eqn.(18), we obtain:

$$ \overline{s}(k+1)\,{}^{I1}\widetilde{\boldsymbol{x}}(k+1) = \overline{s}(k)\,{}^{k+1}\boldsymbol{M}_k\,{}^{I1}\widetilde{\boldsymbol{x}}(k) \tag{21} $$

$$ {}^{k+1}\boldsymbol{M}_k = \boldsymbol{P'}\,{}^{k+1}\boldsymbol{T}_k\,\boldsymbol{P'^{+}}, \quad {}^{k+1}\boldsymbol{M}_k \in \Re^{3\times3}, \tag{22} $$

${}^{k+1}\boldsymbol{M}_k$ represents a transformation matrix by which temporal chaneg of any stationary point can be described.

From eqn.(21), any statinary point of which image coordinate is ${}^{I1}\boldsymbol{x}(k)$ in the camera $c1$ with inverse disparity $\overline{s}(k)$ at the k-th step can be transformed to a point with ${}^{I1}\boldsymbol{x}(k+1) and \overline{s}(k+1)$ by ${}^{k+1}\boldsymbol{M}_k$ at the $(k+1)$-th step. Inversely, if the image position and its disparity at the k and $(k+1)$-th steps of any stationary point in space are given, ${}^{k+1}\boldsymbol{M}_k$ can be obtained.

3.2 Acquisition of VSTIF by stationary landmarks

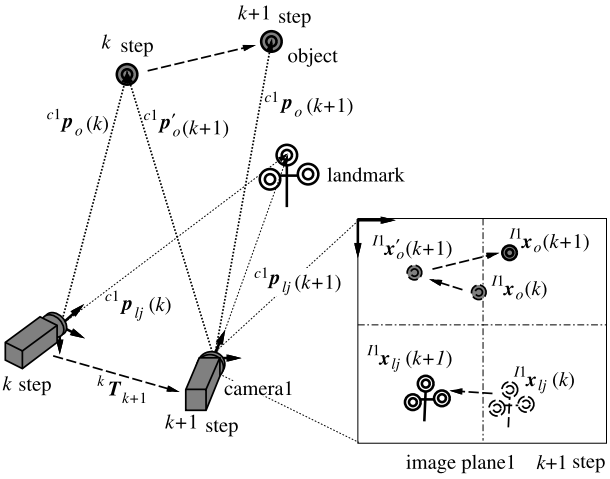


Figure 5: Change of image features for the target and landmarks (camera1)

We can obtain VSTIF by applying ${}^{k+1}\boldsymbol{M}_k$ that can be estimated from the stationary landmarks to unknown moving object. The j-th stationary landmark $lj\,(j = 1,2,3)$ can be obserbed as ${}^{Ii}\boldsymbol{x}_{lj}$ and ${}^{I1}\boldsymbol{x}_{lj}(k+1)$ with their inverse disparities $\overline{s}_{lj}(k) = 1\,/\,\{{}^{I1}x_{lj}(k) - {}^{I2}x_{lj}(k)\}$, $\overline{s}_{lj}(k+1) = 1\,/\,\{{}^{I1}x_{lj}(k+1) - {}^{I2}x_{lj}(k+1)\}$. From these parameters, we obtain ${}^{k+1}\boldsymbol{M}_k$. Since its size is 3 X 3, we need at least three landmarks. By defining $\widetilde{\boldsymbol{t}}_{li}\,(= \overline{s}_{li}\ {}^{I1}\widetilde{\boldsymbol{x}}_{li})$, we obtain ${}^{k+1}\boldsymbol{M}_k$.

$$ \left[\widetilde{\boldsymbol{t}}_{l1}(k+1) \quad \widetilde{\boldsymbol{t}}_{l2}(k+1) \quad \widetilde{\boldsymbol{t}}_{l3}(k+1)\right] \times $$

$$ \left[\widetilde{\boldsymbol{t}}_{l1}(k) \quad \widetilde{\boldsymbol{t}}_{l2}(k) \quad \widetilde{\boldsymbol{t}}_{l3}(k)\right]^{+} = {}^{k+1}\boldsymbol{M}_k. \tag{23} $$

Applying ${}^{k+1}\boldsymbol{M}_k$ to eqn.(21), that is, transforming the image position ${}^{I1}\boldsymbol{x}_o(k)$ and inverse disparity $\overline{s}_o(k) = 1\,/\,\{{}^{I1}x_o(k) - {}^{I2}x_o(k)\}$, we obtain VSTIF ${}^{Ii}\boldsymbol{x}'_o(k+1)$ and its inverse disparity $\overline{s}'_o(k+1)$.

$$ \overline{s}'_o(k+1)\ {}^{I1}\widetilde{\boldsymbol{x}}'_o(k+1) = \overline{s}_o(k)\ {}^{k+1}\boldsymbol{M}_k\ {}^{I1}\widetilde{\boldsymbol{x}}_o(k) \tag{24} $$

$$ {}^{I2}\widetilde{\boldsymbol{x}}'_o(k+1) = \begin{bmatrix} {}^{I1}x'_o(k+1) - 1 \,/\, \overline{s}'_o(k+1) \\ {}^{I1}y'_o(k+1) \\ 1 \end{bmatrix} \quad (25) $$

VSTIF $\boldsymbol{x}'_o(k+1)$ obtained above can be applied to $\dot{\boldsymbol{x}}'_o$ and $\dot{\boldsymbol{x}}'_{r_o}$ in eqn. (7) (see **Fig.5**). That is,

$$ \dot{\boldsymbol{x}}'_o(k+1) = \boldsymbol{x}'_o(k+1) - \boldsymbol{x}_o(k). \quad (26) $$

$$ \dot{\boldsymbol{x}}'_{r_o}(k+1) = \boldsymbol{x}_o(k+1) - \boldsymbol{x}'_o(k+1). \quad (27) $$

4 An Extended AVS system for unknown moving target

We can construct an extended AVS system based on the estimated Jacobian $\widehat{\boldsymbol{J}}_o$ [1]. Applying $\widehat{\boldsymbol{J}}_o$ to eqn.(8),

$$ \dot{\boldsymbol{x}}_o = \widehat{\boldsymbol{J}}_o \; \dot{\boldsymbol{\theta}} + \dot{\boldsymbol{x}}'_{r_o}. \quad (28) $$

Then, the following control low is derived from both the feedback term that converges $\boldsymbol{x}_o$ to $\boldsymbol{x}_{od}$ and the feedforward one that compensates $\dot{\boldsymbol{x}}'_{r_o}$ due to the object motion itself. That is, for the input $\boldsymbol{u}$ to joints:

$$ \boldsymbol{u} = \widehat{\boldsymbol{J}}_o^{+} \; \{\boldsymbol{K}(\boldsymbol{x}_{od} - \boldsymbol{x}_o)\} - \widehat{\boldsymbol{J}}_o^{+} \; \dot{\boldsymbol{x}}'_{r_o}, \quad (29) $$

where $\widehat{\boldsymbol{J}}_o^{+}$ denotes a pseudo invers matrix of $\widehat{\boldsymbol{J}}_o$, and $\boldsymbol{K}$ is an $m \times m$ positive gain matrix. In RHS of eqn.(29), the first term is for the feedback, and the second for the feedforward for $\dot{\boldsymbol{x}}'_{r_o}$ when the Jacobian is sufficiently estimated. **Fig.6** shows a block diagram.

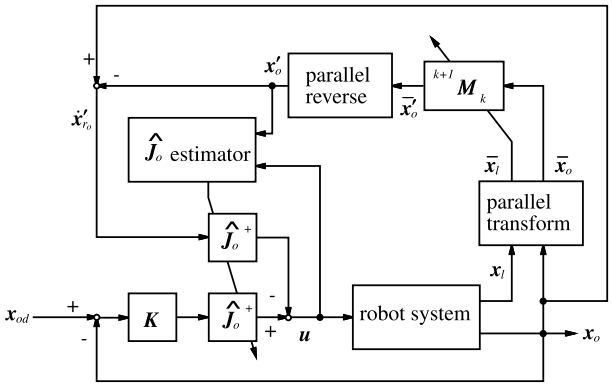


Figure 6: A control block diagram

5 Experimental Results

Figs.7, **8**, and **9** show a block diagram of experimetal system, a photo of the sytem and a sample image pair of the binocular vision system captured by two small ELMO CCD cameras attached at the end of the robot arm MHI PA10, respectively. In addition to three reference markers and one target object that is independently controlled by another robot

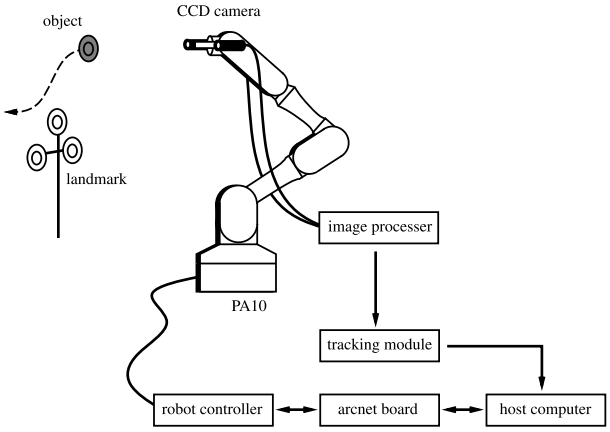


Figure 7: The experimental system

arm KAWASAKI Js-5, five more marks are observed for off-line camera calibration (estimation of transformation matrix between two cameras) in **Fig.9**. All reference markers and the target are prespecified by the programmer. Three reference markers and one target object are visually tracked in realtime (every 33ms) by FUJITSU tracking module based on template matching of image pattern. Distances from the references and the target from the cameras are about 1.4m and 1.5m, respectively. Any system parameters such as kinematic and optic ones are all unknown.

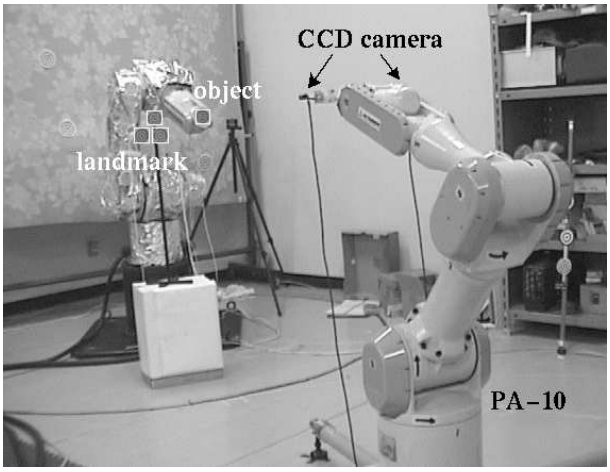


Figure 8: A perspective view of the experimental system

Table **Table 1** shows averaged Euclidean norm of image errors in the cases of the conventional and the proposed AVS systems for a variety of target speeds from 150mm/s to 1200mm/s of which projected speeds onto the image plane of the virtually static camera range from about 32 pixel/s to 259 pixel/s. **Fig.10** shows the error on the left camera

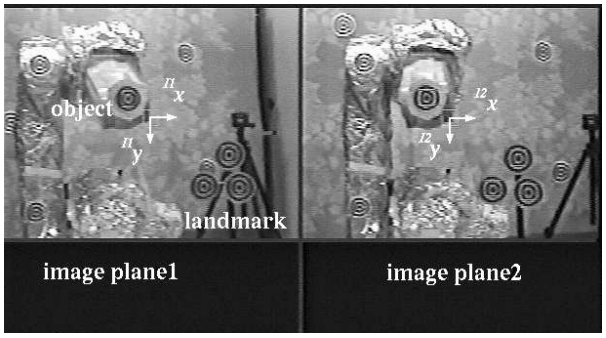


Figure 9: A sample image pair

$({}^{I1}\boldsymbol{x}_o - {}^{I1}\boldsymbol{x}_{od})$ in the cases of the conventional and the proposed AVS systems with the taget velocity 750mm/s.

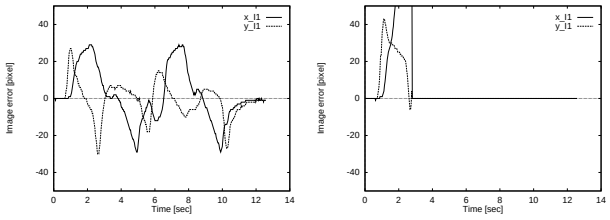
Table 1: Euclidean norm average of image errors

object speed		Error [pixel]	
[mm/s]	[pixel/s]	proposed AVS	conventional AVS
150	32.4	8.301651	12.573951
300	64.8	12.194477	18.274611
600	129.6	20.452076	27.014025
750	162.0	24.027504	85.376846
900	194.4	25.559967	-
1200	259.2	28.600138	-

From these table and figure, we can conclude that the difference between the conventional and the proposed AVS systems are not so much with the speed 150mm/s. However, as the target speed increases, the difference becomes larger and the conventional AVS does not work with the speed faster than 750mm/s while the proposed AVS system does work with about 25 pixel errors in average.

6 Discussion

We have proposed an extension of the AVS system so that it can handle unknown moving objects. Due to the space limit, we skiped the details of the real-time control system which can be found elsewhere. The basic idea is to use binocular vision system with references fixed to the world coordinate system. Intuitively, the binocular vision system can provide the position in space and the references are used to estimate the transformation matrix to predict the image motion of the virtually stationary target. The method does not need any system parameters and geometrical calibration, but only needs references.



(a) Adaptive visual servo with feedforward compensator

(b) Adaptive visual servo without feedforward compensator

Figure 10: Image error (object speed : 750[mm/s])

For more dynamic situations, the visual tracking for stationary points should not be fixed, but adaptive according to camera motion to track moving objects. In that case, the system should have a filter which points can be qualified as references. These are under the investigation.

References

[1] K. Hosoda and M. Asada. Versatile visual servoing without knowledge of true jacobian. In *Proc. of IROS'94*, pages 186–193, 1994.

[2] G.C.Buttazzo, B.Allotta, and F.P.Fanizza. Mousebuster: a robot system for catching fast moving objects by vision. In *Proc. of the IEEE International Conference on Robotics and Automation (1993)*, pages 932–937.

[3] K.Hashimoto, T.Kimoto, T.Ebine, and H.Kimura. Manipulator control with image-based visual servo. In *Proc. of the IEEE International Conference on Robotics and Automation (1991)*, pages 2267–2272.

[4] F.Chaumette, P.Rives, and B.Espiau. Positioning of a robot with respect to an object, tracking it and esitmating its velocity by visual servoing. In *Proc. of the IEEE International Conference on Robotics and Automation (1991)*, pages 2248–2253.

[5] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.