International Journal of Computer Vision 37(1), 7–20, 2000

A Direct Interpretation of Dynamic Images with Camera and Object Motions for Vision Guided Robot Control

KOICHIRO DEGUCHI
Graduate School of Information Sciences, Tohoku University, Aoba-campus 01, Sendai 980-8579, Japan
kodeg@fractal.is.tohoku.ac.jp

Received ; Revised ; Accepted

Abstract. A general scheme to represent the relation between dynamic images and camera and/or object motions is proposed for applications to visual control of robots. We consider the case where a moving camera observes moving objects in a static scene. The camera obtains images of the objects moving within the scene. Then, the possible combinations of the camera and the objects' poses and the obtained images are not arbitrary but constrained to each other. Here we represent this constraint as a lower dimensional hypersurface in the product space of the whole combination of their motion control parameters and image data. The visual control is interpreted as to find a path on this surface leading to their poses where a given goal image will be obtained. In this paper, we propose a visual control method to utilize tangential properties of this surface. First, we represent images with a composition of a small number of "*eigen images*" by using *K-L (Karhunen-Loève)* expansion. Then, we consider to reconstruct the eigen space (the eigen image space) to achieve efficient and straightforward controls. Such reconstruction of the space results in the constraint surface being mostly flat within the eigen space. By this method, visual control of robots in a complex configuration is achieved without image processing to extract and correspond image features in dynamic images. The method also does not need camera or hand-eye calibrations. Experimental results of visual servoing with the proposed method show the feasibility and applicability of our newly proposed approach to a simultaneous control of camera self-motion and object motions.

Keywords: visual servoing, motion images, eigen space method, camera motion control

1. Introduction

There are many works on dynamic image sequence analysis with respect to camera motion. Visual servoing is a main application of the works, where the robot arm mounting a camera is controlled to track a moving object or to move to a goal pose with visual feedback. Most approaches assume that (Hashimoto, 1993; Espiau et al., 1992; Weiss et al., 1987; Hutchinson et al., 1996) (1) the object images are expressed with simple primitive features such as points, lines and conics, and (2) the correspondences of those features are always maintained in the image sequence. Apparently these assumptions do not hold when applied to a long sequence of real images. Furthermore, although points or lines are too simple to express the real objects, accurate extractions and their correspondences in images are not easy to establish. Even when we obtain accurate image features, we need camera and hand-eye calibrations to utilize them.

In this paper, we propose a new approach of image based visual servoing. This belongs to the so-called appearance based one where we do not need any image processing for feature extraction and correspondence search. We applied the visual servoing to more complex situations. Figure 1 shows one typical example, where a target object is picked up by a robot-arm and moved to a certain position, and its motion control is carried out based on the image taken by the camera mounted on another mobile robot-arm. For this purpose also, the camera position should be controlled to get a good observation. Then, for a given goal image, our task in this case is to move a robot-arm mounting a camera to the position where a given goal image was taken and,

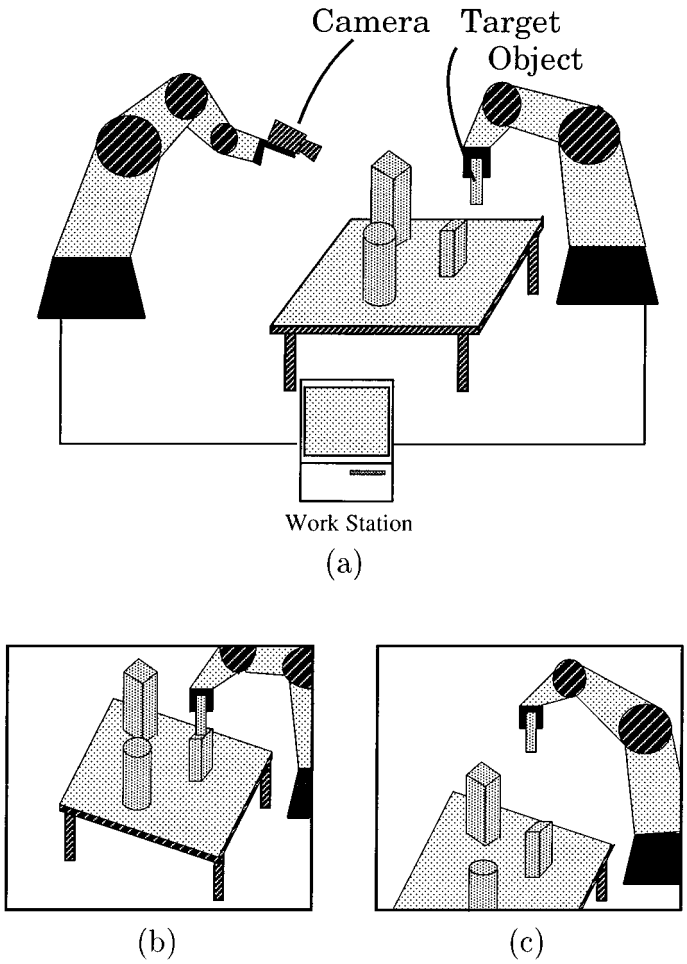

Figure 1. (a) The configuration for the typical example of our visual servoing task. Robot arm mounting a camera is controlled to move to the goal pose where it obtains the goal image (b). At the same time, a target object is moved to a certain position by another robot to the goal position where it has just the same pose as in the goal image. This simultaneous control is carried out by referring to the difference between the current image (c) and the goal image.

at the same time, to move some target objects to the same positions as they were in the goal image.

Furthermore, we consider the case where two cameras observe the target object motion from different viewing angles and the respective camera positions themselves are also controlled.

For these purposes, it is not enough to extract and observe some fixed object features in an image by the camera. Moreover, what kind of image features must be utilized for efficient control in these situations is not clear.

To overcome these difficulties and to achieve flexible control of the motions, we proposed a basic idea of a direct interpretation of the relation between the image in sequence and the camera motion (Deguchi, 1997; Deguchi and Noguchi, 1996). In this paper, we extend the use of the idea for simultaneous controls of motions of the cameras and the target objects to deal with the above, as well as more complex tasks.

The summary of the main idea is as followings: Let us consider that the information contained within an image taken by a camera can be expressed with an N dimensional vector. Usually, N equals the number of pixels in an image. The camera pose and position are determined with F parameters. For the case where the object motion is simultaneously controlled, the F motion parameters also include those of the object motion. The imaging system, including the camera intrinsic

parameters, illumination condition, and so on, is specified with P parameters.

For example, a frame of raw gray scale image with $n \times n$ pixels can be represented as an $N = n \times n$ dimensional vector. If a camera and an object pose has 6 degrees of freedom each, and we are intending to control them, F becomes 12. If we employ two cameras, as shown in one of the experiments later, F increases up to 18, that is 3×6. In this case, the obtained image can be considered to have twice the pixel size of $N = 2 \times n \times n$.

In addition, an event where the camera gets an image of the object in a scene at certain positions under some imaging condition can be represented by a point in the $N + F + P$ dimensional space. In this paper, we fix the imaging condition to be constant ($P = 0$) and consider $N + F$ dimensional space to make the basic principle clear.[1]

For a specific object, the camera cannot obtain any arbitrary image, so that the possible combination of the camera and the object poses and the obtained image should be constrained within a surface in the $N+F$ dimensional space. It must have much smaller dimensionality.

Our approach is based on the analysis of the properties of this constraint surface. More specifically, we utilize its differential or tangential property for visual servoing.

At this point of view, first, the dimension of the image information becomes a key problem[2] because a raw image contains more than a hundred thousand pixels and its number N becomes much larger than F. For the first step, we propose to use the principal component analysis and to represent images with a composition of a small number of "*eigen images*" by using *K-L (Karhunen-Loève)* expansion (Deguchi, 1997). We represent images by using the eigen system obtained by the conventional eigen space method.

Next, we describe how a normal vector of the above mentioned surface is related to the so-called Interaction Matrix, which is used to relate camera motion and image motion in conventional visual servoing techniques.

The second key problem is the efficiency for the control when we utilize the property of this constraint surface. For this problem, we point out that, if the constraint surface is flat with respect to the motion parameters, the image change linearly depends on the motions and the controls will be efficiently realized by simple visual feedback.

According to this idea, we consider the reconstruction of the representation space in which the constraint surface becomes flatest which enables us to represent images more efficiently and to speed up the convergence in the control. This representation can be obtained by constructing a new set of base vectors with weighted summations of the original eigen vectors. We also present an algorithm for this reconstruction.

Finally, we present some experimental results of the proposed method. The first experiment is the simultaneous controls of the camera and the target object motions to return back to their respective goals from arbitrary initial positions. In the second experiment, we show the simultaneous controls of the poses of multiple cameras and the target object. Experimental results of the visual servoing show that complex controls can be achieved with a very simple implementation of our method. The experiments also show the feasibility and applicability of our newly proposed approach.

2. Direct Interpretation of Dynamic Images and Camera Motion

2.1. Image and Camera Motion Representation for Visual Control

As was described, an image taken by a moving camera is expressed with an $N = n \times n$ dimensional vector $\boldsymbol{x}$, whose components are gray levels of each pixels. The camera and the target object poses and positions are determined with an F dimensional vector $\boldsymbol{w}$. If we intend to control only camera motion in a static scene, $F = 6$, and $\boldsymbol{w} = (x, y, z, \omega_x, \omega_y, \omega_z)^\top$, where (x, y, z) is the camera position and $(\omega_x, \omega_y, \omega_z)$ is its pose angle.

For the case where we intend to simultaneously control of the camera and a target object motions in a static scene, we will need up to 12 dimensional vectors, that is, 6 for the camera and 6 for the target object. If we have more cameras and/or more target objects, the number F of the motion parameters to be controlled will become larger.

For a specific scene, an event that the camera takes an image at a position can be represented as a point $(\boldsymbol{x}^\top, \boldsymbol{w}^\top)^\top$ in the $N + F$ dimensional space. This point is the combination of the obtained image and the camera and the target object positions. It must be emphasized that the possible combination of their poses and the obtained image should be restricted within a subspace having much smaller dimensionality. That is, it forms a lower dimensional hyper surface in the $N + F$ dimensional space. The motion images are constrained within this surface (Fig. 2).

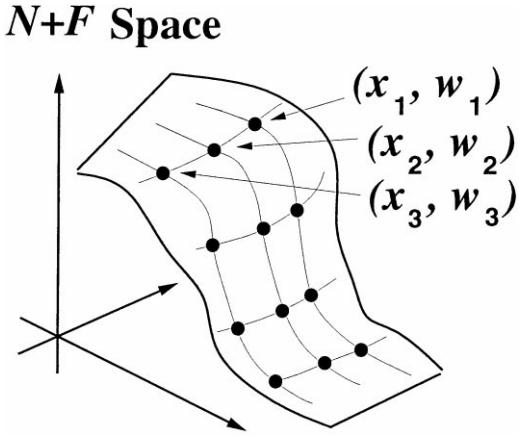

Figure 2. A hypersurface in the $N + F$ dimensional space which constrains the camera and the target object poses and the image combination.

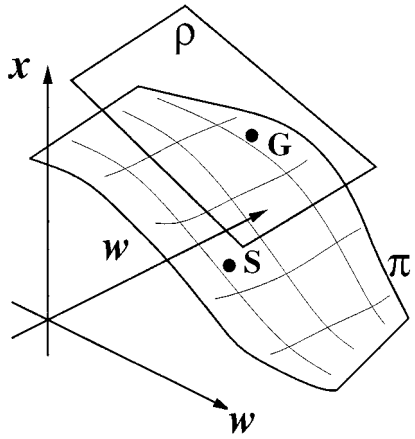

Figure 4. The constraint surface π in $\boldsymbol{w}$-$\boldsymbol{x}$ space is characterized with its tangent plane ρ.

This surface is defined with respect to $\boldsymbol{w}$. That is, for every value of $\boldsymbol{w}$, the surface has one value of $\boldsymbol{x}$. We denote this surface with π.

Then, as shown in Fig. 3, our visual servoing is to control the camera and the target object from initial poses $\boldsymbol{w}_{\mathrm{S}}$, where we have the initial image $\boldsymbol{x}_{\mathrm{S}}$, to their goal poses $\boldsymbol{w}_{\mathrm{G}}$, where its image is supposed to be $\boldsymbol{x}_{\mathrm{G}}$, through a path on this constraint surface π. Of course, we do not know the absolute value of $\boldsymbol{w}$. We can only estimate its relative difference $d\boldsymbol{w}$ and the change of the image $d\boldsymbol{x}$ with respect to $d\boldsymbol{w}$.

At almost every $\boldsymbol{w}$, π can be considered to be continuous and smooth. It implies that a small motion of $d\boldsymbol{w}$ produces a small image change of $d\boldsymbol{x}$ as

$$d\boldsymbol{x} = L^{\top} d\boldsymbol{w} \tag{1}$$

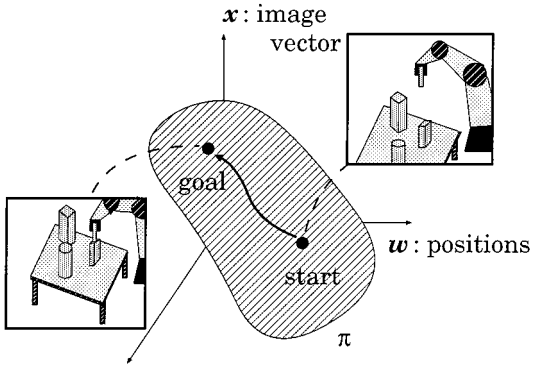

Figure 3. Visual servoing in $\boldsymbol{w}$-$\boldsymbol{x}$ space. The camera and the target object are led to their goal along a path on the constraint surface π.

$L^{\top}$ is called the *Interaction Matrix* or *Image Jacobian*.

This means that the camera can be led to the goal position by forming the visual feedback loop $d\boldsymbol{w} = -\kappa L^{\top +}(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{G}})$, where $L^{\top +}$ is a generalized inverse of $L^{\top}$ and $0 < \kappa < 1$. This $L^{\top}$ can be interpreted as the coefficient matrix of the tangent plane of the surface π at the current camera position (Fig. 4). It should be noted that $L^{\top}$ is given as a function of $\boldsymbol{w}$. If the surface π is flat, which means the positional difference $\boldsymbol{w}$ linearly depends on the image difference $\boldsymbol{x}$, the camera can reach the goal at once with $\kappa = 1$. However, in general, π is not flat, and we must employ iterative feed-back to reach the goal by setting $\kappa < 1$.

We should also emphasize that we cannot obtain the analytical form of π and $L^{\top}$. However, empirically we can obtain an approximation of its derivatives by moving the camera around in small increments and taking images. Conventionally, for the interaction matrix $L^{\top}$, the $\hat{L}^{\top}$ obtained around the goal is commonly used. In Deguchi (1997), we also used this approximation for $L^{\top}$. It worked to bring the camera to its goal, but the trajectories to the goal were not straightforward. In Deguchi and Noguchi (1996), we proposed a dynamic estimation of the $L^{\top}$ to improve the efficiency. It worked well, though it required rather complex online processing, and it may not efficiently apply to the multiple motion controls discussed here.

Therefore, to achieve more efficient control based on our above mentioned approach, in the next subsection we propose a means of reconstructing the constraint surface.

2.2. *Linear Mapping of Images for Optimal Control*

If the constraint surface $\boldsymbol{\pi}$ is flat, the image difference and camera position difference correspond linearly. This means that the Interaction Matrix is constant within an extent around the goal, which results in a linear trajectory of the camera movement to the goal position by $d\boldsymbol{w} = -\kappa L^{\top +}(\boldsymbol{x} - \boldsymbol{x}_{\mathrm{G}})$.

To realize such an effective control, first, we must reduce the dimensionality of the image. This is because the original image dimensionality is too large to evaluate the optimality and difficult to reconstruct the constraint surface directly. Next, we reconstruct the representation space which makes the constraint surface mostly flat around the goal. We achieve both steps by a linear mapping of the image

$$\boldsymbol{x} \longrightarrow \boldsymbol{d}$$

as shown in Fig. 5.

The three step strategy to find such a linear mapping of the image is as follows:

1. For a given goal pose, we take several sample images around the goal. The set of sample images are taken by moving the camera (and the object, for the case to control also the object motion) in small increments.
2. From the image set, by using the eigen space method, we reduce the image dimensionality.
3. From the eigen vectors, we construct the linear mapping which makes the distribution of the sample motion images to be optimally flat with respect to the motion parameters.

When we obtain the optimal image transformation $\boldsymbol{d}$ and good estimate of the Interaction Matrix $\hat{L}^{\top}$ for the $\boldsymbol{d}$, we control the camera and the object movements to their respective goals by iterating the next *look and move* steps:

1. From the image $\boldsymbol{x}$ at current camera and object poses, obtain its eigenspace representation $\boldsymbol{d}$. **(look)**
2. Estimate the difference $\Delta\boldsymbol{w}$ of the current camera pose to the goal pose by
$$\Delta\boldsymbol{w} = \hat{L}^{\top +}(\boldsymbol{d} - \boldsymbol{d}_{\mathrm{G}}) \tag{2}$$
3. Move camera and object by $-\kappa\Delta\boldsymbol{w}$ (where κ is a constant and $0 < \kappa < 1$). **(move)**
4. If $\boldsymbol{d} - \boldsymbol{d}_{\mathrm{G}}$ is not sufficiently small, return back to 1. If it becomes small, the camera and the object are just at their respective goals.

It is known that, if $\hat{L}^{\top +}$ is a good estimate of $L^{\top +}$ in the sense that $L^{\top}\hat{L}^{\top +}$ is always positive definite, the above iteration leads the camera and the object to their goal poses (Espiau et al., 1992).

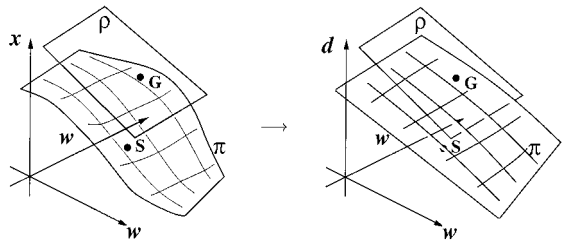

Figure 5. Realization of an effective control by reducing the dimensionality of images and reconstructing the representation space which make the constraint surface mostly flat around the goal.

3. Construction of The Effective Image Transformation for Visual Controls

3.1. *Reduction of the Image Dimensionality by the Eigen Space Method*

First, we briefly describe the eigen space method employed here. We reduce the image dimensionality by the eigen space method. For visual servoing, the dimension reduced representations of the image must maintain properties that are sensitive to small changes in the original image, easy to compute for real-time use, and enable direct and straightforward interpretation to the camera and the object motions. The eigen space method is one feasible technique in this sense (Murase and Nayar, 1993). The image $\boldsymbol{x}$ obtained by the camera is approximated with a linear combination (*K-L expansion*) of M major principal component images in N dimensional data space, where $M \ll N$. These components are given as first M ortho-normal eigen vectors which correspond to M largest eigen values of the $N \times N$ covariance matrix of the sample image set.

We assume a given set of m images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ which are taken around the goal. From this image set, we compute the mean image $\boldsymbol{c}$ as

$$\boldsymbol{c} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i \tag{3}$$

Then, we define the matrix $\boldsymbol{X}$ as

$$\boldsymbol{X} = [\boldsymbol{x}_1 - \boldsymbol{c}, \boldsymbol{x}_2 - \boldsymbol{c}, \ldots, \boldsymbol{x}_m - \boldsymbol{c}], \tag{4}$$

and the covariance matrix $\boldsymbol{Q}$ of the image set of $\{\boldsymbol{x}_i\}$ as

$$\boldsymbol{Q} = \frac{1}{m}\boldsymbol{X}\boldsymbol{X}^\top \tag{5}$$

We denote the first M eigen vectors of $\boldsymbol{Q}$ with $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M$. An effective algorithm to obtain eigen vectors from a set of large dimensional vectors was proposed in Murakami and Vijaya Kumar (1982). An image $\boldsymbol{x}$ within this image set will be well approximated by

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{M} a_i \boldsymbol{e}_i + \boldsymbol{c} \tag{6}$$

where $\boldsymbol{c}$ is the mean vector of the set of $\boldsymbol{x}$.

We call each $\boldsymbol{e}_i$ an *eigen image*, and the set of weight coefficients $\boldsymbol{a} = (a_1, a_2, \ldots, a_M)^\top$ a *reduced image* of $\boldsymbol{x}$, whose components are given as

$$a_i = (\boldsymbol{x} - \boldsymbol{c})^\top \boldsymbol{e}_i \tag{7}$$

Using this formula, the image data dimension M becomes much smaller than N.

3.2. Reconstruction of the Eigen System for Effective Visual Controls

We proposed previously to use the eigen space method itself for visual servoing (Deguchi, 1997; Deguchi and Noguchi, 1996). In those papers, we used the reduced image $\boldsymbol{a}$ of (7) for the original image $\boldsymbol{x}$. It was shown that this introduction of the eigen space method enabled the visual servoing technique without image feature extractions and correspondences. The experimental results showed the proposed method achieved simple and effective robot control which does not need complex image processing procedures. However, the experiments also showed that, in many cases, the pose errors did not decrease monotonically to reach the goal position.

The dominant reason is that the constraint surface $\boldsymbol{\pi}$ for the $\boldsymbol{w}$-$\boldsymbol{a}$ combination is not sufficiently flat around the goal. This is because the eigen system $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M\}$ has not been related to any positional information. They were constructed only from a set of sample images.

If the constraint surface is flat, the image difference and position difference are linearly corresponding, which results in a linear trajectory of the camera movement to the goal position. To realize such an effective control, we must rearrange the eigen system by relating each image to its position where it was obtained around the goal.

Now we introduce a new orthogonal system $\{\phi_1, \phi_2, \ldots, \phi_K\}$ instead of $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M\}$ ($K < M$), where every vector ϕ_j is a linear combination of $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M\}$, to obtain a more effective reduced image $\boldsymbol{d}$ for the control. That is, we construct the system $\{\phi_1, \phi_2, \ldots, \phi_K\}$ as

$$\begin{aligned}\phi_j &= g_{j1}\boldsymbol{e}_1 + g_{j2}\boldsymbol{e}_2 + \cdots + g_{jM}\boldsymbol{e}_M \\ &= E\boldsymbol{g}_j\end{aligned} \tag{8}$$

where $E = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M]$ and $\boldsymbol{g}_j = (g_{j1}, \ldots, g_{jM})^\top$ so that the new reduced image $\boldsymbol{d} = (d_1, \ldots, d_K)$ given by

$$d_j = (\boldsymbol{x} - \boldsymbol{c}, \phi_j), \quad (j = 1, \ldots, K). \tag{9}$$

forms a mostly flat surface in the $\boldsymbol{w}$-$\boldsymbol{d}$ space.

Once, the new eigen system is established, the reduced image $\boldsymbol{d}$ of a newly obtained image $\boldsymbol{x}$ will be obtained easily by (9) in real-time and on-line processing.

However, it should also be noted that, to reconstruct the new eigen system, employing a criterion to evaluate the flatness of the new constraint surface is not enough. This is because we may have a nonsense system that makes all $g_{ij} = 0$. We must also consider the sensitivity to the small change of images. In the next section, we present an algorithm to rearrange the eigen system in order to satisfy these requirements.

4. Optimal Interaction Matrix

4.1. Minimization of Position and Pose Estimation Errors

Now, the problem is to find $\{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_K\}$ to construct $\{\phi_1, \phi_2, \ldots, \phi_K\}$ by (8), which makes the most effective image feature vectors for the control. Before obtaining the orthogonal system, we discuss the optimality of the Interaction Matrix again.

We have a set of sample images $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ taken at the corresponding camera positions $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n\}$. We are using the image feature vector $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iK})^\top$ instead of the $\boldsymbol{x}_i$ for the control. For this case, the optimal Interaction Matrix for this set

of sample images is defined as the matrix which estimates the difference between obtained image and the goal image most accurately from the difference of the camera and the object poses. Here, we discuss how to establish the minimization of the estimation error. We will also discuss the sensitivity of the small difference of the motions and images in the next subsection.

Denoting the Interaction Matrix which minimizes the estimation error by $\hat{L}^\top$, and estimating the reduced image by

$$\hat{\boldsymbol{d}}_i = \hat{L}^\top(\boldsymbol{w}_i - \boldsymbol{w}_\mathrm{G}) + \boldsymbol{d}_\mathrm{G} \tag{10}$$

the $\hat{L}^\top$ is defined as the matrix which minimizes the estimation error

$$\mathcal{E}rr_{\mathrm{ave}}(\hat{L}^\top) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{d}_i - \hat{\boldsymbol{d}}_i\|^2 \tag{11}$$

Let us denote the (jk)-th component of the $\hat{L}^\top$ with l_{jk}, and define the j-th row vector of $\hat{L}^\top$ as $\hat{\boldsymbol{l}}_j = (l_{j1}, l_{j2}, \ldots, l_{jF})^\top$. Then, the estimation error is rewritten as

$$\begin{aligned}\mathcal{E}rr_{\mathrm{ave}}(\hat{L}^\top) &= \frac{1}{n}\mathcal{E}rr_{\mathrm{total}}(\hat{L}^\top) \\ &= \frac{1}{n}\sum_{j=1}^{M} \mathcal{E}rr_j(\hat{\boldsymbol{l}}_j)\end{aligned} \tag{12}$$

where $\mathcal{E}rr_j(\hat{\boldsymbol{l}}_j)$ is the error which depends on $\hat{\boldsymbol{l}}_j$.

This $\mathcal{E}rr_j(\hat{\boldsymbol{l}}_j)$ is given as followings:

First we define

$$Y \equiv [\boldsymbol{x}_1 - \boldsymbol{x}_\mathrm{G}, \boldsymbol{x}_2 - \boldsymbol{x}_\mathrm{G}, \ldots, \boldsymbol{x}_n - \boldsymbol{x}_\mathrm{G}] \tag{13}$$

and

$$\boldsymbol{\beta}_j \equiv Y^\top \boldsymbol{\phi}_j \tag{14}$$

On the other hand, we also define the estimation of this $\boldsymbol{\beta}_j$ by using $\hat{\boldsymbol{l}}_j$ as

$$\hat{\boldsymbol{\beta}}_j \equiv W\hat{\boldsymbol{l}}_j \tag{15}$$

where

$$W \equiv \begin{bmatrix} (\boldsymbol{w}_1 - \boldsymbol{w}_\mathrm{G})^\top \\ \vdots \\ (\boldsymbol{w}_n - \boldsymbol{w}_\mathrm{G})^\top \end{bmatrix} \tag{16}$$

Then the above estimation error depending on the $\hat{\boldsymbol{l}}_j$ is given as

$$\mathcal{E}rr_j(\hat{\boldsymbol{l}}_j) = \frac{1}{n}||\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j||^2 \tag{17}$$

Because the sum of the $\mathcal{E}rr_j(\hat{\boldsymbol{l}}_j)$ makes a total error of $\mathcal{E}rr_{\mathrm{total}}(\hat{L}^\top)$, the minimization of (11) is given by minimizing the $\mathcal{E}rr_j(\hat{\boldsymbol{l}}_j)$. This second minimization is given straightforwardly as $\hat{\boldsymbol{l}}_j = (W^\top W)^{-1}W^\top\boldsymbol{\beta}_j$ and the minimum value is given as

$$\begin{aligned}\mathrm{Min}\{\mathcal{E}rr_j\} &= \frac{1}{n}||\boldsymbol{\beta}_j - W(W^\top W)^{-1}W^\top\boldsymbol{\beta}_j||^2 \\ &= \frac{1}{n}\boldsymbol{\beta}_j^\top(I - W(W^\top W)^{-1}W^\top)\boldsymbol{\beta}_j\end{aligned} \tag{18}$$

4.2. Optimal Base Vectors for Visual Servoing

Based on the previous discussion on the optimality of the Interaction Matrix, we obtain the most effective orthogonal base set of $\{\phi_1, \phi_2, \ldots, \phi_K\}(K < M)$. The following algorithm is an extension of the ODV method for pattern recognition (Hamamoto et al., 1991).

Now we reduce the image $\boldsymbol{x}_i$ into the new reduced image $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \ldots, d_{iK})$ by using the set of new eigen vectors as

$$d_{ij} = (\boldsymbol{x}_i - \boldsymbol{c}, \boldsymbol{\phi}_j) \tag{19}$$

As was discussed, minimizing the estimation error and obtaining the most efficient control is equivalent to obtaining the flatest constraint surface $\boldsymbol{\pi}$ for the camera and the object poses and this new $\boldsymbol{d}$.

However, if we are allowed to choose any arbitrary base vector, we get a nonsense solution $\boldsymbol{d}_i = \boldsymbol{0}$ which makes the estimation error always 0. To avoid such a nonsense solution and, at the same time, to obtain the most sensitive images to the difference of poses, that is, to obtain the flatest constraint surface and, at the same time, the most accurate estimation of the position error to the goal, we must minimize not simply $\{\mathrm{Min}\{\mathcal{E}rr_j\}\}$ but $\{\frac{\mathrm{Min}\{\mathcal{E}rr_j\}}{\sigma_j^2}\}$, where σ_j^2 is the variance of the j-th component of the reduced image. This minimization means to obtain $\boldsymbol{\phi}_j$ which minimizes $\mathcal{E}rr_j$ and, simultaneously, maximizes the variance σ_j^2.

This variance is given by using $X = [\boldsymbol{x}_1 - \boldsymbol{c}, \boldsymbol{x}_2 - \boldsymbol{c}, \ldots, \boldsymbol{x}_n - \boldsymbol{c}]$ and $\boldsymbol{\alpha}_j \equiv X^\top\boldsymbol{\phi}_j$ as $\sigma_j^2 = \frac{1}{n}\boldsymbol{\alpha}_j^\top\boldsymbol{\alpha}_j$.

Thus, we must minimize

$$\frac{\mathrm{Min}\{\mathcal{E}rr_j\}}{\sigma_j^2} = \frac{\boldsymbol{\beta}_j^\top (I - W(W^\top W)^{-1}W^\top)\boldsymbol{\beta}_j}{\boldsymbol{\alpha}_j^\top \boldsymbol{\alpha}_j} \quad (20)$$

This means that the optimal orthonormal system of $\{\phi_1, \phi_2, \ldots, \phi_K\}$ is given by employing the above $\{\boldsymbol{g}_j\}$ which minimizes

$$\begin{aligned}\frac{\mathrm{Min}\{\mathcal{E}rr_j\}}{\sigma_j^2} &= \frac{\boldsymbol{\phi}_j^\top Y(I - W(W^\top W)^{-1}W^\top)Y^\top \boldsymbol{\phi}_j}{\boldsymbol{\phi}_j^\top XX^\top \boldsymbol{\phi}_j} \\ &= \frac{\boldsymbol{g}_j^\top E^\top Y(I - W(W^\top W)^{-1}W^\top)Y^\top E\boldsymbol{g}_j}{\boldsymbol{g}_j^\top E^\top XX^\top E\boldsymbol{g}_j}.\end{aligned} \quad (21)$$

For a vector $\boldsymbol{g}_j$, the form $\boldsymbol{g}_j^\top A\boldsymbol{g}_j / \boldsymbol{g}_j^\top B\boldsymbol{g}_j$ with matrices A and B is well-known as the Rayleigh quotient, and $\boldsymbol{g}_j$ minimizing this form is given as the eigen vector corresponding to the minimum eigen value of $B^{-1}A$. Therefore, applying this theory, this vector $\boldsymbol{g}_j$ is given straightforwardly as an eigen vector corresponding to the minimum eigen value of the matrix

$$\Lambda^{-1}E^\top Y(I - W(W^\top W)^{-1}W^\top)Y^\top E, \quad (22)$$

where $\Lambda = E^\top XX^\top E = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ and the λ_i is the i-th largest eigen value of $Q = \frac{1}{n}XX^\top$.

Once $\boldsymbol{g}_j$ is obtained, we can construct an orthonormal system $\{\phi_1, \phi_2, \ldots, \phi_K\}$ from $\boldsymbol{g}_j$, because $(\phi_i, \phi_j) = (\boldsymbol{g}_i, \boldsymbol{g}_j)$. This can be done by determine $\boldsymbol{g}_1$ first, which minimizes above in (21), then $\boldsymbol{g}_2$ subject to be orthogonal to $\boldsymbol{g}_1$. By iterating this step, we obtain the coefficients $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots$. That is, $\boldsymbol{g}_j$ is searched within the subspace orthogonal to the space spanned by $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{j-1}\}$, and so on.

We summarize the procedure to construct the optimal feature extraction bases for the control in Fig. 6.

5. Experiments for Effective Control

This section presents the experimental results of the simultaneous control of the camera and object motions by the proposed method. The first experiment was to

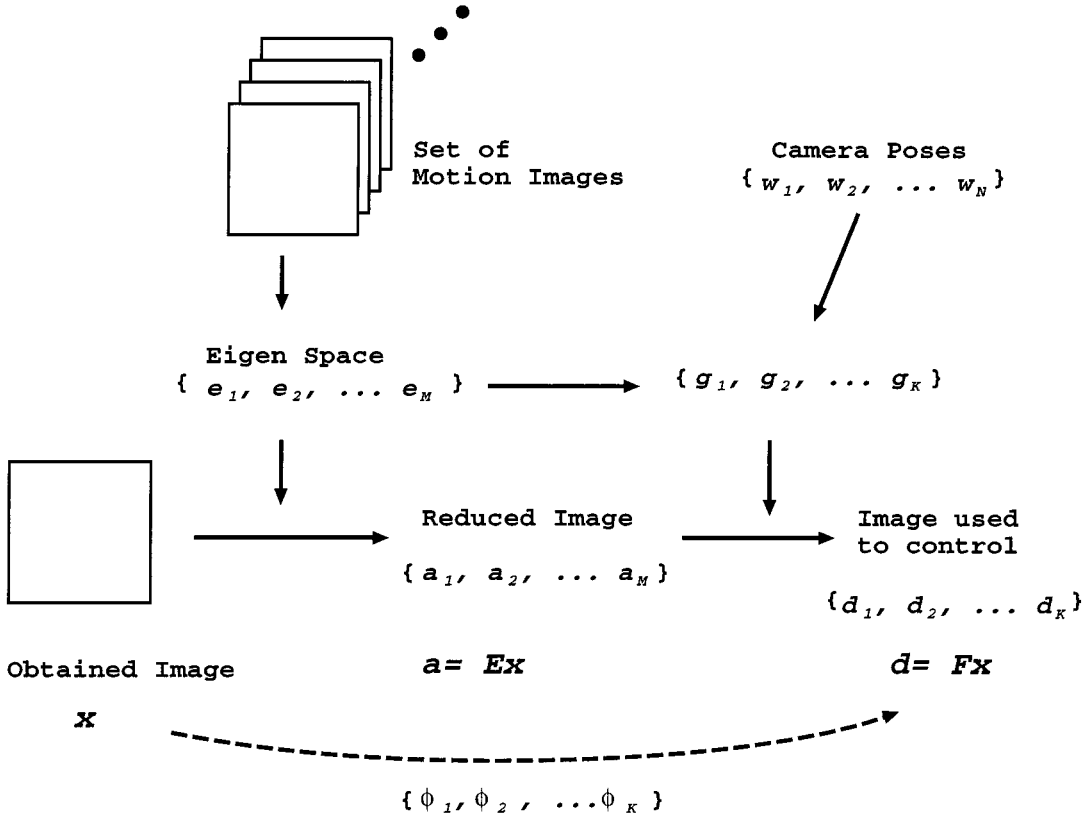

Figure 6. The schematics of the procedure to construct the optimal feature extraction bases for the control.

control the poses of a camera mounted on a robot arm and an object also mounted on another robot arm. The robot arm mounting the camera moved in a scene where the target object was also moving. Other objects in the scene were static.

The second experiment was to control the motions of two cameras and a moving target object in a static scene. The two camera were viewing the target object from different view directions, and two respective goal images were given to specify each camera and the target object goals.

The following experimental results show that our method could be extended for such complex cases.

5.1. *Simultaneous Control of Camera and Object Motions*

The camera was mounted on a robot arm (MITSUBISHI RV-E2), and another robot arm (MITSUBISHI RV-M1) was mounting a target object. The goal image, shown in Fig. 7(a), was taken by the camera at its goal pose. Our objective was to move the camera and the object from their arbitrary initial poses to the poses where the camera could take the same image as the given goal and the target object must be observed just as in the goal image.

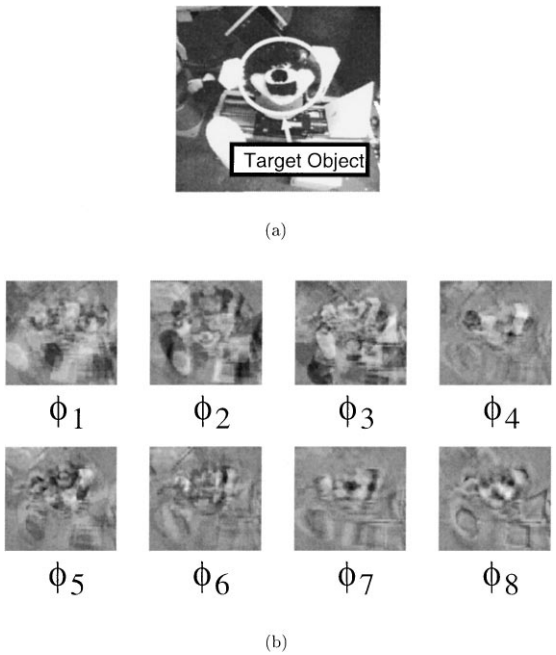

Figure 7. The goal image and the rearranged eigen images for the control of the camera and object motion in the first experiment. (a) Goal image; (b) Rearranged eigen images.

In this experiment, the camera and the object poses each had 3 DOF (x, y, and z positions), So the total DOF was 6. That is, the control parameter vector $\boldsymbol{w}$ was a 6 dimensional vector.

First, we took 61 sample images at different camera positions around and including its goal pose by fixing the target object position at its goal pose. Then, fixing the camera pose at its goal, 60 additional sample images were taken for different object positions around its goal. The total number of sample images was 121.

From these sample images, we obtained 20 eigen images $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{20}\}$ by the eigen space method. Then, using these 20 eigen images, we computed a new set of 8 reconstructed eigen images $\{\phi_1, \ldots, \phi_8\}$. These reconstructed eigen images are shown in Fig. 7.

Figure 8 shows the constraint surface $\boldsymbol{\pi}$ in the $\boldsymbol{d}$-$\boldsymbol{w}$ space. This shows the plottings of (a) component d_1 of the new reduced image $\boldsymbol{d}$ with respect to x coordinate R_x of the camera and x coordinate T_x of the target object, and (b) d_2 with respect to the the robot arm's x and y coordinates R_x and R_y.

Their values were obtained by (9) by applying the set of new eigen images of Fig. 7(b) to images taken at every position of the camera and the object shown in this figure.

This figure shows we had an almost flat surface for the constraint surface $\boldsymbol{\pi}$.

Using these new reconstructed eigen images, we computed the Interaction Matrix, and controlled the camera pose and target object motions with the algorithm described above.

Their motion converged to the goal from a wide range of initial positions. Figure 9 shows the

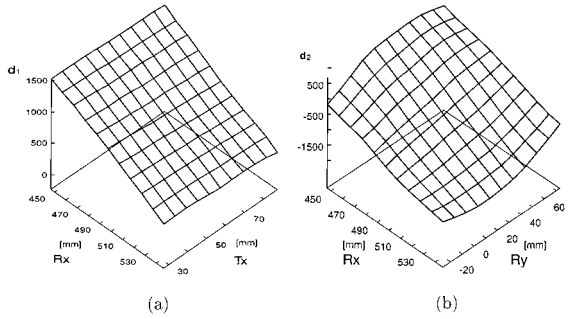

Figure 8. The constraint surface in $\boldsymbol{d}$-$\boldsymbol{w}$ space for the first experiment. (a) d_1 with respect to the camera's x coordinate R_x and the target object's x coordinate T_x; (b) d_2 with respect to the camera's x and y coordinates R_x and R_y.

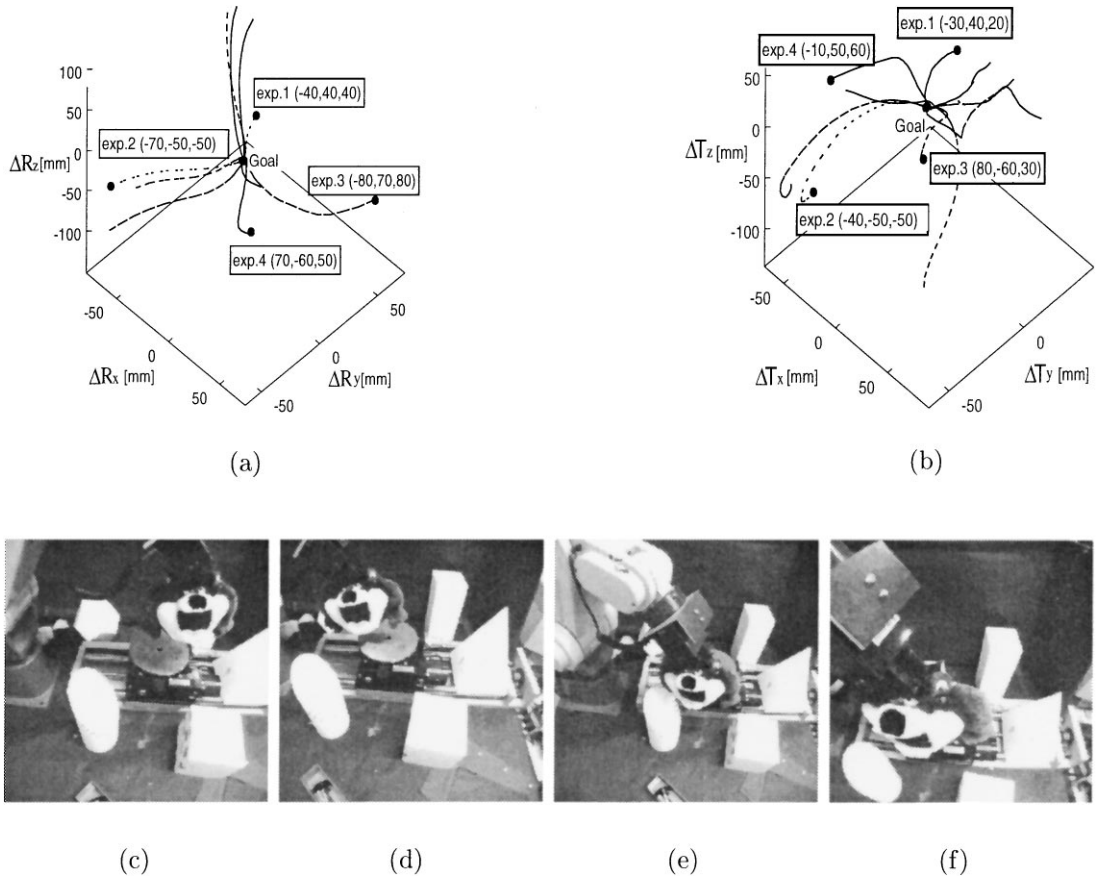

Figure 9. Experimental results starting at several initial positions. (a) and (b) the trajectories of the camera and the target object, respectively, starting at several initial positions, from (c) to (f) four examples of initial images obtained by the camera at the four positions indicated in (a) and (b). ((c) Exp. 1; (d) Exp. 2; (e) Exp. 3; (f) Exp. 4).

trajectories of the camera and the target object starting at several initial positions. In this figure, four examples of initial images obtained by the camera at the four positions indicated in the figures (a) and (b) are shown in from (c) to (f).

5.2. Control of Two Cameras and an Object

Next, we conducted an experiment for a more complicated case (Fig. 10). Here, the target object was moved on a x-ϕ stage which made translation and rotation with 2 DOF. That is, the target object can shift along one direction and rotate in a static scene. This motion was observed by two cameras from different view directions. These cameras were mounted on the robot arms used for the previous experiment. One camera had 6 DOF and the other had 3 DOF in their motions. Therefore, the total number of the control parameters

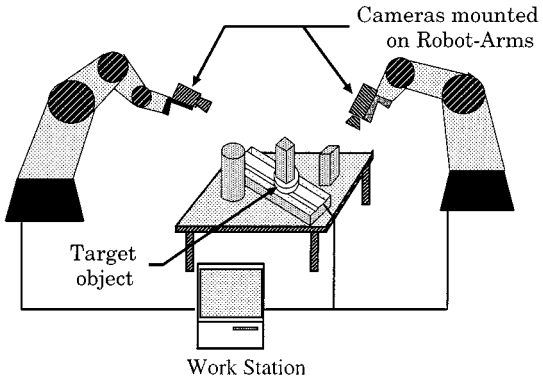

Figure 10. The configuration of the second experiment of two cameras and the object motions.

was 11, which was the dimensionality of $\boldsymbol{w}$ for this case.

We have two images by the two cameras. We combined the two images to form a single image, now twice

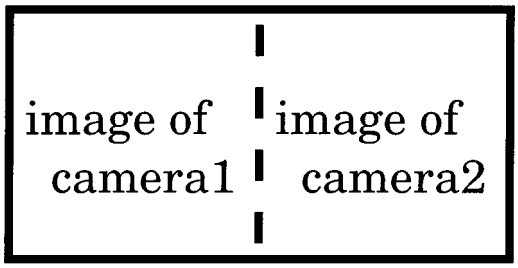

Figure 11. Two images by the two cameras are considered as a single image, twice the original size by combining them.

as large, as shown in Fig. 11. Then, the dimension of the image data N becomes $(2n) \times n$. We treat these combined two images as one observed image.

Figure 12 shows the goal image in this form. It should be noted that, as shown in the goal image, beside the target object, camera 2 itself was also the object for camera 1. This would be too complicated problem for the conventional approach.

For the sample image set, we took 321 images by changing the pose of camera 1 only around its goal pose, 60 images by changing the pose of camera 2 only, 40 images by changing the target object pose, and the goal image. The total number of sample images was 422.

For this case, as the first step, we obtained 70 eigen images by the eigen space method from this sample set. Then, we reconstructed the space for optimal control spanned by 20 newly reconstructed eigen images. In Fig. 13, 9 out of these 20 reconstructed eigen images are shown.

Figure 14 shows an example of the initial image obtained at one of the initial positions of the cameras and the target object. The result of the control from the initial positions of the cameras and the target object is shown in Fig. 15. In this figure, (a) shows the changes of the errors of the positions and poses of camera 1 from its goal with respect to the control steps. (b) shows the errors of camera 2 positions and the position and pose of the object. These plottings show that all the errors converged to 0 monotonically and all positions and poses returned back to their respective goals. (c) shows images taken by the cameras on the way to the goal in this experiment.

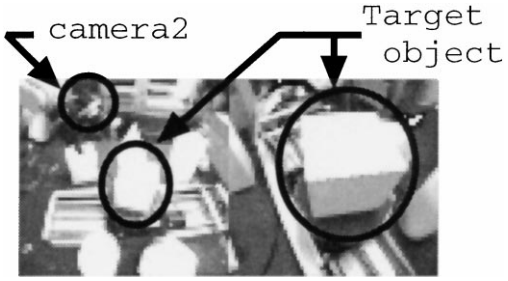

Figure 12. Goal image for the second experiment.

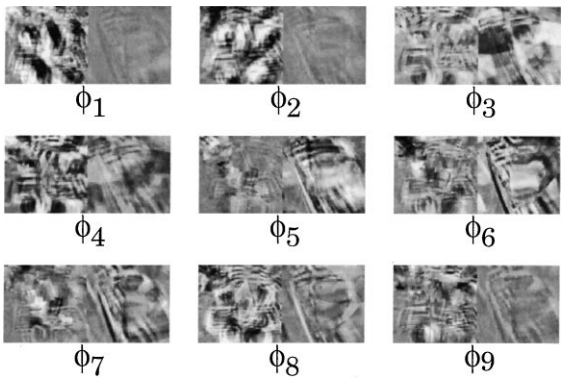

Figure 13. Reconstructed eigen images for the second experiment.

Also, we have made experiments starting from various initial positions in a wide range around the goal, and almost all controls succeeded to return the cameras and the target object to their respective goals. This indicates the feasibility of the proposed method for such complex cases where the simultaneous controls of plural cameras and target objects' motions. Fig. 16 shows other experimental results of the controls starting from four initial positions. (a) shows the trajectories of camera 1, (b) shows the trajectories of camera 2, and from (c) to (f) show the initial images obtained at these four initial positions.

As described above, the goal image of camera 1 included the image of camera 2. Therefore, we also conducted an experiment to simultaneously control the motions of the two cameras and the object by using only the image by camera 1. The comparisons of the final errors for their positions and poses starting from

Figure 14. An initial image.

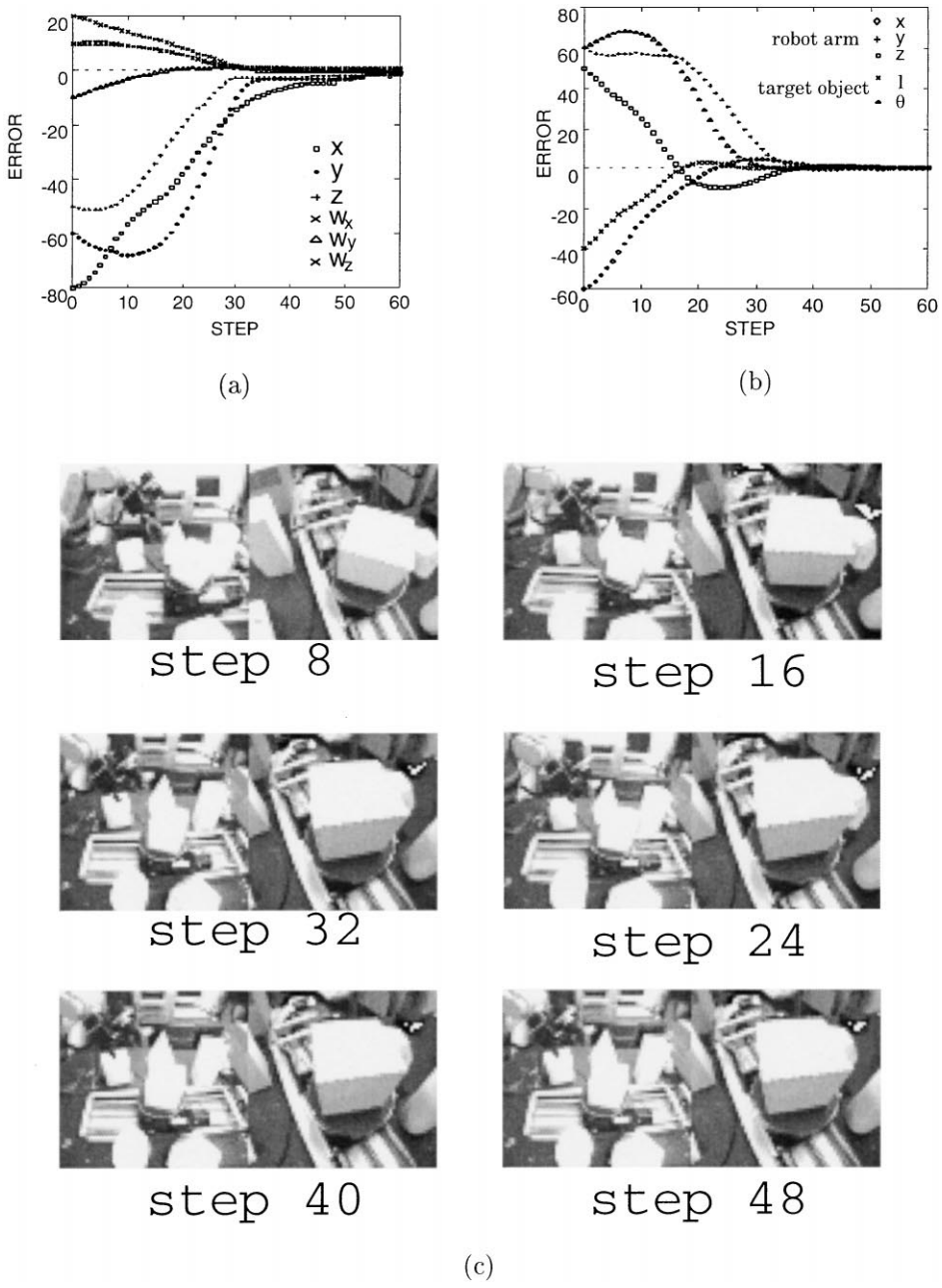

Figure 15. Experimental results of the control from an initial positions where the initial images are shown in Fig. 14. (a) Position and pose errors of camera 1; (b) Position and pose errors of camera 2 and the object; (c) Images taken on the way to the goal.

the same initial positions in Fig. 15 are shown in the Table 1.

This table shows that all the cameras and object could return their respective goals by using only one image, but, by using two cameras, we obtained much higher accurate final poses than in the single camera cases. Of course, we had more and various information from two cameras, and we could utilize this to achieve better control.

However, it should be noted here that such a configuration could not be realized by using the conventional image feature based method. This is due to the fact

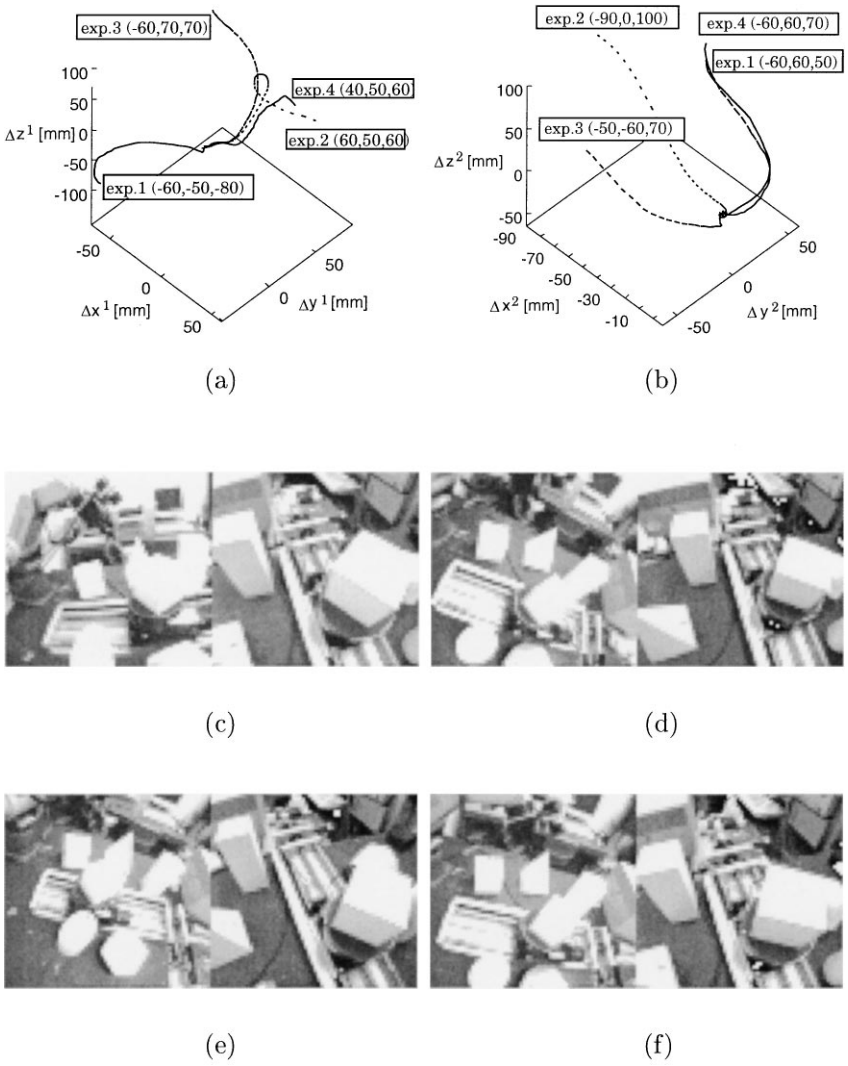

Figure 16. (a), (b) Trajectory of the camera motions from four initial positions. From (c) to (f) are the images taken at the four initial positions shown in (a) and (b). (a) Trajectories of camera 1 motion; (b) Trajectories of camera 2 motion; (c) Exp. 1; (d) Exp. 2; (e) Exp. 3; (f) Exp. 4.

that the Interaction Matrix would be hard to construct properly.

On the other hand, as shown here, all we need to do is to increase the dimensionality of the raw image by a factor of two. We could apply our method to such a complex case and achieve efficient control without any further considerations. This is the main advantage of the proposed method.

Table 1. Comparison of the final errors of the controls using one image and two images (Positions in *mm* and Poses and Rotation in *degree*).

Number of images	Camera 1		Camera 2	Target object	
	Position	Pose	Position	Position	Rotation
2	1.42	0.32	0.45	0.24	0.23
1	5.25	7.25	4.25	12.3	−4.52

6. Conclusions

A general scheme to represent the relation between dynamic images and camera motion was presented. Then, its application to visual servoing was proposed. In the method, image processings to extract image features and to establish their correspondences between dynamic images were not needed. It does

not also require exact camera and hand-eye calibrations.

For the visual servoing we use the eigen space method to reduce the dimensionality of image data. Then, we considered to reconstruct the eigen space to enable more effective control. The result suggests the importance of using proper Interaction Matrices for visual servoing. The results of the experiments promise usefulness of our approach in vision guided robotics.

Acknowledgments

This work was supported by Research Project 10555141, Grant-in-Aid for Scientific Research by Ministry of Education, Japan, and by CDV Project of "Research for the Future" program, The Japan Society for the Promotion of Science.

Notes

1. If the imaging condition is controllable, it can be considered to belong to the components within F dimensional space. If it is only an observed value, it belongs to the N dimensional space.
2. The feasibility of this approach depends on how much this dimension can be reduced. Image feature extraction, where some features characterizing the image are represented with a number of geometrical values can also be considered as a method of this reduction of dimensionality. But, as described above, such an approach still has some difficulties.

References

Deguchi, K. 1997. Direct interpretation of dynamic images and camera motion for visual servoing without image feature correspondence. *Journal of Robotics and Mechatronics*, 9(2):104–110.

Deguchi, K. and Noguchi, T. 1996. Visual servoing using eigenspace method and dynamic calculation of interaction matrices. In *Proc. the 13th International Conference on Pattern Recognition*, Vol. 1, pp. 302–306.

Espiau, B., Chaumette, F., and Rives, P. 1992. A new approach to visual servoing in robotics. *IEEE Trans. on Robotics and Automation*, 8(3):313–326.

Hamamoto, Y., Matsuura, Y., Kanaoka, T., and Tomita, S. 1991. A note on the orthonormal discriminant vector method for feature extraction. *Pattern recognition*, 24(7):681–684.

Hashimoto, K. (Ed.) 1993. *Visual Servoin—Real-Time Control of Robot Manipulators Based on Visual Sensory Feedback*. World Scientific Publishing.

Hutchinson, S., Hager, G.D., and Corke, P.I. 1996. A tutorial on visual servo control. *IEEE Trans. on Robotics and Automation*, RA-12(5):651–670.

Kinoshita, K. and Deguchi, K. 1994. Simultaneous determination of camera pose and intrinsic parameters by visual servoing. In *Proceedings of 12th ICPR*, Vol. A, pp. 285–289.

Murakami, H., Vijaya Kumar, B.V.K. 1982. Efficient calculation of primary images from a set of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 4(5):511–515.

Murase, H. and Nayar, S.K. 1993. Learning and recognition of 3D objects from appearance. In *Proc. of IEEE Workshop on Qualitative Vision*, pp. 39–50.

Weiss, L.E., Sanderson, A.C., and Neauman, C.P. 1987. Dynamic sensor-based control of robots with visual feedback. *IEEE Journal of Robotics and Automation*, RA-3(5):404–417.