2007 IEEE International Conference on
Robotics and Automation
Roma, Italy, 10-14 April 2007

On-line Evolutionary Head Pose Measurement by Feedforward Stereo Model Matching

Wei Song, Mamoru Minami, Yasushi Mae and Seiji Aoyagi

Abstract— **This paper presents a method to estimate the 3D pose of a human's head using two images input from stereo cameras. The proposed method utilizes an evolutionary search technique of genetic algorithm (GA) and a fitness evaluation based on a stereo model matching. To improve the dynamics of recognition, a motion-feedforward method is proposed for the hand-eye system. The effectiveness of the method is confirmed by the experiments where the motion of the hand-eye camera compensated for the relative motion of the object in camera frame, resulting robust recognition against the hand-eye motion.**

I. INTRODUCTION

This work is motivated by our desire to establish a visual system for a patient robot that is used to evaluate the ability of the medical treatments of nurse students, as shown in Fig.1. It is necessary for nurse to pay attention to the condition of the patient during, e.g. injection, to sense tiny sign of patient's state and as a result to avoid medical accidents. What is the most important for nurses is to check the patient's face periodically and carefully to infer their inside conditions. To evaluate this nurse abilities, the patient robot have to contrarily track the nurse's head pose, then the patient robot can judge whether the students can give their patient a good treatment. The behaviors of patient robot to position its head pose relative to the nurse's to observe the nurse's head pose and gazing direction of eyes is one of visual servo to 3D pose.

There is a variety of approaches for head pose estimation, and they can be classified into three general categories: feature-based, appearance-based, and model-based. Feature-based approach is to select a set of feature points like the corners of the eyes or mouth, which are matched against the incoming video to update the estimation pose, [1], [2]. Detection of facial features is not accurate and often fails because it is affected by other parameters depending on identity, distance from the camera, facial expression, noise, illumination changes, and occlusion. Appearance-based (also template-based) approach attempts to define the face as a whole to deal with aspect changes and occlusions, [3], [4]. In [3], the image is compared with a set of reference key-frames from several views to determine which one most closely matches the image. The head pose is evaluated by that of the key-frame, which needs a learning process to be registered offline. The author improved the accuracy by interpolating new key-frames between the predefined key-frames. The third method is to use a 3D solid model to search a target head in the image, and the model is composed based on how the target object can be seen in the input image [5], [6]. Our method is included in this category. The matching degree of the model to the target can be estimated by a function, whose maximum value represents the best matching and can be solved by GA, using the matching function as a fitness function. An advantage of our method is that we use a 3D solid model which enables it to possess six degree of freedom (DOF), both the position and orientation, without following hindrances. In other methods like feature-based recognition, the pose of the target object should be determined by a set of image points, which makes it need a very strict camera calibration. Moreover, searching the corresponding points in Stereo-vision camera images is also complicated and time consuming,e.g., [7].

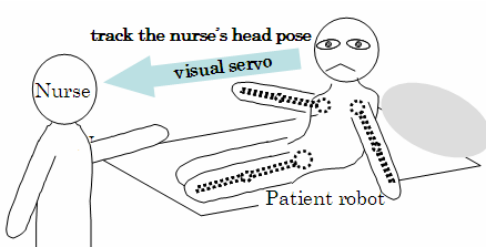


Fig. 1. visual system for a patient robot

GA is well known as a method for solving parameter optimization problems [9]. The GA-based scene recognition method described here can be designated as "evolutionary recognition method", since for every step of the GA's evolution, it struggles to perform the recognition of a target in the input image. To recognize a target input by CCD camera in real-time, and to avoid time lag waiting for the convergence to a target, we used GA in such manner that only one generation is processed to newly input image, which we called "1-Step GA". In this way, the GA searching process and the convergence to the target does not consist in one image but the convergence is achieved in the sequence of the input image to recognize it in the continuously input images. The adaptive searching is also considered by global/local

This is a product of research which was financially supported by the Kansai University Grant-in-Aid for the Faculty Joint Research Program, Feasibility Study Program of Fukui prefecture, Key Research Program in University of Fukui and Incubation Laboratory Factory of University of Fukui, 2006-.

Wei Song, Mamoru Minami and Yasushi Mae are with Graduate School of Engineering, University of Fukui, Fukui, 910–8507, Japan { songwei, minami, mae } @rc.his.fukui-u.ac.jp

Seiji Aoyagi is with Faculty of Engineering, Kansai University, Osaka, 564-8680 Japan aoyagi@iecs.kansai-u.ac.jp

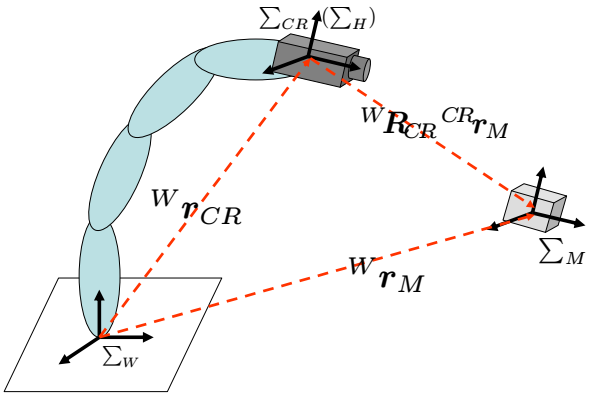


Fig. 2. coordinate system

searching [10], which is switched depending on the matching degree of the target in the images and the model created. During the process of the local searching, gazing operation is suggested to shorten recognition time and raise the accuracy, which is inspired from gazing action of human.

II. Motion-Feedforward Compensation

Most visual servo systems use an eye-in-hand configuration, having the camera mounted on the robot's end-effector. In this case, the motion of the target in the camera coordinate will be effected by both the motion of the target in real world and the motion of the camera itself. Here what we are interested in is how to predict the target velocity based on the motion of the camera. For an eye-in-hand manipulator, the question is how to predict the target velocity based on the joint velocity of manipulator. This can be considered the same as human's action. As human, we can predict the target pose caused by the motion of ourselves. To apply such an intelligence into a manipulator, we propose an robust recognition method, called a motion-feedforward recognition, in which the target velocity is predicted based on the joint velocity of manipulator to compensate the influence from the motion of the camera itself.

A. Kinematics of Hand-Eye

We explain how to describe such a relationship between a target and a moving camera in a mathematical formulation.

First, we establish relations among relative velocities of three frames, world coordinate system Σ_W, target coordinate system Σ_M and camera coordinate systems as Σ_{CR}, shown in Fig.2. Take Σ_W as the fixed reference frame. Denote the vector from O_W (the origin of Σ_W) to O_{CR} expressed in Σ_W as ${}^W\boldsymbol{r}_{CR}$, the vector from O_W to O_M expressed in Σ_W as ${}^W\boldsymbol{r}_M$, and the vector from Σ_{CR} to Σ_M expressed in Σ_{CR} as ${}^{CR}\boldsymbol{r}_{CR,M}$. We define robot's end-effector coordinate system as Σ_H, which is considered same as Σ_{CR} since the camera is mounted on the robot's end-effector. So the rotation matrix ${}^W\boldsymbol{R}_{CR}$ is a function of the joint vector $\boldsymbol{q}$. Then the following relations hold:

$$ {}^{CR}\boldsymbol{r}_{CR,M} = {}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\boldsymbol{r}_M - {}^W\boldsymbol{r}_{CR}(\boldsymbol{q})). \quad (1) $$

Differentiating Eq.1 with respect to time

$$ {}^{CR}\dot{\boldsymbol{r}}_{CR,M} = {}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\dot{\boldsymbol{r}}_M - {}^W\dot{\boldsymbol{r}}_{CR}) + \boldsymbol{S}({}^{CR}\boldsymbol{\omega}_W)\,{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\boldsymbol{r}_M - {}^W\boldsymbol{r}_{CR}(\boldsymbol{q})). \quad (2) $$

where $\boldsymbol{S}(\cdot)$ is the operator performing the cross product between two (3×1) vectors. Given $\boldsymbol{\omega} = [\omega_x, \omega_y, \omega_z]^T$, $\boldsymbol{S}(\boldsymbol{\omega})$ takes on the form

$$ \boldsymbol{S}(\boldsymbol{\omega}) = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}. \quad (3) $$

Similarly, the angular velocities of Σ_{CR} and Σ_M with respect to Σ_W are ${}^W\boldsymbol{\omega}_{CR}$ and ${}^W\boldsymbol{\omega}_M$, respectively, and the angular velocity of Σ_M with respect to Σ_{CR} is ${}^{CR}\boldsymbol{\omega}_{CR,M}$. Then the following relations hold:

$$ {}^{CR}\boldsymbol{\omega}_{CR,M} = {}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\boldsymbol{\omega}_M - {}^W\boldsymbol{\omega}_{CR}). \quad (4) $$

The camera velocity, which is considered as the end-effector velocity, can be expressed using the Jacobian matrix $\boldsymbol{J}(\boldsymbol{q}) = [\boldsymbol{J}_p^T(\boldsymbol{q}), \boldsymbol{J}_o^T(\boldsymbol{q})]^T$,

$$ {}^W\dot{\boldsymbol{r}}_{CR} = \boldsymbol{J}_p(\boldsymbol{q})\dot{\boldsymbol{q}}, \quad (5) $$

$$ {}^W\boldsymbol{\omega}_{CR} = \boldsymbol{J}_o(\boldsymbol{q})\dot{\boldsymbol{q}}, \quad (6) $$

$$ \begin{aligned} \boldsymbol{S}({}^{CR}\boldsymbol{\omega}_W) &= -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{S}({}^W\boldsymbol{\omega}_{CR})\,{}^W\boldsymbol{R}_{CR}(\boldsymbol{q}) \\ &= -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{S}(\boldsymbol{J}_o(\boldsymbol{q})\dot{\boldsymbol{q}})\,{}^W\boldsymbol{R}_{CR}(\boldsymbol{q}). \end{aligned} \quad (7) $$

A detailed deduction of Eq.7 is shown in Appendix A.

The target velocity in Σ_{CR} represented by ${}^{CR}\dot{\boldsymbol{\phi}}_{CR,M}$ is defined as

$$ {}^{CR}\dot{\boldsymbol{\phi}}_{CR,M} = \begin{bmatrix} {}^{CR}\dot{\boldsymbol{r}}_{CR,M} \\ {}^{CR}\boldsymbol{\omega}_{CR,M} \end{bmatrix}, \quad (8) $$

where the translation velocity of Σ_M with respect to Σ_{CR} ${}^{CR}\dot{\boldsymbol{r}}_{CR,M}$ is given in Eq.2, the angular velocity ${}^{CR}\boldsymbol{\omega}_{CR,M}$ is given in Eq.4.

Substituting Eqs.5, 6, 7 to Eqs.2, 4, the target velocity ${}^{CR}\dot{\boldsymbol{\phi}}_{CR,M}$ can be described by a mathematical formulation using $\boldsymbol{a}\times\boldsymbol{b} = -\boldsymbol{b}\times\boldsymbol{a}$, that is, $\boldsymbol{S}(\boldsymbol{a})\boldsymbol{b} = -\boldsymbol{S}(\boldsymbol{b})\boldsymbol{a}$:

$$ \begin{aligned} {}^{CR}\dot{\boldsymbol{\phi}}_{CR,M} &= \begin{bmatrix} {}^{CR}\dot{\boldsymbol{r}}_{CR,M} \\ {}^{CR}\boldsymbol{\omega}_{CR,M} \end{bmatrix} \\ &= \begin{bmatrix} -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{J}_p(\boldsymbol{q}) + {}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) \\ \boldsymbol{S}({}^W\boldsymbol{R}_{CR}(\boldsymbol{q})\,{}^{CR}\boldsymbol{r}_{CR,M})\boldsymbol{J}_o(\boldsymbol{q}) \\ -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{J}_o(\boldsymbol{q}) \end{bmatrix}\dot{\boldsymbol{q}} \\ &\quad + \begin{bmatrix} {}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) & 0 \\ 0 & {}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) \end{bmatrix}\begin{bmatrix} {}^W\dot{\boldsymbol{r}}_M \\ {}^W\boldsymbol{\omega}_M \end{bmatrix} \\ &= \boldsymbol{J}_m(\boldsymbol{q})\dot{\boldsymbol{q}} + \boldsymbol{J}_n(\boldsymbol{q})\,{}^W\dot{\boldsymbol{\phi}}_M. \end{aligned} \quad (9) $$

The relationship $\boldsymbol{J}_n(\boldsymbol{q})$ given by above describes how target pose change in Σ_{CR} with respect to the pose changing of itself in real word. The relationship $\boldsymbol{J}_m(\boldsymbol{q})$ in the same equation describes how target pose change in Σ_{CR} with respect to changing manipulator pose which influences the

recognition from the relative motion of the camera to the object.

In this paper, the target is considered static so we can rewrite Eq.9 as

$$^{CR}\dot{\boldsymbol{\phi}}_{CR,M} = \boldsymbol{J}_m(\boldsymbol{q})\dot{\boldsymbol{q}}. \quad (10)$$

Using Eq.10, we can predict the target velocity in Σ_{CR} based on the joint velocity of manipulator $\dot{\boldsymbol{q}}$.

B. Prediction of Object's Pose

In this paper, the target orientation is expressed by roll, pitch, yaw angles, represented by ϕ, θ, ψ respectively. So the position/orientation of the target based on Σ_{CR} can be expressed by a six-parameter representation $^{CR}\boldsymbol{\psi}_M = [^{CR}\boldsymbol{r}^T_{CR,M}, ^{CR}\boldsymbol{\epsilon}^T_M]^T$, where $^{CR}\boldsymbol{r}_{CR,M} = [t_x, t_y, t_z]^T$, $^{CR}\boldsymbol{\epsilon}_M = [\phi, \theta, \phi]^T$.

The target's position/orientation volecity is defined as

$$^{CR}\dot{\boldsymbol{\psi}}_M = \begin{bmatrix} ^{CR}\dot{\boldsymbol{r}}_{CR,M} \\ ^{CR}\dot{\boldsymbol{\epsilon}}_M \end{bmatrix}, \quad (11)$$

where the time derivation of target's position $^{CR}\dot{\boldsymbol{r}}_{CR,M}$ is given by Eq.9. The relation between the time derivative of $^{CR}\boldsymbol{\epsilon}_M$ and the body angular velocity $^{CR}\boldsymbol{\omega}_{CR,M}$ is given by the inverse of matrix $\boldsymbol{J}_c$

$$^{CR}\dot{\boldsymbol{\epsilon}}_M = \boldsymbol{J}_c{}^{-1}{}^{CR}\boldsymbol{\omega}_{CR,M}, \quad (12)$$

where

$$\boldsymbol{J}_c = \begin{bmatrix} 0 & -sin\phi & cos\phi cos\theta \\ 0 & cos\phi & sin\phi cos\theta \\ 1 & 0 & cos\psi cos\theta \end{bmatrix}. \quad (13)$$

The body angular velocity $^{CR}\boldsymbol{\omega}_{CR,M}$ is also given by Eq.9.

Then the position/orientation of the target in time $t + \Delta t$ can be predicted from the current end-effector motion, presented by

$$^{CR}\hat{\boldsymbol{\psi}}_M(t + \Delta t) = ^{CR}\boldsymbol{\psi}_M(t) + ^{CR}\dot{\boldsymbol{\psi}}_M \Delta t. \quad (14)$$

$^{CR}\dot{\boldsymbol{\psi}}_M \Delta t$ is the changing extent from the current pose to the next. We consider that the recognition ability will be improved by using Eq.14 to predict the future pose of the target based on the relative motion from the camera to the object. And the recognition will be robust to the motion of manipulator itself.

III. Evolutionary Recogniton

A. Kinematics of Stereo-Vision

We utilize a perspective projection as projection transformation. The coordinate systems of left and right cameras and object (here we take a solid head model as an example) in Fig.3 consist of world coordinate system as Σ_W, model coordinate system as Σ_M, camera coordinate systems as Σ_{CR} and Σ_{CL}, image coordinate systems as Σ_{IR} and Σ_{IL}. A point i on a solid model of the target head can be described using these coordinates and homogeneous transformation matrices. At first, a homogeneous transformation matrix from Σ_{CR} to Σ_M is defined as $^{CR}\boldsymbol{T}_M$. And an arbitrary point i

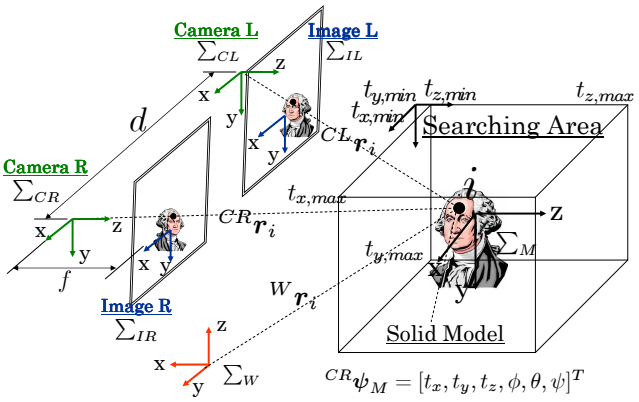


Fig. 3. Coordinate systems

on the target object in Σ_{CR} and Σ_M is defined $^{CR}\boldsymbol{r}_i$ and $^{M}\boldsymbol{r}_i$. Then $^{CR}\boldsymbol{r}_i$ is,

$$^{CR}\boldsymbol{r}_i = {}^{CR}\boldsymbol{T}_M \, {}^{M}\boldsymbol{r}_i. \quad (15)$$

The position vector of i point in right image coordinates, $^{IR}\boldsymbol{r}_i$ is described by using projection matrix $\boldsymbol{P}$ of camera as,

$$^{IR}\boldsymbol{r}_i = \boldsymbol{P} \, {}^{CR}\boldsymbol{r}_i. \quad (16)$$

Using a homogeneous transformation matrix of fixed values defining the kinematical relation from Σ_{CL} to Σ_{CR}, $^{CL}\boldsymbol{T}_{CR}$, $^{CL}\boldsymbol{r}_i$ is,

$$^{CL}\boldsymbol{r}_i = {}^{CL}\boldsymbol{T}_{CR} \, {}^{CR}\boldsymbol{r}_i. \quad (17)$$

By the same way as we have obtained $^{IR}\boldsymbol{r}_i$, $^{IL}\boldsymbol{r}_i$ is described by the following Eq.18 through projection matrix $\boldsymbol{P}$.

$$^{IL}\boldsymbol{r}_i = \boldsymbol{P} \, {}^{CL}\boldsymbol{r}_i \quad (18)$$

Then position vectors projected in the Σ_{IR} and Σ_{IL} of arbitrary point i on target object can be described $^{IR}\boldsymbol{r}_i$ and $^{IL}\boldsymbol{r}_i$. The position and orientation of Σ_M based on Σ_{CR} has been defined as $^{CR}\boldsymbol{\psi}_M$, here, we abbreviate $^{CR}\boldsymbol{\psi}_M$ to $\boldsymbol{\psi}$. So Eqs.16, 18 are rewritten as,

$$\begin{cases} ^{IR}\boldsymbol{r}_i = \boldsymbol{f}_R(\boldsymbol{\psi}, \ ^{M}\boldsymbol{r}_i) \\ ^{IL}\boldsymbol{r}_i = \boldsymbol{f}_L(\boldsymbol{\psi}, \ ^{M}\boldsymbol{r}_i). \end{cases} \quad (19)$$

This relation connects the arbitrary points being predefined and fixed on the object and projected points on the left and right images with the variables $\boldsymbol{\psi}$, which is considered to be unknown in this paper. Then, $^{IR}\boldsymbol{r}_i$ and $^{IL}\boldsymbol{r}_i$ are thought to be moved by $\boldsymbol{\psi}(t)$ and the 3D solid model is also. This measurement problem of $\boldsymbol{\psi}(t)$ in real time will be solved by consistent convergence of a matching model to the target object by a "1-Step GA" which will be explained in section III C. When evaluating each point above, the matching problem of corresponding point in left and right images mentioned in the introduction is arisen. Therefore, to avoid this problem, the 3D model-based matching that treats the points of the object model as a set, is chosen instead of point-based corresponding.

The 3D model for head located in Σ_M is shown in Fig.4. The set of coordinates of head's surface is depicted as S_{in},

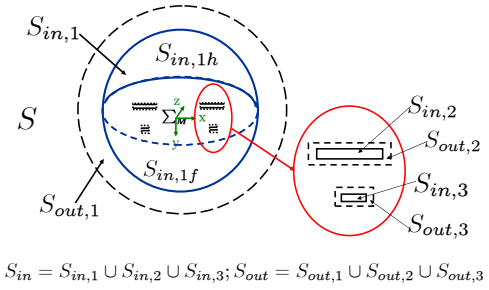
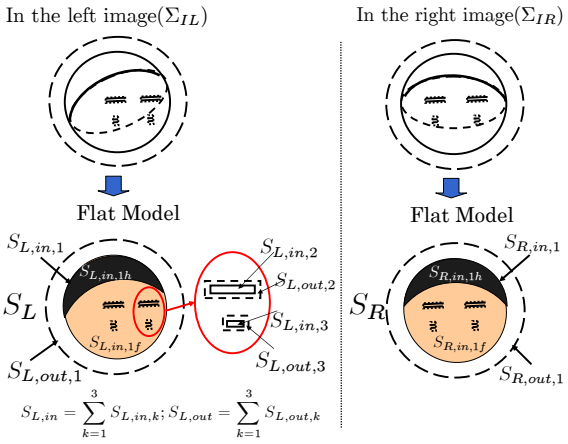


Fig. 4. 3D Solid Model



(a) Left searching model (b)Right searching model

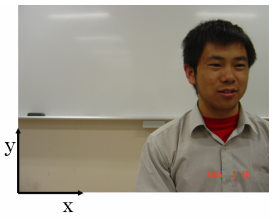
Fig. 5. Searching model

$S_{in} = S_{in,1} \bigcup S_{in,2} \bigcup S_{in,3}$ where the inside surface of head called $S_{in,1}$, the inside surface of eyebrows $S_{in,2}$ and the inside surface of eyes $S_{in,3}$. The outside space enveloping S_{in} is denoted as S_{out}, consisting of $S_{out,1}, S_{out,2}, S_{out,3}$ corresponding to $S_{in,1}, S_{in,2}, S_{in,3}$ respectively. The combination of S_{in} and S_{out} is named as S. $S_{in,1}$ is divided into two parts, one is hair area called $S_{in,1h}$, the other is face area called $S_{in,1f}$. Then, the set of the points of solid searching model S consisted of S_{in} and S_{out} are projected onto 2D coordinates of left camera, expressed as

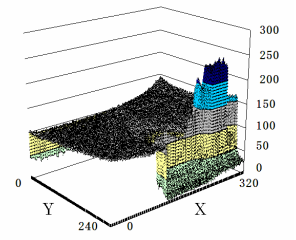
$$S_{L,in}(\boldsymbol{\psi}) = \sum_{k=1}^{3} S_{L,in,k} = \sum_{k=1}^{3} \{ {}^{IL}\boldsymbol{r}_i \in \Re^2 \mid {}^{IL}\boldsymbol{r}_i = f_L(\boldsymbol{\psi}, {}^{M}\boldsymbol{r}_i), {}^{M}\boldsymbol{r}_i \in S_{in,k} \in \Re^3 \} \quad (20)$$

$$S_{L,out}(\boldsymbol{\psi}) = \sum_{k=1}^{3} S_{L,out,k} = \sum_{k=1}^{3} \{ {}^{IL}\boldsymbol{r}_i \in \Re^2 \mid {}^{IL}\boldsymbol{r}_i = f_L(\boldsymbol{\psi}, {}^{M}\boldsymbol{r}_i). {}^{M}\boldsymbol{r}_i \in S_{out,k} \in \Re^3 \} \quad (21)$$

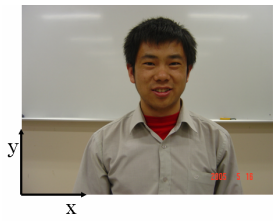
The inside surface of the model in the left camera is called $S_{L,in}$ and the contour-strips is called $S_{L,out}$. $S_{L,in,1}$ also includes two parts $S_{L,in,1h}$ and $S_{L,in,1f}$ corresponding to the hair and face in the left image. The left searching model projected to left camera coordinates is shown in Fig.5(a). The area composed of $S_{L,in}$ and $S_{L,out}$ is named as S_L. The



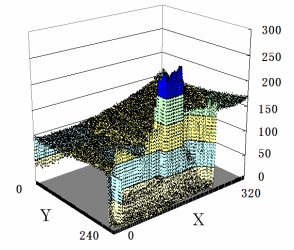
(a)Left image



(b)Left brightness



(c)Right image



(d)Right brightness

Fig. 6. Input image and brightness distribution

above defines only the left-image searching model, the right one is defined in the same way and the projected searching model is shown in Fig.5(b).

B. Definition of Evaluation Function

Here, we define evaluation function to estimate how much the moving solid model S defined by $\boldsymbol{\psi}$ lies on the target being imaged on the left and right cameras. The input images will be directly matched by the projected moving models, S_L and S_R, which are located by only $\boldsymbol{\psi}$. Therefore, if the camera parameters and kinematical relations are completely accurate, and the solid searching model describes precisely the target object shape, then $S_{L,in}$ and $S_{R,in}$ will be completely lies on the target reflected on the left and right images.

The 2D raw images of a target human are shown in Fig.6(a) and (c), their corresponding 3D plot are shown in Fig.6(b) and (d). In these figures, the vertical axis represents the brightness values, where we define 255 as black and 0 as white, and the horizontal axes represent the image coordinates. In order to search for the head in the input images, the searching model shown in Fig.4 and its position calculated by Eqs.20, 21 are used. Take the left image in Fig.6(a) as an example. The brightness distribution of input image lying on the area of searching model is expressed as $p({}^{IL}\boldsymbol{r}_i), \boldsymbol{r}_i \in S_L(\boldsymbol{\psi})$, then the evaluation function of the moving searching model using brightness values is given as,

$$F_{L,gray}(\boldsymbol{\psi}) = \frac{1}{H_1} \Big(\sum_{{}^{IR}\boldsymbol{r}_i \in S_{L,in}(\boldsymbol{\psi})} p({}^{IL}\boldsymbol{r}_i) - \sum_{{}^{IR}\boldsymbol{r}_i \in S_{L,out}(\boldsymbol{\psi})} p({}^{IL}\boldsymbol{r}_i) \Big), \quad (22)$$

where H_1 represents the value of the searching points in $S_{L,in}$ multiplied by 255 (maximum value). It is a scaling factor that normalized $F_{L,gray} \leq 1$. In case of $F_{L,gray}(\boldsymbol{\psi})<0$, $F_{L,gray}(\boldsymbol{\psi})$ is given to zero.

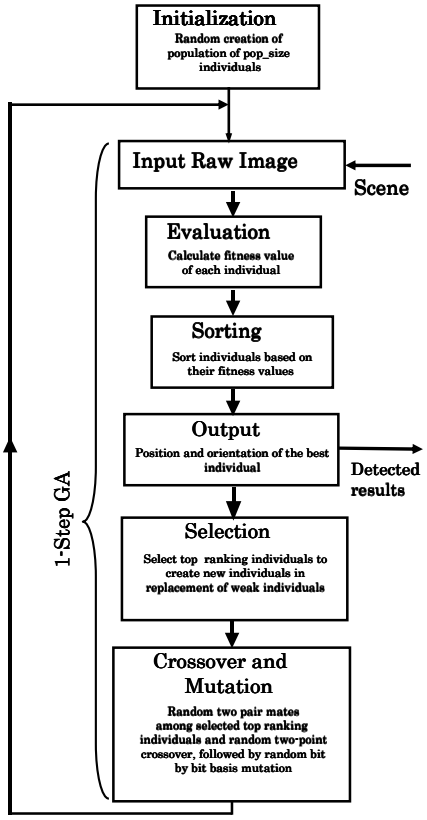


Fig. 7. Flow chart of 1-Step GA recognition

In our previous research, we have confirmed that the robustness of recognition against the background can be improved by adding extraction of the skin color which is one of human characteristic to the evaluation function [8]. It is easy to understand that the color of skin can be limited only by hue value in HSV parameters. Let $h({}^{IL}\boldsymbol{r}_i)$ denote the hue value at the image position ${}^{IL}\boldsymbol{r}_i$. The pose estimation of the searching models by color is given as

$$F_{L,color}(\boldsymbol{\psi}) = \frac{1}{H_2}\Big(\sum_{{}^{IR}\boldsymbol{r}_i \in S_{L,in,1f}(\boldsymbol{\psi})} \boldsymbol{a}({}^{IL}\boldsymbol{r}_i) + \sum_{{}^{IL}\boldsymbol{r}_i \in S_{L,in,1h}(\boldsymbol{\psi})} \boldsymbol{b}({}^{IL}\boldsymbol{r}_i)\Big), \quad (23)$$

where

$$\boldsymbol{a}({}^{IL}\boldsymbol{r}_i) = \begin{cases} 1 & 0 < h({}^{IL}\boldsymbol{r}_i) < 30 \\ 0 & otherwise \end{cases} \quad (24)$$

$$\boldsymbol{b}({}^{IR}\boldsymbol{r}_i) = \begin{cases} 1 & p({}^{IR}\boldsymbol{r}_i) > 220 \ \ and \ \ h({}^{IR}\boldsymbol{r}_i) = 0 \\ 0 & otherwise \end{cases} \quad (25)$$

where H_2 represents the number of the searching points in $S_{L,in,1f}$. It is a scaling factor that normalized $F_{L,color} \leq 1$. In case of $F_{L,color}(\boldsymbol{\psi})$<0, $F_{L,color}(\boldsymbol{\psi})$ is given to zero.

Adding Eqs.22 and 24, the left estimation function is given as

$$F_L(\boldsymbol{\psi}) \quad = \quad F_{L,gray}(\boldsymbol{\psi}) + F_{L,color}(\boldsymbol{\psi}). \quad (26)$$

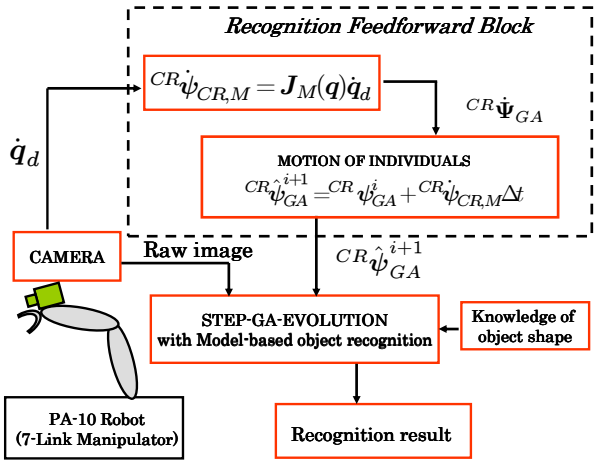


Fig. 8. Feedforward recognition system

The right one is defined in the same way, so the entire matching function is

$$F(\boldsymbol{\psi}) \quad = \quad (F_L(\boldsymbol{\psi}) + F_R(\boldsymbol{\psi}))/2. \quad (27)$$

Eq.27 is used as a fitness function in GA process. When the moving searching models fits to the target head being imaged in the right and left images, then the fitness function $F(\boldsymbol{\psi})$ gives maximum value. Therefore the problem of head pose recognition can be converted to searching problem of $\boldsymbol{\psi}$ such that maximizes $F(\boldsymbol{\psi})$. To recognize the target object in a short time, we solve this optimization problem by GA whose gene representing ${}^{CR}\boldsymbol{\psi}_{GA}$ is defined as,

$$\overbrace{01\ldots01}^{t_x}\overbrace{00\ldots01}^{t_y}\overbrace{11\ldots01}^{t_z}\overbrace{01\ldots01}^{\phi}\overbrace{01\ldots01}^{\theta}\overbrace{01\ldots01}^{\psi}.$$
$$\underbrace{\quad}_{12bit}\ \underbrace{\quad}_{12bit}\ \underbrace{\quad}_{12bit}\ \underbrace{\quad}_{12bit}\ \underbrace{\quad}_{12bit}\ \underbrace{\quad}_{12bit}$$

The 72 bits of gene refers to the range of the searching area: $-150 \leq t_x, t_y \leq 150$, $900 \leq t_z \leq 1200[mm]$, $-20 \leq \phi, \theta, \psi \leq 20[deg]$.

C. On-line Evolutionary Recognition

Although GA have been applied to a number of robot control systems [11], it has not been yet applied to a robot manipulator control system to track a target in 3D space with unpredictable movement in real time, since the general GA method costs much time until its convergence. So here, for real-time visual control purposes, we employ GA in a way that we denoted as "1-Step GA" evolution. This means that the GA evolutional iteration is applied one time to the newly input image. While using the elitist model of the GA, the position/orientation of a target can be detect in every new image by that of the searching model given by the top gene in the GA. That is, the evolving speed to the solution in the image should be faster than the speed of the target object in the successively input images, for the success of real-time recognition by "1-Step GA". The flow chart of the 1-step GA process is shown in Fig.7. We exploit this on-line results of the GA in every newly input image for the feedback signal to the manipulator's controller. Thereby real-time visual servo can be performed. Our previous research

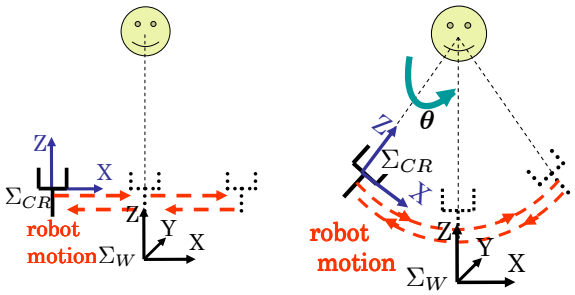


Fig. 9. (a) Shuttle motion of end-effector in x axis of Σ_W (from $x = 120$ to $-120[mm]$). (b) Shuttle motion of end-effector to see the face from $\theta = -7[deg]$ to $\theta = 7[deg]$.

has confirmed the 2D recognition method enabled a eye-in-hand robot manipulator to catch a swimming fish by a net equipped at the hand [10].

However, as the searching space extending to 3D, the time of each GA process will become longer since the parameters are increased to six, three for position and three for orientation. So it becomes more difficult for a robot manipulator to track a target in 3D space in real-time even by using "1-Step GA" method. The proposed motion-feedforward recognition method can help us to conduct such a task since it can predict the motion of the target observed from the cameras by using the known motion of the robot. Using Eq.14, the pose of the individuals ${}^{CR}\hat{\psi}_{GA}$ in the next generation can be predicted from the current end-effector motion, presented by

$$ {}^{CR}\hat{\psi}_{GA}^{i+1} = {}^{CR}\psi_{GA}^{i} + {}^{CR}\dot{\psi}_{CR,M}\Delta t. \qquad (28) $$

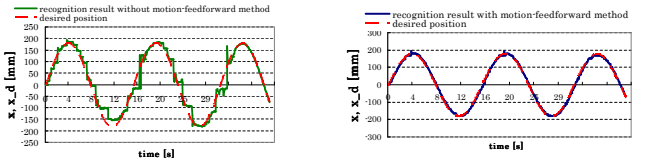
The recognition system of the proposed method is shown in Fig.8. We consider that the recognition ability will be improved by using Eq.28 to move all the individuals to compensate the influence of the motion of the camera. So the recognition is expected to be robust to the motion of robot itself.

IV. Experiment of Recognition

To verify the effectiveness of the proposed motion-feedforward recognition, we have conducted the experiment to recognize a static human head pose with two cameras which are mounted on the robot end-effector. The image processing board, CT-3001, receiving the image from the CCD cameras in real time (30[fps]), is connected to the DELL Optiplex GX1 (CPU: Pentium2, 400 MHz) host computer. Here, we used a doll as the target to eliminate the natural shake of a human being. Two kind of motion has been given to the robot end-effector while recognizing the doll's head pose. We will show effectiveness of the proposed motion-feedforward recognition method by comparing with the recognition result without using motion-feedforward under two robot's motions respectively as follows.

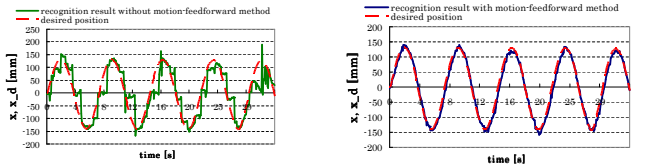
(1) Recognition under translational motion "A": given position changing of end-effector (shown in Fig.9(a))

In this case, the shuttle motion in x axis of Σ_W (from $x = 120$ to $-120[mm]$) is given to the robot end-effector. Here, we fixed the orientation of searching models to true



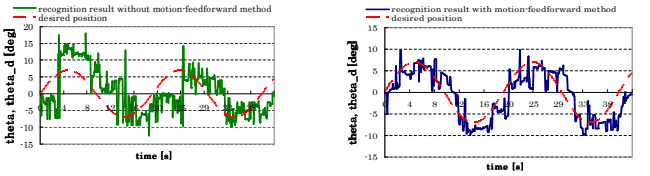
(a) (b)

Fig. 10. Recognition under motion "A" with period $T = 15[s]$. (a) Recognition result of position x without using motion-feedforward method compared with the desired position in Σ_{CR}. (b) Recognition result with motion-feedforward method compared with the desired position in Σ_{CR}.



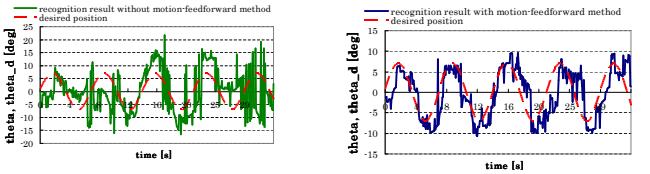
(a) (b)

Fig. 11. Recognition under motion "A" with period $T = 7[s]$. (a) (b) is the same meaning as that in Fig.10.



(a) (b)

Fig. 12. Recognition under motion "B" with period $T = 20[s]$. (a) Recognition result of orientation θ without using motion-feedforward method compared with the desired position in Σ_{CR}. (b) Recognition result of orientation θ with motion-feedforward method compared with the desired position in Σ_{CR}.



(a) (b)

Fig. 13. Recognition under under motion "B" with period $T = 10[s]$. (a) (b) is the same meaning as that in Fig.12.

values as $(\phi, \theta, \psi) = (0, 0, 0)[deg]$, so the searching area is $-150 \leq t_x, t_y \leq 150$, $900 \leq t_z \leq 1200[mm]$. Fig.10 shows the recognition under motion "A" with period $T = 15[s]$. Fig.10(a) is the recognition result of position x without using motion-feedforward method compared with the desired position x_d in Σ_{CR}. Fig.10(b) is x and x_d with motion-feedforward method. The recognition result of position y and z is not shown because the target was not moving in those directions, so it is easy to track and the errors are almost 0. The recognition error of x without using the motion-feedforward method got larger, however, it could be constrained to the extent of $15[mm]$ with motion-feedforward method. It means

that the model can match the target better when using the motion-feedforward recognition method.

When the hand motion period is shorten to $7[s]$, the recognition error got much bigger without using the motion-feedforward method as shown in Fig.11(a). Tracking of the target for GA became more difficult when the speed of the end-effector got quicker, which caused GA's convergence speed is not faster than the target speed relative to the camera. In comparison, the recognition error can still keep in the extent of $15[mm]$ using motion-feedforward recognition method, shown in Fig.11(b). From this experiment, we can say that our proposed motion-feedforward recognition method are robust to the position motion of robot.

(2) Recognition under translational and rotational motion "B": given orientation changing of end-effector (shown in Fig.9(b))

Here, the orientation changing of end-effector is defined as the motion in a circle with a fixed distance to the target, keeping the eye-line (z axis of Σ_{CR}) passing the center of the target. The shuttle motion looking the target from the left side to the right side from $-7[deg]$ to $7[deg]$ is given to the robot end-effector. Here, we fixed (ϕ, ψ) of searching models to true values as $(0, 0)[deg]$, so the searching area is $-150 \leq t_x, t_y \leq 150$, $900 \leq t_z \leq 1200[mm]$, $-20 \leq \psi \leq 20[deg]$.

In the same way as the previous experiment, we gave the shuttle motion with different period time $20[s]$ and $10[s]$, in which the corresponding velocity of the end-effector become quicker. Figs.12(a) ($20[s]$) and 13(a) ($10[s]$) is the recognition result of orientation θ without using motion-feedforward method compared with the desired orientation θ_d in Σ_{CR} (the erorrs of other parameters is not shown since they are almost 0). Figs.12(b) and 13(b) are the results of θ and θ_d with motion-feedforward method. The error of the recognition result without motion-feedforward method changed much bigger than that with the motion-feedforward method. It confirms the effectiveness of the motion-feedforward method, the recognition can be robust to the motion of robot itself because it compensates the influence of the motion of the camera.

V. CONCLUSIONS AND FUTURE WORKS

We have proposed a 3D head pose measurement method which utilizes a genetic algorithm (GA) and model-based matching. The head pose evaluation is based on a fitness function which is composed of head (including facial feature: eyes and eyebrows) detection and pose estimation by color. We have proposed an motion-feedforward recognition method, which is confirmed to be robust by the experiments since it can make a good prediction to compensated for the relative motion of the object in camera frame.

As future research, we will continue to work on improving the accuracy and speed of the recognition. Try to build a stable visual servo system (6DOF) to human face.

VI. ACKNOWLEDGMENTS

This is a product of research which was financially supported by the Kansai University Grant-in-Aid for the Faculty Joint Research Program, Feasibility Study Program of Fukui prefecture, Key Research Program in University of Fukui and Incubation Laboratory Factory of University of Fukui, 2006-.

REFERENCES

[1] A.H.Gee and R.Cipolla. Determining the gaze of face in images. Technical Report CUED/F-INFENG/TR 174, Trumpington Street, Cambridge CB2 1PZ, England, 1994.

[2] R.Yang and Z.Zhang, "Model-based head pose tracking with stereovision", Proc.5th IEEE Int.Conf.on Automatic Face and Gesture Recogniton, 2002.

[3] L.Vacchetti, V.Lepetit and P.Fua, "Stable Real-Time 3D Tracking Using Online and Offline Information", IEEE Transactions on Pattren Analysis and Maching Intelligence, Vol.26, No.10, October 2004,

[4] F.Jurie and M.Dhome, "Real-Time 3D Template Matching", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01), 2001.

[5] S.Yamane, M.Izumi, K.Fukunaga, "A Method of Model-Based Pose Estimation", IEICE, Vol.J79-D-2, No.2, pp.165-173, Feb, 1996.

[6] F.Toyama, K.Shoji, J.Miyamichi, "Pose Estimation from a Line Drawing Using Genetic Algorithm", IEICE, Vol.J81-D-2, No.7, pp.1584-1590, July, 1998.

[7] Y.Maeda, G.Xu, "Smooth Matching of Feature and Recovery of Epipolar Equation by Tabu Search", IEICE, Vol.J83-D-2, No.3, pp.440-448, 1999.

[8] H. Suzuki, M. Minami *Real-time face detection using hybrid GA based on selective attention*, IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS2004).

[9] D. E.Goldberg, *Genetic algorithm in Search, Optimization and Machine Learning. Reading*, Addison-Wealey, 1989.

[10] H.Suzuki, M.Minami :"Visual Servoing to Catch Fish Using Global/Local GA Search", IEEE/ASME Transactions on Mechatronics, Vol.10, Issue 3, 352-357 (2005,6).

[11] T.Nagata, K.Konishi and H.Zha: *Cooperative manipulations based on Genetic Algorithms using contact information*, Proceedings of the International Conference on Intelligent Robots and Systems, pp.400-5, 1995

APPENDIX

A. A proof of Eq.7

Consider two orthogonal coordinate frame Σ_A and Σ_B, and let ${}^A\boldsymbol{\omega}_B$ denote the angular velocity of Σ_B with respect to Σ_A, ${}^B\boldsymbol{\omega}_A$ denote the angular velocity of Σ_A with respect to Σ_B. The relation of ${}^A\boldsymbol{\omega}_B$ and ${}^B\boldsymbol{\omega}_A$ will be derived here.

The rotation matrix ${}^A\boldsymbol{R}_B$ satisfies

$$ {}^A\boldsymbol{R}_B {}^B\boldsymbol{R}_A = \boldsymbol{I}, \tag{A.1} $$

the time derivative of Eq.(A.1) is given by

$$ \frac{d}{dt}({}^A\boldsymbol{R}_B){}^B\boldsymbol{R}_A + {}^A\boldsymbol{R}_B\frac{d}{dt}({}^B\boldsymbol{R}_A) = \boldsymbol{0}. \tag{A.2} $$

For an arbitrary vector ${}^B\boldsymbol{p}$ expressed in Σ_B, we have

$$ \begin{aligned} \frac{d}{dt}({}^A\boldsymbol{R}_B){}^B\boldsymbol{p} &= {}^A\boldsymbol{\omega}_B \times ({}^A\boldsymbol{R}_B {}^B\boldsymbol{p}) \\ &= \boldsymbol{S}({}^A\boldsymbol{\omega}_B){}^A\boldsymbol{R}_B {}^B\boldsymbol{p}. \end{aligned} \tag{A.3} $$

$$ \frac{d}{dt}({}^A\boldsymbol{R}_B) = \boldsymbol{S}({}^A\boldsymbol{\omega}_B){}^A\boldsymbol{R}_B. \tag{A.4} $$

Similarly, we can obtain

$$ \frac{d}{dt}({}^B\boldsymbol{R}_A) = \boldsymbol{S}({}^B\boldsymbol{\omega}_A){}^B\boldsymbol{R}_A. \tag{A.5} $$

Input Eq.(A.4) and Eq.(A.5) to Eq.(A.2), we have

$$ \boldsymbol{S}({}^A\boldsymbol{\omega}_B) = -{}^A\boldsymbol{R}_B\boldsymbol{S}({}^B\boldsymbol{\omega}_A){}^B\boldsymbol{R}_A. \tag{A.6} $$