ON-LINE HEAD POSE ESTIMATION WITH BINOCULAR HAND-EYE ROBOT BASED ON EVOLUTIONARY MODEL-BASED MATCHING

Fujia Yu, Mamoru Minami, Wei Song, Jianing Zhu, Akira Yanou

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 1-Step genetic algorithm (1-Step GA) being improved to adapt for real-time recognition in dynamic images and a fitness evaluation based on a stereo model matching. Here, the head position and orientation are detected simultaneously using the brightness distribution and color information of the input images, evaluating the facial features, eyebrows and eyes. Moreover, to improve the dynamics of recognition, feedforward model-based matching is proposed for hand-eye visual servoing. The effectiveness of the method is shown by the experiments to compensate the motion of the hand-eye camera against the relative motion of the object in camera frame, having resulted in the robust recognition against the hand-eye motion.

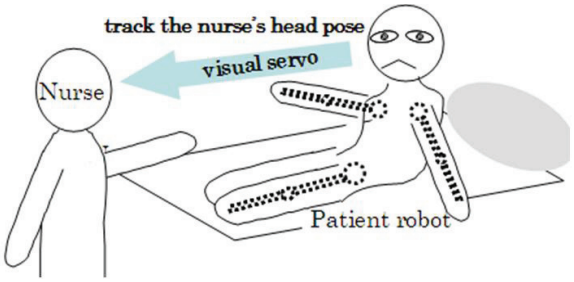
INTRODUCTION

This work is motivated by our desire to establish a visual system for a patient robot that is used to evaluate an ability of the medical treatments of nurse students, as shown in Fig.1. It is important for nurse to pay attention to the condition of the patient during, e.g. injection, to sense tiny sign of patient's state so as 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 servoing to 3D pose.

There is a variety of approaches for face representation of poses, and they can be classified into three general categories: feature-based, appearance-based, and model based. Feature-based approaches use local features like points, line segments, edges, or regions. The main idea of this method is to select a set of feature points, which are matched against the incoming video to update pose estimation. In a feature-based approach, estimation based on the relationship between human facial features [1], [2] relies heavily on the accuracy of the facial feature detection schemes. 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 [3], [4]. Appearance-based approaches attempt to capture and define the face as a whole. The image is compared with various templates to determine which one most closely matches the image, resulting in wasting time to recognize. This approach has received lot of attention recently. In one technique [5], templates representing facial feature are used in determining head position and orientation. The image is compared with various templates to determine which template most closely matches the image, resulting in wasting time to recognize. On the other hand, Model-based approaches is to use a model to search a target object in the image, and the model is composed based on how the target object can be seen in the input image. For recognizing a face and detecting its pose, 3-D models of face are mapped on input images to estimate the head pose [6], [7]. The recognition method developed in our paper is in this category. An

advantage of our method is that our 3D model is set up in a solid coordinate which enables it possess 6 degree of freedom (both the position and orientation). 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

Fig.1. Visual system for a patient robot.



corresponding points in Stereo-vision camera images is also complicated and time consuming [8].

Here, head is extracted using the model-based matching which is based on a given knowledge of the shape. An objective function is defined to evaluate the extent of how much the head model matches with the head being imaged, by changing the recognition problem into an optimization problem. Therefore, we employ a Genetic Algorithm (GA) because of its high performance of optimization.

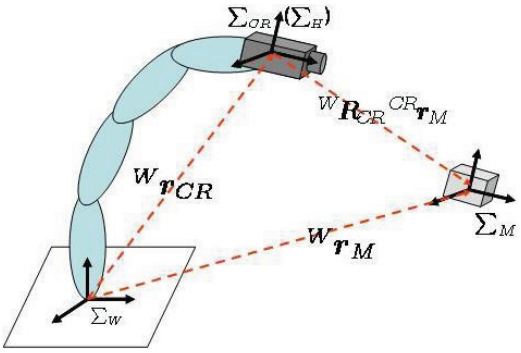
GA is well known as a method for solving parameter optimization problems [9]. Since contrary to the traditional search techniques, GA

can efficiently explore the search space, without need to scan the whole solution space, but only need to check the model-based fitness function—this function represents correlation distribution between target object's shape in a camera image and predefined model shape over a searching space of full pose, i.e., three positions and three orientations— in the selected areas designated by genes, that is possible solutions. Hence, GA will provide faster recognition performance to the vision system. Moreover, in order to realize the real-time nature and extract the target's pose from the consecutively input images, GA is used in such a way that every input image is evaluated at least only one time by model-based fitness function, and the convergence of GA is realized in the sequences of dynamic images, which is named "1-Step GA"[10]. Comparing with the other recognize approaches using Tayler expansion [11], [12], this method can find the target object through the whole search space, and recover the recognition error itself. As our previous research, we have confirmed that this method enabled a hand-eye robot system to catch a swimming fish by a net equipped at hand [13].

MOTION-FEEDFORWARD RECOGNITION

As we know, the merit of using GA to search is the efficiency for global search in the beginnings and the fast convergence when some individuals are close to the target object. Since here the individuals of

Fig.2. coordinate system



GA are set in the camera coordinate represented as Σ_{CR}, the GA search method is very effective when the camera is fixed in the workspace because the object is usually moving successively in real world. However, in practice, cameras are often moving together with the mobile robot. For example, in our work, two cameras are set as the eyes of the patient robot, which will be moved suddenly when the patient feels painful. And also, most visual servo systems use an

eye-in-hand configuration, having the camera mounted on the robot's end-effector. In such cases, the motion of the target in the camera coordinate will be effected by both the motion of the target in real world and the one of the camera. Take the eye-in-hand robot as an example, we will explain how to describe such a relationship between a target and a moving camera in a mathematical formulation.

Kinematics Of Hand-Eye Next we will establish relations among relative velocities of three moving 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 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:

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

Differentiating Eq.1 with respect to time

$$\begin{aligned} {}^{CR}\dot{\boldsymbol{r}}_{CR,M} = {}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\dot{\boldsymbol{r}}_M - {}^W\dot{\boldsymbol{r}}_{CR}) + [{}^{CR}\boldsymbol{\omega}_W\times] \\ ({}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\boldsymbol{r}_M - {}^W\boldsymbol{r}_{CR}(\boldsymbol{q}))) \end{aligned} \tag{2}$$

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:

$$ {}^{CR}\boldsymbol{\omega}_{CR,M} = {}^{CR}\boldsymbol{R}_W(\boldsymbol{q})({}^W\boldsymbol{\omega}_M - {}^W\boldsymbol{\omega}_{CR}) \tag{3}$$

Thus, Eq.2 gives the translation velocity of Σ_M with respect to Σ_{CR}, Eq.3 gives the orientation velocity. The camera velocity, which is considered as the end-effector velocity, can be expressed using the Jacobian matrix $\boldsymbol{J}(\boldsymbol{q}) = [\boldsymbol{J}_p^T(\boldsymbol{q}),\ \boldsymbol{J}_o^T(\boldsymbol{q})]^T$.

$$ {}^W\dot{\boldsymbol{r}}_{CR} = \boldsymbol{J}_p(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{4}$$

$$ {}^W\boldsymbol{\omega}_{CR} = \boldsymbol{J}_O(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{5}$$

$$\begin{aligned} [{}^{CR}\boldsymbol{\omega}_W\times] &= -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})[{}^W\boldsymbol{\omega}_{CR}\times]{}^W\boldsymbol{R}_{CR}(\boldsymbol{q})^1 \\ &= -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})[\boldsymbol{J}_O(\boldsymbol{q})\dot{\boldsymbol{q}}\times]{}^W\boldsymbol{R}_{CR}(\boldsymbol{q}) \end{aligned} \tag{6}$$

In this paper, the roll, pitch, yaw angles are used to describe the target orientation, as $\boldsymbol{\psi} = [\Phi, \theta, \psi]^T$. The relation between ${}^{CR}\dot{\boldsymbol{\psi}}_{CR,M}$ and ${}^{CR}\boldsymbol{\omega}_{CR,M}$ is given by the inverse of matrix $\boldsymbol{J}_C$:

$$ {}^{CR}\dot{\boldsymbol{\psi}}_{CR,M} = (\boldsymbol{J}_C^{-1}){}^{CR}\boldsymbol{\omega}_{CR,M} \tag{7}$$

Substituting Eq.4, 5, 6 to Eq.2, 7, the target velocity in Σ_{CR}, represented by ${}^{CR}\dot{\boldsymbol{\Psi}}_{CR,M}$, can be described by a mathematical formulation:

$$ {}^{CR}\dot{\boldsymbol{\Psi}}_{CR,M} = \begin{bmatrix} {}^{CR}\dot{\boldsymbol{r}}_{CR,M} \\ {}^{CR}\dot{\boldsymbol{\psi}}_{CR,M} \end{bmatrix} $$

$$ = \begin{bmatrix} -{}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{J}_p(\boldsymbol{q}) + {}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) \\ [({}^W\boldsymbol{R}_{CR}(\boldsymbol{q}){}^{CR}\boldsymbol{r}_{CR,M})\times]\boldsymbol{J}_O(\boldsymbol{q}) \\ -(\boldsymbol{J}_C^{-1}){}^{CR}\boldsymbol{R}_W(\boldsymbol{q})\boldsymbol{J}_O(\boldsymbol{q}) \end{bmatrix} \dot{\boldsymbol{q}} $$

$$ + \begin{bmatrix} {}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) & 0 \\ 0 & (\boldsymbol{J}_C^{-1}){}^{CR}\boldsymbol{R}_W(\boldsymbol{q}) \end{bmatrix} \begin{bmatrix} {}^W\dot{\boldsymbol{r}}_M \\ {}^W\boldsymbol{\omega}_M \end{bmatrix} $$

$$ = \boldsymbol{J}_m(\boldsymbol{q})\dot{\boldsymbol{q}} + \boldsymbol{J}_n(\boldsymbol{q}){}^W\dot{\boldsymbol{\Phi}}_M \tag{8}$$

The relationship $\boldsymbol{J}_n(\boldsymbol{q})$ in Eq.8 describes how target pose change in Σ_{CR} with respect to the pose changing of themselves in real word. The relationship $\boldsymbol{J}_m(\boldsymbol{q})$ in Eq.8 describes how target pose change in Σ_{CR} with respect to changing manipulator pose which causes the negative influence to recognition from the motion of the camera.

[1]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

$$ {}^A\boldsymbol{R}_B\,{}^B\boldsymbol{R}_A = \boldsymbol{I} \tag{9}$$

the time derivative of Eq.7 is given by

$$ \frac{d}{dt}({}^A\boldsymbol{R}_B){}^B\boldsymbol{R}_A + {}^A\boldsymbol{R}_B\frac{d}{dt}({}^B\boldsymbol{R}_A) = \boldsymbol{0} \tag{10}$$

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

$$\begin{aligned} \frac{d}{dt}({}^A\boldsymbol{R}_B){}^B\boldsymbol{p} &= {}^A\boldsymbol{\omega}_B \times ({}^A\boldsymbol{R}_B\,{}^B\boldsymbol{p}) \\ &= [{}^A\boldsymbol{\omega}_B\times]\,{}^A\boldsymbol{R}_B\,{}^B\boldsymbol{p} \end{aligned} \tag{11}$$

Hence

$$ \frac{d}{dt}({}^A\boldsymbol{R}_B) = [{}^A\boldsymbol{\omega}_B\times]\,{}^A\boldsymbol{R}_B \tag{12}$$

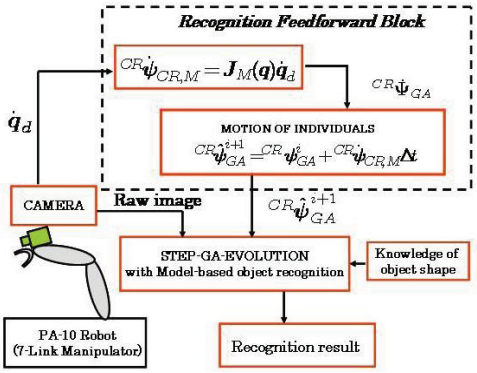
Similarly, we can obtain

$$ \frac{d}{dt}({}^B\boldsymbol{R}_A) = [{}^B\boldsymbol{\omega}_A\times]\,{}^B\boldsymbol{R}_A \tag{13}$$

Input Eq.12 and Eq.13 to Eq.10, we have

$$ [{}^A\boldsymbol{\omega}_B\times] = -{}^A\boldsymbol{R}_B[{}^B\boldsymbol{\omega}_A\times]\,{}^B\boldsymbol{R}_A \tag{14}$$

Fig.3. Feedforward recognition system



Motion-Feedforward Recognition Method In this paper, we do not discuss about $\boldsymbol{J}_n(\boldsymbol{q})$, so we rewritten Eq.8 without it.

$$ {}^{CR}\dot{\boldsymbol{\Psi}}_{CR,M} = \boldsymbol{J}_m(\boldsymbol{q})\dot{\boldsymbol{q}} \tag{15}$$

Suppose that the motion of the camera is known, like human knows his moving, and can also predictive the target pose changing along with his moving. Based on this human ability, a robust recognition has been proposed, which we named as a motion-feedforward recognition method, since it uses the motion information of the camera to move the whole individuals to compensate the influence from the camera itself.

Using Eq.15, the pose of the individuals in the next generation can be predicted from the current end-effector motion, presented by

$$ {}^{CR}\boldsymbol{\Psi}^{i+1}_{GA} = {}^{CR}\boldsymbol{\Psi}^{i}_{GA} + \Delta t\,{}^{CR}\dot{\boldsymbol{\Psi}}_{CR,M} \tag{16}$$

$\Delta t\ {}^{CR}\dot{\Psi}_{CR,M}$ is the moving distance from the current individuals to the next generation's. The recognition system of the proposed method is shown in Fig.3. We consider that the recognition ability will be improved by using Eq.16 to move all the individuals to compensate the influence of the motion of the camera. So the recognition will be very robust to the motion of robot itself.

EVOLUTIONARY RECOGNITION

This section will introduce the evolution recognition method in our experiments.

Model Setting Up

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.4 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 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,

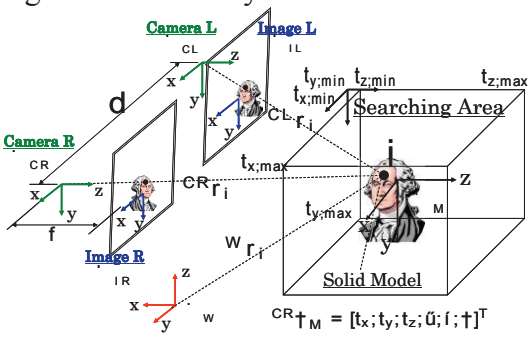
$$ {}^{CR}\boldsymbol{r}_i = {}^{CR}\boldsymbol{T}_M \, {}^{M}\boldsymbol{r}_i \tag{17}$$

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,

$$ {}^{IR}\boldsymbol{r}_i = \boldsymbol{P} \, {}^{CR}\boldsymbol{r}_i \tag{18}$$

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

Fig.4. Coordinate systems



$$ {}^{CL}\boldsymbol{r}_i = {}^{CL}\boldsymbol{T}_{CR} \, {}^{CR}\boldsymbol{r}_i \tag{19}$$

By the same way as we have obtained ${}^{IR}\boldsymbol{r}_i$, ${}^{IL}\boldsymbol{r}_i$ is described by the following Eq.20 through projection matrix $\boldsymbol{P}$.

$$ {}^{IL}\boldsymbol{r}_i = \boldsymbol{P} \, {}^{CL}\boldsymbol{r}_i \tag{20}$$

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 ${}^{IL}\boldsymbol{r}_i$. Here, position and orientation of the origin of Σ_M based on Σ_{CR} are represented as $\underline{\boldsymbol{\phi}} = [\underline{t_x}, \underline{t_y}, \underline{t_z}, \underline{\Phi}, \underline{\theta}, \underline{\psi}]^T$, in which $\underline{\Phi}$, $\underline{\theta}$, $\underline{\psi}$ are roll, pitch and yaw angles, $\underline{\Phi}$ represents roll, $\underline{\theta}$ represents pitch, $\underline{\psi}$ represents yaw, and then Eq.18, 20 are rewritten as,

$$ \begin{cases} {}^{IR}\boldsymbol{r}_i = f_R(\underline{\phi}, \, {}^{M}\boldsymbol{r}_i) \\ {}^{IL}\boldsymbol{r}_i = f_L(\underline{\phi}, \, {}^{M}\boldsymbol{r}_i). \end{cases} \tag{21}$$

This relation connects the arbitrary points on the object and projected points on the left and right images with the variables $\underline{\boldsymbol{\phi}}$, which is considered to be unknown in this paper. 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.

3D solid model. The 3D model for head located in Σ_M is composed of two circles, shown in Fig.5. One circle that passes through the plane XOY is named as C_v. C_v represents the outline of head. The set of coordinates inside of C_v is depicted as S_{in} and the outside space enveloping S_{in} is denoted as S_{out}, and the combination is named as S. The other circle that passes through the plane ZOX is named as C_h. C_h represents the line of demarcation between skin region and hair region. Here, we describe approximately that the head is a sphere, and the hair occupies hemisphere above C_h while the face is the other hemisphere below C_h.

The set of the points of solid searching model consisted of S_{in} and S_{out}, which are projected onto the 2D coordinates of right camera are expressed as,

$$ \begin{cases} S_{R,in}(\underline{\phi}) = \{ {}^{IR}\boldsymbol{r}_i \in \Re^2 \mid {}^{IR}\boldsymbol{r}_i = f_R(\underline{\phi}, {}^{M}\boldsymbol{r}_i), \\ \qquad\qquad {}^{M}\boldsymbol{r}_i \in S_{in} \in \Re^3 \} \\ S_{R,out}(\underline{\phi}) = \{ {}^{IR}\boldsymbol{r}_i \in \Re^2 \mid {}^{IR}\boldsymbol{r}_i = f_R(\underline{\phi}, {}^{M}\boldsymbol{r}_i), \\ \qquad\qquad {}^{M}\boldsymbol{r}_i \in S_{out} \in \Re^3 \} \end{cases} \tag{22}$$

where the right searching model projected to right

camera coordinates is shown in Fig.6(b). The area composed of $S_{R,in}$ and $S_{R,out}$ is named as S_R. The left one is defined in the same way and the projected searching model is shown in Fig.6(a) (the equations and explanations about the left one will be abridged in the following parts of this paper). Notice

Fig.5. 3D Solid Model

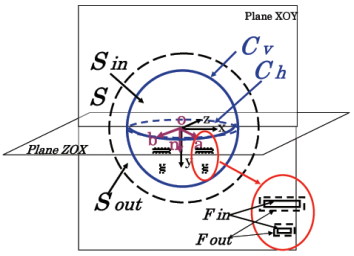
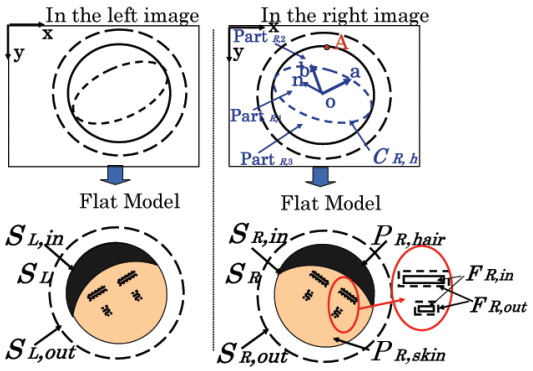


Fig.6. Searching model



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

that C_v is the outline of head, it only serves the recognition of head position.
In the same way, the set of the points of feature models consisted of F_{in} and F_{out}, which are projected to 2D are depicted as $F_{R,in}$ and $F_{R,out}$. Here, $\underline{\phi}$ possesses both the position and orientation information.
The points of C_h projected onto 2D coordinates of right camera are expressed as,

$$C_{R,h}(\underline{\phi}) = \{ {}^{IR}\boldsymbol{r}_i \in \Re^2 \mid {}^{IR}\boldsymbol{r}_i = f_R(\underline{\phi}, {}^{M}\boldsymbol{r}_i), {}^{M}\boldsymbol{r}_i \in C_h \in \Re^3 \} \quad (23)$$

$C_{R,h}$ will divide the inside area of $S_{R,in}$ into three parts. After deciding the color of all the three parts, we name the parts with yellow color as $P_{R,skin}$ and the parts with black color as $P_{R,hair}$.

Definition of evaluation function

Head Detection. The 2D raw images of a target human are shown in Fig.7(a) and (c), their corresponding 3D plot are shown in Fig.7(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 this research, the input images are directly matched by the projected moving models, S_L and S_R, which are located by only $\underline{\phi}$ as described in Eq.22 that includes the kinematical relations of the left and right camera coordinates.
Here, we define evaluation function to estimate how match the moving solid model S_{in} defined by $\underline{\phi}$ lies on the head being imaged on the left and right cameras. In order to search for the head in the gray scale image, the surface-strips model shown in Fig.6 and its position calculated by Eq.21 are used. The brightness distribution of input image lying on the area of searching model is expressed as $\boldsymbol{p}({}^{IR}\boldsymbol{r}_i)$, $\boldsymbol{r}_i \in S_R(\underline{\phi})$, then the evaluation function $F_{site}(\underline{\phi})$ of the moving surface-strips model is given as,

Fig.7. Input image and brightness distribution

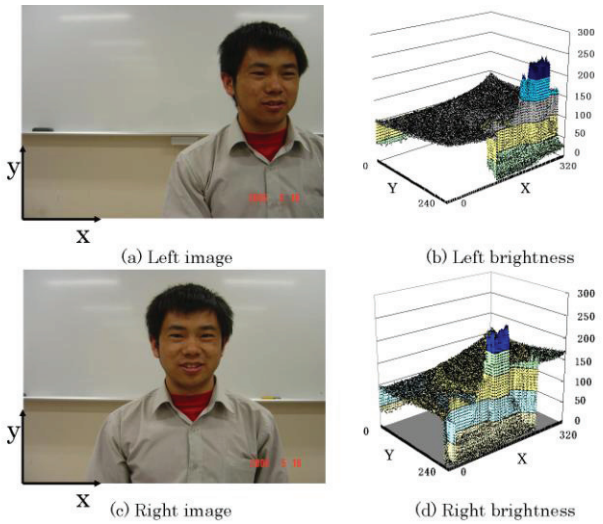
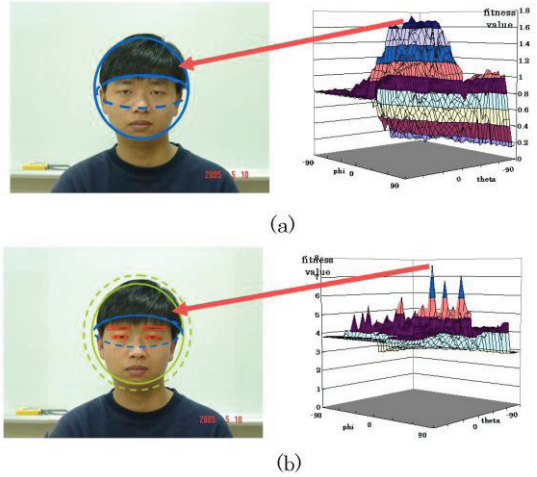


Fig.8. (a) Correlation graph by using the model without eyebrows and eyes. (b) Correlation graph by using the model with eyebrows and eyes.



$$F_{R,site}(\underline{\phi}) = \sum_{{}^{IR}\boldsymbol{r}_i \in S_{R,in}(\underline{\phi})} p({}^{IR}\boldsymbol{r}_i) - \sum_{{}^{IR}\boldsymbol{r}_i \in S_{R,out}(\underline{\phi})} p({}^{IR}\boldsymbol{r}_i) \quad (24)$$

$$F_{site}(\underline{\phi}) = (F_{R,site}(\underline{\phi}) + F_{L,site}(\underline{\phi}))/2 \quad (25)$$

where, in case of $F_{site}(\underline{\boldsymbol{\phi}}) \le 0$, $F_{site}(\underline{\boldsymbol{\phi}})$ is given to zero. This function expresses the brightness difference between the one of the internal surface and the one of the contour-strips of the surface-strips model, and used as a fitness function in GA process.
When the moving searching model fits to the target object being imaged in the right and left images, then the fitness function $F_{site}(\underline{\boldsymbol{\phi}})$ will give maximum value, then the recognition problem is converted into the optimization one.

Pose Estimation By Color. Here the hair color is defined as black so it can be represented by RGB color system. However, for the color of skin, there is great difference among the RGB value at the various situations even the same human, so the HSV color system is considered. And it is easy to understand that the color of skin can be limited only by hue value in HSV parameters [14]. Take the right image as an example, for skin region $P_{R,skin}$, we define

$$h({}^{IR}\boldsymbol{r}_i) = \begin{cases} 1 & 0 < H < 30 \\ 0 & otherwise \end{cases} \quad (26)$$

For hair region of 2D images $P_{R,hair}$, we define

$$b({}^{IR}\boldsymbol{r}_i) = \begin{cases} 1 & p({}^{IR}\boldsymbol{r}_i) > 220 \ and \ h({}^{IR}\boldsymbol{r}_i) = 0 \\ 0 & otherwise \end{cases} \quad (27)$$

Then the evaluation function $F_{color}(\underline{\boldsymbol{\phi}})$ of the searching models is given as

$$F_{R,color}(\underline{\phi}) = \sum_{{}^{IR}\boldsymbol{r}_i \in P_{R,skin}(\underline{\phi})} h({}^{IR}\boldsymbol{r}_i) + \sum_{{}^{IR}\boldsymbol{r}_i \in P_{R,hair}(\underline{\phi})} b({}^{IR}\boldsymbol{r}_i) \quad (28)$$

$$F_{color}(\underline{\phi}) = (F_{R,color}(\underline{\phi}) + F_{L,color}(\underline{\phi}))/2, \quad (29)$$

This function expresses the extent of recognition accuracy. When both the hair region and the skin region of the searching models fit to the head being imaged in the right and left images, then the fitness function $F_{color}(\underline{\boldsymbol{\phi}})$ gives maximum value.

Pose Estimation by Facial Features: Since model of the facial features, eyebrows and eyes, have the same surface-strip structure as the model of the head outline, shown in Fig.6, the recognition method for them is the same as head detection. So we do not give detailed explanation here. The evaluation is shown as follows:

$$F_{R,feat}(\underline{\phi}) = \sum_{{}^{IR}\boldsymbol{r}_i \in F_{R,in}(\underline{\phi})} p({}^{IR}\boldsymbol{r}_i) - \sum_{{}^{IR}\boldsymbol{r}_i \in F_{R,out}(\underline{\phi})} p({}^{IR}\boldsymbol{r}_i) \quad (30)$$

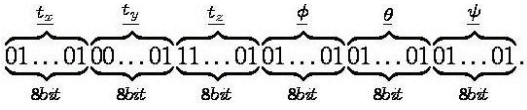
$$F_{feat}(\underline{\phi}) = (F_{R,feat}(\underline{\phi}) + F_{L,feat}(\underline{\phi}))/2, \quad (31)$$

The effectiveness of pose recognition can be improved while adding the evaluation of eyebrows and eyes. We confirmed this by comparing the F_{feat} distribution with F_{color} distribution under the condition ($\underline{t_x}$, $\underline{t_y}$, $\underline{t_z}$, $\underline{\psi}$) = (0, 0, 1000,-90) using the input images. The image from the right camera is shown in Fig.8(a). Fig.8(b) shows F_{color} distribution. We can find that a number of combinations of ($\underline{\Phi}$, $\underline{\theta}$) will give fitness values which are close to the maximum value. In Fig.8(c), the mountain of fitness function values is sharpened by using F_{feat} that will improve the speed and accuracy of GA recognition.

Our approach is to evaluate the head position and orientation simultaneously using the whole fitness function defined as

$$F(\underline{\phi}) = F_{site}(\underline{\phi}) + F_{color}(\underline{\phi}) + F_{feat}(\underline{\phi}) \quad (32)$$

Therefore the problem of head pose recognition can be converted to searching problem of $\underline{\phi}$ such that maximizes $F(\underline{\phi})$. To recognize the target object in a short time, we solve this optimization problem to search for $\underline{\phi}$ to maximize $F(\underline{\phi})$ by GA whose gene representing $\underline{\phi}$ is defined as,

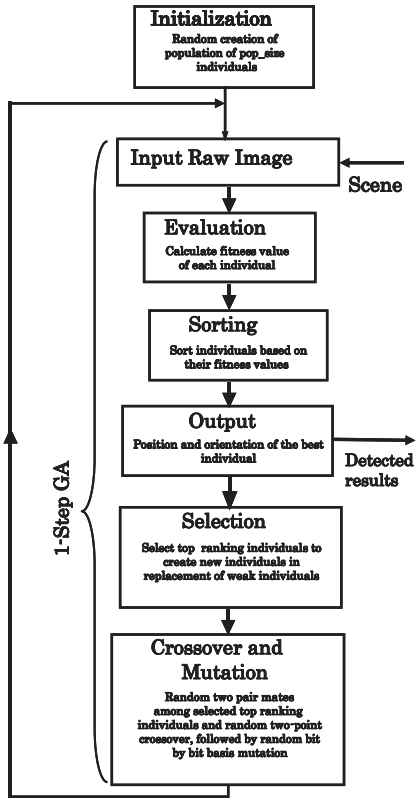


The 48 bits of gene refers to the range of the searching area: $-100 \le t_x, t_y \le 100$, $900 \le t_z \le 1200[mm]$, $-20 \le \Phi, \theta, \psi \le 20[deg]$. Any target object in this searching area can be recognized by our method, whatever the original position and orientation is. Furthermore, because new genes are created in every generation, the system can recover the recognition error itself.

On-Line Evolutionary Recognition. Although GA have been applied to a number of robot control systems [15], 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 cost much time until its convergence.

So here, for real-time visual control purposes, we

Fig.9. Flow chart of 1-Step GA recognition



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. In addition, this feature happens to be favorable for real-time visual recognition. We exploit this fact to output the current results of the GA in every newly input image, to be used as command value to the manipulator's controller. Thereby real-time visual servoing can be performed. The flow chart of the 1-step GA process is shown in Fig.9. The effectiveness of the proposed 2D recognition method is confirmed by the experiment of catching a fish by visual servoing of a robot equipped with a camera and a net at the hand [10], as shown in Fig.10. Fig.9 shows that the image inputting process is included in the GA iteration process seeking for the potential solution, i.e., toward the target. That is, the evolving speed to the solution in the image should be faster than the swimming speed of the fish in the successively input images, for the success of real-time recognition by "1-Step GA." However, as the searching space extending to 3D, the time of each GA process will become longer since the parameters is 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 seeing from the cameras by using the motion of the robot which is considered as known.

Definition Of Recognition Error. Recognition error is used to evaluate the accuracy extent of our recognition method. It defined as the difference between the real target's position/orientation and the recognition result given by the GA genes which shows the maximum fitness function value.

We define the error of position as $\Delta^{CR}\boldsymbol{r} = ({}^{CR}\boldsymbol{r})_d - {}^{CR}\boldsymbol{r}^i$, where $\boldsymbol{r} = [x, y, z]^T$, and $({}^{CR}\boldsymbol{r})_d$ is the true position and ${}^{CR}\boldsymbol{r}_i$ is the best gene of i-th generation.

Since addition can not be used to calculate the error of orientation between $({}^{CR}\boldsymbol{\psi})_d$ and ${}^{CR}\boldsymbol{\psi}^i$, which is defined as $\Delta^{CR}\boldsymbol{\psi}$, where $\boldsymbol{\psi} = [\Phi, \theta, \psi]^T$ and $({}^{CR}\boldsymbol{\psi})_i$ is the true orientation, ${}^{CR}\boldsymbol{\psi}^i$ is the best gene of i-th generation. The following equations show how to deduce $\Delta^{CR}\boldsymbol{\psi}$ using the rotation matrix of two poses.

Fig.10. Fish catching by hand-eye visual servo

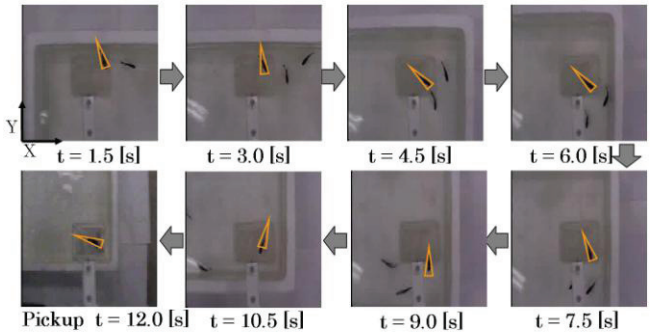
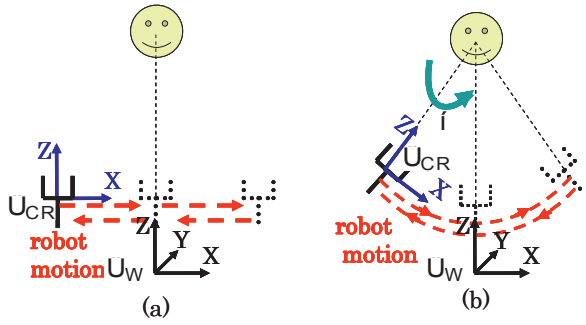


Fig.11. (a) shuttle motion of end-effector in x axis of Σ_W (from x = 150 to -150[mm]). (b) shuttle motion of end-effector to see the face from $\theta = -10[deg]$ to $\theta = 10[deg]$.



Here, the matrix defined as $\boldsymbol{R}_d$ expresses the true orientation $({}^{CR}\boldsymbol{\psi})_d$, and the matrix defined as $\boldsymbol{R}^i$ expresses the orientation ${}^{CR}\boldsymbol{\psi}^i$ which possess the best genes, and $\Delta\boldsymbol{R}_{(d,i)}$ is defined as the changing rotation matrix between the two poses. Then we obtain the relation

$$\boldsymbol{R}^i = \boldsymbol{R}_d \Delta\boldsymbol{R}_{(d,i)} \tag{33}$$

From Eq.15 we have

$$\Delta\boldsymbol{R}_{(d,i)} = (\boldsymbol{R}_d)^T \boldsymbol{R}^i = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \tag{34}$$

We define $\Delta^{CR}\boldsymbol{\psi} = [\Delta\Phi, \Delta\theta, \Delta\psi]^T$, using Eq.34 $\Delta\Phi, \Delta\theta, \Delta\psi$ is given by

$$\begin{aligned} \Delta\phi &= atan2(\mp R_{21}, \mp R_{11}), \\ \Delta\theta &= atan2(-R_{31}, \pm\sqrt{R_{32}^2 + R_{33}^2}), \\ \Delta\psi &= atan2(\mp R_{32}, \mp R_{33}). \end{aligned} \tag{35}$$

Here, the previous calculation deduction is shortly written as $\Delta^{CR}\boldsymbol{\psi} = \boldsymbol{D}(\boldsymbol{R}_d, \boldsymbol{R}^i)$.

To sum up, the recognition error is defined as

Fig.12. Recognition under position shuttle motion of end-effector 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}.

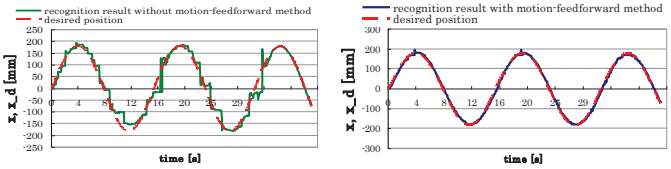
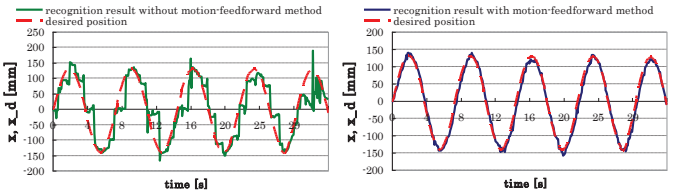


Fig.13. Recognition under position shuttle motion of end-effector with period T = 7[s]. (a) recognition result of position x without using motion-feedforward method compared with the desired position in Σ_{CR}. (b) recognition result with motion-feedforward method compared with the desired position in Σ_{CR}.



$$\Delta^{CR}\Psi = \begin{bmatrix} \Delta^{CR}\boldsymbol{r} \\ \Delta^{CR}\boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} ({}^{CR}\boldsymbol{r})_d - {}^{CR}\boldsymbol{r}^i \\ \boldsymbol{D}(\boldsymbol{R}_d, \boldsymbol{R}^i) \end{bmatrix}. \tag{36}$$

EXPERIMENT OF RECOGNITION FOR MOBILE ROBOT

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 use 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 separately as follows.

Recognition Under Given Position Changing Of End-effector. In the case of shuttle motion in x axis of Σ_W (from x = 150 to -150[mm]) is given to the robot end-effector, the static target in Σ_W will seem moving inversely in Σ_{CR}, as shown in Fig.11 (a). Fig.12 shows the recognition under position shuttle motion of end-effector with period T = 15[s]. Fig.12 (a) is the recognition result of position x without using motion-feedforward method compared with the desired position in Σ_{CR}. Fig.12 (b) is the recognition result

with motion-feedforward method compared with the desired position in Σ_{CR}. The recognition error without using the motion-feedforward method get larger, however, the recognition error of the motion-feedforward method is reduced to the extent of 15[*mm*]. 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*], for our hand-eye robot such high speed can lead to a little shaking which can influence the recognition, so the recognition error get much bigger without using the motion-feedforward method as shown in Fig.13 (a). In comparing, the recognition error can still keep in the extent of 15[*mm*] using motion-feedforward recognition method, shown in Fig.13 (b). From this

Fig.14. Recognition under orientation shuttle motion of end-effector
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}.

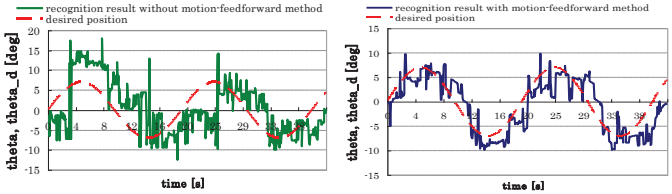
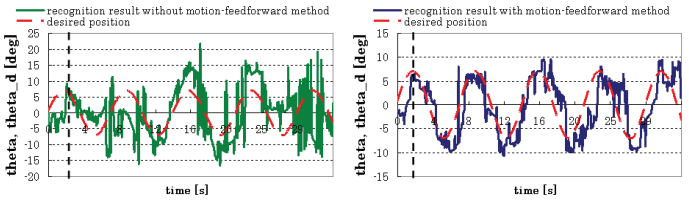


Fig.15. Recognition under orientation shuttle motion of end-effector
with period T = 10[*s*]. (a) recognition result of orientation θ without
using motion-feedforward method compared with the desired position
in Σ_{CR}. (b) recognition result of orientation θ with motion-feedforward
method compared with the desired position in Σ_{CR}.



experiment, we can say that our proposed motion-feedforward recognition method is very robust to the positon motion of robot.

Recognition Under Given Orientation Changing Of End-Effector. Here, the orientation changing of end-effector is defined as the motion in a circle with a fixed distance to the target and keeping the eye-line (z axis of Σ_{CR}) passes the center of the target. The shuttle motion that looking the target from the left side to the right side from -10[*deg*] to 10[*deg*] is given to the robot end-effector, as shown in Fig.11 (b). In the same way as the previous experiment, we make the shuttle motion within different convergent time 20[*s*] and 10[*s*], the corresponding velocity of the end-effector become larger and larger. Fig.14 (a) and Fig.15 (a) is the recognition result of orientation θ without using motion-feedforward method compared with the desired position in Σ_{CR}. Fig.14 (b) and Fig.15 (b) is the recognition result with motion-feedforward method compared with the desired position in Σ_{CR}. 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 very robust to the motion of robot itself because it compensates the influence of the motion of the camera. Also in both experiments with and without motion-feedforward method, we can see that as an example.

In the first 2 seconds the system cannot recognize the target object, so the error became bigger and bigger till "1-step GA" found the best gene. But when the system found the best gene, which can express the target object, the recognition result changed to the correct one immediately after the time which is marked by dot line in Figure 15.

CONCLUSION

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 includes head detection, pose estimation from color and facial feature (eyes) detection. We have proposed an motion-feedforward recognition method, which is confirmed very robust since it can move the whole individuals to compensate the influence from the camera itself. 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.

REFERENCES

Roberto Brunelli, (1994) "Estimation of pose and illuminant direction for face processing," Technical Report AIM-1499.

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

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

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

Niyogi.S and Freeman.W.T, (1996.) "Example-based head tracking," Technical Report TR96-34, MERL Cambridge Research.

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

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

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

Goldberg.D. E. (1989) "Genetic algorithm in Search, Optimization
and Machine Learning," Addison-Wealey.

Minami. M, Agbanhan. J and Asakura. T, (1999) "Manipulator Visual Servoing and Tracking of Fish using Genetic Algorithm,"Int. J. of Industrial Robot, 29-4, pp.278-289,.

Drummond. T, Cipolla. R, (1999)"Visual tracking and control using Lie algebras," Proceedings of IEEE International Conference
on Computer Vision and Pattern Recognition, pp. 652 - 657 Vol. 2.

Keller. Y, Averbuch. A, (2004) "Fast motion estimation using bidirectional gradient methods," IEEE, Transactions on Image Processing, pp.1042 - 1054 Vol.13, No.8.

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

Suzuki. H, Minami .M ,(2004) "Real-time face detection using hybrid GA based on selective attention," IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), pp.1329 - 1334.

Nagata. T, Konishi. K and Zha. H, (1995) "Cooperative manipulations based on Genetic Algorithms using contact information,"
Proceedings of the International Conference on Intelligent Robots and Systems, pp.400-5.