Face-to-face interactive humanoid robot

Takayuki Kanda[1], Nicolas Miralles[1], Masayuki Shiomi[1,2], Takahiro Miyashita[1],
Ian Fasel[1,3], Javier Movellan[1,3], and Hiroshi Ishiguro[1,2]

[1]ATR Intelligent Robotics and Communication Laboratories, [2]Osaka University, and [3]University of California, San Diego
E-mail: kanda@atr.co.jp

Abstract—**This paper reports progress on the development of a humanoid robot designed for realtime face-to-face interaction with humans. An essential component for face-to-face interaction with humans is being able to find faces, tracking them and smoothly move back and forth between gazing to a face and other objects of interest. In this paper we propose a system that integrates peripheral vision and foveal vision in a principled manner using particle filters. The developed system generates hypotheses about face position by using peripheral vision and verifies them by integrating peripheral vision and foveal vision. Even though a face may not be present in foveal vision, while the robot is gazing at another object, it keeps plausible hypotheses about the location of the human face with peripheral vision, and restarts the face-following by verifying the hypothesis later.**

Keywords-component; human-robot interaction, face-tracking, condensation

I. INTRODUCTION

Over the past several years, many humanoid robots have been developed. We believe that in the near future humanoid robots will interact with humans in our daily life. Their human-like bodies enable humans to intuitively understand their gestures and cause people to unconsciously behave as if they were communicating with humans [1]. That is, if a humanoid robot effectively uses its body, people will naturally communicate with it. This could allow robots to perform communicative tasks in human society such as route guides.

Previous research works proposed various kinds of communicative behaviors made possible by humanoid robots. In particular, the eye (head orientation) is a very important body part. Humans utilize their eye-gaze to convey their attention about certain objects to other humans while maintaining eye-contact, which is widely known as the joint-attention mechanism [2]. Scassellati developed a robot with a joint-attention mechanism that recognizes another's gaze in order to share attention [3]. Kozima and his colleagues also developed a robot with a joint-attention mechanism [4]. As Imai and his colleagues showed, it is important to convey a robot's intention to humans through eye-contact [5]. Otherwise, humans would not understand the robot's utterance. In addition, humans can successfully determine a robot's intention from its head orientation [6]. We believe that these gestures produced by changes in head orientation are essential functions for humans and robots to engage in natural social interaction.

Since tracking of a human face is an essential function for interactive humanoid robots, several researchers have developed face-tracking mechanisms for robots. Nakadai and his colleagues developed a robot that can track a speaking person by integrating visual and auditory information [7, 8]. Matsusaka and his colleagues also developed such a robot that uses eye contact [9]. Doi and his colleagues integrated tracking of face and body with peripheral vision [10]. However, little previous research utilized eye and head orientation in making gesture with tracking the human face for eye-contact. Shibata and his colleagues developed a stabilization function of tracking eyes based on biological knowledge [11].

For human-robot communication, it is difficult to utilize only foveal vision (narrow spatial range with high-resolution vision) for eye-contact because a robot sometimes needs to look away to gaze at other objects to show its attention. Moreover, it is difficult to just use peripheral vision (wide spatial range with low-resolution vision) for eye-contact because the accuracy of the eye-contact will be limited with this approach. In addition, when a robot makes a gesture, humans remain silent, so it is difficult to rely on only auditory-based tracking. Several research works have employed an approach of integrating peripheral and foveal vision [3, 11, 12]. We also integrate peripheral vision and foveal vision to keep face-tracking even when a robot is gazing at other objects.

On the other hand, it is also important to reduce the calculation costs of the sensor processing system for autonomous mobile robots. For humanoid robots in particular, very limited space and energy are available for computer resources. A previous work on computer vision has proposed a probabilistic method that has advantages in both tracking performance (by generating multiple hypotheses) and calculation cost, known as *particle filter* [13].

In this paper, we propose a robust face-following mechanism based on the integration of peripheral vision and foveal vision with a *particle filter*-based method. Our approach consists of the following three steps: generating hypotheses about face position (finding potential human faces) with peripheral vision, verifying the hypotheses with integrating peripheral and foveal vision, and providing robust and accurate face-following with foveal vision. Even though foveal vision is unavailable in cases such as when it is gazing at other objects, it continues generating the hypotheses about face position with peripheral vision and restarts the face-following by verifying the hypotheses later. We report experimental results showing that the system works well indoors environments with difficult illumination and background conditions.

This research was supported by the Telecommunications Advancement Organization of Japan.

II. DEVELOPMENT OF FACE-FOLLOWING FUNCTION

A. Interactive humanoid robot Robovie

Figure 1 displays the interactive humanoid robot "Robovie," which is characterized by its human-like body expression and various sensors. The human-like body consists of eyes, a head and arms, which components generate the complex body movements required for communication. In order not to alarm humans, we decided on a size of 120 cm. The diameter is 40 cm. The robot has two arms (4*2 DOFs), a head (3 DOFs), and a mobile platform. It has a speaker for talk in a synthesized voice.

The robot has two types of vision sensors: an omnidirectional vision sensor (considered as peripheral vision) and a stereo vision sensor (considered as foveal vision). It also has various sensors: 16 skin sensors covering the major parts of the robot, 10 tactile sensors around the mobile platform, a microphone for listening to humans, and 24 ultrasonic sensors. Further, the robot satisfies the mechanical requirements of autonomy. It includes all computational resources needed for processing the sensory data and for generating behaviors. It has two PCs on board: a Pentium4 PC (2.4 GHz) for image processing and speech recognition and a Pentium III PC (933 MHz) for processing other sensory data and controlling the motors that generate interactive behaviors.

B. Integration of foveal vision and peripheral vision

For the interactive humanoid robot, it is important to perform eye-contact (looking at human face) and to show its attention (look at other objects). Thus, the robot should be able to

- accurately look at a human face so that the human believes he/she has established eye-contact with the robot, and
- robustly re-look at the face after looking at something else (even if the human moves or does not react).

To achieve these capabilities, we integrated peripheral vision and foveal vision with the particle filter method. It consists of the following three steps. First, it uses peripheral vision to generate hypotheses about the position of face-like objects. Similar to humans, peripheral vision is useful for finding a moving object. Second, it uses foveal vision to verify the hypotheses generated with peripheral vision so that it can accurately look at a human face. Finally, it integrates these two kinds of sensor information by updating the probability distribution corresponding to the peripheral vision.

1) Hypotheses generation with peripheral vision

A human's peripheral vision is usually utilized to obtain motion information as well as color information. However, spatial resolution is limited, so it is difficult to recognize detailed shape with peripheral vision. By imitating this human mechanism, we retrieve color and motion information from peripheral vision to generate hypotheses on the position of face-like object(s). In addition, a *particle filter* mechanism controls attention to reduce the amount of calculation.

Figure 2 represents the relationships among variables in the system as a graphical model. In the figure, single circles represent the variables related to the sensor data, and double circles represent the variables of generated hypotheses. The solid lines represent the internal dependencies among the variables, which

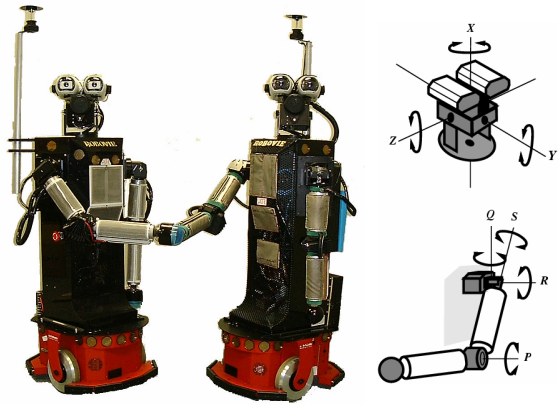

Fig. 1: Interactive humanoid robot "Robovie"

are used for estimating the face position. The broken lines represent the external dependencies. For example, with the variable "human-like moving object," the robot controls its gazing direction, which affects the variables "color" and "shape" in foveal vision.

The left side of the Figure 2 represents the process related to the peripheral vision. The system retrieves color information (second upper picture in left column) and temporal difference information (third upper picture) from the peripheral vision and integrates them to generate hypotheses on "human-like moving regions (bottom picture)". These hypotheses are preserved in each pixel as a probability $p(x_t)$ (equation 1), which is also utilized for attention control. As *particle filter* mechanism, it only checks the neighbors of current hypotheses $q(x_t)$ to reduce calculation (equation 2).

$$p(x_t) = \begin{cases} \pi(x_t) & \dots \ if\ q(x_t) > q_{th} \\ 0 & \dots \ otherwise \end{cases} \quad (1)$$

$$q(x_t) = \max_{\forall y} \{ p(x_{t-1} - y) + \lambda(y) \} \quad (2)$$

$$\pi(x_t) = diff(x_t) \bullet \alpha_d + color(x_t) \bullet \alpha_c + face(x_t) \bullet \alpha_f \quad (3)$$

where x_t represents the pixel in foveal image $\mathbf{x}_t$ at time t, q_{th} is the threshold for attention control, y is the two-dimensional vector that satisfies $(x_{t-1} - y) \in \mathbf{x}_{t-1}$, and $\lambda(y)$ is a function that decreases along with the increase of $|y|$. For instance, the function $\lambda(y) = w^y / 2^{|y|}$ satisfies this requirement, where w^y is a standard normal random variate. In equation 3, $diff(x_t)$ and $color(x_t)$ represent the likelihood of a human face at each pixel x_t, and these are calculated based on the temporal difference and color features, respectively. We will describe $face(x_t)$ later, which is calculated based on frontal face detection in foveal vision. The coefficients α_d, α_c, and α_f stand for internal parameters for retrieving information from measured features. We do not incorporate expectation of human movements into $\lambda(y)$, since it is difficult to estimate human movement in communication (for example, humans do not constantly move with certain velocity or acceleration when they communicate with others). Instead, the particle filter method provides the random sampling using MCMC, which estimates such random movements as humans make.

2) Frontal-face detection in foveal vision

Foveal vision is used to verify the hypotheses generated with peripheral vision. The right half of Figure 2 represents the

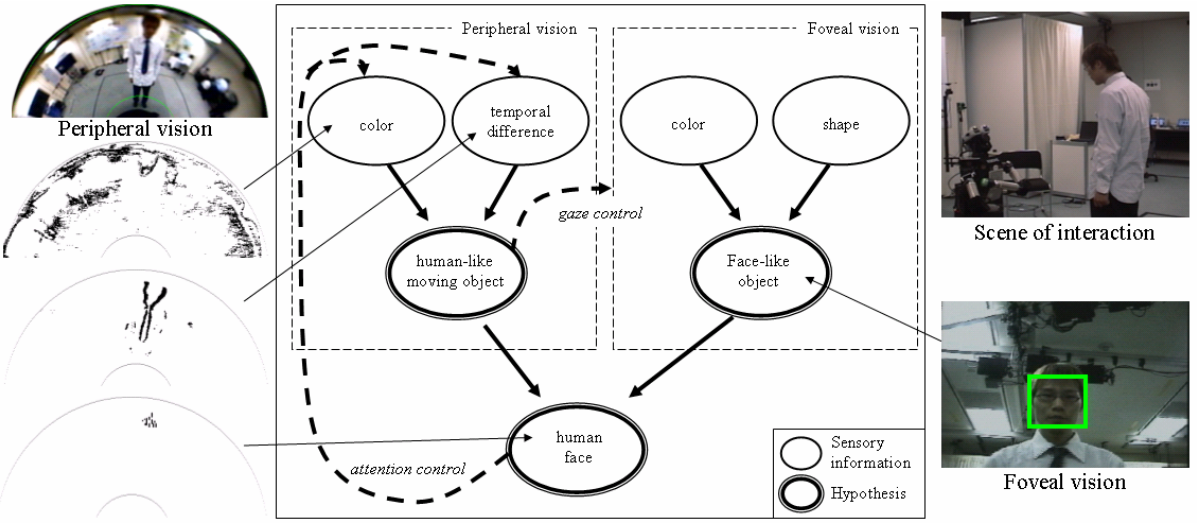

Figure 2: Relationships among internal probabilistic variables (internal and external dependencies)

process related to foveal vision. The system controls head orientation to look at the face-like region obtained from peripheral vision (upper picture in right column of Figure 2) and then tries to find "shape" information (human frontal face) with foveal vision. Frontal face detection was based on the MPISearch system available at http://kolmogorov.sourceforge.net [14,15]. This calculation is slightly expensive (currently it runs at about 3.5 fps with this mechanism), while the found face is also tracked with "color" information in realtime. By integrating the "shape" and "color" information, the hypotheses "face-like object" is generated in the foveal vision (bottom picture in right column).

3) Integration of periheral and foveal vision

First, the hypotheses on the frontal face in foveal vision ("face-like object") are compared with the hypotheses on the face-like region in peripheral vision ("face-like moving object"). If these hypotheses match, the system assumes that the hypotheses are verified. The verification is formalized as:

$$verify(\mathbf{x}_t') = \underset{\forall x_t : \mu(x_t) \in \mathbf{x}_t'}{average} \; (p(x_t)) \tag{4}$$

where $\mathbf{x}_t'$ includes every pixel in a detected region as a frontal face in the foveal vision, and $\mu(x)$ is a coordination conversion function from the pixel position in peripheral vision x to the pixel position in foveal vision. If the verification is successful (that is, $verify(\mathbf{x}_t') > v_{th}$), it updates the probability distribution in peripheral vision. We calculate $face(x_t)$ as follows, where v_{th} is a certain threshold for the verification.

$$face(x_t) = \begin{cases} 1 \ldots if\ \mu(x_t) \in \mathbf{x}_t'\ and\ verify(\mathbf{x}_t') > v_{th} \\ -1 \ldots if\ \mu(x_t) \in \mathbf{x}_t'\ and\ verify(\mathbf{x}_t') \le v_{th} \\ 0 \ldots otherwise \end{cases} \tag{5}$$

4) Look away mode

There are three modes in the face-following system. "Finding" mode is usually chosen when the system does not have a sufficient hypothesis on face-like regions. "Following" mode is usually chosen for following a human face with this mechanism. "Look away" mode is only chosen by the upper layer module, which allows the robot to intentionally look away, such as for gazing at an object during joint-attention. After looking away, the robot looks back toward the face-like region to find the human face, which enables the robot to reestablish eye-contact after showing its attention by gazing at the other object. This mode transition is accomplished simply by changing the parameters in equation (4). **Table 1** indicates the parameters (α_d, α_c, and α_f) for these three modes.

	α_d	α_c	α_f
Finding	0.75	0.25	0
Following	0.43	0.14	0.43
Look away	0.4	0.6	0

Table 1: System parameters for three tracking modes

III. EVALUATION

We performed experiments to verify the performance of the face-following system. This section reports the results of the experiments on face-following performance. (Videos showing the scene and internal status can be seen at http://www.irc.atr.co.jp/~kanda/ft/)

A. Performance of face-following

First, we measured the face-following performance for one human who is walking around the robot and interacting with the robot. **Figure 3** shows the results of the experiment. In the figure, "tracking status" (uppermost graph in the left half ("face-following")) shows whether the robot found the frontal face (denoted as "foveal") and whether it found the human in peripheral vision (denoted as "peripheral"), which correspond to the generated hypotheses "human-like moving object" and "frontal face" in Figure 2. "Face orientation" (center graph in left half) stands for the orientation of the human face from the robot (0 degree indicates that the human face is toward the robot head, and 90 degrees indicates that the human face is at a right angle to the robot head. In the "accuracy" graph (bottom graph in left half), "human" represents the direction of human face position from the robot's head, and "robot" stands for the orientation of the robot's head (0 degree represents the front direction of the robot). These movements of the human and the robot were measured by a motion capturing system. "Internal status" (right half of the figure) shows the internal status corresponding to scenes 1 to 5, which is denoted in the bottom of

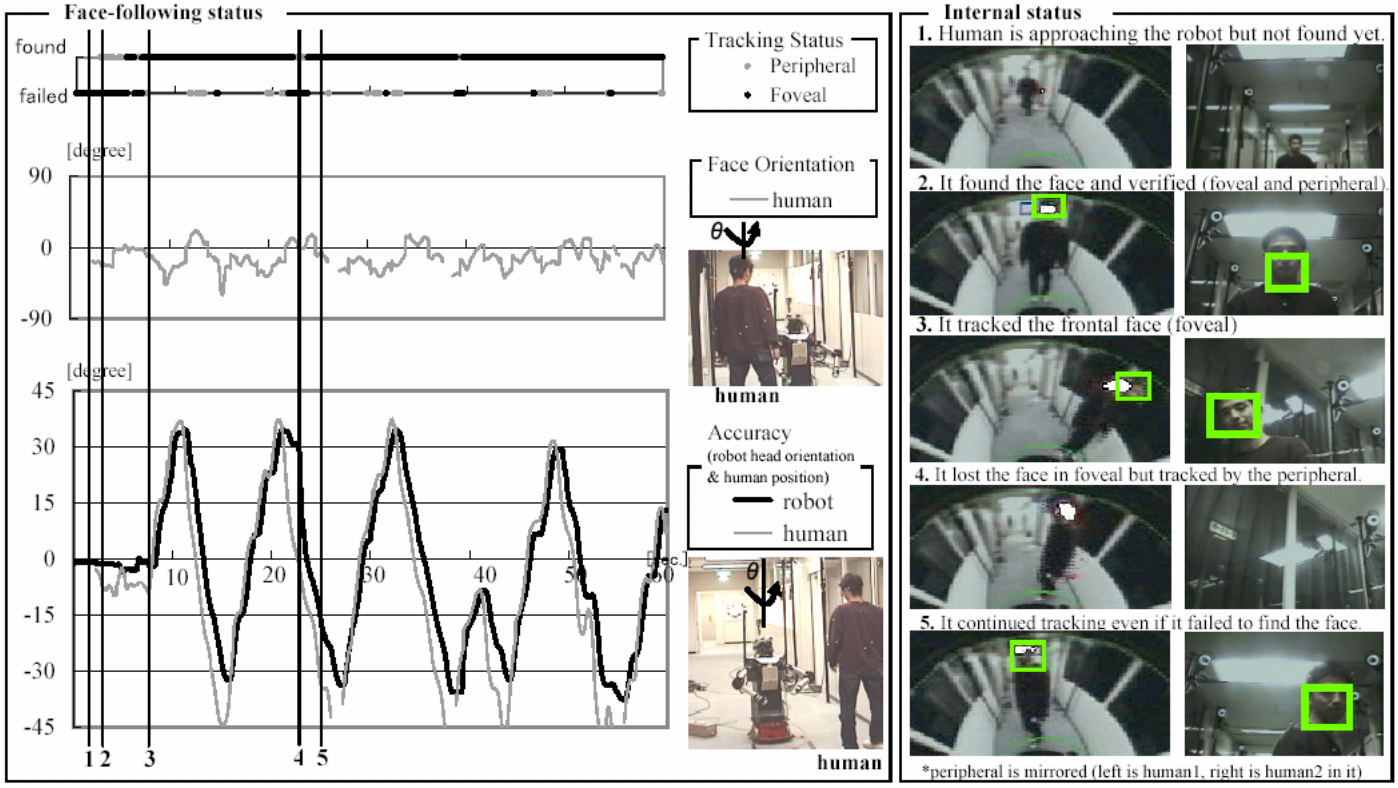

Figure 3: Experimental results for finding and tracking a human face

"Internal status"t shows the internal status corresponding to the scenes 1 to 5, which are denoted at the bottom of the left half of the figure. For the internal status, each picture on the left denotes the probability $p(x_t)$ *as white colored regions on the image of peripheral vision, and each picture on the right denotes the detected human face as a square on the image of foveal vision.*

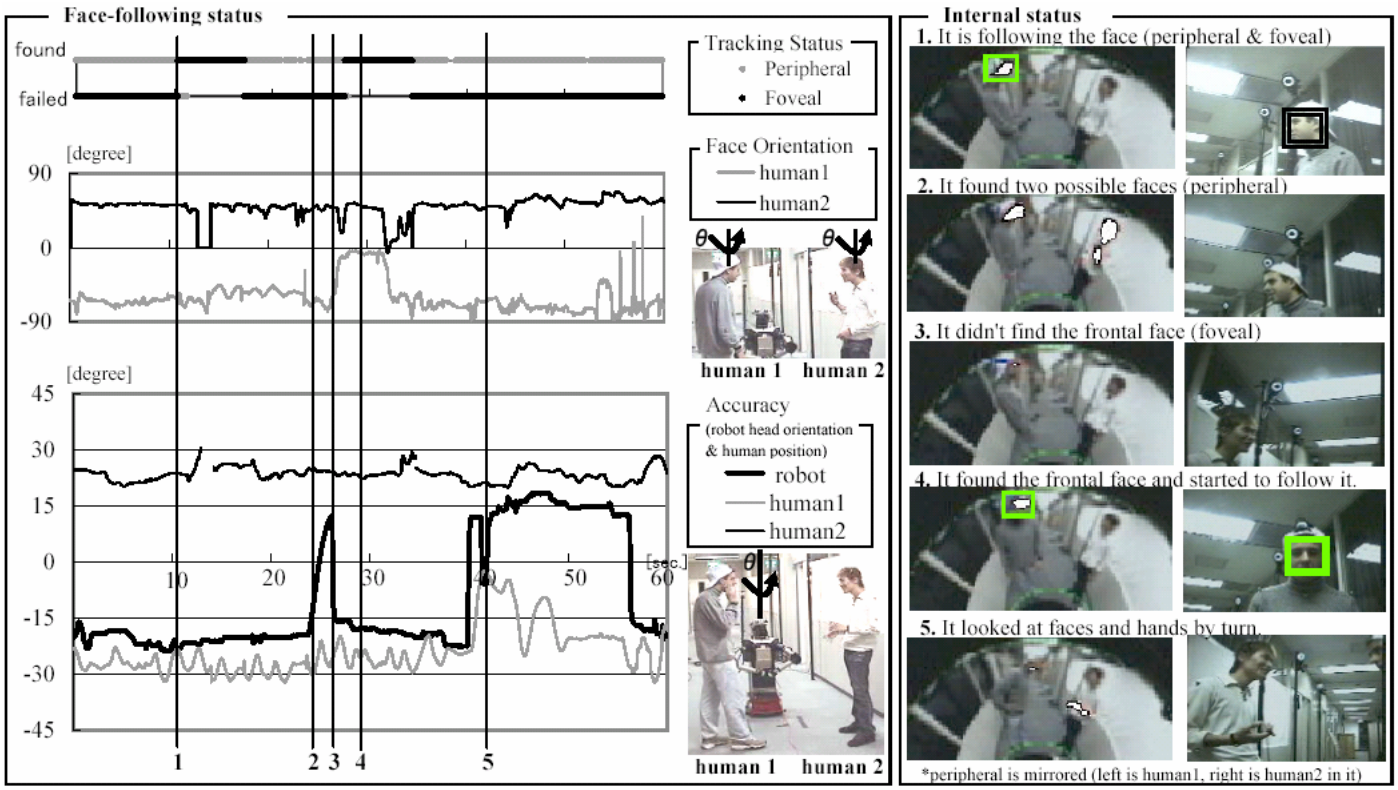

Figure 4: Experimental results for tracking of multiple humans

left half of the figure. For the internal status, each picture on the left denotes the probability $p(x_t)$ as white colored region(s) on the peripheral vision image, and each picture on the right denotes the detected face as squares on the foveal vision image.

As shown in the figure, the robot stably tracked the human face. Scene 1 is the initial status of the robot. The human was approaching the robot while the robot was in "finding mode." There is no meaningful probability for human face existence ($p(x_t)$) at that time. At scene 2, the human came in front of the robot, and then the robot immediately found the human and looked at the human's face. In scenes 3, 4, and 5, it was able to follow the human face when the human moved around. In particular, in scene 4, it lost the frontal face once because the human moved too fast, but it found the frontal face again by using peripheral vision. We believe these results show the robustness and stability of the face tracking function.

During this one-minute experiment, the average error of tracking (difference of the angle between the direction to human face position and orientation of the robot head) was 13.5 degrees (standard deviation was 5.82, which also suggests the stability). As the figure shows (shown in the lower-left as "face-following"), it stably tracked the human face even though it sometimes failed to follow the face with foveal vision. The system successfully found the face in 91% of the frames in foveal vision after the human came in front of the robot.

We also analyzed the system performance during interaction with two humans. As a result, the system is capable of preserving multiple hypotheses (potential human faces). **Figure 4** indicates the result of the experiment. In the experiment, two people talked near the robot, and one of them sometimes interacted with the robot. In the figure, "human 1" and "human 2" stand for the direction of each person's face (in "face orientation") and position to the robot (in "accuracy").

The system followed the face of human 1 (scene 1), found human 1 and human 2 (scene 2), looked at human 2 but did not find the frontal face, which decreased the probability (scene 3), started to maintain eye-contact with human 1 when he started to face the robot (scene 4), and looked at human 2's hand (scene 5) but did not find the frontal face. As a result, the robot successfully detected the potential faces, successfully followed the frontal faces when the humans faced the robot, and did not misunderstand human hands as human faces. (The system is designed to only track a face owing to the integration of peripheral and foveal vision.)

B. Performance of looking back

We evaluated the looking back performance after the robot looked away ("look away" mode). The robot gazed at certain directions (looked away) for *N* second(s) and then looked back to the human face. Meanwhile, the human in front of the robot *followed the gaze* or *stayed still*. Since the robot senses motion and color information in peripheral vision, it is easy to observe a human who moves around it (we expect perfect performance for this condition). However, it is difficult to detect a human who does not move much. Thus, we prepared these two conditions. In the "stay" condition, the human did not move at all. In the "follow gaze" condition, the human only moves his head to follow the robot's gaze.

N \ human	Stay	Follow gaze
1 [s]	1.00	0.95
3 [s]	0.90	0.90
5 [s]	0.85	0.90
7 [s]	0.80	0.90

Table 2: Look-back performance (successful rate)

	Peripheral [ms]	Foveal [ms] ([fps])
Peripheral only (**with** search limit)	8.0	—
Foveal only	—	169.5 (5.9)
Integrated, **w/o** search limit	9.7	322.6 (3.1)
Integrated, **with** search limit	6.4	285.7 (3.5)

Table 3: Comparison of calculation costs of face-following

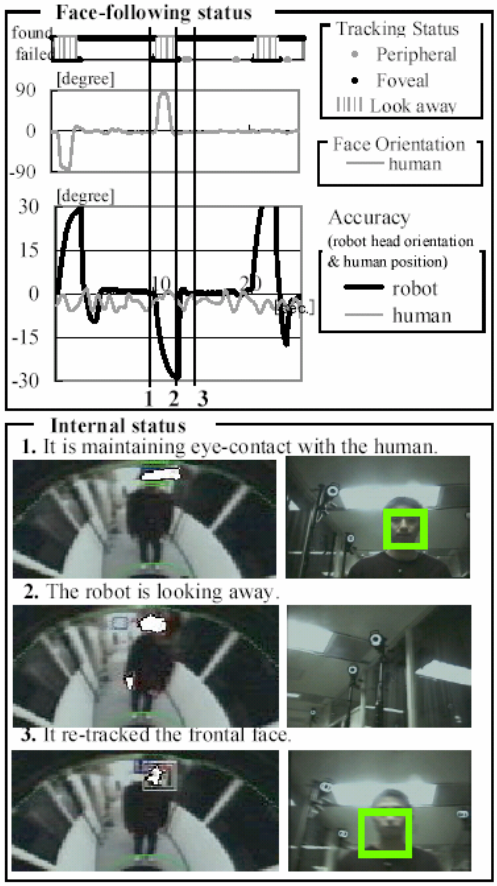

Figure 5: Transition of internal status in "look away mode"

Figure 5 indicates the internal states of the robot during one of the trials. In the "tracking-status" graph (uppermost in the figure), "look away" indicates that the robot was looking away, which corresponds to the "face-following" graph in the figure where the robot's head direction becomes about ±30 degrees. As the "internal status" shows, it was able to preserve the hypotheses on a potential face when it was looking away and continued the eye-contact after that.

Table 2 indicates the rate of successful looking back after looking away for *N* second(s). The human *followed its gaze* or *stayed to gaze at the robot*. In each condition, we performed 20 trials with the robot's looking away and checked the successful rate of looking back. As a result, the performance is quite good when looking away duration *N* is small. In the

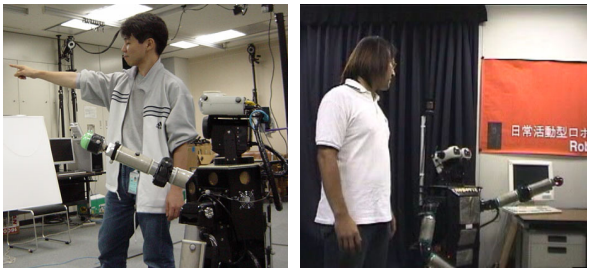

Figure 6: Expression of intention by gazing and pointing

"stay" condition, the performance becomes worse when N increases, since it did not obtain motion information in peripheral vision and thus the preserved region of the human face became a little vague. Even if it failed to look back immediately, it rapidly found the face again.

C. Evaluation of calculation time

We evaluated the calculation time of our integration method. **Table 3** shows the calculation time of peripheral vision and foveal vision. In the table, "foveal only" and "peripheral only" stands for the cost where the system only processed the foveal or peripheral vision. In the "Integrated, w/o search limit" condition, the system searches for faces around every pixel in peripheral vision (equivalent with "$q_{th} = 0$" in equation 1), while "peripheral only" and "Integrated, with search limit" utilized certain q_{th}. Since the system processed peripheral vision with a higher priority than foveal vision, the calculation time of peripheral vision affected that of foveal vision (the amount of calculation in foveal vision is constant). Peripheral vision works in real-time in any experiment condition, and foveal vision is processed 3.5 frames per second in our proposed method. (In the experiment, "foveal" represents the process of retrieving "shape (frontal face)" information in foveal vision, and the process of retrieving "color" information in foveal vision works in realtime). In addition, "Integrated, with search limit" is faster than "peripheral only," since it utilizes foveal vision, consequently, the search limit with $q(x_t)$ was more effective. We believe these results demonstrate the advantage of our approach.

IV. Discussion and Conclusion

We implemented a robust face-following mechanism for an interactive humanoid robot. It is based on integration of peripheral vision and foveal vision and a particle filter method. This mechanism reduces the vagueness of sensor information with this integration method and tracks human face robustly in a realistic daily environment in realtime. Moreover, even when the robot is gazing at other objects, it continues tracking a human face with peripheral vision, and restarts the face-following after that. We conducted several experiments to confirm the performance, and, as a result, the robot successfully maintained eye-contact with a human even after it looked away.

This function of face-to-face interaction is essential for interactive humanoid robots. A robot needs to maintain eye-contact with humans as well as to show its intention by gazing at objects (**Figure 6**). Moreover, it has been found that cooperative body movements such as eye-contact correlates with subjective evaluation of the robot [17]. Furthermore, this eye-contact ability allows robots to perform richer interaction. For example, face-to-face interaction (accomplished by maintaining eye-contact) allows robots to understand human emotion from the face [18], to detect human intention by checking the frontal face, and to effectively gather human utterances by using a unidirectional microphone.

Acknowledgement

We wish to thank Daniel Eaton for his help in this research.

References

[1] T. Kanda, H. Ishiguro, T. Ono, M. Imai and R. Nakatsu, `` Development and Evaluation of an Interactive Humanoid Robot "Robovie"," *IEEE Int. Conf. on Robotics and Automation (ICRA 2002),* pp.1848-1855, 2002.

[2] C. Moore and P. J. Dunham eds., *Joint Attention: Its Origins and Role in Development,* Lawrence Erlbaum Associates, 1995.

[3] B. Scassellati, Investigating Models of Social Development Using a Humanoid Robot, *Biorobotics,* MIT Press, 2000.

[4] H. Kozima and E. Vatikiotis-Bateson, Communicative criteria for processing time/space-varying information, *IEEE Int. Workshop on Robot and Human Communication (ROMAN 2001)*, 2001.

[5] M. Imai, T. Ono, and H. Ishiguro, ''Physical relation and expression: joint attention for human-robot interaction,'' *IEEE Int. Workshop on Robot and Human Communication (ROMAN 2001),* pp. 512 –517, 2001.

[6] T. Kanda, H. Ishiguro, and T. Ishida, ``Psychological analysis on human-robot interaction," *IEEE Int. Conf. on Robotics and Automation (ICRA 2001),* pp. 4166-4173, 2001.

[7] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-Time Auditory and Visual Multiple-Object Tracking for Robots," *Int. Joint Conf. on Artificial Intelligence (IJCAI2001),* pp. 1425-1432, 2001.

[8] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," *Int. Conf. on Spoken Language Processing (ICSLP-2002),* pp. 193-196, 2002.

[9] Y. Matsusaka, S. Kubota, T. Tojo, K. Furukawa, and T. Kobayashi, "Multi-person conversation robot using multi-modal interface," *World Multiconf. on Systems, Cybernetics and Informatics,* Vol. 7, pp. 450-455, 1999.

[10] M. Doi, M. Nakakita, Y. Aoki, and S. Hashimoto, "Real-time Vision System for Autonomous Mobile Robot," *IEEE Int. Workshop on Robot and Human Communication* (ROMAN 2001), pp. 442-449, 2001.

[11] T. Shibata and S. Schaal, "Biomimetic Gaze Stabilization based on Feedback-Error-Learning with Nonparametric Regression Networks," *Neural Networks*, 14(2), pp. 201-216, 2001.

[12] A. Ude, C. G. Atkeson, and G. Cheng, "Combining Peripheral and Foveal Humanoid Vision to Detect, Pursue, Recognize and Act," *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2003),* pp. 2173-2178, 2003.

[13] M. Isard and A. Blake, "CONDENSATION -- conditional density propagation for visual tracking," *Int. J. Computer Vision,* 29, 1, pp. 5-28, 1998.

[14] Javier R. Movellan, Bret Fortenberry, and Ian Fasel, (under review) "A Generative Approach for Real Time Object Detection, Object Location and Object Recognition," MPLAB TR 2003.02 http://mplab.ucsd.edu

[15] Ian Fasel, Bret Fortenberry and Javier R. Movellan, "A Generative Framework for Boosting with Applications to Real-Time Eye Coding," INC MPLab TR 2003.01.

[16] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Trans. on Neural Networks*, 13(6), pp. 1450-64, 2002.

[17] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Body Movement Analysis of Human-Robot Interaction," *Int. Joint Conference on Artificial Intelligence (IJCAI 2003),* pp.177-182, 2003.

[18] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Realtime face detection and facial expression recognition: development and applications to human computer interaction," *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction,* 2003.