Embodied Proactive Human Interface "PICO-2"

Ryo Kurazume, Hiroaki Omasa, Seiichi Uchida, Rinichiro Taniguchi, and Tsutomu Hasegawa
Kyushu University
6-10-1, Hakozaki, Higashi-ku, Fukuoka, Japan
kurazume@is.kyushu-u.ac.jp

Abstract

We are conducting research on "Embodied Proactive Human Interface". The aim of this research is to develop a new human-friendly active interface based on two key technologies, an estimation mechanism of human intention for supporting natural communication named "Proactive Interface", and a tangible device using robot technology. This paper introduces the humanoid-type Two-legged robot named "PICO-2", which was developed as a tangible telecommunication device for the proactive human interface. In order to achieve the embodied telecommunication with PICO-2, we propose new tracking technique of human gestures using a monocular video camera mounted on PICO-2, and natural gesture reproduction by PICO-2 which absorbs the difference of body structure between the user and the robot.

1 Introduction

With rapid and widespread diffusion of the high speed Internet, a wide range of personal and social activities has been performed though the computer system in recent years. However, as expressed with the word of "digital divide", an invisible gap between certain groups who can access information or not induces a variety of serious problems in modern society. The underlying causes of the information disparity might be summarized as the following two reasons.

1. The most of activities through the computer system is performed through a keyboard, a mouse, and a display. However, these actions tend to lose touch with the reality in the real world.

2. To fully utilize the computer system, a user has to give precise instructions to the computer system explicitly and perfectly.

To overcome the above problems and break down communication barriers existing between the user and the computer system, we propose a new framework of a human interface connecting the user and the computer system named "Proactive Human Interface". This framework is characterzed by the following technologies; i) new principle of computer operation named "Proactive System" in which precise instruction by a user is not indispensable, and ii) the use of robotics technology to realize a tangible device instead of ordinary virtualized interfaces.

An example of the proposed proactive human interface is illustrated in Fig.1. Let's suppose the telecommunication

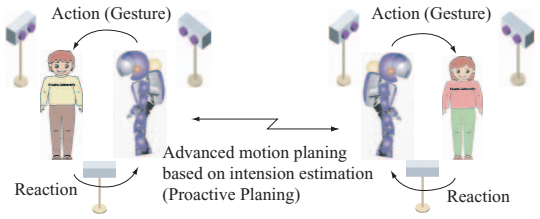


Figure 1. Concept of proactive interface

system consisting of the video phone and a humanoid robot. In case that two persons away from each other communicate with this system, the humanoid robot act as an avatar and a variety of information including not only verbal but also non-verbal information such as gestures, eye sight, and facial expression is communicated through the robot. By giving the sense of the reality to the other through the tangible robot, the performance of communication is augmented comparing with ordinary virtualized interfaces.

In addition, if it is possible to estimate the user's intention from conversation or a record of gestures, and reproduce the appropriate gesture by the robot according to the estimated intension, it becomes a powerful measure to communicate personal intention more precisely even if the reproduced gesture is not actually performed by the user.

There are a large number of studies in the related field such as transmission of special skill using robotics technology [15] or motion support system based on the estimation of human intension [8]. The characteristic of this research is the combination of new principle of computer operation based on the estimation of human intension and an embodied interface which gives the sense of the reality.

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

We have already proposed the prediction system of human intention using the gesture network and continuous DP matching [14]. This paper introduces a prototype humanoid robot of the proactive human interface named "PICO-2", which is designed for an embodied telecommunication system shown in Fig.1. In addition, we propose a new tracking technique of human gestures using a monocular video camera mounted on PICO-2, and a natural gesture generation method which absorbs the difference of body structure between the user and the robot.

2 Humanoid-type proactive interface, "PICO-2"

As a new tangible device of the embodied telecommunication system shown in Fig.1, we developed a humanoid-type proactive interface robot named "PICO-2 (Proactive Interface for Communication)". PICO-2 is designed based on the humanoid robot, HOAP-2 (Fujitsu Automation Ltd.), and is additionally equipped with a LCD, a speaker, a microphone, and an IEEE 1394 digital video camera on its head (Fig.2).

The LCD and the digital video camera are used to record and display user's facial expression, additional information associated with the contents of the conversation, or the conversation itself. Therefore, PICO-2 enables to communicate and reproduce verbal and nonverbal information such as the conversation, the facial expression, the eye sight, and the whole body motion including gestures by hands and even walking or dancing motion.

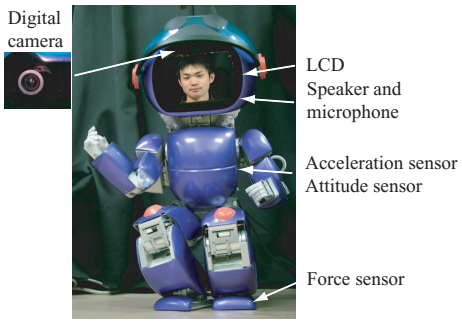


Figure 2. Embodied proactive human interface, "PICO-2"

3 Tracking and reproduction of human gestures using monocular camera

This section introduces the tracking and reproduction techniques of human gestures using a monocular video camera (Fig.2) mounted on PICO-2. The proposed technique utilizes the 3D geometric model of human body and a silhouette on the video image. The distance map [10] from the contour line of the silhouette is constructed and used to align the 3D model and the silhouette in order to estimate human gestures.

3.1 Tracking of human gestures using silhouette and distance map

Tracking and analysis of human motion using a monocular camera is an ill-posed problem since no depth information is obtained such as a stereo camera or a marker-based tracking system. Thus the 3D poses of hands and arms have to be determined by introducing some conditions of constraint. Many researchers have attempted to measure hand motion from 2D video images [21],[13],[1]. The proposed techniques can be classified into two groups, image based approach and model based approach. Image based approach captures and stores a plenty of images before hand, and uses them to recognize human gestures by comparing with an acquired image [20],[12]. Since coarse images are sampled and compressed using PCA, HMM, or other data compression techniques, precise poses of hands and arms can not be retrieved. In addition, since the computational cost is quite high, it is not suitable for real-time applications. On the other hand, model based approach tries to create a 3D model of a human [7] using simple skeleton model[4][17] or precise 3D geometrical model of the body [6],[5],[3], and the motion is tracked by finding an appropriate posture of the 3D model which matches with the image.

Our technique belongs to the model based approach. We utilizes a distance map [10] for the comparison of a 2D silhouette and a synthesized silhouette of the 3D geometrical model of a human body. The processing flow is summarized as follows;

1. At first, a contour line of the boundary of the body (silhouette) on the 2D image is detected using active contour model (ex. Snakes or the Level Set Method [19], [9]).

2. Next, the distance map from the contour line is constructed using the Fast Marching Method [19] (Fig.3).

3. The 3D geometric model of the body (Fig.4) is placed with arbitrary pose and a projected silhouette on the 2D image plane is synthesized. The position of the torso is assumed to be aligned roughly to the 2D silhouette image beforehand. We created a whole 3D model of human body using the laser range finder, VIVID700.

4. Contour points of the projected silhouette and their corresponding patches on the 3D model are identified.

5. Force $\mathbf{f}_i$ is applied to the selected patch i of the 3D model in 3D space according to the distance value obtained from the distance map. The force $\mathbf{f}_i$ is the vector perpendicular to the line of sight $\mathbf{v}$, and the projection of the $\mathbf{f}_i$ onto the 2D image plane coincides with

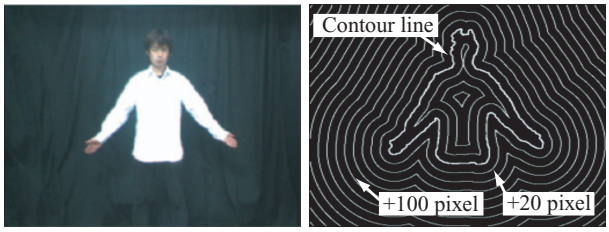
the $\mathbf{f}_{D_i}$, which is the vector toward the direction of the steepest descent of the distance map (Fig.5). We assume that the magnitude of the $\mathbf{f}_{D_i}$ is proportional to the value at the projected contour point of the 3D model on the distance map.

$$\mathbf{f}_{D_i} = D_{s,t} \frac{\nabla D_{s,t}}{|\nabla D_{s,t}|} \quad (1)$$

6. The moment around the elbow and the shoulder joints of the 3D model is determined as shown in Fig.5. The details are explained in Section 3.2.

7. The poses of the arm and the torso of the 3D model is changed according to the total force and moment.

8. Repeat from step 1 to 7 until the projected image of the 3D model and the 2D image coincide each other.



2D image Distance map from the silhouette

Figure 3. Distance map from the contour of the silhouette

A number of registration techniques of 2D image and 3D model using silhouettes have been proposed so far [11],[16],[5],[18],[3]. In contour-based approaches, the error is usually computed as the sum of distances between points on a contour line in a 2D image and on a projected contour line of a 3D model [5]. However, since a number of point correspondences between both contours have to be determined for calculating the registration error, these algorithms are computationally expensive. On the other hand, we adopt a distance map for evaluating a registration error instead of the point correspondence [10]. Therefore, once the distance map is created, our algorithm runs faster than the conventional point-based approach. In addition, since the distance map can be created quite rapidly using the Level Set Method named the Fast Marching Method [19], our method is able to track an object in real-time even if the object moves.

3.2 Calculation of joint motion

In step 6 of the above procedure, the moment applied to the elbow and the shoulder joints are calculated from the force applied to the forearm and the upper arm.

Figure 4. 3D Geometric model of human

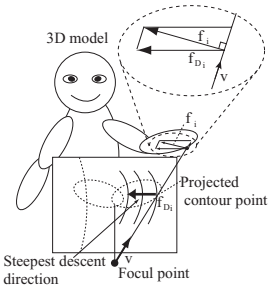


Figure 5. Force applied to each patch.

Firstly, the 3D models of the forearm and the upper arm are projected on the 2D image, and the 3D patch i corresponding to the contour of the synthesized silhouette image is determined. Here we denote the 3D patches corresponding to the upper arm and the forearm as P_{u_i} and P_{f_j}, the vector from the shoulder to the patch P_{u_i} as $\mathbf{r}_{u_i}$, the vector from the elbow to the patch P_{f_j} as $\mathbf{r}_{f_i}$, and the vector from the shoulder to the elbow as $\mathbf{r}_{se}$, respectively. Then the moment around the shoulder and the elbow is calculated according to the following equations.

$$\mathbf{M}_u = \sum_{i \in P_u} \rho(\mathbf{r}_{u_i} \times \mathbf{f}_{u_i}) + g_1 \sum_{j \in P_f} \rho\{(\mathbf{r}_{se} + \mathbf{r}_{f_j}) \times \mathbf{f}_{f_j}\} \quad (2)$$

$$\mathbf{M}_f = \sum_{j \in P_f} \rho(\mathbf{r}_{f_j} \times \mathbf{f}_{f_j}) \quad (3)$$

where g_1 is a suitable gain and $\rho(z)$ is a particular estimate function. To overcome the occlusion problem of the contour, we utilize the robust M-estimator and introduce the estimate function $\rho(z)$. In our implementation, the Lorentzian function is used as the estimation function. In addition, the hand position extracted by skin color is also utilized to determine the moment around the shoulder and the elbow in the following experiments.

Next, minute displacement of the joint angle $\Delta\phi$ is obtained by calculating an inner product of the moment around the joint and its rotation axis. Since the shoulder joint of the developed model has 3 DOFs and the elbow joint has 1 DOF, minute displacement of each joint angle is shown as the following equation.

$$\Delta\phi_{r(p,y,e)} = \mathbf{M}_u \cdot \mathbf{s}_{r(p,y,e)} \quad (4)$$

where the three rotation axes of the shoulder joint are denoted as $\mathbf{s}_r$,$\mathbf{s}_p$, and $\mathbf{s}_y$, and the rotation axis of the elbow joint as $\mathbf{s}_e$.

Therefore, by adding the minute displacement of the joint angle given from Eq.4 to the current joint angle, human gestures can be tracked by the 3D model and the joint angles are estimated.

3.3 Reproduction of human gestures by PICO-2

Since the length of links and the configuration of joint axes are different between human and PICO-2, it is not suitable to apply the measured joint angle of human gestures to PICO-2, directly. In order to achieve human-like motion with a humanoid robot [2], our method is to reproduce the appearance of human gestures so that the directions of vectors $\mathbf{r}_{eh}$ and $\mathbf{r}_{sh}$ of human and PICO-2 coincide with each other.

Let's denote the length of the upper arm of PICO-2 as L_1, the length of the forearm as L_2, the length from the shoulder to the hand as k as shown in Fig.6. The normal vectors of $\mathbf{r}_{eh}$ and $\mathbf{r}_{sh}$ are denoted as $\mathbf{n}_{eh}$ and $\mathbf{n}_{sh}$, respectively. The vector from the shoulder to the elbow $\mathbf{P}_{el}$ is

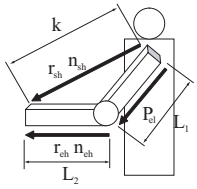


Figure 6. Gesture reproduction by directing hand and forearm directions.

expressed as the following equation.

$$\mathbf{P}_{el} = k\mathbf{n}_{sh} - L_2\mathbf{n}_{eh} \quad (5)$$

Since $\mid \mathbf{P}_{el} \mid = L_1$, k is determined as

$$k = L_2\mathbf{n}_{sh}\cdot\mathbf{n}_{eh} + \sqrt{{L_2}^2(\mathbf{n}_{sh}\cdot\mathbf{n}_{eh})^2 - ({L_2}^2 - {L_1}^2)} \quad (6)$$

¿From Eqs.5 and 6, the elbow position is determined and the appearance of human gestures can be reproduced so that the directions of the hand and the forearm are matched between the human and PICO-2.

4 Experiments

We conduct experiments of the tracking and reproduction of human gestures using PICO-2 and the monocular video camera mounted on PICO-2. The human gesture is estimated and the appearance of the gesture is reproduced in real-time using the proposed techniques.

4.1 Motion tracking by monocular video camera using distance map

Firstly, the experiments of motion estimation by the monocular video camera are carried out. The left column of Fig.7 shows the acquired image by the monocular camera and the right column shows the estimated pose by the 3D geometrical model. It is clear that the poses of the 3D geometrical model are almost same as the input images and the human gesture can be estimated by the proposed distance-map-based method. Estimation of 3D motion of whole body is also performed as shown in Fig.8.

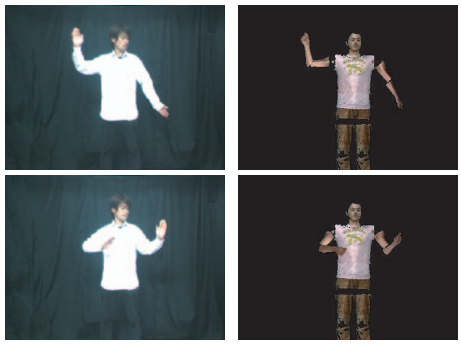


Figure 7. Imitation of human motion by 3D model

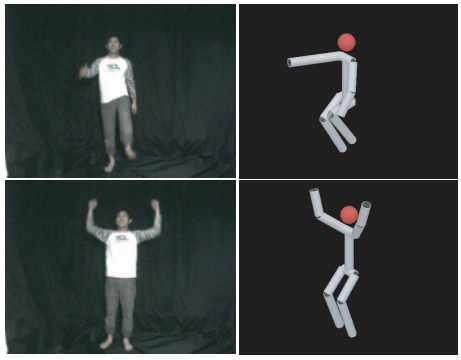


Figure 8. Whole body motion tracking by a monocular camera

4.2 Motion reproduction by PICO-2

Next, we conducted the experiments of appearance-based motion reproduction using PICO-2. The human gesture is estimated by the above technique and the appearance of the human gesture is reproduced so that the directions of the hand and the forearm are matched between the human and PICO-2. The experimental results are shown in Fig.9. Processing time from the image acquisition to the calculation of the motor command is less than 30 ms. ¿From these results, it is verified that the human gesture is estimated and the gesture with similar appearance is reproduced by PICO-2 in real-time.

5 Conclusion

This paper introduces the concept of the "Embodied Proactive Human Interface", and the humanoid-type 2

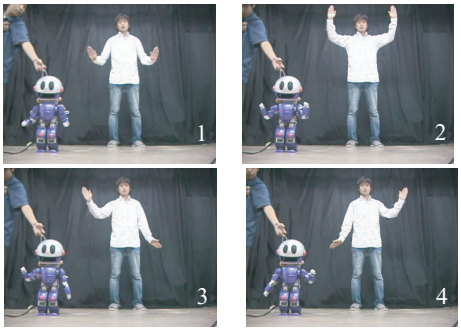


Figure 9. Imitation of human motion by PICO-2

legged human interface robot named "PICO-2". The aim of this research is to produce a natural and user-friendly human interface for many different kinds of people utilizing i) new principle of computer operation named "Proactive System" in which precise instruction by an operator is not indispensable, and ii) the use of robotics technology to realize a tangible device instead of ordinary virtualized interfaces, for breaking down communication barrier existing between the user and the computer system.

In this paper, we propose new distance-map-based tracking technique of the human gesture using a monocular video camera mounted on PICO-2, and natural gesture reproduction by PICO-2 which absorbs the difference of body structure between PICO-2 and the user.

Acknowledgment

This research was supported in part by the 21st Century Center of Excellence Program under the title of "Reconstruction of Social Infrastructure Related to Information Science and Electrical Engineering", and by the Ministry of Public Management, Home Affairs, Posts and Telecommunications of Japan under the Strategic Information and Communications R&D Promotion Programme (SCOPE).

References

[1] J. K. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding: CVIU*, 73(3):428–440, 1999.

[2] T. Asfour and R. Dillmann. Human-like motion of a humanoid robot arm based on a closed-form solution of the inverse kinematics problem. In *Proceedings of International Conference on Intelligent Robots and Systems*, pages 1047–1412, 2003.

[3] J. Carranza, C. Theobalt, M. Magnor, and H. Seidel. Free-viewpoint of human actors. In *Proceedings of SIGGRAPH 200*, pages 569–577, 2003.

[4] Z. Chen and H. J. Lee. Knowledge-guided visual perception of 3-d human gait from a single image sequence. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(2):336–342, 1992.

[5] Q. Delamarre and O. Faugeras. 3d articulated models and multi-view tracking with silhouettes,. In *In Proc. IEEE International Conference on Computer Vision*, pages pp.716–721, 1999.

[6] D. M. Gavrila and L. S. Davis. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In *International Conference on Automatic Face and Gesture Recognition*, pages 272–277, 1995.

[7] Y. Guo, G. Xu, and S. Tsuji. Tracking human. body motion based on a stick-figure model. *Journal of Visual Communication and Image. Representation*, 5(1):1–9, 1994.

[8] Y. Inagaki, H. Aisu, H. Sugie, and T. Unemi. A study of a method for intention inference from humanfs behavior. In *Proceedings of the IEEE International Workshop on Robot and Human Communication*, pages 142–145, 1993.

[9] Y. Iwashita, R. Kurazume, K. Hara, and T. Hasegawa. Robust motion capture system against target occlusion using fast level set method. In *Proc. IEEE International Conference on on Robotics and Automation*, 2006.

[10] Y. Iwashita, R. Kurazume, K. Konishi, M. Nakamoto, M. Hashizume, and T. Hasegawa. Fast 2d-3d registration for navigation system of surgical robot. In *Proc. IEEE International Conference on on Robotics and Automation*, pages pp.909–915, 2005.

[11] H. Lensch, W. Heidrich, and H.-P. Seidel. Automated texture registration and stitching for real world models. In *In Pacific Graphics '00*, pages 317–326, 2000.

[12] J. Martin and J. Crowley. An appearance-based approach to gesture recognition. In *Proceedings of the International Conference on Image Analysis and Processing*, pages 340–347, 1997.

[13] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268, 2001.

[14] A. Mori, S. Uchida, R. Kurazume, R. Taniguchi, T. Hasegawa, and H. Sakoe. Early recognition and prediction of gestures. In *Proc. International Conference on Pattern Recognition*, 2006.

[15] S. Nakaoka, A. Nakazawa, K. Yokoi, and K. Ikeuchi. Leg motion primitives for a dancing humanoid robot. In *Proceedings of the IEEE 2004 International Conference on Robotics and Automation*, pages 610–615, 2004.

[16] P. J. Neugebauer and K. Klein. Texturing 3d models of real world objects from multiple unregistered photographic views. In *Computer Graphics Forum 18*, pages 245–256, 1999.

[17] C. Orrite-Urunuela, J. del Rincon, J. Herrero-Jaraba, and G. Rogez. 2d silhouette and 3d skeletal models for human detection and tracking. In *Proceedings of the 17th International Conference on Pattern Recognition*, pages 244–247, 2004.

[18] R. Plaenkers and P. Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding: CVIU*, 81(3):285–302, 2001.

[19] J. Sethian. *Level Set Methods and Fast Marching Methods, second edition*. Cambridge University Press, UK, 1999.

[20] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *International Workshop on Automatic Face and Gesture Recognition*, pages 189–194, 1995.

[21] L. Wan, W. Hu, and T. Tan. Recent developments in human motion analysis. *Pattern Recognition*, 36:585–601, 2003.