Behavior learning of a face robot using human natural instruction

Fumiya Iida [(1)], Harumi Ayai [(2)], Fumio Hara [(2)]

iida@ifi.unizh.ch, {haru, hara}@hafu0103.me.kagu.sut.ac.jp

[(1)]AILab, Dept. of Information Technology, Univ. of Zurich

Winterthurerstr 190, CH-8057 Zurich, Switzerland

[(2)]Dept. of Mechanical Engineering, Science Univ. of Tokyo

1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

Keywords: Man-machine interaction, Reinforcement learning, Q-learning, Robot behavior learning, Human instruction, Face Robot

1. Introduction

Interaction between human and robot that has the function not only to achieve its task but also to take friendly action to humans has recently taken an attention very much [1][2]. It is important to develop such an interacting robot, because future robot is expected to work much more in human society.

Interaction between robot and human can be realized in the situation that the robot has a specific sensory-motor coordination in which the sensory inputs are human actions to the robot and motor outputs are the responding action to humans. The sensory-motor coordination of an interacting robot should be at least well understandable or something to create preferable behaviors for human partners, thus it is also significant to investigate how human partners are impressed through observing the robot behaviors. The impression of such interacting robot behaviors that are evaluated by human observers is strongly dependent on the sensory-motor coordination of the robot [3]. However we have no strategy to design the sensory-motor coordination of such robot to behave friendly for human partners, because it seems to be too complicated since the robot has to have a large number of sensory inputs and action outputs needed for the interaction.

From this perspective, one of the solutions to design the sensory-motor coordination of the interacting robot, could be autonomous-organization of sensory-motor coordination through the robot's experience of interaction with human partners. The most primary method is thought that a human partner teaches the robot by means of "reward for robot friendly behaviors": i.e. the human partner gives the instruction to the robot by evaluating robot's behavior.

Contrary to this, in conventional approaches of such instructions, a human teacher shows the robot one of the best behaviors [4][5], and it is difficult to make the robot learn such ambiguous human behavior that is only understood by human, or that cannot be shown as one of the best examples for the robot.

During the robot-instruction experiments using reward-based instruction [6], some problems have been found. Namely we need many number of rewarding process to organize the robot sensory-motor coordination, which causes human fatigue, and furthermore, it is expected that the robot cannot organize its sensory-motor coordination in the case when the robot has many degrees of freedom in its action.

In this paper, based on the concept "Learning Oriented Interaction (LOI)" as a basic strategy of organizing the sensory-motor coordination for the interacting robot, we first point out the problems of the conventional approach to realize LOI, i.e. "Direct instruction method". Then we propose an alternative approach, i.e. "Natural Instruction Method", and we conduct the experiments using a Face Robot system in order to evaluate effectiveness of the new method in real robot-interaction with human.

In the next section, we explain briefly the concept of "Learning Oriented Interaction", and point out its problems. In section 3, we propose the alternative idea for Learning Oriented Interaction, i.e. "Natural Instruction Method". In section 4, we explain the experiment using a Face Robot implemented with Natural Instruction Method, and evaluate its performance.

2. Learning Oriented Interaction

2.1 Robot-human interaction

For friendly interaction between a human partner and a robot, we think that it is necessary for the robot to interact with the human partner with not only verbal, but also non-verbal

modalities, such as facial expressions, gestures, and so on. Generally, the behavior of a robot is determined by means of sensory-motor coordination, and in the framework of interaction between human and robot, the sensory inputs are thought to be human partner's states, and the motor outputs to be the robot actions to the human partner. However, if it is assumed that the robot has a large number of sensory inputs and motor outputs that are needed for friendly interaction, the sensory-motor coordination would be too complicated to be determined by a designer. In addition, not only each human partner's state that is recognized by the robot but also robot's action have different meanings for individual human partner in certain situations, which makes it more difficult to determine the sensory-motor coordination by a designer.

From this point of view, we employ the concept of "Learning Oriented Interaction (LOI)", in which an interacting robot learns its sensory-motor coordination through the interaction experiences with a human partner. To make the robot organize its sensory-motor coordination through learning, human partner needs to give "rewards" to the robot as instructional information. In this learning, there are two methods to give the robot rewards, i.e. Direct Instruction Method and Natural Instruction Method. In the Direct Instruction Method, a human partner gives the robot reward in a fixed manner; i.e. a designer determines beforehand how the human partner gives the reward to the robot, in which the human partner needs to know how to teach the robot before the interaction. In the Natural Instruction Method, on the other hand, the robot autonomously understands the values of human states by means of its experience of interaction, and then the values are used as reward for its behavior learning.

2.2 Direct instruction method

In the “Direct Instruction Method (DIM)” (Fig. 1), the robot takes a certain action corresponding to recognized current human state, and then a human partner gives a positive/negative reward to the robot on the basis of human partner’s evaluation of the robot action in the interaction.

This interaction is achieved by following procedure.

The robot recognizes a human state, and then take an action corresponding to the human state for one step.

The action robot takes at the step is determined by means of Q-table, which consists of Q-values for each combination of state **s** and action **a**, i.e. **Q(s,a)**, where **s** is a human state ID

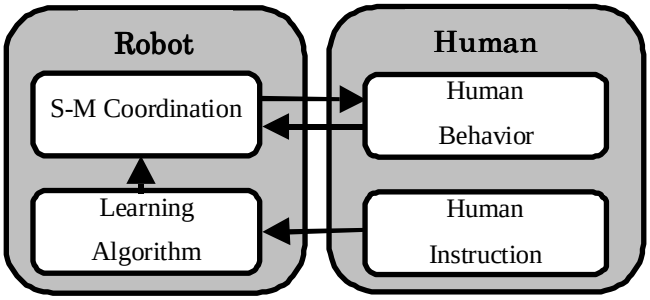

Fig. 1 Direct Instruction Method

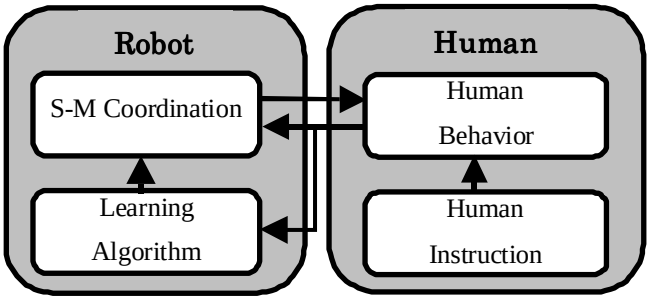

Fig.2 Natural Instruction Method

number and **a** is a robot action ID number. The robot decides an action to take in each step by comparing probability **P(s,a)** calculated from following formula (1).

$$P(a/s) = e^{\frac{Q(s,a)}{T}} \Big/ \sum_{b \in A} e^{\frac{Q(s,b)}{T}} \qquad (1)$$

where T, Boltzmann temperature, is the variable to control “exploitation and exploration trade-off”, i.e. the robot takes the action randomly if T is set to be high.

Human partner gives the reward as a scalar value **r** for each step. The robot changes its Q-values by means of following formula (2).

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha r \qquad (2)$$

where α is learning rate.

In this method, the delayed-reward mechanism is not taken into account, thus the reward given by the human partner is directly reflected to the Q-values through the robot interaction.

We have conducted interaction experiments with human test subjects [6], but the following problems are found. First, the human partner has to have sufficient number of rewarding to make the robot learn only one action, thus it is expected that a human partner cannot teach complicated robot behavior that needs a large number of combination between human states

and robot's actions. Second, since a human partner is forced to give rewards in a fixed manner designed by the designer, he/she cannot give precise instructional information to the robot.

3. Behavior learning through human natural instruction

3.1 Idea

To solve these problems stated in 2.2, we propose "Natural Instruction Method (NIM)" (Fig. 2). This method is derived from the fact that human partner's behaviors responding to the robot usually include a lot of instructional information which may be taken advantage of behavior learning of the robot. Thus we call such human behavior "human natural instruction".

In order to utilize instructional information in human natural instruction, the robot needs to know the values of human states, which is used for the robot behavior learning. In this method, therefore, a human partner does not need to give any rewards explicitly, but just responds to each action of the robot: i.e. the robot autonomously understands the values of human partner's states and then organizes sensory-motor coordination using such values. This method has the following advantages. First, a variety of instruction methods can be acceptable, because the robot can recognize the values of individual states through the interactions and utilize such values as instructional information. Secondly, compared with DIM, the robot can recognize more detailed and more instructional information, because the values of all of the human states that are learned through interaction will be used for other behavior learning. Finally, owing to the second advantage, the robot can learn faster.

3.2 Natural instruction learning algorithm

To make the robot learn the values of human states, we employ Q-learning algorithm with delayed reward [7]. The most significant feature of Q-learning is the delayed reward mechanism, in which Q-values are propagated by means of state transition. In the Q-learning, the following formula (3) is used for calculation of **Q(s,a)**, instead of formula (2) in DIM.

$$Q(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha(r_t+\gamma\max_{b\in A}Q(s_{t+1},b)) \quad (3)$$

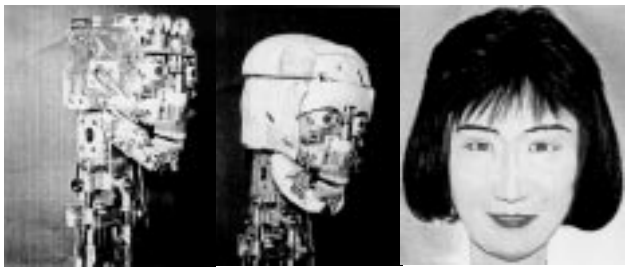

Fig.3 Appearance of face robot

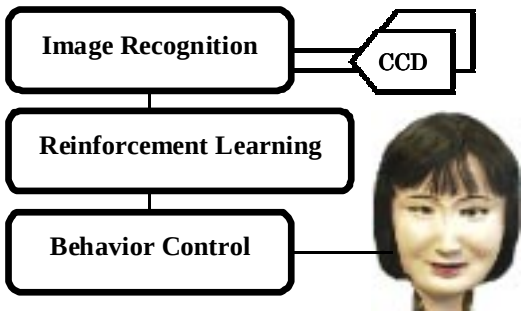

Fig.4 Structure of Face Robot system

Where γ is a discount rate, and $\max_{b\in A} Q(s_{t+1},b)$ is the maximum Q-value in the state recognized in **t+1**. Thus, for example, as a human state $\mathbf{s_t}$ transits to $\mathbf{s_{t+1}}$, $Q(s_t,a_t)$ is changed if $Q(s_t,a_t)$ is not the same value as $\max_{b\in A} Q(s_{t+1},b)$.

In this method, once we determine initial Q-values, i.e. set some values but zero to some of the Q-values before interaction, Q-values are changed by means of human state transition, which means that the robot learns the values of human states through the interaction.

4. Experiment

4.1 Experimental procedures

The purpose of this experiment is to illustrate how NIM works in real robot-human interaction and analyze the robot behavior.

As a platform of the experiment, we employ the Face Robot system [1], since the Face Robot system has many interaction channels. The main functions of the Face Robot system are as follows:

Table 1 Action elements and action choices

Action element	Choice 1	Choice 2	Choice 3
Facial expression	neutral	happy	
Face direction	looking toward human partner	looking away from human partner	
Voice	"come on"	"go away"	nothing

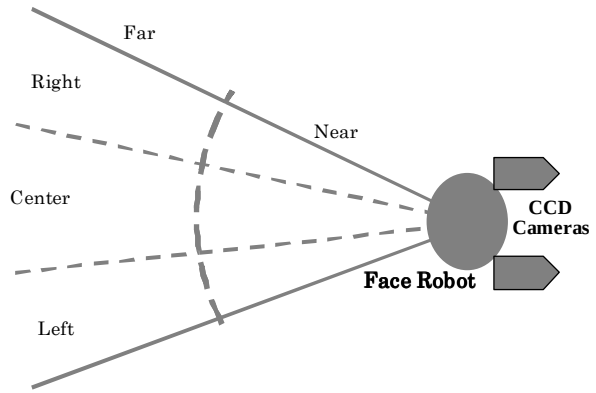

Fig.5 Experiment area

(1) Face Robot is equipped with 18 small pneumatic actuators at proper locations inside the face structure to generate 6 typical facial expressions such as anger, surprise, fear, disgust, sad, and happy. Besides the Face Robot has 2 degrees of freedom in the eyeball rotation and 3 degrees of freedom in the head rotation.

(2) Face Robot is implemented with a real-time vision system to identify the location of Face Robot's human partner or his/her facial expressions.

(3) Face Robot is equipped with a conventional voice synthesizing system.

The size of the face robot was almost 120 % of a normal human head and her appearance is a young woman face as shown in Fig. 3.

Fig. 4 shows the structure of Face Robot system that consists of 3 parts, i.e. behavior control part, image recognition part, and reinforcement learning part. Behavior control part controls the action of the Face Robot. In this experiment the Face Robot is implemented with the actions shown in Table 1. The Face Robot executes an action in one step, which includes 3 action elements. In image recognition part, the Face Robot system recognizes human position, i.e. direction and distance from the Face Robot by using 2 CCD cameras, and then evaluates which region the human position is in the prespecified fan-like 6 regions shown in Fig. 5.

We conduct 2 types of experiment in DIM and NIM, in order to evaluate the effectiveness of NIM and compare that with DIM. In the experiment of DIM, a test subject not only changes his/her position, but gives a positive/negative reward to the Face Robot in each step based on the evaluation of a Face Robot action in that step. In the formula (3) of the learning algorithm of the Face Robot system is set to γ=0.0, thus Q-table obtained in this experiment is directly reflected by the reward pattern given by test subject. In the NIM experiment, a test subject moves from one region to another and gives no reward. Once the Face Robot recognizes that the human position has been changed, the Face Robot proceeds its learning step to the next one automatically. In this experiment, γ in the learning algorithm is set to 1.0, in which the Face Robot can propagate Q-value by means of human state transition.

In the each experiment, 5 test subjects are used for human partner and the learning takes 100 steps. Parameters for the learning algorithm are $\alpha = 0.5,\ T = 0.1$ for the DIM, and $\alpha = 0.5,\ T = 100$ for the NIM experiment.

4.2 Results

First, we define the destination Q-table for the learning algorithm of the Face Robot system in order to evaluate NIM. In this paper, as a destination Q-table, we employ the average of Q-table obtained in the DIM experiments, because the Q-tables obtained in the DIM reflect human positive/negative rewards directly.

Then we define the error of learning to measure the difference between the one obtained in the NIM and the destination Q-table. The error of learning is calculated as follows: (1) translate Q-values obtained in both DIM and NIM experiments into probability using formula (1), (2) subtract each probability in the NIM from corresponding probability in the DIM, and sum up all differences. Namely the smaller the error of learning, the better the Face Robot system learns the destination Q-table correctly.

The error of learning is strongly dependent on the transition frequency of human states, which indicates that each test subject inclines to take the same actions after the same Face Robot action. Thus we calculate the standard deviation of state

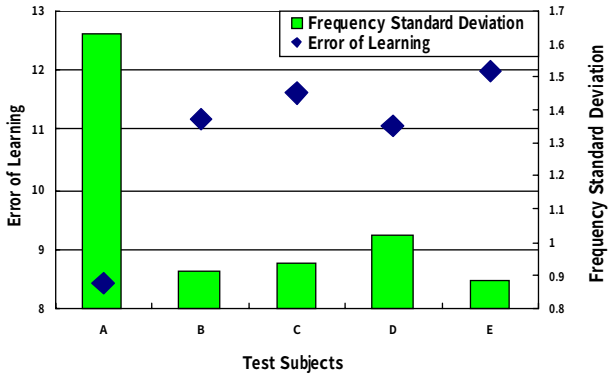

Fig. 6 Correlation between error of learning and standard deviation

transition frequency in the following manner: (1) sum up the number of states to which test subjects move after each sequence of state and Face Robot action, (2) calculate the standard deviation of the sums obtained from (1). Therefore the greater the standard deviation of transition frequency, the more human state transitions have the inclination to taking the same action after the same Face Robot action.

Fig.6 shows the error of learning and standard deviation of state transition frequency. The higher standard deviation of state transition frequency, the smaller the error of learning, which indicates that the Face Robot can learn destination Q-table if test subjects have inclinations in their state transitions.

Fig. 7 shows the average standard deviation for all test subjects with respect to 2 action elements of facial expressions and voice. In this figure, there are differences between 2 choices of action elements, which indicates that the state transition of test subjects corresponding to each Face Robot action is different. The differences in standard deviations suggest that the state transition can be changed depending on the Face Robot actions, which affects the learning performance of the Face Robot.

From these results, it is clarified that the behavior learning of the Face Robot through human-robot interaction converges differently from each action, more specifically, the Face Robot can learn easily the meaning of such action that has a strong feature for human to like happy facial expression or the voice "Come". The reason the Face Robot learns such actions seems to be due to that human can easily understand the meaning of the Face Robot actions and take the same responsive action as the action of the Face Robot.

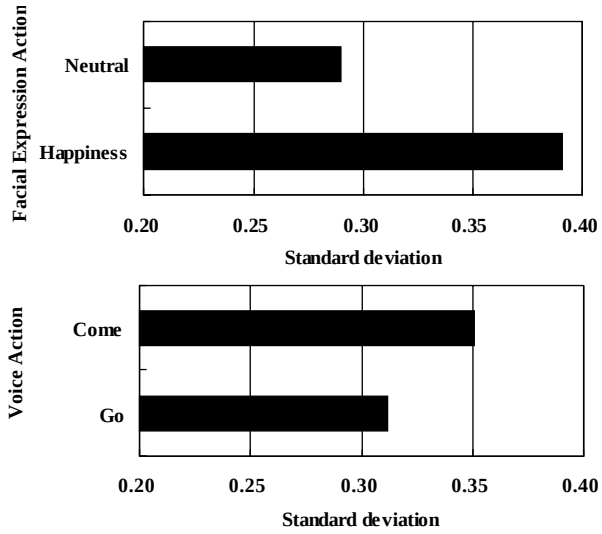

Fig.7 Standard deviation for action choices in each action element

5. Conclusions

In this paper we have developed "Natural Instruction Method", based on the concept of LOI, a strategy to design such human friendly robot that interacts with human in both verbal and non-verbal communication modalities. Furthermore, based on this proposed method, we have conducted the experiments using Face Robot system to analyze characteristics of the proposed method.

This paper shows that the interacting robot can understand the value of human states autonomously by using delayed reward mechanism of Q-learning, and then organize its behavior. In addition, it is clarified that there is difference in the robot behavior learning with respect to the meaning of the action for human partners.

Since these results are strongly depenent on initial Q-values and structure of clustered human states and robot actions, we need to analyze the statistics of NIM systematically in terms of these whole factors, which will be done in the future works.

Acknowledgement

This work was supported by the JSPS Research for the Future program P9600803 and the thanks go to all of the people involved in the program.

References

[1]F.Hara, H.Kobayashi: State-of-the Art in Component Technology for an Animated Facerobot- Its Component Tech-

nology Development for Interactive Communication with Humans, The Intl. Jour. of the Robotics Society of Japan Advanced Robotics, Vol.11, No.6, pp.585-604, 1997

[2]Tatsuya Harada, Tomomasa Sato, Taketoshi Mori: Psychological Effect of Contact Interaction Robot, Journal of the Robotics Society of Japan Vol. 16 No.5, pp122-128, 1998 (in Japanese)

[3] F. Iida, M. Tabata, F. Hara: Generating Personality Character in a Face Robot through Interaction with Human, Proc. of 7th IEEE International Workshop on Robot and Human Communication, pp481-486 (1998)

[4] Long-Ji Lin: Programming Robots Using Reinforcement Learning and Teaching, Proc. of 9th National Conference on Artificial Intelligence, pp.781-786, 1991

[5] J.A.Clouse, P.E.Utgoff: A Teaching Method for Reinforcement Learning, Proc. of 9th International Conference on Machine Learning, pp.92-101, 1992

[6] Fumiya IIDA, Fumio HARA, Harumi AYAI: Face Robot behavior learning based on the characteristics of human instruction, Proc. of 4th Robotics Symposia, pp.38-43, 1999 (in Japanese)

[7]Watkins, C. J. C. H., Dayan, P.: Q-learning, Machine Learning 8(3), pp279-292, 1992