Robot Behavior Adaptation for Human-Robot Interaction based on Policy Gradient Reinforcement Learning*

Noriaki Mitsunaga[1], Christian Smith[1,2] Takayuki Kanda[1],
Hiroshi Ishiguro[1,3], Norihiro Hagita[1]

[1]*ATR Intelligent Robotics and Communication Laboratories*
[2]*Royal Institute of Technology, Stockholm*
[3]*Graduate School of Engineering, Osaka University*
mitunaga@atr.jp

***Abstract*— In this paper we propose an adaptation mechanism for robot behaviors to make robot-human interactions run more smoothly. We propose such a mechanism based on reinforcement learning, which reads minute body signals from a human partner, and uses this information to adjust interaction distances, gaze meeting, and motion speed and timing in human-robot interaction. We show that this enables autonomous adaptation to individual preferences by an experiment with twelve subjects.**

***Index Terms*— policy gradient reinforcement learning (PGRL), human-robot interaction, behavior adaptation, proxemics.**

I. Introduction

When humans interact in a social context, there are many factors apart from the actual communication that need to be considered. Previous studies in behavioral sciences have shown that there is a need for a certain amount of personal space [2] and that different people tend to meet the gaze of others to different extents [3]. For humans, this is mostly subconscious, but when two persons interact, there is an automatic adjustment of these factors to avoid discomfort. When a conversational partner stands too close, we tend to move away, and when we are stared at, we tend to avert our eyes [9].

There have been a few studies of how to apply this sort of behavior to robots, e.g. [7] [10] [8], but so far, none of them have addressed the problem of adapting to individual preferences in these matters. In order to facilitate human-robot interaction, a robot must be able to read small discomfort signals in humans and adjust its behavior accordingly, just like a human would.

We also note that it is important for a robot to be able to perform this kind of adaptation autonomously. A system that requires tuning by a specialist is not only cumbersome in the start-up phase, but for a robot that would meet new people on a daily basis, it would be very unreasonable to require personal adjustment before interaction could start. Autonomous adjustment would be especially important for robots working in a public environment, for example performing guide tasks.

Incremental learning of behavior decision through interaction with a human [6] has been proposed. They used a kind of teaching method which requires conscious answers from a human user. This kind of learning will be important for a task completion. However, an adaptation method, which reads subconscious responses from the human, is required for smooth human-robot communication.

The matters are further complicated by the fact that human preferences seem to be interdependent. The discomfort of personal space invasion is lessened if gaze-meeting is avoided [9]. Where human-robot interaction is concerned, studies also show that a person's feeling of comfortable distance for a robot varies with how menacing the robot's actions are perceived, i.e. the robot's movement speed [7]. This means that in order to construct a system that adapts to personal preferences, several parameters have to be considered simultaneously, resulting in a large multi-dimensional space that needs to be searched in order to find the optimum.

Another requirement for a system for this kind of adaptation is that it performs reasonably well during the adaptation process. Especially in a public setting, for example in a school or a museum, it is very important that the robot does not exhibit antisocial behavior during the initial phases of adaptation.

In this paper, we propose a behavior adaptation system based on reinforcement learning to solve this problem. Using small discomfort signals from the human partner as input to the reward function, it finds the behavior that minimizes these signals, and thereby also minimizes the actual discomfort experienced by the human. In the following, we first show the proposed behavior adaptation system. Then we show the experimental setup and results. Finally we give the discussions and conclusions.

II. The Behavior Adaptation System

A. Behavior adaptation

Our robot behavior adaptation system monitors subconscious responses from a human partner that indicate discomfort, and changes the robot's behavior in order to minimize these signals. The signals that are monitored are the amount of body repositioning done and the time spent averting gaze

*This research was supported by the National Institute of Information and Communications Technology of Japan

by the human. These values are then used as the reward that is to be minimized by the system. The system consists of the *policy gradient reinforcement learning* (PGRL) algorithm, that minimizes this reward by changing its policy, which in turn determines the robot's behavior. In PGRL, the policy is directly adjusted to minimize the reward. Other reinforcement learning techniques, such as Q-learning, learn an action-value function which the policy will be extracted from. We used a policy that consists of six adapted parameters, three classes of interaction distances and three interaction parameters. The following sections will explain in detail the adapted parameters, the PGRL algorithm, and the reward function.

B. Adapted Parameters

In this experiment, we used six different parameters. These were the interaction distance for each of three classes of interaction (detailed explanation in III-A), the extent to which the robot would meet a human's gaze, waiting time between utterance and action, and the speed at which motions were carried out. We chose these parameters since they seem to have large impacts on interaction and low implementation costs and we could keep the number of parameters small and thereby the dimensionality of the search space at a minimum.

The distances were measured as the horizontal distance between robot and human foreheads. For gaze-meeting, the robot would meet human gaze, and then look away, with a cycle length of 5 seconds, which is the average cycle length in human-human interaction, according to [3]. The gaze-meeting parameter is the proportion of the cycle spent meeting the human subject's gaze. The speed of the robot was divided into two parameters. One controlled how long the robot would wait between utterance and action (e.g. the waiting time between saying "let's shake hands" and reaching out with its right arm), and the other controlled the actual speed of the motion itself.

C. The PGRL Algorithm

The learning algorithm used in this experiment is *policy gradient reinforcement learning* (PGRL), a reinforcement learning method that directly adjusts the policy without calculating action value functions (detailed descriptions can be found in [1] and [5]). Figure 1 shows the algorithm [5] in pseudo code. The $\mathbf{\Theta}$ indicates the current policy which has n parameters. A total of T perturbations of $\mathbf{\Theta}$ are generated, tested with a person, and the reward function is evaluated. Perturbation $\mathbf{\Theta}^t$ of $\mathbf{\Theta}$ is generated by randomly adding ϵ_j, 0, or $-\epsilon_j$ to each element θ_j in $\mathbf{\Theta}$. The step sizes ϵ_j are set independently for each parameter.

When all T pertubations have been run, the gradient $\mathbf{A}$ of the reward function in the parameter space is approximated by calculating the partial derivatives for each parameter. Thus, for each parameter θ_j, the average reward when ϵ_j is added, no change is done and when ϵ_j is subtracted are calculated. The gradient in dimension j is then regarded as 0 if the reward is greatest for the unperturbed parameter, and regarded as the difference between the average rewards for the perturbed parameters otherwise. When the gradient $\mathbf{A}$ has been calculated, it is normalized to overall step size η and for the individual step sizes ϵ in each dimension. The parameter set $\mathbf{\Theta}$ is then adjusted by adding $\mathbf{A}$.

```
$\mathbf{\Theta} = \{\theta_j\} \leftarrow$ Initial parameter set vector of size $n$
  $\mathbf{\Theta^t} = \{\theta_j^t\}$ : Perturbed vector derived from $\mathbf{\Theta}$
$\epsilon \leftarrow$ parameter step size vector of size $n$
$\eta \leftarrow$ overall step size
while (not done)
    for $t = 1$ to $T$
        for $j = 1$ to $n$
            $r \leftarrow$ unbiased random choice
                from $\{-1, 0, 1\}$
            $\theta_j^t \leftarrow \theta_j + \epsilon_j * r$ ,
        Run system using parameter set $\mathbf{\Theta}^t$,
        evaluate reward
    for $j = 1$ to $n$
        $Avg_{+\epsilon,j} \leftarrow$ average reward for all $\mathbf{\Theta}^t$
            with positive perturbation in dimension $j$
        $Avg_{0,j} \leftarrow$ average reward for all $\mathbf{\Theta}^t$
            with zero perturbation in dimension $j$
        $Avg_{-\epsilon,j} \leftarrow$ average reward for all $\mathbf{\Theta}^t$
            with negative perturbation in dimension $j$
        if $(Avg_{0,j} > Avg_{+\epsilon,j})$ AND
            $(Avg_{0,j} > Avg_{-\epsilon,j})$
            $a_j \leftarrow 0$
        else
            $a_j \leftarrow ( Avg_{+\epsilon,j} - Avg_{-\epsilon,j} )$
    $\mathbf{A} \leftarrow \frac{\mathbf{A}}{|\mathbf{A}|} * \eta$
    $a_j \leftarrow a_j * \epsilon_j, \forall j$
    $\mathbf{\Theta} \leftarrow \mathbf{\Theta} + \mathbf{A}$
```

Fig. 1. PGRL Algorithm

This method is guaranteed to converge towards a local optimum given a static reward function (for a more stringent analysis of the algorithm, see [1]). Given a stochastic value function, that possibly evolves over time, PGRL will continuously move towards the vicinity of the current local optimum, but of course, no convergence guarantees can be given. The main advantage over other reinforcement learning methods, is that little knowledge of the dynamic model behind the system is required (as compared to dynamic programming), and that since large tables, e.g. those of Q-learning, need not be calculated, learning should be considerably quicker.

D. Reward Function

The reward function was based on the amount of movement of the subject and the propotion of the time spent gazing directly at the robot in one interaction,

$$R = 0.2 \times (movement[mm]) + 500 \times (gazing\ prop.). \quad (1)$$

Figure 2 shows the block diagram. These two features were chosen as they are typical discomfort signals in interaction, which are easy to measure. When we feel that a conversational

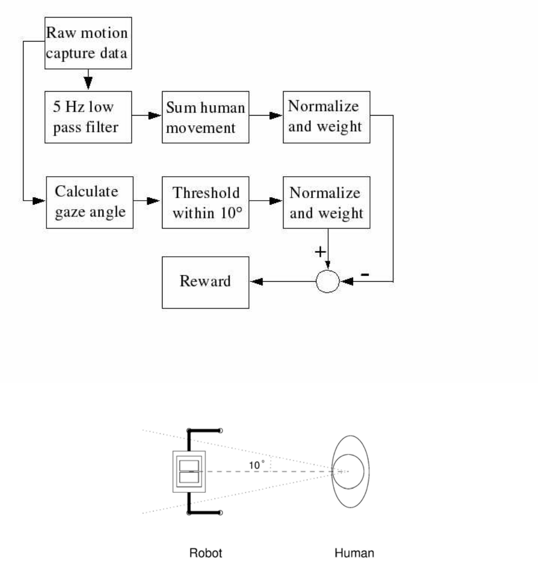

Fig. 2. The reward function and the angular interval determined as human gazing at robot

partner is standing too close for current interaction, we have tendency to move away, and when we are stared at too much, we tend to avert our eyes [9]. We also assume that we avert our eyes from the partner when the interaction is boring or unrecognizable, for example, when a robot's behavior is too slow or too fast. Human body movement analysis [11] also reports that the evaluation from subjects had positive correlation with the length of the gazing time and negative correlation with the distance which the subject moved. The weights that balanced them were chosen so that in a typical case the contributions of the two factors were of equal size. The movement factor was given a negative weight, since it represents a sign of discomfort.

The movement was measured as the translation of the subject's forehead in the horizontal plane. The movement measure was first filtered with a lowpass filter in order to get rid of high frequency noise. This means that only movements on the scale of 5 Hz or slower were considered. The gazing factor was calculated as the percentage of time that the subject's face was turned towards the robot, with an allowance of ±10 degrees (Figure 2). Face direction was chosen over actual eye direction as it is easier to measure accurately, and as it tends to follow eye direction closely in face-to-face communication.

III. The Experiment

A. The environment and the robot

This experiment was conducted in a space approximately 3.5 × 4.5 meters in the middle of a room, the limits being set by the area that can accurately be perceived by the motion capture system used for sensoring. Figure 3 shows the experimental setup. The robot was initially placed in the middle of this area, and the subjects were asked to stand in front of the robot and to interact with it in a relaxed, natural way. Apart from this, and an explanation of the limits of the motion capture system, the subjects were not told to behave or react in any particular way.

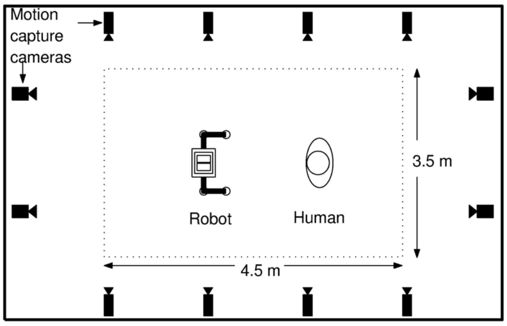

Fig. 3. Experimental setup

For the purpose of this study, we used the humanoid communication robot RobovieII (see Figure 4), developed at ATR [4]. Its size is approximately that of a human child, with a height of 120 cm. It is fitted with a two-wheeled base that uses differential drive for locomotion. The robot has two arms, each with four degrees of freedom, that can be moved in most ways that a human arm can. The head is mounted with three degrees of freedom, and the two camera eyes mounted on it have two degrees of freedom each, enabling the robot to direct its eyes in any frontal direction, making it capable of meeting the gaze of a human subject. The robot's head is fitted with a speaker and microphone, in order to produce and recognize speech.

The software platform used in RobovieII consists of several separate behaviors. Each behavior consists of a single action or set of actions that naturally fit together to form an interaction unit. The behaviors that we used were grouped into three different categories, according to their interaction class: intimate, personal and social (Table I). These are the terms used by Hall to describe the three closest types of interaction distances in his work on proxemics [2]. The sorting into classes was determined by a prestudy in which we exposed 8 subjects to the behaviors, and let the subjects choose what distance they were comfortable with for each of these. In normal human interaction, casual conversation is usually classed as *social*, but our prestudy showed that the subjects preferred a closer distance, equaling that of the touch-based interactions found in the *personal* group. This is mainly due to limitations of the robot's speech capabilities. There are of course variations of preferences for each person within the same class, but these tend to be small.

The measurements for this experiment were done using a motion capture system with 12 cameras. This allowed for acquiring of position data of robot and subject alike in millimeter accuracy. Head and shoulder positions where used to calculate locations of the foreheads of the robot and the

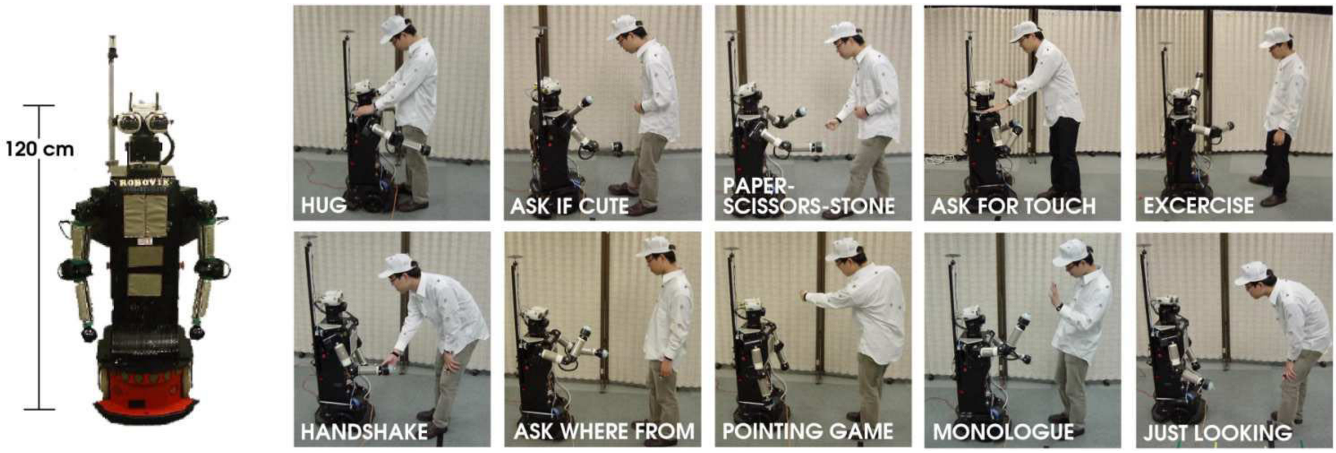

Fig. 4. RobovieII and the behaviors

TABLE I
BEHAVIOR CLASSES

Class	Name
intimate	Hug
personal	Shake hands
personal	Ask where person comes from
personal	Ask if robot is cute
personal	Ask person to touch robot
social	Play paper-scissors-stone
social	Play pointing game
social	Perform arm-swinging exercise
social	Hold “thank you” monologue
social	Look at human without speaking

TABLE II
PARAMETER VALUES AND STEP SIZES

#	Parameter	Initial value	Step size ϵ
1	intimate distance	50 cm	15 cm
2	personal distance	80 cm	15 cm
3	social distance	100 cm	15 cm
4	gazing ratio	0.7	0.1
5	waiting time	0.17 s	0.3 s
6	speed factor	1.0	0.1

subject. We assumed that the averaged gazing direction equals to the direction of the foreheads. Then, we used the direction of the foreheads as the gazing direction. The motion capture data were forwarded to the robot’s onboard computer via TCP/IP, resulting in data lags of at most 0.1 seconds in this experiment, ensuring sufficient response speed.

B. Experimental procedure

The interaction experiment lasted for 30 minutes, during which the robot executed the behaviors at random. It repeatedly proposed different typical Japanese children’s games, asking questions or demanding to be touched, hugged, or shake hands. This was accompanied by motions of the robot’s arms and head. For example, when the robot demands to be hugged, it reaches out with open arms, and closes the arms in an embracing motion if the subject is in a position within reasonable distance in front of the robot. During this time, the adaptation system was running on the robot in real-time, adapting to the subject’s reactions. The initial values and the search step sizes used for the different parameters can be seen in Table II.

The reward function was calculated once per executed action of the robot, or roughly once every ten seconds. A total of ten different parameter combinations were tried before the gradient was calculated and the parameter values upgraded. This means that an iteration of the algorithm took 1 minute and 40 seconds on an average.

After this session, the subjects were asked their impression of the robot’s movements and general behavior. More detailed measurements followed, in which the subjects were asked to stand in front of the robot, at the distance they felt was the most comfortable for a representative action for each of the three distances studied: intimate, personal and social. They were also asked to indicate how close the robot could come without the interaction becoming uncomfortable or awkward, as well as how far away the robot could be without disrupting the interaction. These distances were measured using the same motion capture system.

Furthermore, each subject was shown the robot’s behavior performed in turn with three different values - one low, one average and one high - for each of the remaining parameters, gaze-meeting, timing and motion execution speed. The remaining parameters were at this time set to fixed values. The subjects were asked to indicate which of the three shown behaviors they felt comfortable with. A few subjects indicated several values for a single parameter, and some indicated preferences between or outside the shown values.

A total of 15 subjects were used in this experiment: all except one were japanese, and all understood the spoken utterances of the robot. The subjects were of ages 20-35, most in the range 20-25. Six of the subjects were female. All subjects were employees or interns at ATR, meaning that they

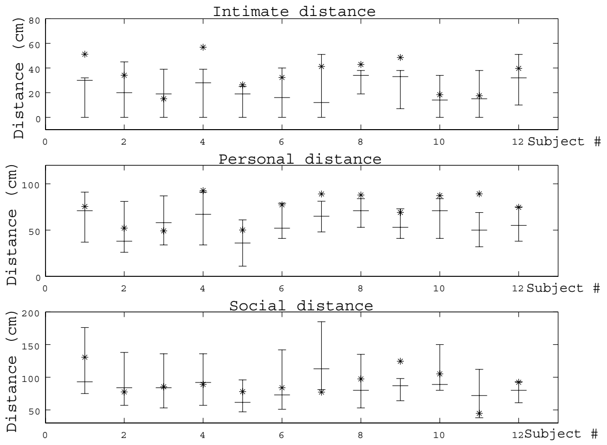

Fig. 5. Learned distance parameters and preferences for 12 subjects. Asterisks show the learned parameters. The shorter bars show the interval for acceptable distance and longer bars show the preferred value.

had a certain familiarity with RobovieII. Three of the subjects did not react in the way we had anticipated. They neither averted gaze nor shifted position however inappropriate the robot's behavior got, but showed their discomfort in words and facial expression. The system did not have any success in adapting to these subjects, but as these reaction types do not fit our model of friendly interaction with the robot, we do not include these results in our evaluation.

IV. Experimental Results

A. The results

For most of our subjects, the parameters reached reasonable convergence to stated preferences within 15-20 minutes, or approximately 10 iterations of the PGRL algorithm.

In Figure 5 we show the learned values for the distances as compared to the stated preferences for 12 subjects. The learned distance is here calculated as the average parameter value during the last fourth (about 7.5 minutes) of each experiment run, since the algorithm keeps searching the optimum value. The bars show the interval for acceptable distance and the preferred value, and the asterisks are the learned values. Figure 6 shows the remaining three parameters, where circles show what values the subjects indicated as preferred. Some subjects indicated a preference in-between two values, these cases are indicated with a triangle showing that preferred value. The asterisks again show the learned values as the mean values for the last quarter of the experiment runs.

As can be seen, there is a large difference in success rate between different parameters. This is due to the fact that all parameters are not equally important for successful interaction. It is a typical trait for PGRL that parameters that have a larger impact on the reward function are adjusted faster, while parameters that have a lesser impact will be adjusted at a slower rate.

When we compared the results of the learning with the comments given by our subjects in connection to the experiments, we found that some of the subjects to whose preferences the robot made the best adaptation were also the ones on which the robot's adaptation made the least impression. We see this as a well-adapted behavior being considered as natural by the subjects.

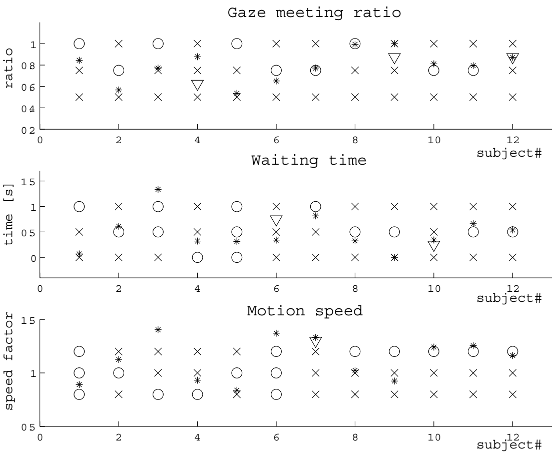

Fig. 6. Learned gaze and speed parameters and indicated preferences for 12 subjects. Circles show what parameter settings the subject indicated as preferable, 'x's show the non-indicated settings. In the cases where values outside the given settings were indicated, a triangle shows the preferred value.

TABLE III

Average deviation from preferred value (normalized to stepsize units)

parameter	average deviation	initial deviation
intimate	0.9	1.8
personal	1.3	1.6
social	1.3	1.2
gaze	1.0	2.4
wait	0.8	1.5
speed	1.1	1.4

In order to get an overview of the general performance, we calculated the average deviation for each parameter over all subjects (Table III). The rightmost column of this table shows how much the initial values deviated from the stated preference, as a reference. All values have been normalized for step size.

Most parameters converged to within one step size, the exceptions being the personal and social distance parameters. It should be noted, that for these parameters, the average stated tolerance (the difference between the closest comfortable distance and the farthest) was of a size corresponding to several step sizes. For example, for personal distance, the average stated tolerance was 3.02 step sizes and for social distance it was 5.01. As Figure 5 shows, for all subjects but one, the learned social distance parameter values falls within the stated acceptable interval.

That the parameters converge to a value within one step size of the learning algorithm is an indication of successful learning. In order to achieve better results, it would be necessary to decrease the step size, either as a fixed value, or

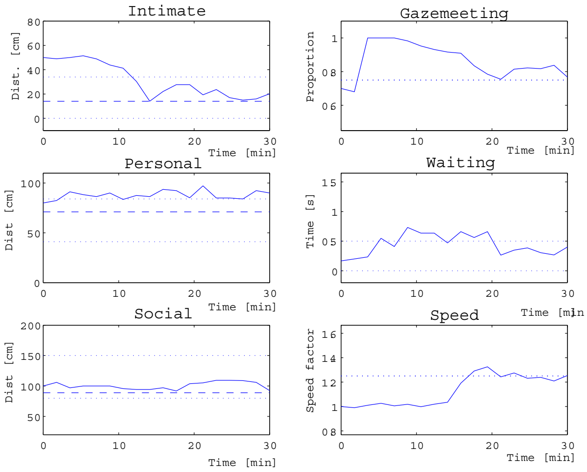

Fig. 7. Results achieved for subject 10. The dotted lines represent the measured preferences of the subject. The dashed lines represent the most preferred values.

to let the step size become smaller as the parameters start to converge.

As can be seen in Table III, the initial values are not far (in step size units) from the preferred values. That the system takes in the order of 10 iterations or more to converge is due to the stochastic behavior of the human subjects. As can be seen in Figures 5 and 6, for some subjects some values did not converge to the desired values for the entire duration of the experimental runs. If the subjects always reacted in an unambiguous and consistent manner, convergence should theoretically be reached in 3 to 4 iterations.

In the following, we show more details of how the system behaved for different groups of subjects. We have divided the results into five groups, successful, partial success with content subjects, successful but discontent subjects, partially successful, and unsuccessful runs.

B. *Successful experimental runs*

There were three subjects for whom the system performed very well. Not only were the subjects themselves content with the performance, but all parameters had a good convergence to their stated preferences. Common for all of them was a tendency to be very interested in interaction with the robot, and they had a very positive interaction pattern, much as when interacting with another human.

The tenth subject (Figure 7) was impressed by the robot's behavior and said that it quickly became much better. The plots support this, as all parameters are adjusted to well within stated preference, except *personal* distance, which is but slightly farther. This subject stated a preference for an interval of *waiting* times, hence the two lines in the plot showing the borders of this interval.

C. *Partial success - content subject*

The next group consists of two subjects that were content with the robot's behavior, even though analysis of the results show that some parameters were far from stated preference.

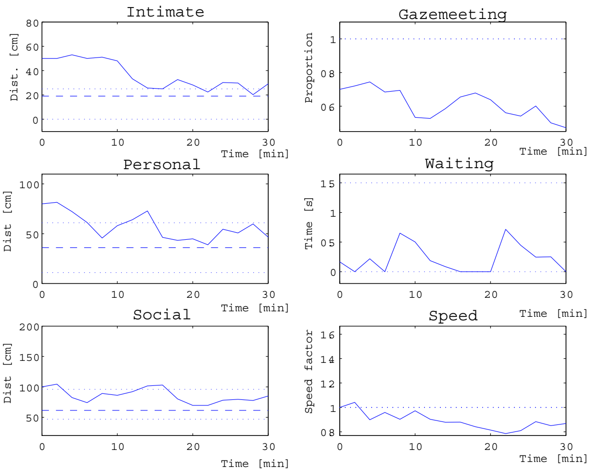

Fig. 8. Results achieved for subject 5. The lines represent the measured preferences of the subject as in Figure 7.

The experimental run for the fifth subject (Figure 8) resulted in good adaptation for the distance parameters, but less successful adaptation for the remaining parameters. This subject stated that all shown values for *waiting* were equally good, so this parameter cannot be evaluated in this case.

Interestingly though, this subject stated content with the *gaze-meeting* results, even though it is obvious from the plots that these were far from his stated preference. He was also satisfied with the *speed* parameter, which is as much as 20% off from specified preference. It is possible that this discrepancy can be explained by different conditions during the experimental run and when measuring the preferences afterward, or simply the fact that the subject actually accepted a fairly wide range of parameter values.

This subject showed a slightly different behavior pattern than the others. He preferred touching the robot even during normal speech interaction and robot monologues. This resulted in a shorter *social* distance than the other subjects, but the system did not seem to have any problems to adapt to this unexpected behavior.

D. *Successful run - discontent subject*

One subject, the seventh (Figure 9), was discontent with the robot even though the values seem to converge to her stated preference. She described her first impression of the robot's behavior as "tentative", but that it became more active as time passed. She also stated that she thought that it tended to get too close, which is a bit surprising when actual distances are compared to preference in the plots. A guess as to why this is so is that during the experiment the robot raised its movement speed in accordance with the subject's preference, while the preference measurements were conducted at default speed. The subject might actually prefer a farther distance when the speed is higher.

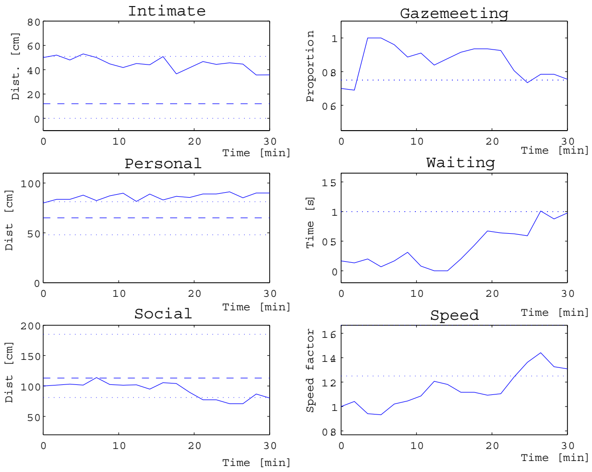

Fig. 9. Results achieved for subject 7. The lines represent the measured preferences of the subject as in Figure 7.

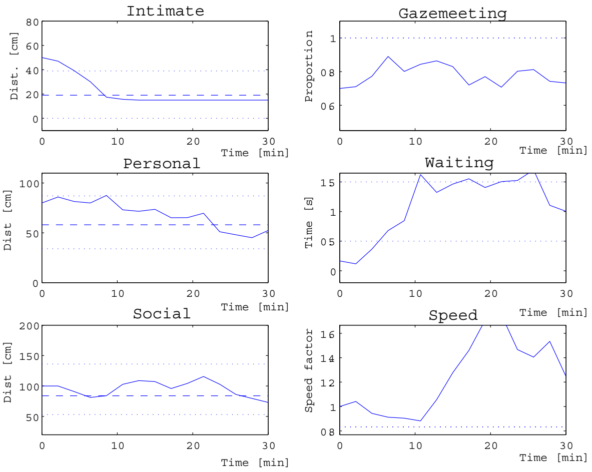

Fig. 11. Results achieved for subject 3. The lines represent the measured preferences of the subject as in Figure 7.

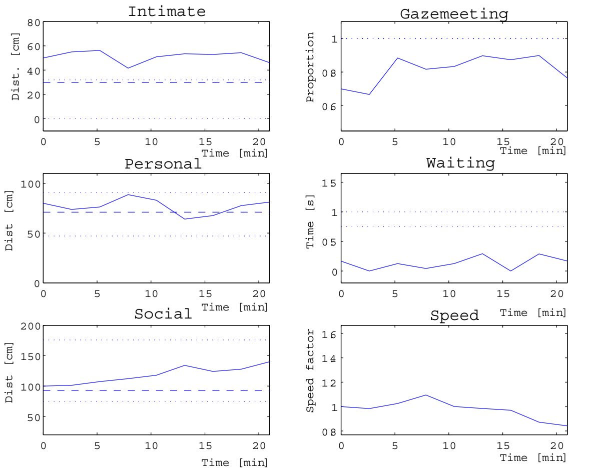

Fig. 10. Results achieved for subject 1.

Most parameters can be said to eventually converge acceptably, though the adaptation is slow, taking approximately 25 minutes, which accounts for the subjects discontent.

E. Partially successful runs

There were five subjects for which the system only performed partially well. These subjects were content with the aspects that worked, and discontent with the ones that did not.

The experiment with the first subject (Figure 10) was aborted after 21 minutes due to technical problems, and as such might be less significant than then the other runs that lasted the entire planned 30 minutes.

The distances for *personal* and *social* interaction stay well within the stated preferences, whereas the distance for *intimate* interaction never enters the stated preference interval. Since all distances tested in this case are outside the preferred interval, the algorithm gains little or no information on the reward gradient for this parameter. The subject stated discontent with the *intimate* distance.

The *gaze-meeting* parameter stays around 80%, which should be deemed as acceptable as the subject indicated 100% as being preferable to 50% or 75%. The *waiting* parameter is however fairly far from the indicated preference. This particular subject did not show much of an interest for speed and timing issues, and could not indicate any speed as being preferable to any other.

The third subject (Figure 11) stated a discontent with the *personal* distance for the first half of the experiment, which correlates well with the plot, as the value is initially at the outer limit of what was indicated as acceptable. She was also content with the *gaze-meeting*, even though the actual values achieved were closer to 75% than her specified preference of 100%.

The plot of the *intimate* distance shows the lower limit of 15 cm that the system has for safety reasons. This subject stated a fairly wide preference interval for the *waiting* parameter, thus making the results for this more difficult to evaluate.

The only other remarkable result of this experimental run is that the speed parameter is far away from the stated preference, something the subject also complained about. Observations of the actual experiment showed that as the robot increased its movement speed, the subject seemed to watch the movements carefully and fix her gaze at it. This is a weakness of the reward function.

F. Unsuccessful run

The results attained for the ninth subject (Figure 12) are not very good. As can be seen, apart from the *personal* distance and *gazing* parameters, the results are far from stated preference. There were no observable problems with this subject's behavior, so the reason for these poor results are

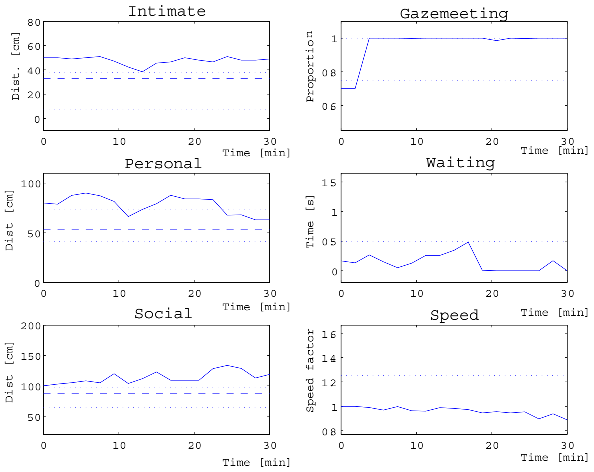

Fig. 12. Results achieved for subject 9. The lines represent the measured preferences of the subject as in Figure 7.

still unclear. It is possible that this subject was not clear in showing his dislike when the robot behaved in an unwanted manner. It is also noteworthy that the subject said that he felt as if the robot didn't like him, but forced itself to be somewhat polite and talk to him anyway.

V. Discussions and conclusions

We have shown that the robot adapted to individual preferences for the most of the subjects in the experiment. Here we conclude that a robot can adapt its behavior parameters to individuals by policy gradient reinforcement learning. We also note that the comments by the subjects suggested that well-adapted robot behavior is perceived as natural. This is an important step towards making robots that are as easy to interact with as humans.

We found several issues to be solved however. First, it is very difficult to measure true preferences since the parameters are interdependent as we stated earlier. The fifth subject was content with the learned parameters even though they are far from the stated preferences. On the contrary, the seventh subject claimed that the robot got too close though the distance were near to her stated preference. Although this issue may not influence the adaptation itself, it makes it difficult to evaluate how well the system works for a person.

Second, the method could not find the gradient and cannot find the direction to the local optimum for some parameters of some subjects. The reason is that the behaviors of the subject did not show any difference to the small perturbed values if the current parameter is too far from the preferred values.

Third, there were subjects whose behaviors were different from our expectation. The third subject had tendency to fix her gaze to the robot when the motion speed was higher than her preference. The ninth subject did not show his preference in his behaviors. Then the robot could not have any preference measure from his movements. We need different reward functions for people who have different reactions.

To overcome the second and third issues, grouping of the people will be a solution. By grouping, it will be possible to start the adaptation from parameters nearer to optimum ones, adapt faster, and select the appropriate reward function for the person. Of course, we have many open questions. How many and what kind of reward functions are needed? How does the robot know what reward function is appropriate for a person? How many groups should there be? How does the robot know a person is in a certain group? How many parameters does the robot have to adapt for humans? It will also require a way to identify individuals, as the system needs to switch between different parameter sets as different people approach the robot.

For other future work, it would be interesting to take this experiment out of the laboratory and give it a field try in a public setting. In order to do this, all measurements, including the face direction detection, will have to be done with the robot's onboard systems, which have a lower accuracy than the motion capture system presently used. This may require a new reward function. In the public setting, faster adaptation and adaptation in multiple interactions of shorter period will be required. A small-scale prestudy indicates that our subjects tend to keep the same preferences over time. This supports the possibility of dividing the learning process over multiple sessions. This would remove the need for lengthy interaction sessions in this experiment, and enable the robot to "get to know" people over time.

References

[1] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[2] E. T. Hall. *The Hidden Dimension*. DoubleDay Publishing, 1966.

[3] S. Duncan jr. and D. W. Fiske. *Face-to-Face Interaction: Research, Methods, and Theory*. Lawrence Erlbaum Associates, Inc., Publishers, 1977.

[4] T. Kanda, H. Ishiguro, M. Imai, T. Ono, and R. Nakatsu. Development and evaluation of an interactive humanoid robot - "Robovie". In *Proc. of International Conference on Robotics and Automation*, pp. 1848–1855. IEEE, 2002.

[5] N. Kohl and P. Stone. Policy gradient reinforcement learning for fast quadrupedal locomotion. In *Proc. of International Conference on Robotics and Automation*, vol. 3, pp. 2619–2624. IEEE, 2004.

[6] T. Inamura, M. Inaba, and H. Inoue, Acquisition of Probabilistic Behavior Decision Model based on the Interactive Teaching Method. In *Proc. of the 1999 International Conference on Advanced Robotics*, 523–528, 1999.

[7] K. Nakajima. *Research regarding personal distances between people and a moving robot (in Japanese)*. PhD thesis, Kyuushuu Institute of Design, 1998.

[8] Y. Nakauchi and R. Simmons. A social robot that stands in line. *Autonomous Robots*, 12:313–324, 2002.

[9] E. Sundstrom and I. Altman. Interpersonal relationships and personal space: Research review and theoretical model. *Human Ecology*, 4(1), 1976.

[10] T. Tasaki, S. Matsumoto, K. Komatani, T. Ogata, H. G. Okuno. Dynamic communication of humanoid robot with multiple people based on interaction distance. In *Proc. of International Workshop on Robot and Human Interaction (Ro-Man-2004)*, pp. 81-86, IEEE, 2004.

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