Adapting Robot Behavior for Human–Robot Interaction

Noriaki Mitsunaga, Christian Smith, Takayuki Kanda,
Hiroshi Ishiguro, and Norihiro Hagita

***Abstract*—Human beings subconsciously adapt their behaviors to a communication partner in order to make interactions run smoothly. In human–robot interactions, not only the human but also the robot is expected to adapt to its partner. Thus, to facilitate human–robot interactions, a robot should be able to read subconscious comfort and discomfort signals from humans and adjust its behavior accordingly, just like a human would. However, most previous research works expected the human to consciously give feedback, which might interfere with the aim of interaction. We propose an adaptation mechanism based on reinforcement learning that reads subconscious body signals from a human partner, and uses this information to adjust interaction distances, gaze meeting, and motion speed and timing in human–robot interactions. The mechanism uses gazing at the robot's face and human movement distance as subconscious body signals that indicate a human's comfort and discomfort. A pilot study with a humanoid robot that has ten interaction behaviors has been conducted. The study result of 12 subjects suggests that the proposed mechanism enables autonomous adaptation to individual preferences. Also, detailed discussion and conclusions are presented.**

***Index Terms*—Behavior adaptation, human–robot interactions, policy gradient reinforcement learning (PGRL), proxemics.**

I. Introduction

When humans interact in a social context, there are many factors that are adjusted in order to make communication smooth. Previous studies in behavioral sciences have shown that there is a need for a certain amount of personal space [1] and that different people tend to meet the gaze of others to different extents [2]. For example, when a conversational partner stands too close, we tend to move away, and when we are stared at, we tend to avert our eyes [3].

As Reeves and Nass [4] point out, humans tend to subconsciously treat nonpersonified objects such as computers and televisions like they would treat other humans. When observing human–robot interactions, we notice that most people show the same behaviors when interacting with a robot as they would when interacting with a human. Therefore, we believe that it would be natural for people to expect the same type of adaptation to one another from robots as they are used to in human–human interactions.

Several behavior adaptation systems for human–robot and human–agent interactions have been proposed. Inamura *et al*. [5] have proposed incremental learning of decision-making rules for a mobile robot that moves in a corridor through the interaction with a human. The system learns the user's preferences regarding the robot's behavior by reflecting the user's bias in his/her commands. Isbell *et al*. [6] have implemented a virtual agent that chats with people over a computer network. The agent learns how to behave according to the conversations of other participants based on reinforcement learning. Participants have to consciously give rewards to the agent by pressing appropriate buttons. However, in human–human interactions, people subconsciously adapt to each other, meaning it is difficult for people to consciously provide feedback. Thus, an adaptation method, which reads subconscious responses from the human, is required for smooth human–robot communication.

Meanwhile, there have been a few studies about personal space and robot's behaviors [7]–[9], but so far, none of them have addressed the problem of adapting these factors to individual preferences.

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 [3]. Where human–robot interactions 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 to be, i.e., the robot's movement speed [7]. This means that a system that adapts to personal preferences has to consider several parameters simultaneously.

In this paper, we propose a behavior adaptation system based on *policy gradient reinforcement learning* (PGRL). Using comfort and discomfort signals from the human partner as input for the reward function, it simultaneously searches for the behavioral parameters that maximize the reward, thereby also maximizing and minimizing, respectively, the actual comfort and discomfort experienced by the human. We use a reward function that consists of the human's movement distance and gazing period in human–robot communication [10]. The system adapts six behavioral parameters: three parameters that determine interaction distance/personal space [1] and one parameter each to determine the period for which the robot looks at the human's face, the delay after which the robot starts a gesture after an utterance, and the speed of the gestures.

In the following, we first explain the proposed behavior adaptation system. Then we show the setup of the pilot study and results. Finally, we present our discussion and conclusions.

II. Behavior Adaptation System

A. Adapted Parameters

We adopted six parameters to be adapted by the system. These were three interaction distances (*intimate*, *personal*, and *social* distances) for three classes of proxemics zones, the extent to which the robot would meet a human's gaze (*gaze-meeting-ratio*), waiting time between utterance and gesture (*waiting-time*), and the speed at which gestures were carried out (*motion-speed*). We chose these since they seem to have a strong impact on interaction and low implementation costs, allowing us to keep the number of parameters small, and thereby, the dimensionality of the search space.

B. Reward Function

The reward function is based on the *movement* distance of the human and the proportion of time spent *gazing* directly at the robot in one interaction. An analysis of human body movement [10] in human–robot interactions reports that the evaluation from subjects had a positive correlation with the length of the *gazing* time and a negative correlation with the distance that the subject moved.

Manuscript received July 2, 2007; revised February 20, 2008. This paper was recommended for publication by Associate Editor C. Laschi and Editor H. Arai upon evaluation of the reviewers' comments. This work was supported by the Ministry of Internal Affairs and Communications of Japan.

N. Mitsunaga is with the Department of Robotics, Kanazawa Institute of Technology, Ishikawa 921-8501, Japan, and also with the Advanced Telecommunications Research (ATR) Intelligent Robotics and Communication Laboratories, Kyoto 619-0288, Japan (e-mail: mitunaga@neptune.kanazawa-it.ac.jp).

C. Smith is with the School of Computer Science and Communication, Royal Institute of Technology, Stockholm SE-100 44, Sweden (e-mail: ccs@nada.kth.se).

T. Kanda and N. Hagita are with the Advanced Telecommunications Research (ATR) Intelligent Robotics and Communication Laboratories, Kyoto 619-0288, Japan (e-mail: kanda@atr.jp; hagita@atr.jp).

H. Ishiguro is with the Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan and also with the Advanced Telecommunications Research (ATR) Intelligent Robotics and Communication Laboratories, Kyoto 619-0288, Japan (e-mail: ishiguro@ams.eng.osaka-u.ac.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TRO.2008.926867

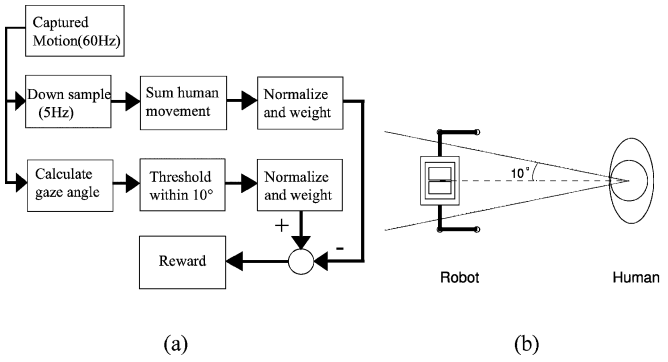


Fig. 1. (a) Block diagram to calculate the reward function. The 3-D-position data captured by a motion capture system at 60 Hz are down-sampled to a 5 Hz sampling rate. They are used to calculate the human *movement* distance and *gazing* period. They are then normalized, weighted, and summed. (b) Angular interval determined as human *gazing* at the robot is $\pm 10^\circ$.

Fig. 1(a) shows a block diagram of reward calculation. The foreheads' positions and directions of the human and the robot are measured by a 3-D motion capture system at a sampling frequency of 60 Hz. They are then projected onto the horizontal plane and downsampled to 5 Hz. The humans' *movement* distances are the sum of the distances that he/she moved in all sampling periods (200 ms) of a behavior. The *gazing* factor was calculated as the percentage of time that the subject's face was turned toward the robot in the interaction behavior, with an allowance of $\pm 10^\circ$ [Fig. 1(b)] in horizontal direction. The reward function R is defined as

$$R = -0.2 \times (\text{movement distance (millimeters)}) + 500 \times \frac{(\text{time human spent looking at robot})}{(\text{time spent for the interaction behavior})}.$$

The weights in this equation were determined with a prestudy.[1] The contributions of the two factors were of equal size as a result. Note that this process did not require severe tuning.

C. PGRL Algorithm

In general, there are two possible learning approaches that make use of a reward signal but do not need teaching signals. These are genetic algorithms and reinforcement learning. The problem at hand requires that the system runs real-time but does not perform unacceptably during the learning phase. This rules out the genetic algorithms and ordinal reinforcement learning methods, such as Q-learning [11]. Kohl and Stone [12] suggest hill climbing and PGRL as such. PGRL is a reinforcement learning method that directly adjusts the policy without calculating action value functions [13], [14].

The main advantage over other reinforcement learning methods is that the learning space could be considerably reduced by using human knowledge when we prepare the policy function. Meanwhile, it is difficult to measure how small the search space could potentially be since it highly depends on the design of the learning system.

Fig. 2 shows the algorithm we adopted [15]. The variable $\mathbf{\Theta}$ indicates the current policy or the values of n behavioral 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.

[1]We recorded the human's and the robot's movements, and measured parameters that the human preferred. We then ran the PGRL algorithm with different weights using the recorded values and tuned weights so that the behavioral parameters quickly and stably converged to the preferred values.

```
1  Θ ← Initial parameter set vector of size n
2  ε ← parameter step size vector of size n
3  η ← overall step size
4  while (not done)
5      for t = 1 to T
6          for j = 1 to n
7              r ← unbiased random choice
                   from {−1, 0, 1}
8              θ_j^t ← θ_j + ε_j * r , where Θ^t is
           perturbed parameter set of same size as Θ
9      for t = 1 to T
10         Run system using parameter set Θ^t,
           evaluate rewards
11     for j = 1 to n
12         Avg_{+ε,j} ← average reward for all Θ^t
               with positive perturbation in dimension j
13         Avg_{0,j} ← average reward for all Θ^t
               with zero perturbation in dimension j
14         Avg_{−ε,j} ← average reward for all Θ^t
               with negative perturbation in dimension j
15         if (Avg_{0,j} > Avg_{+ε,j}) AND
               (Avg_{0,j} > Avg_{−ε,j})
16             a_j ← 0
17         else
18             a_j ← ( Avg_{+ε,j} − Avg_{−ε,j} )
19     A ← A/|A| * η
20     a_j ← a_j * ε_j, ∀ j
21     Θ ← Θ + A
```

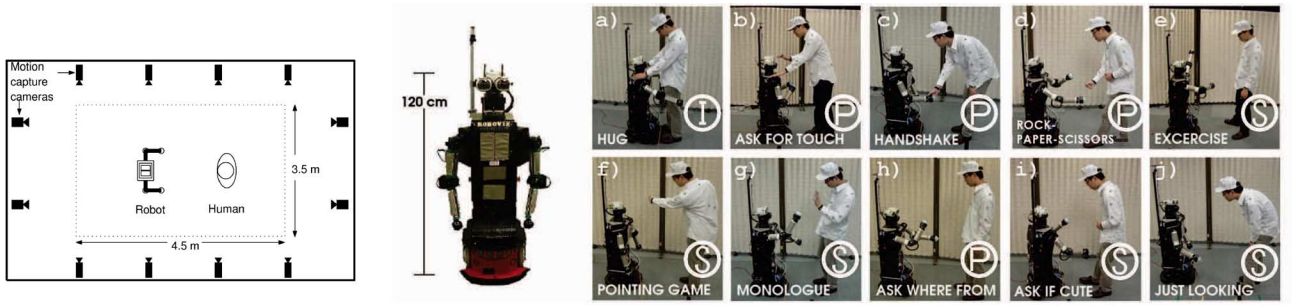
Fig. 2. This is the PGRL algorithm that we adopted for the adaptation system.

The robot tests each policy $\mathbf{\Theta}^t$ with an interaction behavior and then receives the reward. Note that the interaction behavior can be different for each test since we assume that the reward is not dependent on the behaviors but on the policy only. When all T perturbations 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 performed, and cases 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 is considered to be 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 $\mathbf{A}$.

III. Pilot Study

A. Environment and the Robot

The study was conducted in a room equipped with a 3-D motion capture system comprising 12 cameras [Fig. 3(a)]. The system can capture 3-D positions of markers attached to the human and the robot at a sampling rate of 60 Hz. We used a space measuring 3.5 m $\times$ 4.5 m in the middle of the room, the limits were set by the area that



(a) (b)

Fig. 3. (a) Pilot study was conducted in a room equipped with a 3-D motion capture system comprising 12 cameras. The area used for the study was the area measuring 3.5 m × 4.5 m in the middle of the room.(b) Robovie II and ten interaction behaviors used for the study. Robovie II's height is about 1.2 m. Ten interaction behaviors are: (a) ask to hug it (*hug*); (b) ask to touch it (*ask for touch*); (c) ask for a handshake (*handshake*); (d) rock-paper-scissors; (e) exercise; (f) pointing game; (g) say "Thank you. Bye-bye. See you again" (*monologue*); (h) ask where the human comes from (*ask where from*); (i) ask if it is cute (*ask if cute*); and (j) just look at the human (*just looking*). The I, P, and S indicate intimate, personal, and social distance classes of the behaviors, respectively.

can be perceived by the system at millimeter-level accuracy. The data from the system were forwarded to the robot via Ethernet, resulting in data lags of at most 0.1 s, ensuring sufficient response speed.

B. Interaction Behaviors of the Robot

Fig. 3(b) shows the Robovie II [16] and ten interaction behaviors used in the study. Each behavior took about 10 s to run. The robot always initiates the interaction like a child who asks to play since its concept is a child-like robot.

During the interaction, the robot tries to keep the interaction *distance* of the category to which the behavior belongs. The *distance* was measured as the horizontal distance between robot's and human's foreheads. We classified interaction behaviors into Hall's three interaction categories [1]: *intimate* (0–0.45 m); *personal* (0.45–1.2 m); and *social* (1.2–3.6 m) by another prestudy[2] as in Fig. 3(b).

It also meets the human's gaze in a cyclic manner, where the robot meets and averts the human's gaze in cycles that last 0–10 s (randomly determined, average 5 s), as this is the average cycle length for gaze meeting and averting in human–human interactions [2]. The parameter *gaze-meeting-ratio* is the portion of each cycle spent meeting the human subject's gaze.

The *waiting-time* controlled how long the robot would wait between utterance and action. When it performs behaviors from (a) *hug* to (g) *monologue* that require motion on the part of the human, the robot starts actions after it makes an utterance (like "Please hug me," "Let's play rock–paper–scissors," etc.) and *waiting-time* has passed.

The *motion-speed* controlled the speed of the motion. If *motion-speed* is 1.0, the motion is carried out at the same speed as the gesture is designed to do. As for *gaze-meeting-ratio*, *waiting-time*, and *motion-speed*, the same values are used for all interaction behaviors.

C. The Study

1) Subjects: A total of 15 subjects (nine males and six females) were used in this study. The subjects were of ages 20–35. All subjects were employees or interns of our laboratory. However, they were not familiar with the setup of the study, and most had no prior experience of the type of interaction used in the study. None of them had taken part in the prestudy.

[2]We exposed eight subjects to the behaviors, and let them choose what distance they were comfortable with for each of these. The robot did not move its wheels in the prestudy. There are of course variations of preferences for each person within the same class, but these were so small to change categories. Note that in normal human interaction, casual conversation is usually classed as "*social*," but the subjects preferred closer distances for behaviors (h) and (i), equaling that of the touch-based interactions found in the *personal* group. This is mainly due to limitations of the robot's speech capabilities.

TABLE I
INITIAL VALUES AND STEP SIZES OF BEHAVIORAL PARAMETERS

#	Parameter	Initial value	Step size ϵ_j
1	intimate distance	0.50 m *	0.15 m
2	personal distance	0.80 m	0.15 m
3	social distance	1.0 m	0.15 m
4	gaze-meeting-ratio	0.7	0.1
5	waiting-time	0.17 s	0.3 s
6	motion-speed	1.0	0.1

* We used 0.50 m distance although it is outside of the intimate range.

2) Interaction: The robot was initially placed in the middle of the measurement area, and the subject was asked to stand in front of the robot and interact with it in a relaxed, natural way. Apart from this, and an explanation of the limits of the motion capture system, the subject was not told to behave or react in any particular way.

The robot randomly selected one of the ten interaction behaviors. After one behavior finished, it randomly selected the next one. The interaction lasted for 30 min. Except for controlling the selection not to repeat the same behavior twice in a row, we did not pay any special attention to the randomness of the selection, such as whether to evenly select behaviors or behaviors in the same distance classes.

3) Adaptation: During the interaction, the adaptation system was running on the robot in real time. Table I shows the initial values and the search step sizes. The initial values were set slightly higher than the parameters that the subjects in the prestudy preferred.

For the duration of each interaction behavior, or the test of a policy, the robot kept the interaction distance and other parameters according to Θ^t. The reward function was calculated for each executed interaction of the robot using the accumulated motion and gaze meeting percentage for the duration of the behavior. The duration starts from just after the robot selected the behavior or before it utters and it ends at the end of the behavior or just before the next behavior selection. A total of ten different parameter combinations were tried before the gradient was calculated and the parameter values updated ($T = 10$). The subject did not notice the update of the policy during interaction since the calculation was done instantaneously.

4) Preferences Measurement (Interaction Distances): The subject was asked to stand in front of the robot, at the distance he/she felt was the most comfortable for a representative action for each of the three distances studied by using behaviors (a) *hug*, (c) *handshake*,

and (g) *monologue*, respectively. Other parameters, *gaze-meeting-ratio*, *waiting-time*, and *motion-speed* were fixed to 0.75, 0.3 s, and 1.0, respectively.We also asked the subject to indicate acceptable limits, i.e., 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.

5) Preferences Measurement (Gazing, Waiting, and Speed): Each subject was shown the robot's behavior performed in turn with three different values—low, moderate, and high—for each of the parameters, *gaze-meeting-ratio*, *waiting-time*, and "*motion-speed*." The parameters that were not measured at this time were fixed to *moderate* values (same for all subjects). 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. We recorded as such if he/she said so.

The *moderate* value for gazing, 0.75, was based on the average preference in the prestudy, the *high* value was set to continuous gazing at 1.0, and the *low* value was set equidistantly from the moderate value at 0.5. For motion speed, the preprogrammed motions were assumed to be at a moderate speed, and the *high* and *low* values were set to be noticably faster and slower, respectively, than this value. A time of 0 s was chosen as the *low* value for waiting, the *moderate* value of 0.5 s was chosen to be a noticable pause, and the *high* value of 1 s was chosen to be noticably longer.

We used (g) *monologue* to measure *gaze-meeting-ratio* and *motion-speed*. The interaction distance was fixed to 1.0 m. We used (a) *hug* to measure *waiting-time* and asked the subject to stand at a comfortable distance since the *intimate* varied among the subjects. Other parameters were fixed to the averaged values that the subjects in the prestudy preferred.

6) Impression: We interviewed the subjects for their impressions of the robot's movements, interaction distances, gaze meeting, and general behavior and their changes. We also encouraged them to freely offer other impressions.

IV. Study Results

A. Adaptation Results

For most of the subjects, at least part of the parameters reached reasonable convergence to stated preferences within 15–20 min, or approximately ten iterations of the PGRL algorithm. We have excluded the results of three subjects who neither averted their gaze nor shifted position however inappropriate the robot's behavior became, but showed their discomfort in words and facial expression to the experimenter. These study runs had to be aborted early as safe interaction could not be guaranteed, so no usable data have been collected from them.

Fig. 4 shows the learned values for the distances as compared to the stated preferences for 12 subjects excluding three aforementioned subjects. The intimate distance converged in the acceptable range for 8 out of 12 subjects. The personal and social distances converged in acceptable range for seven and ten subjects, respectively. The learned distance is calculated here as the average parameter value during the last quarter (about 7.5 min) of each study run since the algorithm keeps searching for the optimum value. The bars show the interval for acceptable distance and the preferred value, and the asterisks (*) are the learned values. Fig. 5 shows the remaining three parameters, where circles (◯) indicate what values the subjects indicated as preferred. Some subjects indicated a preference in between two values, and these cases are denoted with a triangle (▽) showing that preferred value. The asterisks again show the learned values as the mean values for the last quarter of the study runs. The *gaze-meeting-ratio*, *waiting-time*, *motion-speed* converged to the values near to selected values for 7, 6, and 5 out of 12 subjects, respectively.

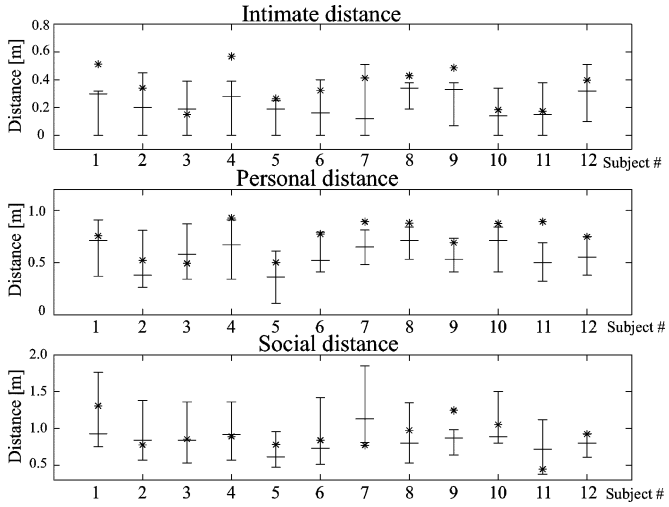


Fig. 4. Learned distance parameters and preferences for 12 subjects. Asterisks (*) show the learned parameters. The shorter bars denote the interval for acceptable distances and longer bars represent the preferred values.

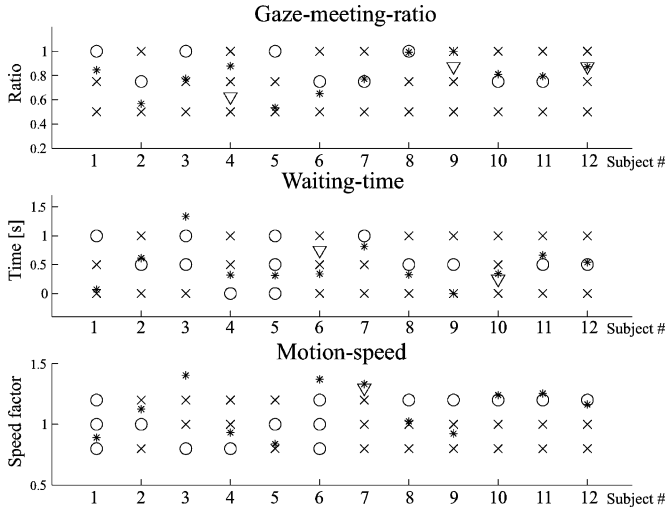


Fig. 5. Learned *gaze-meeting-ratio*, *waiting-time*, and *motion-speed*, and indicated preferences of them for 12 subjects. Circles (◯) show what parameter settings the subject indicated as preferable, "x"s denote the settings shown to the subjects but not selected. In the cases where subjects preferred values other than given settings, a triangle (▽) indicates the preferred value. If the subject said he/she preferred the value between two of the shown values, ▽ is marked at the middle of them. If the subject stated all values were too slow (subject 7), we marked ▽ above the highest shown value.

As can be seen, most of the parameters converged in the acceptable ranges; however, there are large differences in the success rate between different parameters. This is due to the fact that not all parameters are equally important for successful interaction. It is a typical trait for PGRL that parameters having a greater impact on the reward function are adjusted faster, while those with a lesser impact will be adjusted at a slower rate.

Table II shows the average deviation for each parameter over all subjects at the initial and during the last quarter of the study runs. All values have been normalized for step size ϵ_j. Most parameters converged to within one step size, the exceptions being the *personal* and

TABLE II
AVERAGE DEVIATIONS FROM PREFERRED VALUES (NORMALIZED TO STEP SIZE UNITS)

Parameter	Initial deviation	Average deviation (learned)
intimate distance	1.8	0.9
personal distance	1.6	1.3
social distance	1.2	1.3
gaze-meeting-ratio	2.4	1.0
waiting-time	1.5	0.8
motion-speed	1.4	1.1

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 3.0 step sizes, and for *social* distance, it was 5.0. As Fig. 4 shows, for all subjects except one, the learned *social* distance parameter values fall within the stated acceptable interval.

A statistical test (t-test) has been conducted to compare the initial and final average distances between the preferred and learned values. As a result, there were no statistically significant differences between the initial and final distances for the six parameters. However, this does not mean that the system failed to adapt. Some subjects were content with the learned values even though they were different from the stated preferences, and also, there are difficulties in measuring the preferences. We show such examples as case studies in the following sections.

B. Case Studies: Adaptation Results and Subjects' Impressions

In the following, we show how the system behaved for different groups of subjects. According to the adapted values and the subjects' impressions, we have divided the subjects into five groups: 1) successful; 2) partial success with content subjects; 3) successful but discontent subjects; 4) partially successful but discontent subjects; and 5) unsuccessful runs.

1) Successful Runs—Content Subjects: There were three subjects for whom the system performed very well. Not only were the subjects themselves content with the performance, but also all parameters displayed 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.

Subject 10 (Fig. 6) was impressed by the robot's behavior and said that it quickly became much better. The plots support this, as all parameters are adapted to the preferences, except *personal* distance, which is only slightly farther. This subject stated a preference for an interval of *waiting-time*, hence the two lines in the plot showing the borders of this interval.

2) Partially Successful Run—Content Subjects: The next group consists of two subjects who were content with the robot's behavior, even though analysis of the results shows that some parameters were far from stated preferences.

The study run for subject 5 (Fig. 7) 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-time* were equally good, so this parameter can be any value in this case. Interestingly though, this subject stated that he was content with the *gaze-meeting-ratio* results, even though it is obvious from the plots that these were far from his stated preference. He was also satisfied with the *motion-speed* parameter, which is as much as 20% off from his specified preference.

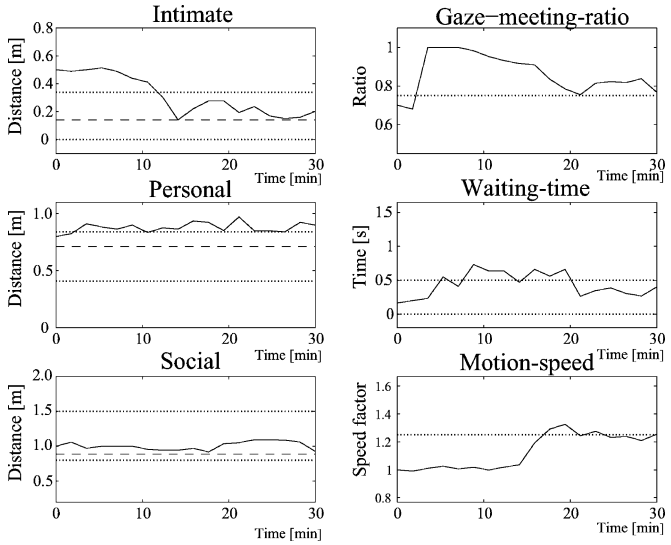


Fig. 6. Successful run—content subjects: results achieved for subject 10. The dotted lines represent the measured preferences of the subject. The dashed lines represent the most preferred values.

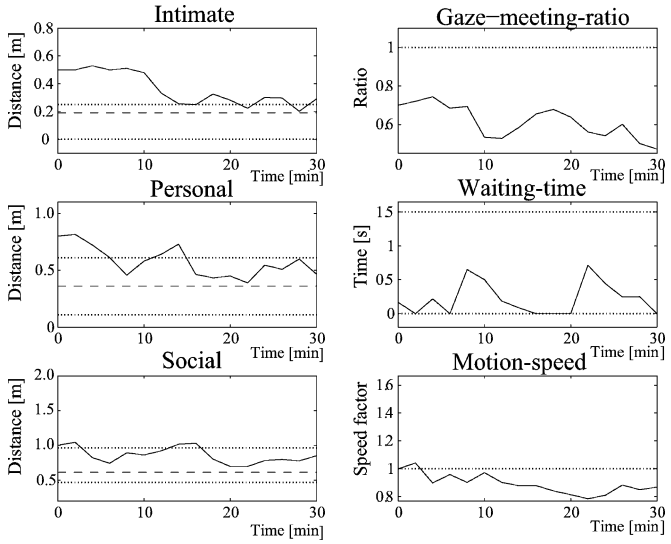


Fig. 7. Partially successful run—content subject: results achieved for subject 5. The lines represent the measured preferences of the subject as in Fig. 6.

3) Successful Run—Discontent Subject: Subject 7 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 it tended to get too close, even though actual *intimate* and *personal* distances converged to her farther acceptable limits.

4) Partially Successful Runs—Discontent Subjects: 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.

In case of subject 3, most of the parameters converged to preferred values or in the acceptable ranges at least. However, *motion-speed* parameter was far away from the stated preference, something the subject also complained about. Observations of the actual study showed

that as the robot increased its *motion-speed*, the subject seemed to watch the movements carefully and fix her gaze at it.

5) Unsuccessful Run—Discontent Subject: The results attained for subject 9 were not very good. Apart from the *personal distance* and *gaze-meeting-ratio*, the results were far from stated preferences. There were no observable problems with this subject's behavior, so the reason for these poor results are still unclear. It is also noteworthy that the subject said that he felt as if the robot did not like him, but forced itself to be somewhat polite and talk to him anyway.

V. Discussion

We found several issues to be solved in the study. First, it is very difficult to measure true preferences. For example, subject 5 (Fig. 7) was content with the learned parameters even though they were far from the stated preferences. On the contrary, subject 7 claimed that the robot got too close, even though the distances were farther than her stated preferences.

Second, for some subjects, the method could neither find the gradient of some parameters nor the direction to the local optimum. The reason is that the behaviors of the subject did not display any difference for policies $\boldsymbol{\Theta}^t$ if the current parameter was too far from the preferred values. This suggests that the adaptation should be done in an appropriate search space, where subjects behave as the reward function expects. This would further enforce the design goal "to not behave unacceptably during adaptation."

Third, there were subjects whose behaviors were different from our expectation. Subject 3 had a tendency to fix her gaze to the robot when the *motion-speed* was higher than her preference. Thus, we need different reward functions for people who have different reactions.

We have simplified the human model by making the following two assumptions. First, people will neither stand still nor stare at the robot to show discomfort or confusion since the interactions are simple to understand and the adapted parameters start from points where they do not produce antisocial behavior. Second, we have only chosen actions that contain a direct interaction and communication between the robot and the subject in our study. It remains a future issue how to prepare an appropriate reward function based on gaze information when these assumptions are not fulfilled. There are also many other possible human behaviors that can be used for rewarding as Tickle-Degnen and Rosenthal [17] suggest.

VI. Conclusion

We have proposed a behavior adaptation system based on PGRL for a robot to interact with a human. We have shown that the robot has successfully adapted at least part of the learning parameters to individual preferences for 11 out of the 12 subjects in the study by reading subconscious signals. Although there are still issues to be solved, we believe that this is an important step toward building robots that are as easy to interact with as humans.

References

[1] E. T. Hall, *The Hidden Dimension*. New York: DoubleDay, 1966.

[2] S. Duncan, Jr. and D. W. Fiske, *Face-to-Face Interaction: Research, Methods, and Theory*. Hillsdale, NJ: Lawrence Erlbaum, 1977.

[3] E. Sundstrom and I. Altman, "Interpersonal relationships and personal space: Research review and theoretical model," *Human Ecol.*, vol. 4, no. 1, pp. 47–67, 1976.

[4] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[5] T. Inamura, M. Inaba, and H. Inoue, "Acquisition of probabilistic behavior decision model based on the interactive teaching method," in *Proc. 1999 Int. Conf. Adv. Robot.*, pp. 523–528.

[6] C. Isbell, C. R. Shelton, M. Kearns, S. Singh, and P. Stone, "A social reinforcement learning agent," in *Proc. 5th Int. Conf. Auton. Agents*, J. P. Müller, E. Andre, S. Sen, and C. Frasson, Eds. Montreal, QC, Canada: ACM Press, 2001, pp. 377–384.

[7] K. Nakashima and H. Sato, "Personal distance against mobile robot," *Jpn. J. Ergonom.*, vol. 35, no. 2, pp. 87–95, 1999.

[8] T. Tasaki, S. Matsumoto, K. Komatani, T. Ogata, and H. G. Okuno, "Dynamic communication of humanoid robot with multiple people based on interaction distance," in *Proc. Int. Workshop Robot Human Interact. (Ro-Man 2004)*, pp. 81–86.

[9] Y. Nakauchi and R. Simmons, "A social robot that stands in line," *Auton. Robots*, vol. 12, no. 3, pp. 313–324, 2002.

[10] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Body movement analysis of human–robot interactions," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI 2003)*, pp. 177–182.

[11] C. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.

[12] N. Kohl and P. Stone, "Machine learning for fast quadrupedal locomotion," in *Proc. 19th Nat. Conf. Artif. Intell.*, 2004, pp. 611–616.

[13] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems,* vol. 12, Cambridge, MA: MIT Press, 2000, pp. 1057–1063.

[14] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *J. Artif. Intell. Res.*, vol. 15, pp. 319–350, 2001.

[15] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2004, vol. 3, pp. 2619–2624.

[16] T. Kanda, H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu, "Development and evaluation of an interactive humanoid robot "Robovie"," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2002, pp. 1848–1855.

[17] L. Tickle-Degnen and R. Rosenthal, "The nature of rapport and its nonverbal correlates," *Psychol. Inquiry*, vol. 1, no. 4, pp. 285–293, 1990.