学術・技術論文

方策勾配型強化学習によるロボットの対人行動の個人適応

光永法明[*1]　クリスチャン スミス[*1*2]　神田崇行[*1]
石黒浩[*1*3]　萩田紀博[*1]

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] *and* Norihiro Hagita[*1]

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 and that different people tend to meet the gaze of others to different extents. For humans, this is mostly subconscious, but when two persons interact, there is an automatic adjustment of these factors to avoid discomfort. In this paper we propose an adaptation mechanism for robot behaviors to make human-robot interactions run more smoothly. We propose such a mechanism based on policy gradient reinforcement learning, that 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 the experiment with twelve subjects.

Key Words: policy gradient reinforcement learning (PGRL), human-robot interaction, behavior adaptation, proxemics.

1. はじめに

人間同士は，その対話において相手との間に適度な空間（パーソナルスペース）をとる [1]．また，対話中には適度に視線を合わせる [2]．たとえば，相手が近過ぎると感じれば少し離れ，見つめ合い過ぎていると感じれば視線をそらす [3]．人間は，こういった適応行動を無意識にとり，対話における快適さを保っている．ここで重要なのは，この空間の形状や大きさは，対話の内容や相手により異なり，また，対話の際に視線を合わせる頻度も人により異なることである．人間は相手に応じてこれらの行動を適応的に調整している．

これは，人間と対話するロボットにおいても重要な機能である．しかしながら，このような個人適応の仕組みを，人と対話するロボットに持たせた研究は数少ない．これまでにも，人と関わるロボットにパーソナルスペースを考慮させて行動させる研究が行なわれている [4]〜[6] が，個人差についてはほとんど考慮されていない．

原稿受付
[*1]株式会社 国際電気通信基礎技術研究所
[*2]スウェーデン王立ストックホルム工科大学
[*3]大阪大学大学院工学研究科
[*1]Advanced Telecommunications Research Institute International
[*2]Royal Institute of Technology, Stockholm
[*3]Graduate School of Engineering, Osaka University

ヒューマンインタフェース分野でも，利用者に明示的なカスタマイズを許すインタフェースや，統計情報に基づき個人に合わせた情報を自動的に選択するものなどが研究されているが [7]，これらは人と対話するロボットが扱う実空間での行動を伴う問題とは異なる．対話における学習の研究としては，実空間で，ユーザとの対話に基づいた行動決定モデルの獲得により，人のタスク実行における好みを，ロボットの行動に反映する研究が報告されている [8]．また，バーチャル空間において，人と関わるエージェントの参加者に合わせた行動の学習も試みられている [9]．しかし，前者は教示に基づくため，ユーザが何をすべきかを指示する必要がある．後者は強化学習を用いており，行動を逐一指示する必要はないが，参加者が意識的にエージェントに対し報酬を与える必要がある．人が無意識に相手に適応している部分については，そもそも明示的に表現できるものではなく，人がロボットに適切な値を指示することは難しい．また，個人毎に適切な値をあらかじめ様々な対話場面で計測しておくことも繁雑である．つまり，多くの人とコミュニケーションするロボットは，自律的に対話相手に適応する仕組みを持つ必要がある．

さらに，人の対話中の快適さには，複数の要素が相互に関係している．たとえばパーソナルスペースに他者が入ってきても，視線が合わなければ不快さは少ない [3]．人と対話するロボットに関しても，人が適当と感じるロボットとの距離は，ロボット

の移動速度に関係することが指摘されている [4]．このように，ロボットが個人適応する際には，複数の要素を総合して考慮する必要がある．

このような背景のもと，本論文では，人の行動に無意識に表れる快・不快の表現を基に，対話においてロボットが，個人に適応する機能を実現し，その有効性を検証する．具体的には，ロボットは人からの快表現を増加し，不快表現を減少するように，強化学習により対話相手に適応する．我々は，従来研究 [10] をもとに，快・不快表現が人の対話中の移動距離とロボットの顔を見る時間の長さに表れると考え，それらに基づく報酬関数を構成した．適応するパラメータは，3 種の対話距離カテゴリ（親密距離，個体距離，社会距離）[1] 毎の保つべき対人距離，人の顔の方向にカメラを向ける時間の長さ，発話から動作開始までの遅れ時間，動きの速さとする．ロボットは報酬関数を最大化するようパラメータを学習し，対話相手に個人適応する．なお，人との対話においては，対話相手に同じようなやり取りを何度も行わせることは困難であるため，適応における学習の収束はある程度速いことが重要である．また，対話相手に適切なパラメータを指定させることは難しいため，学習手法として教師なし学習が必要である．これらを考慮し，学習アルゴリズムに方策勾配型強化学習を用いる．

以下では，まず提案する個人適応システムの，適応するパラメータ，報酬関数，方策勾配型強化学習について述べる．次に実験環境とロボット，実験手順について述べ，詳細な実験結果を報告する．そして，最後にまとめと今後の課題を述べる．

2. 個人適応システム

2.1 適応するパラメータ

円滑なコミュニケーションには，パーソナルスペース（対話距離）[1]，視線を合わせる頻度 (以下，視線頻度) [2] [3]，動きの速さ [4]，テンポ (速度や間隔)，発声，発話の音量，しぐさなど様々な非言語的要素の適応が必要である．中でも対話距離，視線頻度，テンポは重要である．適応するパラメータを多くすると学習に時間がかかるため，対話に重要かつ，実装が容易な以下の 6 つのパラメータを適応するパラメータとする．まず視線頻度を，人の顔の方向にカメラ（顔）を向ける時間の長さ (「注視時間」) で合わせる．次にテンポを合わせるため，発話から動作開始までの遅れ時間 (「遅れ時間」)，動きの速さ（「動作速度」）を適応パラメータとする．そして対話距離として，3 種の対人距離,「親密距離」,「個体距離」,「社会距離」を適応する．

ここでは次のようなロボットを想定している．ロボットには,「握手」,「じゃんけん」,「ばいばい」といった複数の対話行動があらかじめ用意されている．各対話行動は，人に働きかける発話と，体の動きで構成される．対話行動毎に用意した発話と動作の再生により，人に働きかけ，人の反応動作を引き起こし，対話を実現する．また対話行動中に，ロボットはカメラを人の顔へ向け，そらす動作と，対話相手との距離（対人距離）を一定に保つ動作をする．本研究では，人対人の対話における注視の周期 [2] に合わせ，ロボットは 5[s] を 1 周期としてカメラを振る．適応パラメータの「注視時間」は，この 1 周期中にカメラを人の顔へ向けている時間の割合とする．

適応パラメータの「動作速度」は，用意した動作の再生速度とする.「動作速度」は，作成時の速さを 1 とする相対速度で表す.「動作速度」が 1 のとき，ロボットとしての印象が一貫するように各対話行動は作成する．

「遅れ時間」は，発話後に動作を行なう対話行動において，発話から動作再生開始までの時間とする．たとえば,「握手してね」と発話してから，手を出すまでの時間である.「注視時間」,「遅れ時間」,「動作速度」は，すべての対話行動で共通とした．

対話距離の適応については，まずロボットに用意する各対話行動で必要な対人距離が，Hall [1] の対話距離カテゴリの親密 (intimate)，個体 (personal)，社会 (social) の 3 種のいずれの距離に該当するかを調べる．ここで対人距離は，人とロボットそれぞれの額間の水平距離とする．適応パラメータの「親密距離」,「個体距離」,「社会距離」は，それぞれのカテゴリの対話行動を実行中に保つ対人距離である．対話行動毎に異なる対人距離を適応するのではなく，同一カテゴリの対話行動で共通の距離を用いることで学習時間を減少できる．今回の実験に利用した対話行動の詳細と対話距離カテゴリは第 3 章に示す．

2.2 報酬関数

報酬関数は，人が対話中に無意識に行なう動作を利用し構成する．まず，人は相手の位置が対話に近過ぎると感じると相手から離れ，見つめ合い過ぎていると感じると目をそらす傾向がある [3]．また，人ロボット対話において身体動作を解析した研究 [10] によると，ロボットの振舞いに好印象を持つ被験者はロボットに顔を向ける傾向があり，対話中の移動距離も短い傾向が見られている．これは対話相手の動きが緩慢で退屈である場合や，速過ぎて理解できない場合に他に顔を向けると考えると，自然な反応である．そこで，ロボットとの対話において，人の快・不快が無意識に移動距離とロボットに顔を向けている時間に表れると考え，報酬関数を設計した．報酬関数は 1 つの対話行動実行中の人の移動距離とロボットに顔を向けている時間の重みつき和,

$$\begin{aligned}(\text{報酬}) \;=\; & -0.2 \times (\text{移動距離}\ [mm]) \\ & +500 \times \frac{(\text{人がロボットに顔を向けている時間の和})}{(1\ \text{つの対話行動の時間})}\end{aligned}$$

とした．すなわち，対話行動中に人の移動が少なく，人が顔をロボットに向けているほどよい．

実験で使用した報酬計算の流れをブロック図で Fig. 1 に示す．まず人に取り付けたマーカの 3 次元位置をモーションキャプチャシステム (サンプリング周波数は 60[Hz]) で測定し，額の位置と方向を求める．そして，それらを床面に水平な 2 次元平面に投影した座標に変換する．2 次元平面内での額の位置を 5[Hz] でダウンサンプリングした後に，1 時刻前からの移動量を求め，それを積算したものを「移動距離」とする．数フレーム程度以下での細かい身体の揺れを移動距離に含まないようにダウンサンプリングの周波数は実験的に決定している．Fig. 2 のように，人の額とロボットの額の中央を通る直線に対し，人の額の法線ベクトルが± 10 度以内を向いているとき,「人がロボットに顔を向けている」とし，該当する時間を積算して「人がロボットに顔を向けている時間の和」とする．

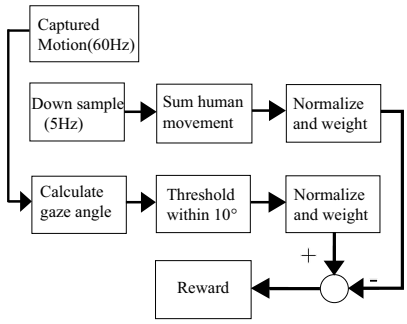


Fig. 1 The reward function

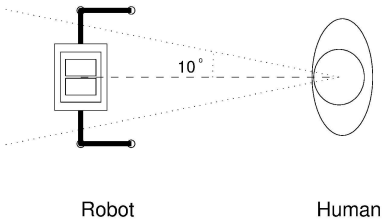


Fig. 2 The angular interval determined as human gazing at robot

報酬計算式の重みは次のようにして決定した．まず予備実験を行ない，人とロボットの動きをモーションキャプチャシステムにより記録するとともに，被験者が適切と感じる各適応パラメータの値を測定した．そしてコンピュータ上で動きを再生し，次章に述べる方策勾配型強化学習を実行し，適応パラメータの学習に伴う挙動を観察し，被験者が適切とじる値に，各パラメータが速く安定して収束するよう，重みを試行錯誤的に求めた．結果，得られた重みでは，それぞれの項の報酬への寄与がほぼ等しくなった．

2.3 方策勾配型強化学習

Q 学習 [11] に代表される多くの強化学習手法では，最適な振舞い (方策) を学習するために，出来るだけ広範囲の空間を探索し，あらゆる方策を試行する．そのため学習結果はグローバルに最適なものが得られるが探索には膨大な試行が必要である．それに対し方策勾配型強化学習 (Policy Gradient Reinforcement Learning，本稿では以後 PGRL とよぶ) [12] [13] では，現在の方策を報酬を得られる方向へ修正していくことで局所最適解を求める．報酬から方策を直接変化させるので，報酬伝播の遅れが少なく学習時間が短い特長がある．

本研究では，対話行動中に適応するパラメータを変化させない PGRL とした [14]．Fig. 3 に学習アルゴリズムを示す．現在の方策 (適応するパラメータ) を $\mathbf{\Theta}$ で表す．学習には，まず現在の方策 $\mathbf{\Theta}$ を少し変動させた T 通りの方策 $\mathbf{R}^i$ を用意する．$\mathbf{R}^i$ は $\mathbf{\Theta}$ の各成分 θ_j にランダムに ϵ_j, 0, $-\epsilon_j$ のいずれかを加えて生成する．変動ステップサイズ ϵ_j はパラメータ毎に異なる値でよい．

次に，それぞれの方策 $\mathbf{R}^i$ にしたがって対話行動を T 回行ない (T 個の異なる方策を試行することになる) 報酬を得る．ここで T 回の対話行動は同一である必要はなく，毎回異なるものでよい．これは，対話行動によらず，方策 (適応パラメータ) によってのみ報酬が異なると仮定しているからである．T 通りの方策すべてについて対話行動を行なった後，報酬関数の $\mathbf{\Theta}$ に対する勾配 $\mathbf{A}$ を近似的に求める．各パラメータ θ_j について，ϵ_j を加えた時の平均報酬，0 を加えた時の平均報酬，$-\epsilon_j$ を加えた時の平均報酬を求める．0 を加えた時の平均報酬が最も大きい場合は，θ_j についての勾配は 0 とする．そうでない場合には，ϵ_j と $-\epsilon_j$ を加えた場合の平均報酬の差を勾配とする．$\mathbf{A}$ を求めたあと，$\mathbf{A}$ を正規化し η を掛けたものに，各成分に ϵ_j の重みをつけ，$\mathbf{\Theta}$ を更新する．この T 回の対話行動とパラメータの更新が学習の 1 ステップである．これを繰り返すことで，パラメータは報酬が極大となる局所最適な値に近付く．

```
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 i = 1 to T
6          for j = 1 to n
7              r ← unbiased random choice
                   from {−1, 0, 1}
8              R^i_j ← Θ_j + ε_j * r , where R^i is
           perturbed parameter set of same size as Θ
9      for i = 1 to T
10         Run system using parameter set R^i,
           evaluate rewards
11     for j = 1 to n
12         Avg_{+ε,j} ← average reward for all R^i
               with positive perturbation in dimension j
13         Avg_{0,j} ← average reward for all R^i
               with zero perturbation in dimension j
14         Avg_{−ε,j} ← average reward for all R^i
               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. 3 PGRL Algorithm

3. 実　　験

3.1 実験環境と使用したロボット

Fig. 4 に実験室の様子を示す．実験室には 12 台のカメラからなるモーションキャプチャシステムが備えられている．このシステムは被測定物に取り付けたマーカの 3 次元位置を 60[Hz] で計測出来る．実験は，位置精度の良い実験室中央 3.5×4.5[m] の範囲で行なう．実験領域内でのマーカの位置測定精度は約 1[mm] である．被験者とロボットの頭と肩に取り付けたマーカの位置から，それぞれの額の位置と方向を計算で求める．額の位置と方向はロボットに LAN を通して送り，ロボットの動作決定と報酬関数の計算に用いる．実験中を通して，通信による遅れは 0.1[s] 以内であり，遅れは無視できた．

実験には コミュニケーションロボット Robovie II [15] を用い

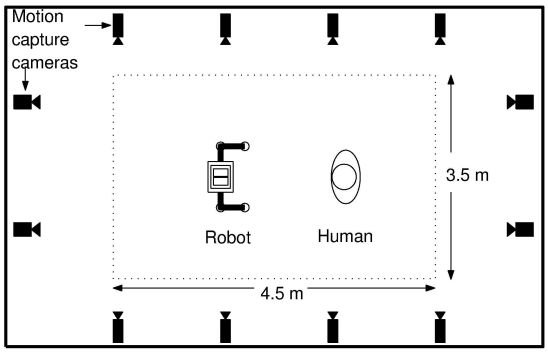


Fig. 4 Experimental setup

た．Robovie II の身長は小学校低学年の児童程度の約 120[cm] である．底部には 2 つの駆動輪があり，上半身には 4 自由度の腕が 2 本ある．頭部には 2 つのカメラがあり，カメラ毎にパン・チルトの 2 自由度と頭部全体を動かすパン・チルト・ロールの 3 自由度がある．これにより，被験者に視線を向けることが出来る．またマイクとスピーカを内蔵しており，被験者に呼びかけることが出来る．

3.2 ロボットの対話行動

用意した対話行動は，a) だっこ (hug)，b) じゃんけん (paper-scissors-stone)，c) 触ってね (ask for touch)，d) 運動 (excercise)，e) 握手 (handshake)，f) あっちむいてほい (pointing game)，g) ありがとうバイバイまたね (monologue)，h) どこから来たの？(ask where from)，i) ロボビーってかわいい？(ask if cute)，j) 相手を見る (just looking) の 10 種類である．Fig. 5 にロボットの外観と用意した対話行動を示す．いずれの対話行動もロボットの腕や頭の動きを伴う．たとえば，a)「だっこ」は，ロボットが「だっこしてね」と発声し，腕を広げ，人がロボット正面の適当な距離に立つと腕で人に抱きつく．これらの対話行動は子供が振舞うような雰囲気で印象が一貫するように作成している．

a) から g) の「だっこ」「じゃんけん」など，人とロボット両者の同時動作が必要な対話行動では，人間の動作を誘う発話（「だっこして」「じゃんけんしようよ」など）をし，「遅れ時間」経過後に腕や頭部の動作を開始する．h)「どこから来たの？」と i)「ロボビーって可愛い？」は，主に音声による反応を求め，j)「相手を見る」では発話をせず，発話後に開始する動作がないため「遅れ時間」は使用しない．

用意した対話行動は，ロボットが主体となって働きかけ，対話を開始する．ロボット主体としているのは，現在の音声認識や画像処理の技術では人の任意の発話や働きかけの認識が難しいという技術的制約のためである．また，実験で用いたロボット Robovie II のコンセプトは子供を模したものであり，子供から大人に対して遊びを誘うことは自然であり，ロボット主体であることはコンセプトと一致している．

これらの対話行動における対人距離を予備実験により測定し，Hall [1] の対話距離カテゴリ，親密距離 (0[m] から 0.45[m])，個体距離 (0.45[m] から 1.2[m])，社会距離 (1.2[m] から 3.6[m]) に分類した．予備実験には 8 名の被験者を集めた．ロボットは移動せずに対話行動を実行し，各被験者にはそれぞれの対話に適していると思う位置に移動してもらい，人ロボット間距離を対人距離として測定した．被験者間で多少の距離の差は見られたが，分散は小さく分類には影響しなかった．予備実験による分類結果を Table 1 に示す．結果として，人同士の対話では社会距離をとる対話行動 (h)「どこから来たの？」，i)「ロボビーってかわいい？」) が，個体距離に分類された．これはロボットの発話の聞きとりにくいためと，人がマイクに近付いて発声することが必要なためと考えられる．

Table 1 Behavior Classes

Class	Name
intimate	Hug
personal	Shake hands
personal	Ask person to touch robot
personal	Ask where person comes from
personal	Ask if robot is cute
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

3.3 実験

[被験者] 当研究所に勤務する 15 名（男性 9 名，女性 6 名）が実験に参加した．1 名を除き日本人で，全員がロボットの発話を聞き取ることが出来た．被験者の年齢は 20 から 35 才，多くは 20 から 25 才であった．全員が，多少は Robovie II について知っていた．

[実験における対話] 実験開始時にロボットは，モーションキャプチャシステムの測定領域の中央におり，被験者はロボット正面に立った状態から，リラックスして自然な気持ちでロボットと対話する (遊ぶ) よう求められた．モーションキャプチャシステムの測定範囲内にいるように求めた以外は，被験者に何も要求していない．被験者はおよそ 30 分間，ロボットと対話した．

対話内容は，ロボットが乱数により 10 種の対話行動から 1 つを選択し実行することで決まる．対話行動 1 つの実行にかかる時間は約 10 秒である．1 つの対話行動が終わると，乱数により次の行動を選択する．乱数が前回と同じ対話行動を選択した場合には，再度乱数により異なる対話行動を選択した．各対話行動，あるいは距離カテゴリの発現回数を均等にするための特別な工夫はしなかった．

[パラメータ学習] Table 2 に各パラメータの初期値と変動ステップサイズ ϵ_j を示す．初期値は予備実験に参加した被験者が適当とした値の平均からやや外れた値とした．報酬はロボットが一つの対話行動を終了する（約 10 秒）毎に計算する．適応の 1 ステップで試行する方策数は $T = 10$ とし，対話行動 10 回毎にパラメータを更新した．前述のように乱数により対話行動は決定しており，適応の 1 ステップ内でも各試行で対話行動は異なる．パラメータの更新に要する時間は十分に短く，更新のために被験者に感じられる対話の中断はない．

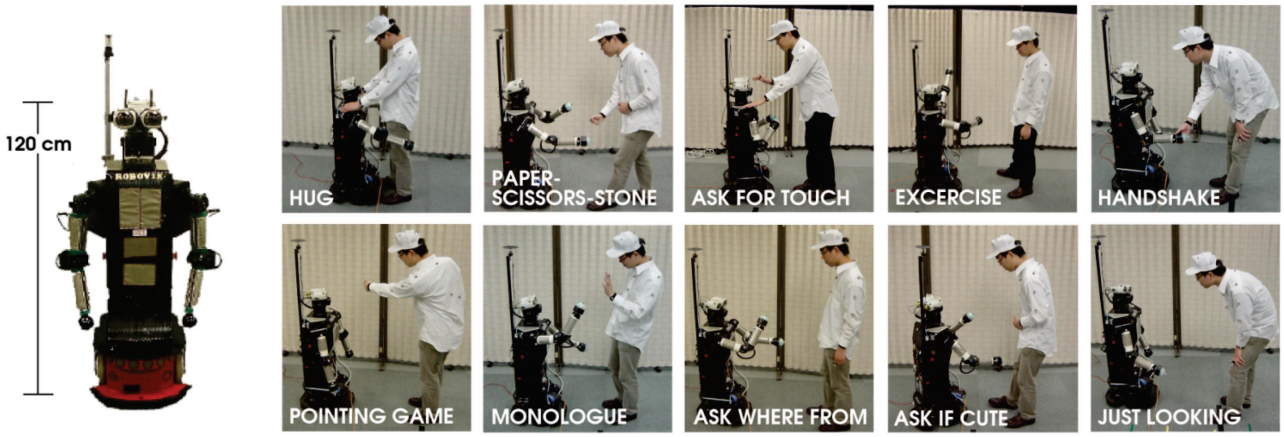


Fig. 5 Robovie II and the behaviors

Table 2 Parameter values and step sizes

#	Parameter	Initial value	Step size ϵ_j
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

[被験者の評価] 実験後，被験者にインタビューし，ロボットの動作，距離，視線の合わせ方の印象と，実験中それらがどのように変化していったかを尋ねた．また，それ以外にも感じたことがあれば自由に答えてもらった．

[被験者の望むパラメータ値の測定 (時間と速度)]
各被験者が望ましいと感じる各適応パラメータの値を実験後に測定した．測定パラメータの値のみ 3 段階に変えた対話行動から，被験者には適当と感じるものを選択してもらった．被験者の中には複数の段階が適当であるとした場合や，中間の値がよいと思うと回答するものがいた．そのような被験者については，その旨を記録した．

適応パラメータ「注視時間」と「動作速度」の測定には，g)「ありがとう」を使い，対人距離は 1.0[m] とした.「遅れ時間」の測定には， a)「だっこ」を使い，距離についてはロボットの移動を止め，被験者に適切と思われる位置に立ってもらった．これは「親密距離」の個人差が大きかったためである．対人距離を除く測定外のパラメータは予備実験の被験者が適当とした値の平均値とした．

[被験者の望むパラメータ値の測定 (対人距離)]
対人距離の測定も実験後に行なった．対話行動しているロボットの正面の適当と感じる位置へ被験者に立ってもらい，モーションキャプチャシステムで距離を測定した.「注視時間」は 0.75,「遅れ時間」は 0.3[s],「動作速度」は 1.0 とした．親密距離は a)「だっこ」，個体距離は e)「握手」，社会距離は g)「ありがとう」の対話行動で測定した．また適当と感じる距離からロボットを近づけた場合，遠ざけた場合に，被験者が適切でないと感じ始める距離を許容限度として測定した．

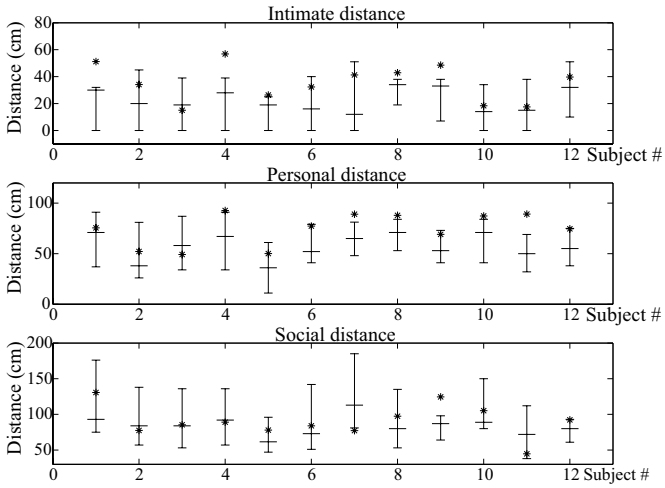


Fig. 6 Learned distance parameters and preferences for 12 subjects

測定外のパラメータを固定した上記の手順で，測定に 15 分を要した．最適なパラメータの組合せを測定することが理想的ではあるが，膨大な時間が測定に必要になると予想されるため，上記の手順とした．

4. 実 験 結 果

学習結果と被験者が望ましいと感じたパラメータを比較し，学習結果の好ましさを評価した．また，本研究の個人適応という問題の性質上，被験者それぞれにどの程度適応できたかの個別の結果も重要であるため，個別の結果を事例として紹介する．

なお，被験者のうち 3 名は我々が想定した行動をとらなかった．ロボットのインタラクションが適当なものであっても，そうでなくても顔の方向や，立ち位置を変えることなく，ロボットあるいは実験者に感想を言葉で述べたり，顔に表出するのみであった．こういった被験者は報酬関数設計の際に前提とした対話反応表現モデルに当てはまらず，システムは正しく動作しない．したがって以下では，これら被験者の結果を除き，残りの 12 名の被験者について結果を報告する．

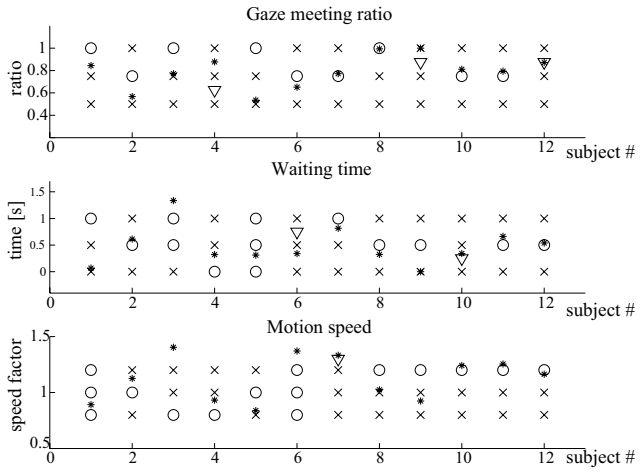


Fig. 7 Learned gaze and speed parameters and indicated preferences for 12 subjects. Circles show what parameter settings the subject indicated as preferable, 'x's show the non-indicated settings. In the cases where values outside the given settings were indicated, a triangle shows the preferred value.

4.1 学習結果

Fig. 6 に 12 名の被験者について，適応の結果得られた対人距離と被験者が適当と判断した距離を示す．適応結果は，対話の最後の 1/4 の期間 (約 7 分半) の平均を示している．これは PGRL が常に局所最適値を探索しているためである．* が適応結果，バーは被験者の許容限度と最適とした値を示す．多くの被験者に対して，15 から 20 分 (約 10 回の PGRL のパラメータ更新) で 適切な値にパラメータが収束した．Fig. 7 に,「注視時間」(gaze meeting ratio),「遅れ時間」(waiting time),「動作速度」(motion speed) についての実験結果を示す．○は被験者が適当と判断した値（複数適当と判断した場合は複数に○がある），*は適応結果である．2 つの値の中間が適当とした被験者については，中間に▽を記している．

これらの図から，被験者の望む値との一致度合はパラメータによって異なるが，多くの場合，適応結果は許容範囲に入ったことが分かる．例えば，動作速度は一致度が低いが，社会距離は 1 人の被験者を除き許容範囲に入っている．一致度合が異なる一因は，それぞれのパラメータの対話への重要性が異なるため，報酬へ寄与の大きいパラメータから収束し，許容範囲の広いパラメータへの収束は遅くなるためと考えられる．

Table 3 に，被験者が適当とした値からの分散を 12 人の被験者について平均したものと，初期値の分散を各パラメータについて示す．値は各パラメータの変動ステップサイズ ϵ_j を 1 として正規化している．また分散は対話の最後の 1/4 の期間についての平均である．表から，学習後の分散の平均は変動ステップサイズの 1 倍前後となったことが分かる．なお，各パラメータの許容範囲の平均は変動ステップサイズよりも大きい．たとえば，個体距離では 3 倍，社会距離では 5 倍である．

これらから PGRL に基づいた個人適応は有効に働き，各パラメータは適切な値に近付いたといえる．より誤差を小さくするには，報酬関数の人の行動への感度を高める，ステップサイズをより小さく，あるいは適応が進むにつれ徐々に小さくするといった工夫が必要である．

Table 3 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

適応パラメータの最適値への収束には 10 回程度のパラメータ更新が必要であったが，Table 3 に示すように初期値も最適値から大きく離れているわけではない．また Fig. 6,7 に示すように，被験者によっては 15 から 20 回の更新でも最適値に収束しないパラメータもあった．しかし，報酬関数を意識しながら，現在のパラメータに応じて，一貫して同じ振舞いをした場合には 4 から 5 回の更新で，報酬関数の重みを求めたシミュレーション上では最短で 3 から 4 回の更新で，最適値へ収束した．したがって，収束までに必要なパラメータ更新回数の多さの一因は，人の動きのバラツキにあると考える．

4.2 事例紹介：被験者の印象とパラメータ適応

インタビューで被験者が述べたロボットの動作に対する印象と，学習結果の最適値への収束度合は，理想的には一致するべきである．しかし，被験者の望むパラメータ値は最適な組合せを測定できておらず，対話時と測定時とで適切と感じる値が違う場合もある．また用意した報酬関数の適切さは被験者によって異なると考えられる．

そこで以下では，学習結果の最適値への収束の良さと，インタビューで得た被験者の印象により，実験結果を 5 つの事例に分け，それぞれ詳細な実験結果を報告し，結果を導いた要因について考察する．

4.2.1 最適値への適応が良好で，被験者の印象もよい場合

個人適応システムは，3 名の被験者に対して，学習結果，インタビュー結果の両面で良好な結果をみせた．各パラメータは最適値へ収束し，被験者は適当と感じるパラメータでロボットが行動したと述べた．被験者に見られた共通点は，ロボットとのインタラクションを楽しみ，ロボットであることを意識せず人に対すると同じように接していたことである．

被験者 10 についてパラメータの変化の様子を Fig. 8 に示す．被験者 10 はロボットの振舞の改善が早かったと感想を述べた．図からも,「個体距離」について若干最適値から離れているのみで，各パラメータは十分に最適値近くに収束しており，感想と一致することが分かる．

また，全般にロボットがよく適応できていた被験者のコメントには，適応パラメータの変化があまり含まれない傾向があった．これはロボットに合わせた行動が不要であれば，当たり前の行動であり，パラメータの適応も意識されなくなる可能性を示唆している．

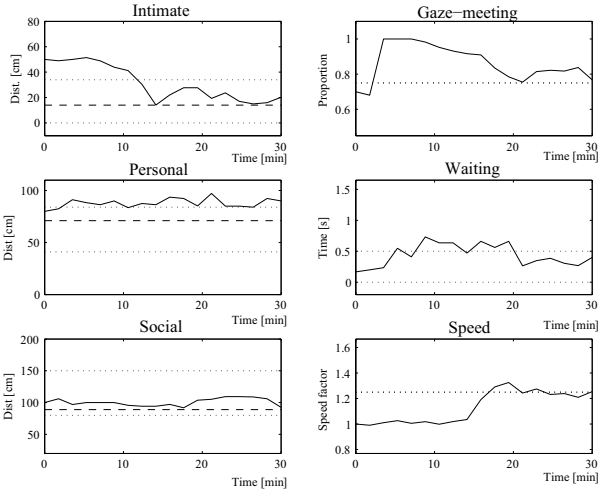


Fig. 8 Results achieved for subject 10. The dotted lines represent the measured preferences of the subject.

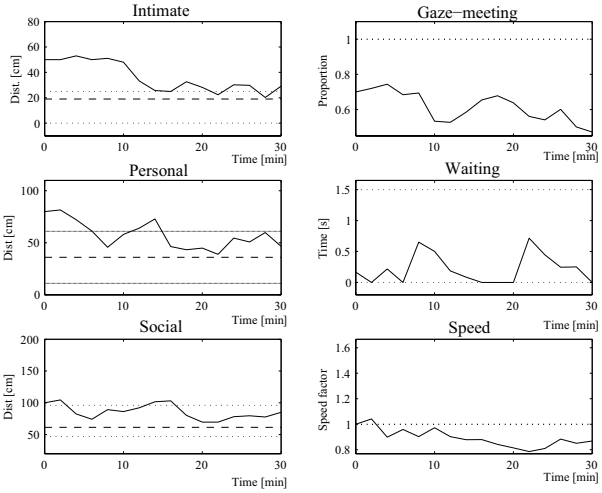


Fig. 9 Results achieved for subject 5. The dotted lines represent the measured preferences of the subject.

4.2.2 一部のパラメータが最適値へ収束していないが，被験者の印象はよい場合

一部の適応パラメータの値が最適値から大きく外れていたが，ロボットの動作について印象がよかったと，2 名の被験者はインタビューに答えた．被験者 5 についてパラメータの変化の様子を Fig. 9 に示す．3 種の対人距離に関しては最適値に収束しており，「遅れ時間」に関して許容範囲が広く適切に学習したといえるが，「注視時間」「動作速度」は最適値から大きく外れている．しかし，被験者はこれらも適当であったと述べた．この原因は，実験中の条件と最適値を測定した条件の違い，あるいは被験者のパラメータの許容範囲が実際には広かったためと考えられる．

また被験者 5 は，他の被験者には見られない行動を行なった．この被験者は「社会距離」に分類される対話行動においても，ロボットの各部を触っていた．その結果，「社会距離」が他の被験者と比較してかなり短くなっている．このような振舞を我々は予期していなかったが，他の被験者と同じ報酬関数により，適応パラメータは被験者が満足する値へ収束した．

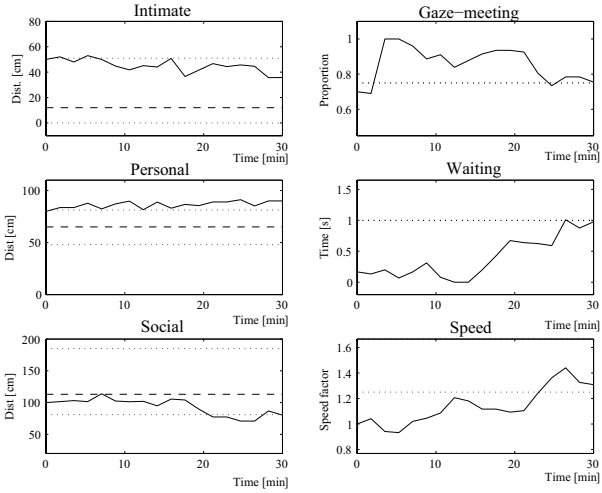


Fig. 10 Results achieved for subject 7. The dotted lines represent the measured preferences of the subject.

4.2.3 最適値へ適応したが，被験者が一部の適応へ不満を持つ場合

被験者 7 についてロボットの各パラメータの変化の様子を Fig. 10 に示す．各パラメータは許容範囲内に収束しているように見られるが，被験者は距離に関して近過ぎたと述べた．しかし，「親密距離」「個体距離」は，被験者が許容する最も遠い距離近くに収束している．また初期の印象は「ためらった感じ」だったが，次第に「活発」になる印象を受けたと述べている．許容範囲内にも関わらず被験者が近すぎると感じたのは，測定時と学習時の「動作速度」の違いが一因であると考えられる．対人距離は「動作速度」1 で測定しているが，被験者の好みに合わせ「動作速度」は学習につれ上昇している．被験者にとっての最適距離は「動作速度」が速い場合には，より遠かった可能性がある．

4.2.4 一部のパラメータが最適値へ収束せず，被験者も一部の適応に不満を持つ場合

5 名の被験者について，一部のパラメータが最適値へ収束せず，被験者もそれらのパラメータの学習結果について不満を述べた．被験者 1 についてパラメータの変化の様子を Fig. 11 へ示す．この実験はトラブルにより他の被験者よりも実験時間が 21 分間と短い．「個体距離」と「社会距離」は最適値近くへ収束したが，「親密距離」は許容範囲へ入らなかった．被験者も「親密距離」が不適当であったと述べている．これはロボットが「親密距離」として試行した距離が常に許容範囲外にあり，いずれの試行に対しても被験者の振舞がほぼ同じであったために，報酬に勾配がつかず，最適な距離に近付くようパラメータを更新できなかったためと考えられる．「注視時間」は，対話の最後の 1/4 の期間で平均して約 0.9 になった．被験者は 1 が最もよく，0.75 から 0.5 程度でもよいと述べたので，適当な値に収束したといえる．被験者は「動作速度」についてはどの値でもよいとし，「遅れ時間」の適応結果は許容範囲外であったが，あまり気にならないと述べた．したがって「親密距離」を除きうまく適

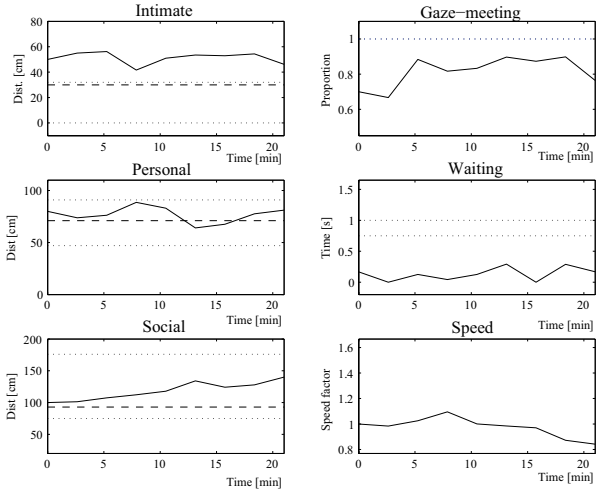


Fig. 11 Results achieved for subject 1.

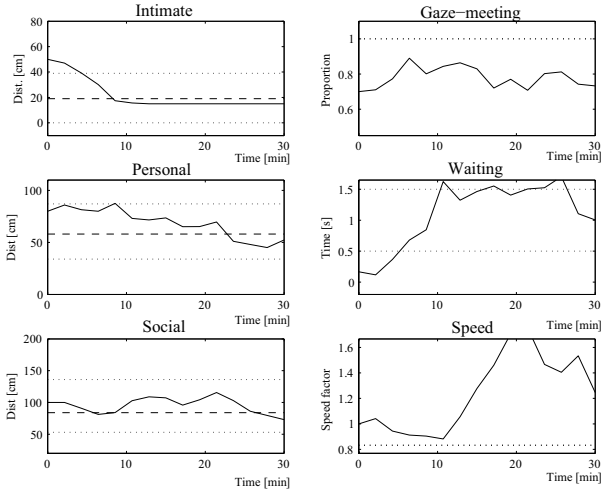


Fig. 12 Results achieved for subject 3. The dotted lines represent the measured preferences of the subject.

応できたといえる．

被験者 3 の適応パラメータの変化の様子を Fig. 12 に示す．被験者 3 は,「個体距離」が実験の前半で不適当であったと指摘した．これは図の「個体距離」の変化，前半は許容限度に近いことと一致している.「親密距離」の学習結果は安全のために設けた下限の 15[cm] になっている.「注視時間」は，学習の結果は約 0.75 であり，最適値は 1 であったが，被験者は十分に満足できたと述べた.「遅れ時間」は許容範囲が広く，学習結果は適当といえる．しかし「動作速度」の学習結果が最適値から大きく離れている．被験者も「動作速度」は不適当だったと述べている．この原因はロボットの「動作速度」が速過ぎる場合に，被験者 3 はロボットをじっと見つめる傾向があり，報酬が誤って大きくなったためと考えられる．

4.2.5 うまく適応できなかった場合

被験者 9 についてパラメータの変化の様子を Fig. 13 に示す．「個体距離」と「注視時間」を除く 4 つの適応パラメータが最適値から大きく外れている．被験者はロボビーに嫌われていて，

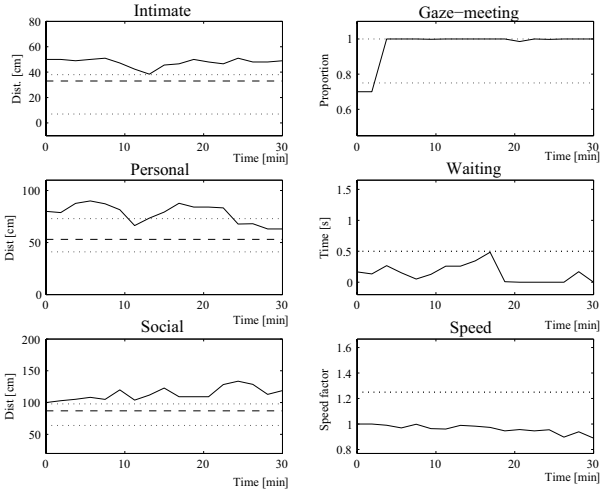


Fig. 13 Results achieved for subject 9. The dotted lines represent the measured preferences of the subject.

ロボビーは，いやいや普通に振舞うよう努力している印象を受けたと述べた．実験中の様子からは大きな問題があるとは観察されなかったが，被験者が率直にロボットに対して反応表現しなかった可能性がある．

5. まとめと今後の課題

本論文では，人とロボットの対話においてロボットが人の行動に無意識に表れる快・不快の表現に基づき適応することを提案した．具体的には，人の「移動距離」と「人がロボットに顔を向けている時間」により報酬を構成し，方策勾配型強化学習により，対人距離や動きの速さといった対話行動に関したパラメータを個人適応する．12 名の被験者への実験により，1 名を除いて複数のパラメータを適切に適応できることを確かめた．また適応したロボットの振舞が自然に見えたという感想を被験者らから得た．個人適応は，より自然に人と対話できるロボットの実現への重要な要素の一つであり，提案手法はその一歩といえる．

一方で，これから解決しなければならない課題も見つかった．1 つ目は，個々人にとって最適なパラメータを測定することの難しさである．例えば，被験者 5 は最適とした値から大きく離れた値であっても適応結果に不満はなかった．一方，被験者 7 は最適とした値近くに学習結果が収束したにも関わらず，距離が近過ぎると感じた．この一因は，パラメータが相互に依存する場合が多いためである．この問題はロボットの適応には影響しないかもしれないが，適応結果，適応システムの評価を難しくする．

2 つ目に，最適値からある程度以上離れたパラメータについては，人が同じ反応表現を返す傾向があるように見られた．このためパラメータが変化したにも関わらず報酬が変化しないために，報酬の勾配を見つけられず，適応できない場合が被験者 1 に見られた．

3 つ目には，1 つの報酬関数ではすべての人の反応表現を正しく扱えない問題がある．本論文の実験でも，報酬関数の設計の

際に想定したモデルに当てはまらない被験者がいた．被験者 3 は，ロボットの動作速度が速過ぎた時にじっと見る傾向があった．報酬関数の設計の際にはじっと見ることは好ましい反応表現であると仮定しているので，ロボットはより速い動作速度が適切であると判断し，その結果，被験者の望まない学習をした．

2 つ目と 3 つ目の問題については，反応表現の違う被験者，あるいは反応表現の違う領域に対して，適切に別の報酬関数を利用する仕組みが必要である．それには，反応表現による人の分類が一つの解決策になるのではないかと考えている．分類により，適した報酬関数の選択と，最適値探索の開始点をより最終値に近いところから開始することが可能になる．結果として適応速度の向上と，適応の安定化が期待される．人を分類するに当たっても多くの課題は残る．どのような報酬関数をいくつ用意すればよいのか．適した報酬関数の選択はどうすれば可能か．何通りに分ければよいのか．人がどの分類に属すかを，どうやってロボットは判断するのか．またどのように素早く判断するのか．個人適応に必要なパラメータの数はいくつなのか．こういった課題を解決していく必要がある．

また一連の対話行動が長い場合や，モーションキャプチャシステムのない環境での適応システムの実現も今後の課題である．会話など長い対話行動の場合には，行動中に複数のパラメータを評価できることが望ましい．そのためには適当な時間的区切りで対話行動を分割できることが必要である．モーションキャプチャシステムを前提としないためには距離や視線 (顔方向) 計測などをロボット上のセンサで測定する技術的課題が残る．またロボット上のセンサの特性を考慮した報酬関数の設計も必要となる．実験室外でのロボットと人の対話は，本実験のように 30 分間連続したものではなく，1 度には 1 ないし 2 程度の対話行動のみの短い対話になると考えられ，複数回の人との接触での対話行動の結果を統合して適応する仕組みが必要になると考えられる．そのためにはロボットが対話する個人を適切に識別し，適切に個人との対話履歴を残す手法も確立する必要がある．

謝辞 本研究は総務省の研究委託により実施したものである．

参 考 文 献

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

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

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

[4] 中島 移動体ロボットに対するヒトの個体距離に関する研究．博士論文, 九州芸術工科大学, 1998.

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

[6] T. Tasaki, S. Matsumoto, H. Ohba, S. yamamoto, M. Toda, K. Komatani, T. Ogata, H. G. Okuno. Dynamic communication of humanoid robot with multiple people based on interaction distance. 人工知能学会誌, vol.20, no.3, pp.209-219, 2005.

[7] 神場, 古関．ひとりひとりにオーダーメイドの情報を—パーソナライゼーション技術—．電子情報通信学会誌, vol.82, no.4, pp.354-359, 1999.

[8] 稲邑, 稲葉, 井上．ユーザとの対話に基づく段階的な行動決定モデルの獲得．日本ロボット学会誌, vol.19, no.8, pp.983-990, 2001.

[9] C. Isbell, C. R. Shelton, M. Kearns, S. Singh, and P. Stone. A social reinforcement learning agent. In *Proc. of the Fifth International Conf. on Autonomous Agents*, pp.377-384, ACM Press, 2001.

[10] 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.

[11] Chris Watkins and Peter Dayan. Q-learning In *Machine Learning*, vol. 8, pp.279-292, 1992.

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

[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 12*, pp. 1057–1063. MIT Press, 2000.

[14] 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.

[15] 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.

光永 法明 (Noriaki Mitsunaga)

1974 年 1 月 5 日生まれ．1997 年大阪大学大学院工学研究科電子制御機械工学専攻博士前期課程修了．同年同学科知能機能創成工学専攻博士後期課程進学．2002 年同退学．同年同大学科学技術振興特任教員．2003 年博士後期課程修了．同年大阪大学大学院工学研究科助手．2004 年より ATR 知能ロボティクス研究所研究員．ロボットの知能に関連した問題に興味を持つ．人工知能学会会員 （日本ロボット学会正会員）

クリスチャン スミス (Christian Smith)

Born Aug. 28, 1975. Intern at the ATR Intelligent Robotics and Communication Laboratories in 2004. Master of Science in engineering physics from the School of Engineering Sciences at the Royal Institue of Technology (KTH), Stockholm, Sweden in 2005. Enrolled in PhD program in Autonomous Systems at the School of Computer Science and Communication at KTH since 2005. Research interest includes human-machine cooperative control of robotic systems.

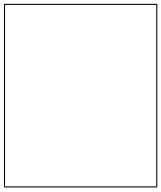
神田 崇行 (Takayuki Kanda)

1975 年 12 月 7 日生．1998 年京都大学工学部情報工学科卒業．2000 年同大学大学院情報学研究科社会情報学専攻修士課程修了．2003 年同専攻博士課程修了．博士 (情報学)．現在，ATR 知能ロボティクス研究所上級研究員．ヒューマンロボットインタラクション，特にロボットの自律対話機構や社会的能力，人間型ロボットの身体を利用した対話に興味を持つ．
（日本ロボット学会正会員）

石黒 浩 (Hiroshi Ishiguro)

1963 年 10 月 23 日生．1991 年大阪大学大学院基礎工学研究科物理系専攻博士課程修了．工学博士．同年山梨大学工学部情報工学科助手，1992 年大阪大学基礎工学部システム工学科助手．1994 年京都大学大学院情報学研究科社会情報学専攻助教授．2001 年，和歌山大学システム工学部情報通信システム学

科教授．現在，大阪大学大学院工学研究科知能・機能創成工学専攻教授，ATR 知能ロボティクス研究所コミュニケーションロボット研究室客員室長．視覚移動ロボット，能動視覚，パノラマ視覚，分散視覚に興味を持つ．人工知能学会，電子情報通信学会，情報処理学会，IEEE，AAAI 各会員． （日本ロボット学会正会員）



萩田 紀博 (Norihiro Hagita)

1978 年 慶應義塾大学大学院工学研究科電気工学専攻修士課程修了．同年 日本電信電話公社 (現 NTT) 武蔵野電気通信研究所に入所．文字認識や画像認識などの研究に従事．NTT 基礎研究所などを経て，現在 ATR 知能ロボティクス研究所所長．工学博士．IEEE，電子情報通信学会，情報処理学会，人工知能学会，各会員．