学術・技術論文

人間の教示特性に基づく顔ロボットの行動学習アルゴリズム

飯田史也[*1] 原 文雄[*2]

Behavior Learning of Face Robot Based on the Characteristics of Human Instruction

Fumiya Iida[*1] *and* Fumio Hara[*2]

We have found a difficulty in the learning of a human-friendly behavior in a face robot by means of human instruction. The difficulty is mainly spotted at the learning algorithm of the robot not taking account of characteristics of human instruction. This paper suggests an effective learning algorithm constructed by the results obtained through analysing characteristics of human instruction, and evaluates its effectiveness by computer simulation and human-face robot interaction experiments.

***Key Words*:** Man Machine Interaction, Reinforcement Learning, Robot Behavior Learning, Human Instruction

1. 背 景

近年, 人間社会で活躍するロボットへの期待が高まっている. 中でも福祉ロボットに代表されるように, 人間と直接触れ合う場で行動するロボットへの要求は徐々に大きくなり, 人間が親和感を感じるロボットについての研究も進められている[1]. 筆者らは人間とロボットのインタラクションを通じて, 人間がロボットについてどのような印象を持つかに注目し, 人間がロボットについて評価する印象を「ロボットの性格」と定義しての実験的分析を行った[2]. その結果ロボットの性格は人間の行動や状態に対するロボットの反応, すなわち「インタラクション特性」に大きな影響を受けることが明らかになった. これまで人間に与えるロボットの印象は, ロボットの外観などの静的特性や動作速度などの動的特性に強く影響されることが指摘されている[1][10]. しかしこの実験の結果, ロボットが認識する状態と行動の対応(以降, 知能と記す)もまた人間に与えるロボットの印象に強く影響することが明らかになった.

ところで, 人間にサービスするロボットなど, 人間と直接触れ合うロボットでは, 人間との感性情報の交流が大切であると考えられるが, その感性は人間の個人によって微妙に異なることを日常経験する. そこで, 人間とロボットとのインタラクションでは, 感性情報のロボットによる表現を設計者が事前に設計することでは, 不十分と考えられる. このような議論から, 本研究の最終的な目標は, 人間と直接触れ合うロボットが人間に快く感じる行動をとることのできる知能を持たせることとする. しかし人間が快く感じるロボットの行動を実現する知能を事前に決定するのは非常に困難であると考えられる. なぜならこのような知能の評価基準となる人間の感性は大きな個人差があり, ロボットが使用される状況によっても大きく変化することが予想されるからである. そこで本研究では学習により人間が快く感じる行動を実現するロボットの知能をロボット自身で人間とのインタラクションを通して獲得していくアプローチを採用し, その方法論の確立を目指す.

ロボットの行動学習に人間の教示を用いる従来の研究報告は, 人間の教示をロボットの行動学習の目標である正解行動の一例として用いることにより, ロボットの行動学習を進めることに重点を置いている[7][8]. しかしこのようなロボットの行動学習へのアプローチは, 人間が正解行動を例として示すことができないロボットの行動学習や, 正解行動が人間の感性に影響され, ロボットを取り巻く状況やロボットのパートナーとしての人間により必ずしも正解行動が正確に定義できないようなロボットの行動学習には適用が困難であると考えられる.

そこで本論文では, ロボットの行動を人間が評価することを「人間によるロボットへの教示」と捉え, この人間の教示によりロボットが知能を自己組織化するという枠組みでロボットの行動学習を議論する. このアプローチでは, 人間の教示により知能の学習が効率的に行われ, さらに人間の感性によって評価されるような曖昧性の強い知能であっても適応的に学習できる可能性があり, 人間が直接触れ合う場で行動するロボットに必要不可欠な, ロボットの柔軟な知能の実現が期待できる. しかし人間が行う教示には特性があり, ロボットの行動学習アルゴリズムにその特性を考慮しないと, 行動の学習が進まないことがある[2]. この事実に基づき, 本論文では人間の教示特性に基づいたロボットの行

原稿受付 1999年2月26日

[*1] チューリヒ大学情報工学科
[*2] 東京理科大学
[*1] University of Zurich, Department of Information Technology
[*2] Science University of Tokyo

動学習アルゴリズムを提案し,その学習アルゴリズムの特性をシミュレーションで分析するとともに,顔ロボットの表情学習実験により学習アルゴリズムの有効性を示す.

第2章では人間の教示によるロボットの行動学習の議論のための枠組みとなる,「学習指向型のインタラクション」というコンセプトを提案し, その枠組みにおける問題点を指摘する. 第3章では問題点の一つであるロボットの学習に大きな影響を及ぼす人間の教示特性を示し, 第４章では人間の教示特性に基づいたロボットの行動学習アルゴリズムを提案する. 第5章ではこれらの行動学習アルゴリズムをシミュレーション実験によりその特性を検証し, 第6章では実際の人間とのインタラクション実験の一例として,顔ロボットの表情学習実験を行うことで学習アルゴリズムの有効性を検証する.

2. 学習指向型インタラクション

本研究で目指すロボットと人間の新しいインタラクションの方法論として,「学習指向型インタラクション」(Learning-oriented interaction)というコンセプトを提案する(**Fig. 1** 参照).

人間と直接触れ合うような場で行動するロボットには,言語的な情報だけでなく,表情やジェスチャーに代表される非言語的な情報である感性情報を伝達し合うモーダリティが必要とされている[3].

このような感性情報をロボットから人間へ伝達するために,ロボットには多様な表現力が必要で,高い行動の自由度が要求される. さらに, ロボットが認識する人間の状態にも多様性が求められるため, これらのロボットの行動と人間の状態を結びつける知能は膨大な組み合わせが考えられるという意味で非常に複雑なものになる. しかしその一方でロボットの知能が複雑化すると, 必然的に個々の人間や場所, 時間などによって, 人間とロボットの間での知能における局所的な共通認識が必要になると予想される. 以上のことから, 将来複雑な知能を持つロボットと人間との間のインタラクションにおいてはロボットの知能に学習機能を持たせることが重要になると考えられる. そこで, このようなインタラクションをするための知能に学習機能を持たせたロボットと人間との間のインタラクションを「学習指向型インタラクション」と定義し, 以降はこの枠組みの中でのロボットの学習機能のあり方を議論する.

学習指向型インタラクションを実現するためのロボットには, 少なくとも以下の機能が必要である.

①ロボットによる人間の言語的, 非言語的行動の認識
②ロボットによる言語的, 非言語的行動の実現
③認識した人間行動からの教示情報の抽出
④抽出した教示情報を用いた知能の学習

①に関しては, 画像処理, 音声処理などの分野で様々な人間の行動を認識する技術が開発されている[9]. また②に関しても, 人間とエージェント間の non-verbal communication が注目され, 様々なモーダリティによるインタラクションデバイスが実現されている[3]. そこで本研究では, これらの要素技術を利用して③,④の機能に注目する.

筆者らは,人間が実際に顔ロボットとのインタラクションを通じて, 顔ロボットの行動の教示を行う際に, 教示方法によっては

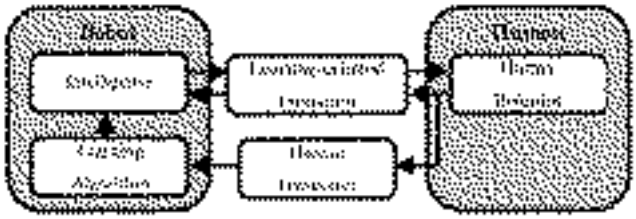


Fig. 1 Learning-oriented Interaction

行動の学習が行われないことがあることを指摘した[2][4]. すなわち人間による教示には特性が存在し,それが原因で従来の強化学習アルゴリズムでは学習が進まないことがある.

そこで以下では,ロボットの行動学習に影響を与える人間の教示の特性を示し,その教示特性に基づいたロボットの行動学習アルゴリズムを提案する.

3. ロボットの行動学習における人間の教示特性

3.1 強化学習法

強化学習法[6]は「報酬」と呼ばれる環境状態の評価情報により,エージェントの置かれる状態と実行行動の対応付けを調整する方法論であり,環境とのインタラクションにより環境から得られる「報酬」を最大化するように学習を進める. 人間とのインタラクションから行動を学習するロボットを扱う本研究では学習段階における人間からの教示を分析するため,追加学習が容易に行える強化学習法を採用する.

強化学習法で用いられる「報酬」という概念は, 通常あらかじめアルゴリズム中にエージェントが認識する環境の状態と環境から得られる報酬値を対応付けておくが,本研究ではエージェントであるロボットに人間が好ましいと感じる行動をさせることを目的とするため, 報酬値は人間により決定されるものとする.

3.2 人間の教示特性

筆者らは学習指向型インタラクションのコンセプトに基づき, 実際に人間の教示に基づく顔ロボットの行動学習の実験を行ったが, 非常に低い自由度の行動学習にもかかわらず, 被験者の教示法によって学習が進まないことがあることを指摘した[2]. この原因は,顔ロボットの学習アルゴリズムに人間の教示特性が考慮されていないことにあることが明らかになった. すなわち, 人間が顔ロボットにある特定の行動を実行させることを望んだ場合, 人間は顔ロボットの学習目標の達成が部分的であってもそれが達成に向かっていれば正の報酬を与えることがある. それに対し, 通常の強化学習アルゴリズムを備えたロボットは正解行動への道筋を示す情報が入っている報酬を活用できずに学習が進まなくなることがある.

一般に,報酬という概念は人間社会における学習行動と密接にかかわっており,特に教育心理学の分野では様々な研究がなされている. J. S. Bruner [5]は人間の学習における本質的特徴として, ①検証手続きの定式化, ②検証手続きの実施, ③検証結果と何らかの基準の比較という手順があることを指摘している. ここで③が行われるときには修正の知識が必要であり, さらに目的の達成に向かっているかどうかという指針があることが望ましく, 教授

者はこれらを学習者に示すことが望ましいと述べている. このように人間社会には段階的に報酬を与えることで学習目標への道筋を示す教示方法が存在することから, このような情報を活用する意義は大きいと考えられる.

そこで本論文ではこのように, 学習目標の達成に向かっているかどうかを基準に段階的に報酬を与える人間の教示方法を人間の教示特性として注目し, この人間の教示特性に基づいた行動学習アルゴリズムを提案し, 実際の人間の教示による行動学習実験でその有効性を示す.

4. 行動学習アルゴリズム

4.1 基本設定

上記の問題を考えるにあたって, *Q*学習の定式化に沿い以下のように問題設定を行う.

①ロボットにはセンサが備えられている. ロボットは1 STEPの間にクラスタリングされた一つの状態をセンサ入力から認識できるものとする.

②ロボットにはアクチュエータが備えられている. このアクチュエータを1自由度ごとにわけ, それぞれを「行動要素」と呼ぶ. それぞれの行動要素は1 STEPに一つの出力値を出力することができるものとし, 1 STEP に出力される行動は各行動要素出力値の論理積の形で表現できるものとする. あるSTEPに出力される行動はそのSTEP に認識された状態 *s* と *Q* 値から式（1）より求まる行動 *a* の選択確率 *P* を用いて確率的に選択される.

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

ここで*A*はロボットが実現できる行動の集合全体を表し, *T*はボルツマン温度と呼ばれる定数で, *T*の値は大きいほど行動はランダムに選択されるようになる. このようにして確率的に決定される行動を「実行行動」と呼び, この実行行動から各行動要素の出力値は一義的に決定できるものとする.

③人間の教示はスカラ値で, ロボットが1 STEPの行動を実行するごとに与えられるものとする.

強化学習の土台となる *Q* 値の更新式は

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

とする. ここで *s* はロボットが認識した環境の状態, *a* は実行した行動, α は学習率, *r* は人間の教示による報酬値とする.

通常の *Q* 学習[6]では報酬値 *r* に遅れ報酬値を加えるが, 本論文では用いないこととする. 前章で述べた人間の教示特性に基づいて人間がロボットに報酬を与えた場合, その報酬は通常の *Q* 学習の観点からは「誤報酬」となる可能性が高い. すなわち, 人間は前回のロボットの行動と比較評価して今回の行動が正解行動に近づいたかどうかという視点で報酬を与える傾向があるため, *Q*学習の観点から, ロボットは一つの状態に対する矛盾した報酬を与えられたと解釈してしまう. そこで本論文で提案する人間の教示特性に基づいた学習アルゴリズムでは, 人間の教示特性を利用して1 STEP当たりに人間がロボットに与える報酬をロボットの行動強化に有効に利用することを検討する. したがって複数STEP間での状態遷移を利用する遅れ報酬は議論の対象外であると考えられるため遅れ報酬は用いないこととする. しかし1 STEPにおける人間からの報酬の有効利用は, 長期にわたる学習STEPに対しても有効であるとともに, *Q*学習などの遅れ報酬のメカニズムを用いた学習アルゴリズムにも適用できると予想される. したがって, 必ずしもマルコフ決定過程として表されない人間とロボットのインタラクションに *Q* 学習を適用するアプローチにとってもこの議論は重要であると考えられる.

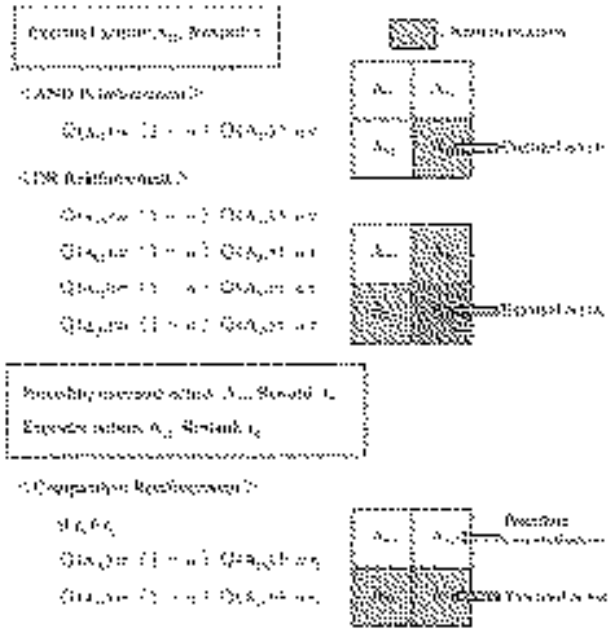


Fig. 2 Proposed reinforcement method for each learning algorithm (case of 2 degrees of freedom)

遅れ報酬を用いない枠組みでは, 各状態に対する行動の強化は他の状態に対する行動の強化とは独立に議論できるため, 本論文では以降, 一つの状態に対する行動の強化について説明することとし, *Q*(*s*,*a*)は状態 s を省略して *Q*(*a*)と記す.

4.2 行動学習アルゴリズム

上記の問題設定のもとで人間の教示特性を考慮に入れた学習アルゴリズムとして,「AND強化型学習アルゴリズム」を比較のための基準アルゴリズムとし,「OR 強化型学習アルゴリズム」および「比較強化型学習アルゴリズム」を提案する. 以下に各学習アルゴリズムの内容を示す.

＜ AND 強化型＞

この学習アルゴリズムは以下に提案する二つのアルゴリズムとの比較のために用意する. 従来の強化学習と同様, 与えられた報酬を実行行動に対応する *Q* 値のみの強化に使う.

＜ OR 強化型＞

与えられた報酬を実行した行動に対応する *Q* 値だけでなく, 実行行動の行動要素を含む行動に対応する *Q* 値も同様に強化を行う. ただし, 実行行動に対応する*Q*値は行動の行動要素数分だけ強化される.

＜比較強化型＞

人間が与える報酬は学習目標の達成に向かっているかどうかという評価に基づいているという教示特性を利用して, 同じ状態の

1 STEP前の実行行動および報酬を記憶しておき，今回の実行行動および報酬と比較して強化を行う．すなわち前回と比較して報酬が変化した場合，1 STEP前と今回の行動要素のうちで変化した行動要素を含む行動を強化する．報酬が1 STEP前と変わらなかったときは今回の実行行動に対応する Q 値のみ強化する．

行動要素数が2で各行動要素の出力値が0と1の2種類の場合を例に挙げる（**Fig. 2**参照）．「AND強化型学習アルゴリズム」では，実行行動がA_{11}（サフィックスは行動要素の出力値を示す）の場合，すなわち行動要素の出力値がともに1の場合，各行動要素出力値の論理積の行動A_{11}に対応するQ値を強化する．それに対し「OR強化型学習アルゴリズム」では，実行行動の行動要素を含む行動，すなわちA_{01}とA_{10}とA_{11}に対応するQ値を強化する．「比較強化型学習アルゴリズム」では，1 STEP前の実行行動がA_{10}，今回の実行行動がA_{11}の場合，1 STEP前と今回の行動要素のうちで変化した行動要素を含む行動，すなわちA_{10}とA_{11}に対応するQ値を強化する．

5. シミュレーション

5.1 シミュレーション方法

提案した学習アルゴリズムの特性をシミュレーションを用いて検証する．シミュレーションでは，4.1の基本設定をもとにした仮想ロボットと，人間の教示特性に基づいた人間のロボット行動評価モデル（以降，行動評価モデルと記す）の間で仮想的なインタラクションを行わせ，学習アルゴリズムの挙動を検証する．仮想ロボットには複数の行動要素が備えられているものとし，各行動要素の出力値は2種類（1または0）であるとする．3種類の強化型学習アルゴリズムに対して，行動要素数は2，6，10の場合を用意した．一方，行動評価モデルが仮想ロボットに与えることができる報酬は○または×の2種類（それぞれ報酬値 ＋1.0，－1.0）であるとし，仮想ロボットが1 STEPの行動を実行するごとに与えられる．与える報酬の判断基準は，[2]での実験における人間の教示特性に基づいて，学習目標である正解行動に対して以下の二つを用意する．

①行動評価モデルA

1 STEP前の行動と比較して，正解行動と同じ出力値の行動要素の数が同じ，または増加した場合は○報酬を与え，減少した場合は×報酬を与える．ただし，すべての行動要素の出力値が正解行動と同じ場合には○報酬とし，逆にすべての行動要素の出力値が正解行動と異なる場合には×報酬とする．

②行動評価モデルB

行動要素の出力値がすべて正解行動と同じときのみ○報酬を与え，それ以外は×報酬を与える．

学習パラメータはボルツマン温度 $T = 0.1$，学習率 $\alpha = 0.1$ としてシミュレーションを行った．

5.2 結果および考察

Fig. 3は行動評価モデルAと行動評価モデルBに対して，学習ステップごとの正解行動に対応するQ値の推移を，行動要素数ごとに示している．すなわちこのQ値が高くなるほど，正解行動の学習が進んだことを示す．それぞれのQ値は10回の試行を平均した値である．

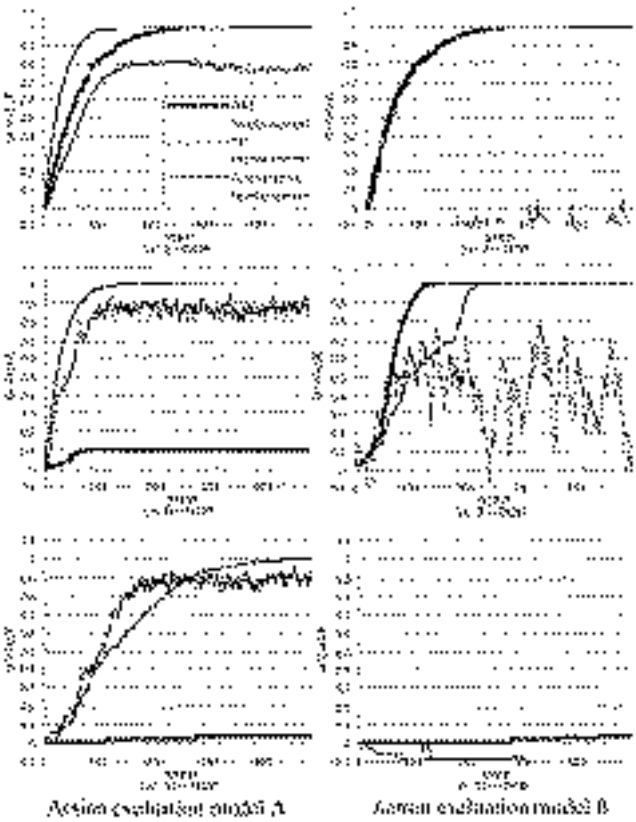


Fig. 3 Result of simulation

行動評価モデルAについては，特にFig. 3の(c)と(e)において，AND強化型と比較してOR強化型と比較強化型の学習アルゴリズムによる正解行動に対応するQ値は学習の早い段階で増加している．これは行動要素数が増えるにつれて正解行動が選択される確率は非正解行動が選択される確率に比べて相対的に低くなることに起因する．すなわちAND強化型では正解行動に対応するQ値が強化されないが，一方でOR強化型と比較強化型による学習では，行動要素の一部が正解行動と同じである，部分的な正解行動が選択されたときに与えられる報酬値が正解行動に対応するQ値の強化に使われ，結果として正解行動の学習が早く収束することを示している．

また，行動評価モデルBに対する学習においては，(b)と(d)と(f)から分かるように，OR強化型の学習では正解行動に対応するQ値は振動的であるか，または負に強化されている．これはOR強化型の学習では正解行動を負に強化してしまう性質によるもので，正解行動が選択されたときには正解行動を強化し，同時に正解行動に近い行動が選択される確率が上がるが，その結果として正解行動に近い行動が選択された場合には，報酬値が負であるために，正解行動も含めて負に強化してしまうため，振動的な学習が行われる．一方，AND強化型と比較強化型については正解行動を負に強化することがないため行動要素数が6までは着実に正解行動の強化が行われている．

このシミュレーションの結果より人間の教示特性に合った学習アルゴリズムを検討する．人間の教示特性には行動評価モデルAと行動評価モデルBの両方がありうることから，両者に対して共

通に速い学習の収束が期待できる比較強化型学習アルゴリズムが適していると考えられる. しかし, 本シミュレーションでは正解行動が厳密に定義されていること, および行動評価モデルによる報酬値には完全に誤りがないことなどを考えると, 本研究で扱うような曖昧性の強い行動の学習には実際の人間とのインタラクションを用いた実験による学習アルゴリズムの検証が必要であると考えられる.

6. 人間とのインタラクション実験

6.1 実験方法

本実験は実際に人間とロボットのインタラクションの中で人間による教示が行われた場合に, 提案した学習アルゴリズムが正解行動の学習に有効であるかどうかを検証することを目的とする. 各学習アルゴリズムの特性の相異はシミュレーション実験により行動要素数が大きくなった場合に顕著に現れるという知見が得られたので, 本実験では行動要素数が大きい, 顔ロボットによる表情の学習を問題として扱う. 本実験は, 人間とのインタラクションからロボットが行動を学習するという枠組みの中の一例として取り上げる.

顔ロボット[3]は, 人間と表情などのnon–verbal communicationを含んだインタラクションを行うことを目的として開発されたロボットで, 人間の頭部動作を模倣できる（**Fig. 4**）. 構造は大きく分けてフレーム部, 頭蓋部, 皮膚部からなり, フレーム部には最大18本のACDISと呼ばれる，それぞれ1自由度の伸縮動作をする空気圧アクチュエータを備えている. このアクチュエータの一端を皮膚と接合し, 伸縮動作をさせることで様々な表情を表出することができる. さらに首と眼球にモータなどのアクチュエータを備え，全部で24自由度を持つ.

本実験ではシミュレーションで検証した10行動要素数までを扱うために，一つのACDISを1行動要素として，10行動要素数分のACDISを選び，表情の学習実験を行う．顔ロボットは1STEPに10個のACDISそれぞれに対し1（伸）または0（縮）のどちらかの出力値を決定する．この10個のACDISの出力値の組み合わせによりそのSTEPにおける顔ロボットの表情が決まる．一方，被験者にはあらかじめ学習目標の表情を伝えておき，STEPごとに表出する表情が目標の表情にみえるかどうかを○，△，×で評価してもらう（報酬値はそれぞれ＋1.0，0.0，－1.0）．本実験では行動要素数を大きくしたことにより, 行動要素数が少ないときよりも正解行動を実行する確率は相対的に小さくなるが, その一方で正解行動に近づいたかどうかが微妙であるような細かい表情の変化が可能である．そのため行動評価モデルAと同じような行動の評価をする被験者が正解行動に近づいたと判断する行動の差が小さい場合に行動評価の違いが良く現れるように, 中間的な評価として△報酬を用意した. 学習目標となる表情は，表情としての特徴が大きい「幸福」の表情に対して行い，学習対象となるACDISの選定にあたっては幸福の表情の表出にかかわるもののほか, 表情の表出に強くかかわるものを用いた. 選んだACDISとその制御点の対応を **Fig. 5** に示す.

実験はAND強化型，OR強化型，比較強化型の各学習アルゴリズムに対してそれぞれ100 STEPずつ，5人の被験者（大学生男女）に対して行った. それぞれの学習パラメータは被験者の負担を小さくするため, シミュレーション実験よりも早く学習が収束するように学習率$\alpha = 0.3$，ボルツマン温度$T = 0.1$とした．本実験ではQ値の更新式として式（2）を用い，状態数は1とする．したがって, 状態遷移による報酬の伝播などは考慮せず, 被験者である人間が目的とする「幸福」表情を被験者からの直接的な報酬によってのみ学習する.

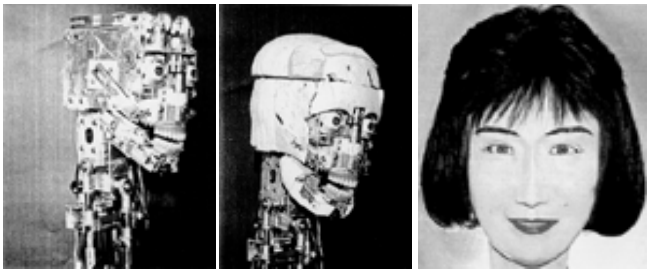


Fig. 4 Appearance of face robot

Fig. 5 Control points of facial expression and their ACDIS numbers

6.2 実験結果および考察

（1）評価方法

人間の教示特性を調べるために, まず各被験者にとっての学習目標である正解行動を調べる．すなわちどのようなACDISの出力値の組み合わせが各被験者にとって幸福の表情に見えるのかを調べ, その出力値の組み合わせを正解行動とする. 本実験では正解行動となるACDISの出力値の組み合わせを，被験者が与えた報酬分布から推測する. まず全実験を通じて各被験者が○報酬を与えたときの各ACDISの出力値を調べ，それを基に各被験者にとっての正解表情を定める. 本実験では，STEP毎にACDISの出力値は0（縮），1（伸）であることから，各ACDISが0を出力したSTEP数と1を出力したSTEP数を，全STEPを通じて数えることができる．そこで，各被験者が○報酬を与えた場合のみの0と1を出力したSTEP数をそれぞれ求め，それを○報酬を与えた全STEP数で割った値を **Table 1** に示す.

Table 1から，各被験者が○報酬を与えた場合の各ACDISの出力値は1を出力した割合の方が大きい場合と，0を出力した割合の方が大きい場合と，どちらでもない場合があることが分かる．各被験者が○報酬を与えた場合に1または0を出力した割合が大

Table 1 Outputs of each control point for +1–rewards given

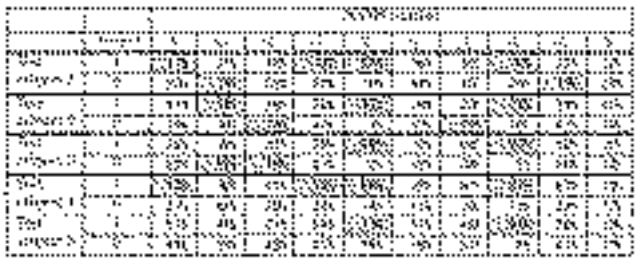


Table 2 Correct facial expression for each test subject

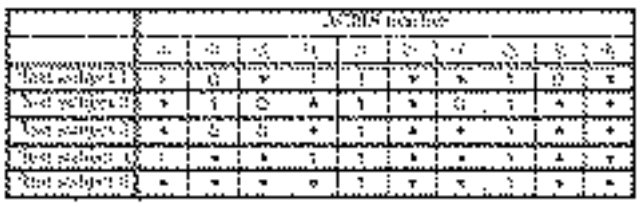


Table 3 +1–reward rate for correct facial expressions and –1–reward rate for incorrect ones

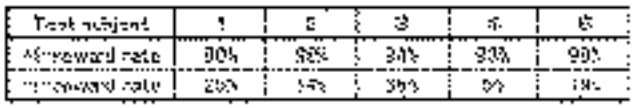


Table 4 Reward distribution in case of approaching to correct facial expression

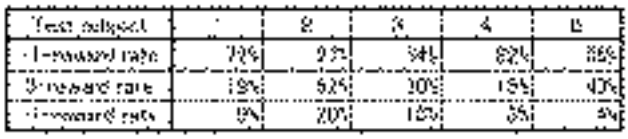
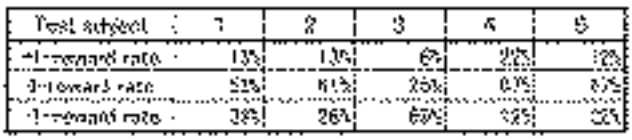


Table 5 Reward distribution in case of leaving from correct facial expression



きいACDISは，顔ロボットが表出した表情を各被験者が正解行動である幸福の表情と判断するために大きな影響を与えるACDISであると考えられる．そこで，Table 1を基に各ACDISの各出力値のSTEP数が70%を超えたACDISの出力値を正解表情と定義し，いずれの出力値も超えなかったACDISに対してはその被験者にとっての幸福の表情の正解表情には影響のないACDISであるとする．**Table 2**は以上から求められた各被験者の正解表情である．＊はその被験者が正解行動とする顔ロボットの幸福の表情には影響を与えない行動要素であることを示し，1または0と示されているACDISは，正解行動となる出力値が定義されているすべてのACDISに対してその出力がなされた場合に正解行動であることを示す．Table 2から被験者によって正解行動の行動要素数は異なり，被験者5の正解行動行動要素数は2，同じく被験者3と4は4，被験者2は5，被験者1は6となった．以降は，これらのACDIS出力値の組み合わせを各被験者の正解行動とする．

（2） 人間の教示特性の検証

提案する行動学習アルゴリズムの有効性を分析するために，まず本論文で指摘する人間の教示特性の検証を行う．

Table 3は顔ロボットが正解行動を実行した際に各被験者が○報酬を与えたSTEP数を顔ロボットが正解行動を実行した全STEP数で割った値と，不正解行動を実行した際に×報酬を与えたSTEP数を顔ロボットが不正解行動を実行した全STEP数で割った値を示す．各被験者とも顔ロボットが正解行動を実行した際には90%以上の割合で○報酬を与えているが，不正解行動を実行した際の報酬には個人差があるものの，正解行動を実行した際の○報酬を与える割合と比較して全体的に低い値となっている．これは非正解行動が実行されたときでも各被験者が○または△報酬を与える割合が多かったことを示しており，本論文で提案する学習アルゴリズムが注目する部分的な正解行動の実行に対しても○または△報酬による教示が行われたことを示している．

次に各被験者が与える○報酬が正解行動に近づいたことがきっかけで与えられているかどうかを検証するために，表情の出力が正解の表情に近づいたときの報酬値の分布を調べる．ここで正解の表情に近づいたかどうかを客観的に測るための基準を以下のように定める．すなわち，それぞれの被験者に対して決めた正解表情における各ACDISの出力値と各STEPにおける各ACDISの出力値を比較し，一致している出力値の個数を「正解表情からの距離」と定義する．本実験では10個のACDISを用いたことから，正解表情を出力したときの正解表情からの距離は0となる．また正解表情からの距離の最大値は10となる．

この基準に基づいて1 STEP前の距離と今回の距離を比較し，距離が小さくなったときの報酬値の分布を調べ，その結果を**Table 4**に示す．Table 4は距離が小さくなった場合に○，△，×の報酬を与えたSTEP数を距離が小さくなったSTEP数全体で割った値である．**Table 5**は同様に距離が大きくなった場合の結果である．被験者1，3，5は距離が小さくなった場合には○報酬を与える割合が高く，距離が大きくなった場合には×報酬を与える割合が高い傾向にあることが分かる．被験者4は全体として○報酬を与える割合が高いことが予想されるが，いずれにせよ距離が小さくなった場合には○報酬を与える割合が高く，距離が大きくなった場合には×報酬を与える割合が高い傾向にある．以上より各被験者とも与える報酬を正解行動からの距離の増減で判断している傾向が強いと考えられ，この顔ロボットの表情学習実験においても本論文で指摘する人間の教示特性が見られることが明らかになった．

（3） アルゴリズムの有効性

次にこれらの人間の教示特性が本研究で提案する行動学習アルゴリズムに対して有効であるかどうかを検討する．学習アルゴリズムの有効性には「学習ができるかどうか」，「学習が早く正解に収束するかどうか」という二つの評価基準が考えられる．本論文ではこの二つの評価基準の両方について有効性を議論することとし，前者の評価を「学習の可能性に関する有効性」と呼び，後者

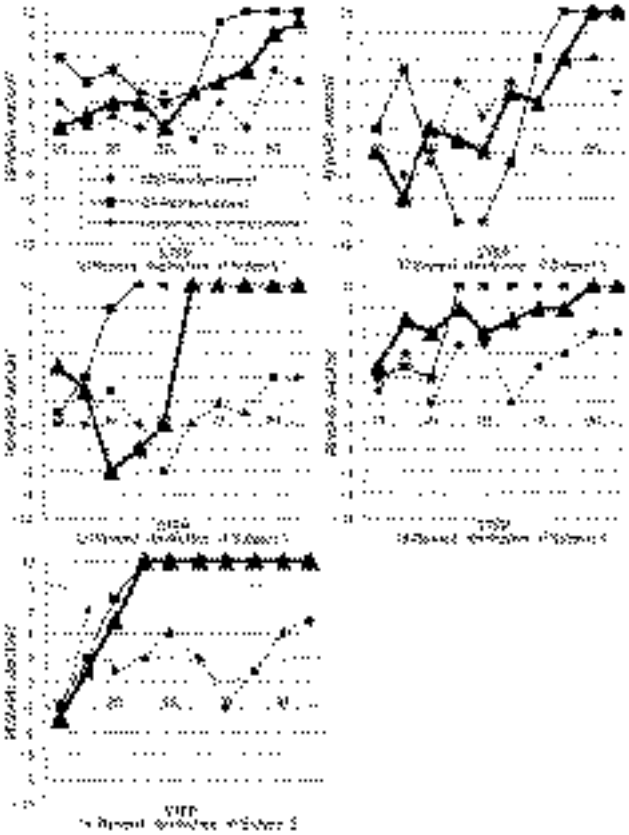


Fig. 6 Transition of rewards given by each test subject

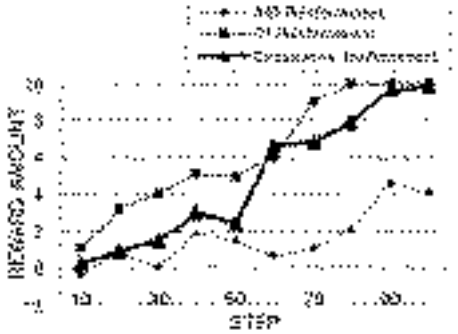


Fig. 7 Average transition of rewards given by all test subjects

Table 6 Average learning efficiency

を「学習の早期収束性に関する有効性」と呼ぶこととする.

Fig. 6 は各アルゴリズムに対する 100 STEP の学習過程で与えられた報酬値を10 STEPごとに和をとったものであり, 報酬値の和が10を示しているのは10 STEP中すべての顔ロボットの実行行動に対して被験者が○報酬を与えたことを示す. いずれの被験者に対してもAND強化型による学習では報酬値の和が10に満たないのに対し, OR強化型, 比較強化型では最終的にすべて10近くに至っている. このことから各被験者がOR強化型と比較強化型の学習実験では最終的に表出した表情を正解表情であると判断したと考えられ, OR強化型と比較強化型の学習アルゴリズムの学習の可能性に関する有効性が示されたといえる.

また学習過程においては, 被験者 2 を除いては, OR 強化型と比較強化型の学習では AND 強化型よりも比較的早い STEP 数で高い報酬値を得ている. 被験者2に関しては, 本論文で提案する学習アルゴリズムが注目する正解行動からの距離に応じた報酬の判断が十分に行われなかったために AND 強化型との差が出なかったものと考えられる.

Fig. 7には全被験者が与えた報酬値を10 STEP毎に和をとり被験者数で割った平均値を示す. また, **Table 6** の上 3 段にはその値を示し, 下 2 段には AND 強化型と比較した OR 強化型と比較強化型の1 STEPあたりに与えられた報酬値の増分を示す. OR強化型, 比較強化型は, AND 強化型と比較して高い報酬値を得ており, 特に 60 STEP 以降ではその差が開き, OR 強化型と比較強化型による学習において最終的には正解行動を実行していると評価していることを示している. 以上よりOR強化型と比較強化型の学習アルゴリズムは全体的にも学習の可能性に関する有効性が認められ, 学習の早期収束性に関する有効性についても, この二つの学習アルゴリズムはAND強化型と比較して有効であることが示された.

OR強化型と比較強化型を比較すると早期収束性という観点でOR強化型のほうが有効であるという結果になった. これは本実験では各被験者が正解行動からの距離の増減を基準に報酬を決めた傾向にあることから, シミュレーション実験で明らかになった行動評価モデルBに対するOR強化型の学習が振動的になる特性が起こりにくかったことが原因であると考えられる.

さらに, 各強化型の学習アルゴリズムに対する実験が終わった直後に被験者に以下のアンケートを実施した. アンケートは5人の被験者全員に対し, 3種類の学習型に対する学習が終了した時点で次に示す対を成す評価基準に従って, 5 段階で評価してもらった.

①顔ロボットは学習をしているように見える―学習をしていないように見える
②顔ロボットは評価に従う―評価に反抗的
③顔ロボットは教えやすい―教えにくい

5段階の評価は上記の三つの項目の左に近い評価から順にそれぞれ＋2, ＋1, 0, －1, －2 点に換算する. 5 人の被験者から得たこの得点の中央値を求め, その結果を**Fig. 8**に示す. AND強化型に対する回答ではいずれの質問項目に対しても低い得点であるのに対し, OR強化型と比較強化型に対する回答では高い得点を得ている. OR強化型と比較強化型を比較するとすべての質問項目に対してOR強化型のほうが得点が高く, 学習の早期収束性に関する有効性の結果と一致している. 各被験者への実験条件は学習アルゴリズムの違い以外は同様であることを考慮に入れると, 「学習の可能性に関する有効性」と「学習の早期収束性に関する有効性」が各被験者の心理的な負担にも反映されたものと考えることができる. このため被験者の心理的にもOR強化型がロ

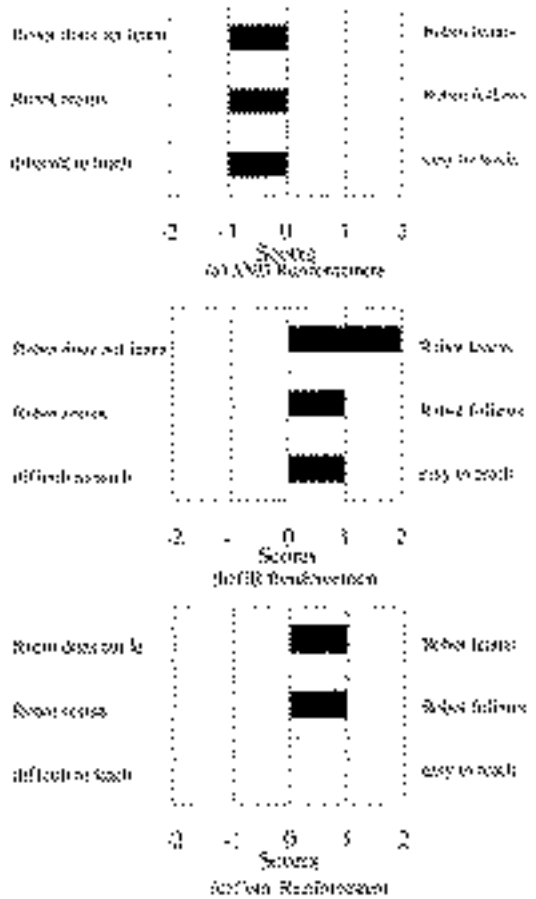


Fig. 8 Medians of scores answered by test subjects

ボットの行動学習アルゴリズムとして好ましいことを示していると考えられる.

7. 結論および今後の展望

本論文ではロボットが人間と, より親密なインタラクションを実現するための基本コンセプトとなる学習指向型インタラクションを提案した. さらに学習指向型インタラクションを実現するためのロボットの行動学習アルゴリズムには人間の教示特性を考慮に入れないとロボットは行動を学習できないことがある事実に基づき, 人間の教示特性を考慮に入れたOR強化型および比較強化型の学習アルゴリズムを提案した. またこれらのアルゴリズムの特性をシミュレーションで検証するとともに, 実際の人間と顔ロボットのインタラクション実験からその有効性を検証した.

人間と顔ロボットのインタラクション実験の結果から, 本論文で提案するOR強化型および比較強化型の学習アルゴリズムは従来のAND強化型学習アルゴリズムと比較して, 学習可能性, 学習早期収束性に関する有効性が実証され, これが人間の教示特性を考慮に入れていることに起因することを定量的に示した.

以上は顔ロボットの表情学習実験を対象に検証された結論であるが, 本実験より人間の教示に基づいてロボットが行動の学習を進める場合には, 人間の教示方法の一般的特性を考慮に入れることで, 行動の評価情報が少なくても, 高い自由度の行動を学習できる可能性があることが明らかになった. 今後は他のロボット行動学習問題に対しても学習アルゴリズムが有効であることの検証が課題として挙げられる. またより多くの被験者に対しての実験から本論文で指摘した人間の教示特性の詳細な検証も今後の課題として挙げられる.

参 考 文 献

[1] 中田, 佐藤, 森, 溝口 : “ロボットの対人行動による親和感の演出”, 日本ロボット学会誌, vol.15, no.7, pp.1068–1074,1997.

[2] F. Iida, M. Tabata and F. Hara: “Generating Personality Character in a Face Robot through Interaction with Human,” Proc. of 7th IEEE International Workshop on Robot and Human Communication, pp.481–486, 1998.

[3] F. Hara and H. Kobayashi: “State–of–the Art in Component Technology for an Animated Facerobot–Its Component Technology Development for Interactive Communication with Humans,” The Intl. Jour. of the Robotics Society of Japan Advanced Robotics, vol.11, no.6, pp.585–604, 1997.

[4] 飯田史也, 原文雄 : “人間の教示特性に基づくロボット行動学習アルゴリズムの提案”, 第16回日本ロボット学会学術講演会予稿集, pp.655–656, 1998.

[5] J.S. ブルーナー : 教授理論の建設. pp.74–78, 黎明書房, 1983.

[6] 浅田稔 :“ 強化学習の実ロボットへの応用とその課題 ”, 人工知能学会誌, vol.12, no.6, pp.831–836, 1997.

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

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

[9] 小林宏, 原文雄 : “ニューラルネットによる人の基本表情認識”, 計測自動制御学会論文集, vol.29, no.1, pp.112–118, 1993.

[10] 原田達也, 佐藤知正, 森武俊 : “触れ合いロボットによる心理効果—接触インタラクションによる安心感の演出と痛みの緩和—”, 日本ロボット学会誌, vol.16, no.5, pp.698–704, 1998.

飯田史也（Fumiya Iida）
1974年5月4日生. 1997年東京理科大学工学部機械工学科卒業. 1999年同大学大学院工学研究科修士課程修了. 同年, チューリヒ大学情報工学科博士課程入学. 現在に至る. マンマシンインタラクション, 機械学習, バイオロボティクスの研究に従事.
（日本ロボット学会学生会員）

原 文雄（Fumio Hara）
1941年4月7日生. 東京大学大学院工学系研究科博士課程修了（工学博士）. 1970年東京大学生産技術研究所講師. 1971年東京理科大学工学部助教授. 1983年同教授（現在に至る）. 1998年同大学工学部第一部学部長. 現在の研究分野や興味を持つテーマ : 知能機械システム学に関する知能と形態の研究, 群ロボットシステムにおける協調行動の創発や形態の自己組織化, クラスターロボットシステムにおける形態と知能の創発特性, 人と顔ロボットとの社会的インタラクションにおける高次知能の獲得など. ASME Fellow, IEEE（正員）, 日本機械学会（正員）, 計測自動制御学会（正員）等に所属.
（日本ロボット学会正会員）