コーチングによる報酬関数の動的生成に基づくエージェントの行動学習

Action Learning from Dynamic Reward Function updated by Coaching

廣川 暢一[1*]　鈴木 健嗣[1]
Masakazu HIROKAWA[1]　Kenji SUZUKI[1]

[1] 筑波大学大学院システム情報工学研究科
[1] Dept. of Intelligence Interaction Technologies, University of Tsukuba

Abstract: This paper describes a novel methodology, namely "Coaching", which allows humans to give a subjective evaluation to the agent in an iterative manner. This is an interactive learning method to improve the reinforcement learning by generating a reward function dynamically according to the learning situation of the agent. We will demonstrate that the agent can learn different reward functions by given instructions such as "good/bad" by human's observation, and can also obtain a set of behavior based on the aquired reward functions through several experiments.

1 はじめに

エージェントによる行動学習は，行動選択→評価→パラメータの調節という一連の流れを繰り返すことで行うが，行動の評価を行う評価関数は基本的にタスク依存であり，その設定によっては学習器の性能が大幅に変わることが知られている．しかしながら，行動選択やパラメータ調節の手法に比べ，適切な評価関数をどのように作成するかについての議論はあまり為されておらず，使用する学習手法やタスクに応じて評価関数を経験的に設定することが多い．

行動学習を実現する学習手法の代表的な例として強化学習がある [1]．強化学習は，時刻 t における状態 $s(t)$ での状態価値 $V(s)$ を，環境から得られる報酬 $r(t)$ に基づいて更新し，最終的に得られる報酬の期待値を最大化するような行動系列を試行錯誤的に求める学習手法である．

一般に強化学習における報酬は，状態を変数とした時不変な関数で与えられ，目標状態において最大になるように設計される．しかしながら実環境での学習を行う際には，常に妥当な条件で報酬関数を設定できるとは限らず，また報酬や環境のダイナミクスが全くの未知な場合，報酬関数が時間的に変化する場合，さらに報酬を得るまでに複雑/冗長な行動系列が必要な場合などにおいて，学習が不可能になる，もしくは学習の初期段階に非現実的な時間が掛かるという問題が指摘されている．

これらの問題に対し，杉本らは変化する環境や報酬に合わせて複数の状態と報酬の予測モデルを適宜切り替えながら学習制御を行う Combinational Model-based RL と呼ばれる手法を提案し，報酬関数が時間的に変化するような環境においても非線形システムの制御則を学習できることを示した [2]．また尾川らは，強化学習において割引率と呼ばれるメタパラメータを,「信頼度」という学習の進度を表わすパラメータによって適当に調節することで，強化学習の学習効率が大幅に改善されることを示した [3]．しかしながら上記のいずれの手法においても，最初の報酬を得るまでに時間が掛かる，あるいは報酬がほとんど与えられないような状況は想定されておらず，実環境での学習に適用するためには解決すべき問題が残っているといえる．

本研究ではこのような問題の解決策として，人間の教示者による報酬関数の操作という観点から，エージェントの学習に対してコーチングという人が適切にエージェントの学習過程に介入する手法を用いる．ここでは人による主観的かつ抽象的な評価に基づいて，エージェントが報酬関数を調節することで学習を誘導・促進する手法を提案する．さらに，単純な強化学習エージェントと組み合わせることで，従来の強化学習では学習困難なタスクを達成できることをシミュレーションおよび実機実験によって示す．

*連絡先：筑波大学システム情報工学研究科
〒 305-0006 茨城県つくば市天王台 1-1-1
E-mail: hirokawa@ai.iit.tsukuba.ac.jp

2　コーチングに基づく報酬関数の動的生成

2.1　コーチング

システムの制御を行うためには制御対象の動特性も含めたモデリング不可欠であり，制御対象が複雑であればあるほど，その作業に多くの時間が費やされる．また，そのような複雑なモデルを制御するために必要な方程式を解くために，システムの自由度や軸配置に関する制約が加えられることも少なくない．強化学習に代表される行動学習アルゴリズムは，そのような設計者の負担を減らすために広く用いられており，状態空間と報酬関数を適切に設定することで，学習の過程を考えることなく実機に実装できることが特徴の一つである．

しかしながら，その場合でも評価関数は設計者の知見に基づき与えられるのみで，その評価関数の妥当性は学習の評価とは切り離して考えられている．そのため評価関数を動的に調節することにより，設計者がエージェントの学習過程にも陽に介入することで，学習効率が改善できるのではないかと考えている．

コーチングとは，人間がエージェントの行動を観察しながら，直接主観的な評価を与えていく対話的な学習手法の枠組みである．そこで，予め学習器に評価関数ではなく教示を解釈する仕組みを組み込み，実際にシステムを動かしながら任意の状態にエージェントの行動学習を収束させることを目的とした，人-ロボットのインタラクションの一形態であるといえる．

中谷らはコーチング法を提案し，ヒューマノイドのバランス制御や歩行動作の獲得を行わせることにより，評価関数の動的な調節による行動学習の可能性を示した．Riley らは人間の主観的であいまいな教示によりヒューマノイドの行動を洗練させる手法を提案している [4][5]．また我々も，コーチングと強化学習を組み合わせることで，事前にモデリングすることなく対象の実機を用いた行動学習が可能であることを報告している [8]．なお教示を与える人間の評価方法の違いによって，ヒューマノイドの学習性能や最終的な学習結果に違いが生じることも，コーチングの興味深い特徴の一つである．

このように人間の主観的な評価を積極的に取り入れる手法として高木らの提案した IEC(Interactive Evolutionary Computation) があるが，IEC では進化計算に必要な評価を常に評価者が与え続けなければならないため，評価者に負担が掛かりすぎることなどが指摘されている [6][7].

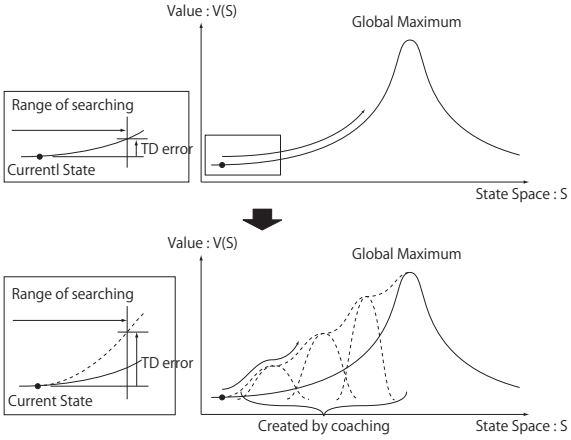

図 1: Concept of proposal method

2.2　報酬関数の動的生成

一般的な強化学習では，エージェントは初期状態からランダムに探索を行い，報酬を得た時点で学習を開始する．しかしながら，初期状態から探索可能な範囲であり，かつ有限時間内に報酬を得られる確率が非常に低い場合，学習効率は著く低下する．そのため，初期状態からでも比較的報酬を得やすいような報酬関数を設計するが，行動系列を学習するための状態価値関数を得るためには，適切な評価関数の設計が不可欠であるといえる．そこで本研究では，図 1 に示すようにコーチングによって状態空間上に短期的な報酬関数を連続的に生成することで，本来報酬の手掛かりを全く持たない状態からでも目標を状態指向する学習を行わせるための手法を提案する．このような報酬関数の動的生成においては，人間の教示を正しく解釈し，かつそれに基づき関数生成を行う枠組みが必要である．その詳細を以下の節で述べる．

2.3　人間の教示特性

本手法では，学習中のエージェントに対して人間がオンラインで教示を与える．そのためエージェントの行動から，その行動に対する教示の入力までの間に必然的に時間遅れが生じる．そこで予備実験として，人間による教示が対象となるエージェントの行動を観測した時点から，どの程度の時間遅れを伴って入力されるのか測定を行った．実験方法はエージェントが行うランダムな行動の中から，ある特定の動作パターンを観察したときに入力を行うこととした．この実験を 4 人の被験者によって行った結果を図 2 に示す．グラフ右端の時刻 0 が被験者による入力が行われたときを示す．またそれぞれのグラフは，対象動作からの時間遅れの平均と分散を表わす．その結果，いずれの被験者においても対象動作の観測から 1 秒以内に教示動作を行っ

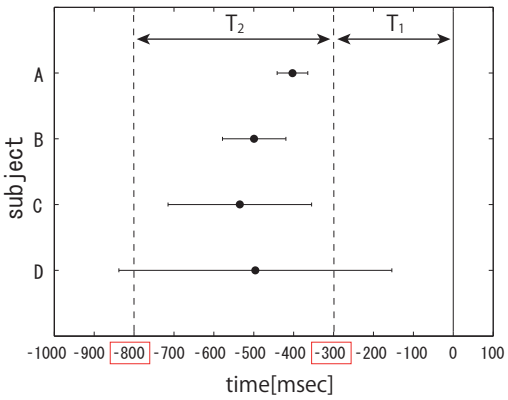

図 2: Delay time of human evaluation

ており，またその分散も 500 ミリ秒程度である．よって，人間の教示特性を表わす時定数として図中 T_1, T_2 で示した値を用いる．

2.4 強化学習への実装

次にコーチングによって実際に報酬関数を作成する方法を述べる．N 次元の状態空間における時刻 t での状態ベクトルを $\boldsymbol{x}(t) = {}^t[x_1(t), x_2(t), \cdots, x_N(t)]$ とする．このとき，時刻 t_0 において教示が与えられたとすると，予備実験で求めた時定数 T_1, T_2 を用いて，教示の対象である行動行列の可能性が高い状態ベクトル

$$\boldsymbol{X} = \{\boldsymbol{x}(t) | t_1 \leq t \leq t_2\}$$
$$t_1 = t_0 - (T_1 + T_2), \quad t_2 = t_0 - T_1$$

を抽出する．その後，この状態ベクトルの集合 $\boldsymbol{X}$ を入力とし，EM アルゴリズムによる混合正規分布モデルを用いた最尤推定を行うことで，得られる確率分布 $P(X|\theta)$ を定め，これを教示によって与えられた評価関数，すなわち新たな報酬関数とみなす．

さらに，人間からの教示が必ずしも正しいと保障されないため，エージェントは与えられた一連の教示に一貫性があるかどうかの判断を行う．もし類似した状態系列において連続して教示が与えらた場合は，それらの教示には一貫性があると判断し，逆にまったく異なる状態において教示が与えられたなら，そこに関連性は無いと判断する．本手法では，過去 C 回分のコーチングにより与えられた教示に基づく確率分布の積集合を取ることで判断するものとした．

$$r_{ev}(\boldsymbol{x}, t) = \min_{i \in C} e^{-\frac{\tau}{T}} P_i(\boldsymbol{x}|\theta) \tag{1}$$

ここで，τ は教示が与えられてからの経過時間，T は減衰時定数を表わす．このように求めた r_{ev} を報酬関数として，従来の強化学習の更新式に外挿する．

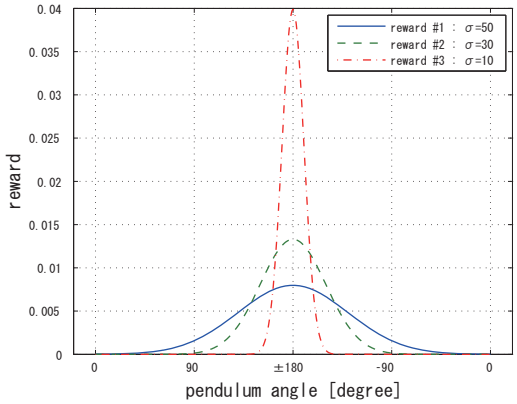

図 3: Reward functions

$$TDerror = r + r_{good} - r_{bad} + \gamma V(s_{t+1}) - V(s_t) \tag{2}$$
$$V(s_t) \leftarrow V(s_t) + \alpha TDerror \tag{3}$$

このようにして環境からの報酬 r が得られない状態においても学習を行うことができる．また，十分な時間経過後には，

$$e^{-\frac{\tau}{T}} P_i(\boldsymbol{x}|\theta) \to 0, \quad (t \to \infty) \tag{4}$$

となるため，学習は予め定めた報酬関数に収束することになる．

3 シミュレーション実験

3.1 実験設定

ここでは 1 リンクの振り子を，鉛直下向きを初期状態として規定時間内に一定角度以上に振り上げるようなタスクを考える．倒立振子の振り上げは非線形制御システムの代表的な課題であり，タスク達成には振り子の周期に合わせた往復運動を行う必要があり，比較的長い行動系列が要求される．また，システムがダイナミクスを持つため遅延報酬による入出力関係の学習が困難な課題である．

今回の実験には，図 3 に示すように 2 種類の正規分布に基づく報酬関数を用意した．これらの報酬関数を用いて，それぞれコーチングを行った場合と行わなかった場合について比較実験を行う．なおコーチングを行う場合は，教示を与えるのは最初のタスク成功時までとし，それ以降は教示を与えないこととする．

3.2 結果

本実験では，振り子の角度を $\theta(t)$，角速度を $\omega(t)$ として，状態変数を $\boldsymbol{x}(t) = {}^t[\theta(t), \omega(t)]$ と定めた．状態

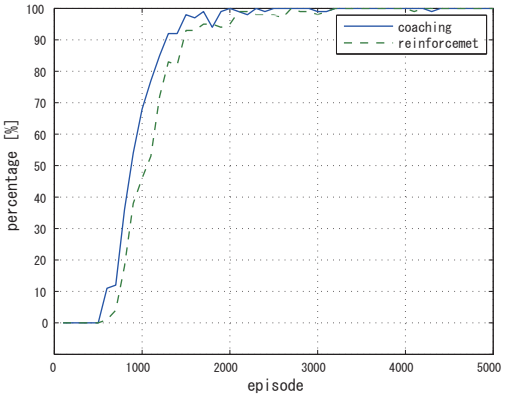

図 4: Success rate : reward # 1

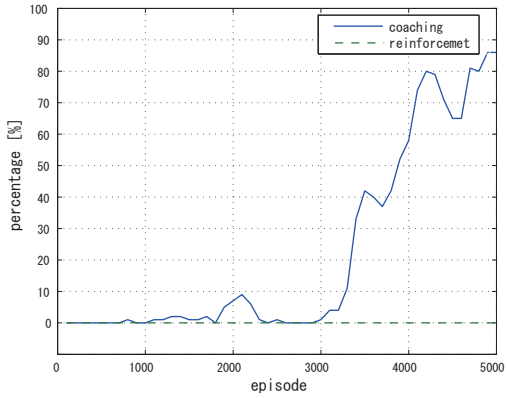

図 5: Success rate : reward # 3

空間は，角度を 36 分割，角速度は 20 分割することで離散化し，状態数は 720 である．また，学習係数は学習率を α，割引率を γ とし，$\alpha = 0.01, \gamma = 0.1$ で時不変とした．これらのパラメータにより，1 回の試行時間を 10 [s]，総試行回数を 5000 回として実験を行った．

はじめに図 3# 1 の報酬関数を用いた実験結果を図 4 に示す．横軸が試行回数，縦軸が 100 試行ごとの成功率を表す．この報酬関数ではいずれの場合も同程度の試行回数でタスクが成功し，その後ほぼ同じ時間で共に学習が収束していることが分かる．これは，# 1 の報酬関数ではタスクの初期状態から目標状態までなだらかに報酬値が与えられるため，どちらも同程度の速度で学習できたものと思われる．

次に，図 3# 3 の報酬関数を用いて実験を行った．この報酬関数は，目標状態である振り子が倒立した状態から，約 ± 30 °の範囲内に到達しない限り報酬が得ることができない．今回の実験設定では運動方程式によって状態の遷移方向が限定され，摩擦によって常に中心方向へ向かう行動が生成されやすい．この状態空間を広く探索するためには，振り子の慣性を利用しながら往復運動を繰り返すような行動政策が必要であるため，ランダムな行動選択による探索可能範囲は非常に限られている．強化学習による行動学習は，学習開始直後はランダムな行動政策によって状態空間の探索を行い，報酬に辿り着いた時点でその情報を手掛かりとして学習を進めていく．そのため今回の報酬関数# 3 のようにランダムな行動選択によって報酬を発見できる可能性が非常に低い場合，行動学習は非常に困難になる．そこで，コーチングによって探索可能な範囲内に短期的な報酬関数を生成し，それによる探索範囲の拡大とともに徐々に目標状態に向かって報酬関数を変更していくことができれば，このような状況においても行動学習が可能となるはずである．

実験結果を図 5 に示す．強化学習のみで学習を試みた場合では，一度もタスクを成功することができなかった．一方コーチングによる教示を行った場合では，約 700 回程度の試行で最初にタスクを達成しており，最終的には成功率が 90% 近くに達した．この結果は，コーチングによって教示者がエージェントの学習状況に合わせて報酬関数を動的に設定することができることを示している．このように本提案手法は，今回実験で設定したような状況において，強化学習のような評価関数ベースの行動学習を行う学習手法を適用するために有効であると考える．

4 考察

4.1 状態価値関数の推移

コーチングによって，どのように学習の誘導が行われているかについて，状態価値関数の遷移から考察を行う．図 6 に，学習時にコーチングを行った場合での状態価値関数の時間変化を示す．各グラフにおける等高線は，その時点での状態価値を表す．

図 6 の上段に示す学習開始直後では，初期状態 (原点) を中心として状態価値が急峻に落ち込んでおり，そのまわりの状態価値が高くなっている．これはコーチングによって初期状態付近で負の評価を与え状態価値を下げるとともに，周囲の状態価値を正の評価によって上げることで，振り子に加振を行う動作を学習させ，状態が原点付近に留まり続けるのを避けることを目的とした人間の意図に依るものである．その後は，エージェントの学習経過に合わせて徐々に等高線の頂上が目標状態の方へと遷移している．また，原点付近の状態価値は時間の経過によって報酬関数が減衰するため徐々になだらかになっていき，図 6 下段に示すように，最終的に予め設定した報酬関数に収束している様子が見て取れる．このことから，今回の実験において図 1 に示したような学習の誘導が行われていると言える．

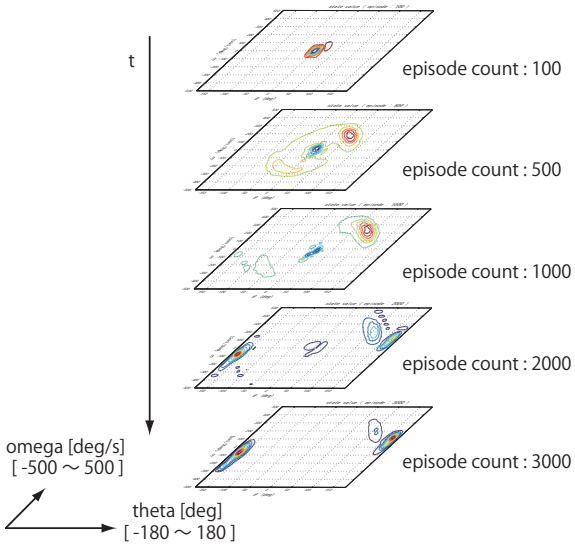

図 6: Transition of state value

4.2 教示入力の推移

続いて，教示者による教示と本手法によって作成された報酬関数についての考察を行う．図 7 左の散布図は状態空間上で教示が与えられた点を表し，右の等高線図はその時に作成された報酬関数を表す．この図から，先ほど述べたように教示者によって振り子の振幅が徐々に大きくなるような報酬関数が作成されていることが分かる．また，複数回の教示から確率密度の積集合をとる処理は，必要以上に報酬関数が広がることを防ぎ，時間経過によって減衰することで以前の教示情報が後の学習の妨げにならず，常に単峰的な学習となっていることが考えられる．

5 実機実験

5.1 実験設定

最後に，本研究で提案した手法を実際のロボットに実装し行動学習を行った．ここでは 6 自由度のロボットアームを用いた (図 8)．学習タスクはシミュレーション同様に倒立振子の振り上げとし，図 8 のようにロボットアームの手先に振り子を取り付け，肘を直角に固定した後，上腕を旋回させることで加振を行う．状態変数，強化学習の学習係数，コーチングにおける時定数およびタスクの試行時間等はシミュレーションと等しくした．なお，学習時間の短縮のため，状態空間の分割数をシミュレーション時の半分とした．

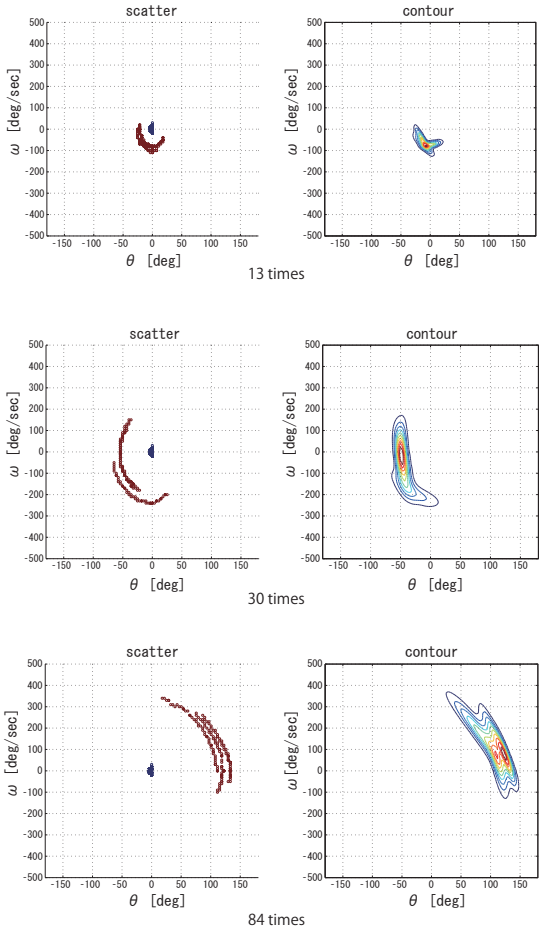

図 7: Transition of reward function

5.2 結果

以上の設定で，図 3:#2 の報酬関数を用いた時の，強化学習と提案手法による比較実験の結果を図 9 に示す．シミュレーションで得られた結果と同様に，強化学習では 100 回の試行において一度もタスクを達成することができなかったが，提案手法ではシステムが行動を獲得できていることが分かる．ここで提案手法において人間が教示を与えた回数は，学習初期段階で 20 回程度である．このように，提案手法は，教示を与える状況や時刻の制限がなく，行動の観察により自由に簡便な支持を与えることができることを示しており，人間の負担を大きく軽減するという目的にも合致していると考えられる．

6 むすび

本研究では，コーチングという，人間の主観的な評価によって学習を行う手法を用いて，人間の教示者が報酬関数を動的に作成することでエージェントの学習を促進させる手法を提案した．それによって，従来の

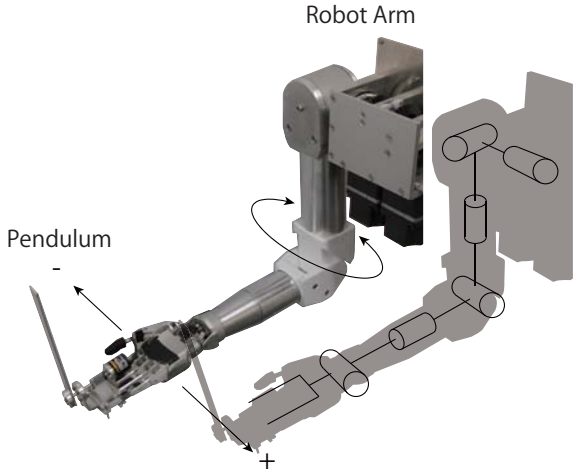

図 8: Experiment environment

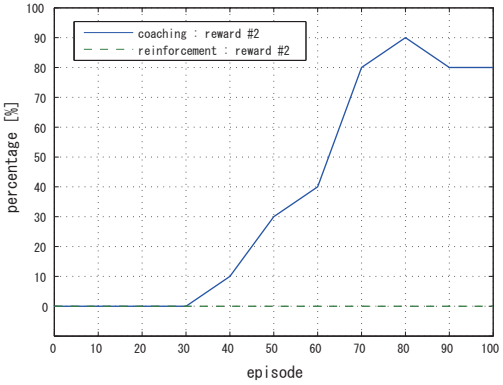

図 9: Success rate : real robot

強化学習手法では学習困難なタスクにおいても，効率的に行動学習が行えることをシミュレーションおよび実機により示した．

本手法におけるコーチングによるエージェントの学習の誘導という行為は，人間の教育者が学習者に対して学習を行わせる過程において，学習者の到達レベルに合わせて学習の難易度を段階的に上げていく行為に似ている．養育者によるタスク難易度の調節によって学習効率が改善されることは，ヒューマノイドが養育者との間に共同注意を獲得するというタスクにおいて，長井らによって示されている [9]．

また，今回の実験において，「振り子の振幅が大きくなるように教示を与える」という政策は人間の教示者によって暗示的に導かれたものである．冒頭で述べたように，評価関数は一般的に設計者によって経験的に設計されることが多い．このことから，少なくともヒューマノイドなど自身と似た身体性を持つエージェントの行動学習において，人間がタスクの達成に効率的な学習戦略を直観的・経験的に導くことができることを示していると考えられる．さらにそれは人間のサブゴールを発見する能力とも密接に関連していると考えられるため [10]，こうした人間の知見を適切に学習過程に反映させる枠組みを用いることで，エージェントによるサブゴールの発見も含めた行動学習が実現できるのではないかと考える．

今後は，異なるタスクや異なる学習手法を用いて本提案手法の有効性を検証したい．また今回は一定としていた時間経過による減衰率や教示の有効回数についても，エージェントが学習状況に応じて調節していくなど，既存の学習手法への適用方法の改善も行っていきたいと考えている．

参考文献

[1] Richard S. Sutton, Andrew G.Barto, “Reinforcement Learning : An Introduction,” *A Bradford Book,* The MIT Press, 1998.

[2] 杉本 徳和, 鮫島 和行, 銅谷 賢治, 川人 光男, “複数の状態予測と報酬予測モデルによる強化学習と行動目標の推定,” *Trans. of the Institute of Electronics, Information and Communication Engineers. D-II J87-D-II(2),* pp.683-694, 2004

[3] 尾川 順子, 並木 昭夫, 石川 正俊, “学習進度を反映した割引率の調整,” *IEICE technical report. Neurocomputing 102(628),* pp.73-78, 2003

[4] M.Nakatani, K.Suzuki, S.Hashimoto, “Subjective-Evaluation Oriented Teaching Scheme for a Biped Humanoid Robot,” *Proc. of IEEE-RAS Intl. Conf. on Humanoid Robots,* 2003.

[5] M.Riley, A.Ude, C.Atkeson, G.Chang, “Coaching: An Approach to Efficiently and Intuitively Create Humanoid Robot Behaviors,” *Proc. of IEEE-RAS Intl. Conf. on Humanoid Robots,* pp. 567-574, 2006.

[6] H.Takagi, “Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation,” *Proc. of IEEE,* Vol89, No.9, 2001.

[7] D.Katagami, S.Yamada, “Real Robot Learning with Human Teaching,” *IEEE International Wortkshop on Robot and Human Interaction,* pp. 258-263, 2000.

[8] 廣川 暢一, 鈴木 健嗣, “コーチングに基づくロボットのオンライン行動学習,” *ROBOMEC,* 2009.

[9] 長井 志江, 浅田 稔, 細田 耕, “ロボットと養育者の相互作用に基づく発達的学習モデルによる共同注意の獲得, ” Trans. of the Japanese Society for Artificial Intelligence : AI 18, pp.122-130

[10] 土井 利忠, 藤田 雅弘, “身体を持つ知能 (インテリジェンス・ダイナミクス),” *Springer Japan,* 2006.