岡山理科大学紀要 第41号 A pp. 79-86(2005)

強化学習ロボットに対する視覚情報の有効性の検証

馬場　安彦・片山　謙吾*・成久　洋之*
岡山理科大学大学院工学研究科情報工学専攻
*岡山理科大学工学部情報工学科
(2005年9月30日受付、2005年11月7日受理)

1. まえがき

現在，人間の代わりとしてロボットに作業させることを目的とした研究が盛んに行われている[8][19]．ロボットのほとんどは設計者によって与えられた制御則に従って行動する．この制御則をロボットに与えることは設計者が環境を熟知していることが前提である．しかし，設計者がその前提を満たすことは困難である場合が多い．そこで設計者がロボットに制御則を与えるのではなく，ロボットが環境に適応した制御則を自律的に獲得する手法として強化学習 (Reinforcement Learning) が注目を集めている[4][9][10][11][16][14][15][20][18]．強化学習は，学習者 (エージェント) が試行錯誤を通して環境に適応する学習制御の枠組みである．

今まで我々はロボットをエージェントとして現実世界とほぼ同様の空間の連続空間になる迷路問題を対象とし研究を行ってきた[1]．問題によってはエージェントが目標と目標に類似した状態を誤認識してしまい，ゴールに到達し難くなるという事態に陥り学習がなされない場合があった．これは，従来使用してきた距離センサだけでは環境を把握するのに情報不足だったのではないかと考えられる．この問題を回避するには，エージェントが目標状態と目標に類似している状態を判別できるようになることが必要不可欠であると考えた．

そこで我々はエージェントに従来使用してきた距離センサに加え視覚情報としてカメラを搭載することで目標類似状態と目標状態を判別できるのではないかと考えた．そして，両状態を判別できればゴールに到達し易くなり学習できるのではないかと考えられる．本論文では，シミュレータ上で，目標類似状態が複数存在する迷路問題を対象として，強化学習エージェントに搭載されたカメラによる視覚情報の有効性を示す．

本論文は，第2章を強化学習の概要，第3章を本研究で扱う状態認識の説明，第4章を実験とその実験結果による考察，第5章をむすびとする．

2. 強化学習

強化学習とは，エージェントが図1のように環境との相互作用を繰り返し，環境に適応する学習制御の枠組である．教師付き学習とは異なり，状態入力に対する正しい行動出力を明示的に示す教師が存在しない．その代わりエージェントは教師のかわりに報酬というスカラーの情報を手がかりに学習する．報酬を与えるだけなので，設計者が問題環境に対して適切な制御則を与えなくても学習を行うことができるというメリットがある．しかし，報酬にはノイズや遅れがあるため，行動を実行した直後の報酬をみるだけでは，エージェントはその行動が正しかったかどうかを判断できないという困難を伴う．

ここでは，強化学習の学習の主体となるエージェントとマルコフ決定過程，強化学習の手法である環境同定型と経験強化型，そして強化学習でよく用いられる行動を選択する手法について説明する．

2.1 エージェント

エージェントは予め環境に関する知識を持たず，状態遷移を繰り返し，やっと目標にたどり着くような段取り的な行動を行う．エージェントは図2のように3つのモジュールにより構成されている．状態認識器はエージェントが現在存在する状態を認識する．そして状態認識器から学習器に現在の状態情報を渡す．学習器は強化学習を適用するモジュールである．学習器には各状態における行動の重みが蓄えられている．そして学習器から行動選択器に重みを渡す．行動選択器はエージェントの次の行動の選択をする．

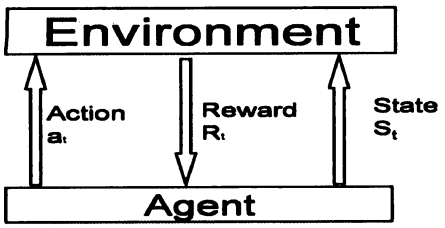

図 1　エージェントと環境の関係

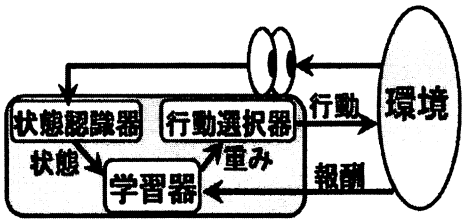

図 2　エージェントの構成

2.2　マルコフ決定過程 (Markov Decision Process, MDPs)

強化学習ではよく環境がマルコフ決定過程 (Markov Decision Process, MDPs) や非 MDPs であることを前提に研究を行う場合が多々ある[2]．MDPs とは，現在の状態 $s(t)$ が一つ前の状態 $s(t-1)$ と行動 $a(t-1)$ にのみ依存し，それ以前の過去の状態と行動に依存しないことである．

MDPs をグリッド環境を例に挙げて説明する．グリッド環境とは図 3 のように格子状 (マス) で区切られた空間のことある．エージェントは図 3 で上下左右に移動できるとする．MDPs の場合，図 4 のようにエージェントが現在いるマスに移動する以前のマスは必ず上下左右のマスのうちのどれかである．しかし，図 5 のようにエージェントが風の影響を受け次の状態が予測できない状況などは非 MDPs である．要するに MDPs ではエージェントの現在の状態とエージェントの行動則から前の状態が予測できるが，非 MDPs ではエージェントに他の影響が加わり現在の状態とエージェントの行動則から前の状態が予測できないということである．また，エージェント自身において摩擦や認識のずれなどがある場合も非 MDPs になる．すなわち，エージェントが移動するような現実世界の問題の多くは非 MDPs であると考えられる．

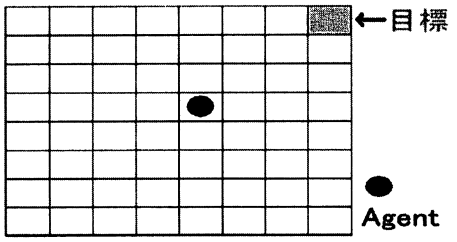

図 3　グリッド環境

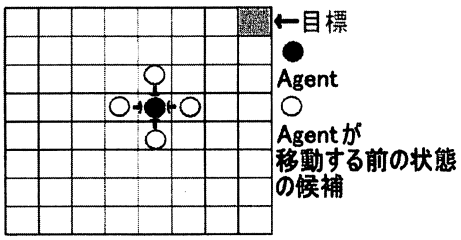

図 4　MDPs

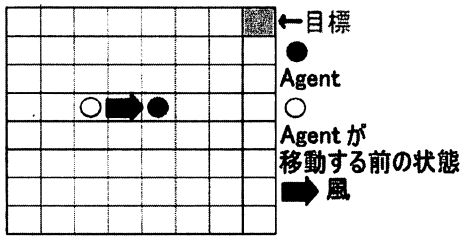

図 5　非 MDPs

2.3　環境同定型

環境同定型は環境をすべて探索することで最適解を導き出す．しかしその前提として MDPs を満たしていなければならない．また最適解を導くには環境すべてを探索する必要があるので学習時間は膨大となる．

環境同定型に属する手法として TD 学習，そして TD 学習を発展させた Q-learning と Actor-Critic などがある．次の節よりそれらの手法について説明する．

TD 学習

TD 学習 (Temporal Difference Learning)[11] は，経験から直接学習し，目標到達しなくても次の状態の行動価値 $V(s_{t+1})$ により現在の行動価値 $V(s_t)$ を更新する．以下の更新式を用いて行動価値 $V(s_t)$ を更新

する．

$$V(s_t) \leftarrow (1-\alpha)V(s_t)+\alpha(r_{t+1}+\gamma V(s_{t+1}))$$

ここで t は現在の時間，s_t は現在の状態，s_{t+1} は次の状態，r_t は環境から得られる報酬，$\alpha\,(0<\alpha\leq 1)$ は学習率，$\gamma\,(0\leq\gamma<1)$ は減衰率である．

Q-learning

Q-learning[17] は，現在の状態行動評価値 $Q(s,a)$ に現在の状態から遷移可能な状態の最大状態行動評価値を減衰した値を反映させ，$Q(s,a)$ を強化する手法である．環境との試行錯誤による相互作用の繰り返しを通して $Q(s,a)$ を推定する．次式を用いて $Q(s,a)$ を更新する．

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha\left(r+\gamma\max_{a'\in A}Q(s',a')\right)$$

ここで s は現在の状態，a は現在の状態における行動，s' は次の状態，a' は次の状態における行動の候補，A は現在の状態 s で行えるすべての行動，r は環境から得られる報酬，$\alpha\,(0<\alpha\leq 1)$ は学習率，$\gamma\,(0\leq\gamma<1)$ は減衰率である．

Actor-Critic

Actor-Critic[7][12] は，行動を司る Actor 部と評価を司る Critic 部に分かれている．Actor 部で行動を選択し，Critic 部で行動の評価を行う．以下の更新式を用いて Actor 部で行動優先度 $P(s_t,a_t)$，Critic 部で状態評価値 $V(s_t)$ を更新する．行動優先度及び状態評価値の更新はエージェントが行動する度に行われる．Actor-Critic の学習モデルは図 6 である．

行動優先度

$$P(s_t,a_t) \leftarrow P(s_t,a_t)+\alpha\ TD-Error$$
$$TD-Error = r_t+\gamma V(s_{t+1})-V(s_t)$$

状態評価値

$$V(s_t) \leftarrow (1-\alpha)V(s_t)+\alpha(r_{t+1}+\gamma V(s_{t+1}))$$

ここで s_t は状態，r_t は環境から得られる報酬，a_t は選択された行動，$\gamma\,(0\leq\gamma<1)$ は減衰率，$\alpha\,(0<\alpha\leq 1)$ は学習率，$P(s_t,a_t)$ は行動優先度，$V(s_t)$ は行動前の状態評価値，$V(s_{t+1})$ は行動後の状態評価値，$TD-Error$ は TD 誤差である．行動優先度は，状態 s_t で行動 a_t のそれぞれを選択する (優先させる) 傾向を与える値である．TD 誤差は，選択された最新の行動 a_t を評価するのに使われる．ある行動に対し TD 誤差が正の場合ならその行動を選択する傾向を強め，負の場合ならその行動を選択する傾向を弱める．

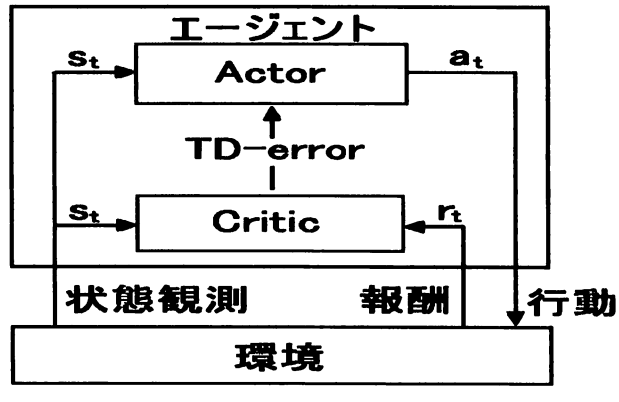

図 6　Actor-Critic のモデル

2.4　経験強化型

経験強化型は，報酬を獲得できる行動を優先して選択するため最適性は保障されない．ただし，環境同定型に比べて非 MDPs の場合でも学習しやすく，学習速度が速い．経験強化型の代表例として，Profit Sharing がよく知られている．次の節より Profit Sharing について説明をする．

Profit Sharing

Profit Sharing[3] は，報酬を得たときにそれまでに使用した状態行動評価値 s_t,a_t を一括して強化する手法である．次式を用いて状態行動評価値 W (s_t, a_t) を更新する．

$$\begin{aligned} W(s_t, a_t) &\leftarrow W(s_t, a_t) + f(t, r_T, T) \\ f(t, r_T, T) &= \beta^{T-t-1} r_T \end{aligned}$$

ここで，r は報酬，$\beta\,(0 \leq \beta \leq 1)$ は減衰率，t は現在時刻，T は報酬が発生した時刻である．f は強化関数であり，状態行動評価値を強化する関数である．

2.5 行動選択法

行動選択法とは，エージェントの行動選択器を司る部分である．上述した Q-learning と Profit Sharing にはよく用いられる行動選択法がある．ここではその行動選択法を説明する．

ε -greedy 選択法

Q-learning の行動選択法として ε -greedy 選択法がよく用いられる．ε -greedy 選択法とは，$\epsilon\,(0 \leq \epsilon \leq 1)$ の確率でランダムに行動を選択し，それ以外の $(1-\epsilon)$ の確率では，現在の状態のおいて最大の評価値を持つ行動を選択する方法である．

ルーレット選択法

Profit Sharing の行動選択法としてルーレット選択法がよく用いられる．ルーレット選択法は，ある状態 s における各状態行動評価値 $W(s, a_i)$ を全状態行動評価値の合計 $\Sigma_a W(s, a)$ で割り，確率を求め，その確率により行動を選択する方法である．

$$P(a_i|s) = W(s, a_i) / \Sigma_a W(s, a)$$

3. 状態認識

本研究では，エージェントの 3 つの構成のうち状態認識器に注目する．文献[1] では，赤外線近接センサによる障害物との距離だけを状態として扱ってきた．図 7 のような障害物の少ない迷路問題では距離情報だけでも学習が見られた．迷路問題はスタートおよびゴールが与えられ，ゴールまでの道には複数の壁が存在する．しかし，図 8 のような目標状態に類似した状態 (目標類似状態) が複数存在するような迷路問題ではエージェントが目標に到達することが困難になり学習がなされなかった．従来使用してきた距離センサだけでは学習するには情報が不十分であったと考えた．そのため目標類似状態を判別する方法が必要であった．

そこで，従来からエージェントに搭載されている距離センサによる距離情報だけでなく，カメラによる視覚情報も状態認識処理に加えることで，目標類似状態と目標状態を区別できるようになるはずである．そして，図 8 のような目標類似状態が複数存在する迷路でも目標状態に到達することが可能になり学習の進行が見られるようになるのではないかと考えられる．

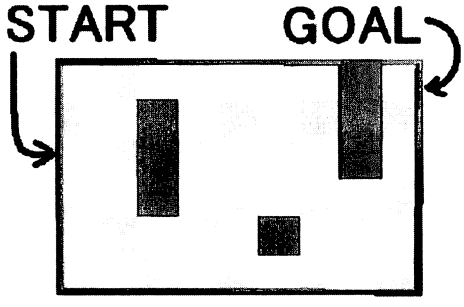

図 7　障害物の少ない迷路

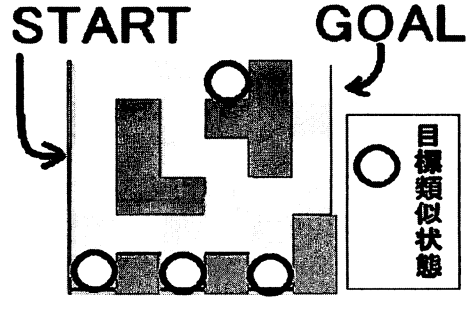

図 8　目標類似状態が複数存在する迷路

4. 実 験

本実験の目的は，文献[1] で良い結果が得られた Profit Sharing を強化学習手法としてエージェントに適用し，目標類似状態が複数存在する迷路問題を対象とし，エージェントにカメラを搭載した場合と搭載しない場合の比較実験を行い，視覚情報の有効性を検証する．本実験はシミュレータで行う．以下では KheperaII の説明，Webots の説明，実験環境 (実験問題)，実験設定，実験結果と考察の順に説明する．

4.1 KheperaII

KheperaII は，強化学習の研究においてよく用いられるロボットである[5][6]．

エージェントとして扱うロボット KheperaII について述べる．KheperaII の外形を図 9 に示す．KheperaII の仕様を以下に示す．

- 直径 70[mm]
- 高さ 30[mm]
- 重さ 80[g]
- CPU モトローラ 68331 プロセッサ 24[MHz]
- RAM512[Kbyte]
- Flash メモリ 512[Kbyte]
- DC モータ (速度 2〜60[cm/sec])2 つ
- 赤外近接センサと光センサが一体化したものを図 10 の 6 箇所に装備
- 赤外線センサの有効範囲は 70[mm]

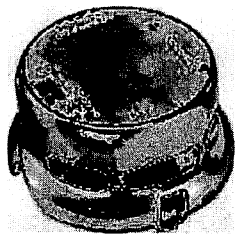

図 9 KheperaII

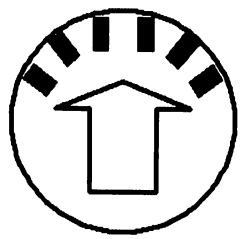

図 10 KhepraII のセンサ位置

4.2 Webots

Webots は，知能ロボット研究者や教育者，技術者のための高機能シミュレータであり，KheperaII のシミュレートでよく用いられる．ロボットの自律動作技術，進化ロボット技術などの知的ロボット技術一般の実験や，コンピュータ視覚系，人工知能技術などの研究に適した研究開発ツールである．

4.3 実験環境 (実験問題)

問題として迷路問題を扱う．実験で用いた迷路は，図 11 のような目標類似状態が複数存在する迷路を用いる．迷路は離散的な環境ではなく連続的な環境となっている．迷路のサイズは縦 60[cm] 横 60[cm] とする．図 11 の S の地点はスタート，G の地点はゴール，X の地点は行き止まりとしている．また，G の地点には青色のマークがあり，X の地点には黒色のマークがある．エージェントはカメラで青色を見つけたらゴール，黒色を見つけたら行き止まりと判断する．

4.4 実験設定

エージェントの設定は，KheperaII の外形及びセンサの設定と同一である．エージェントは 1 回の行動選択につき，図 12 に示すように前進，左に 45 度回転，左に 90 度回転，右に 45 度回転，右に 90 度回転，後退のどれか 1 つを行う．1 回の行動につき 1 ステップとする．スタートからゴールまで到達することを 1 学習とする．また，図 10 の全ての距離センサが障害物を認識した場合，前方に壁があるとし，カメラによる状態認識を行う．このときエージェントは図 11 の X の地点で黒色を見つけたら行き止まりと判断する．また図 11 の G の地点で青色を見つけたらゴールと判断し 1 学習終了とする．Profit Sharing のパラメータ設定は，各初期状態行動評価値を 0.1，減衰率を 0.95，報酬を 10 とする．両学習法の学習回数は 10000 回とする．また，1 学習の上限ステップ数 20000 で実験を行う．

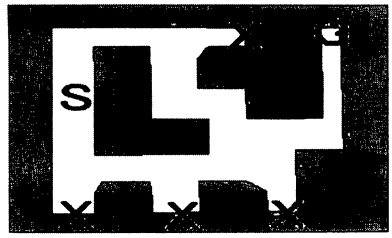

図 11 実験問題の目標類似状態が複数存在する迷路

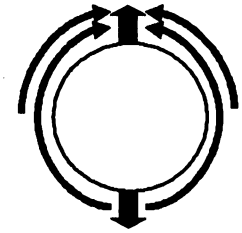

図 12 1 状態における行動の候補

4.5 実験結果と考察

図 13 にカメラ有り，図 14 にカメラ無しの結果を示す．縦軸は 1 回の学習かかったステップ数，横軸は学習回数を示す．図 13 は，上限 20000 ステップを越える事がほとんど無く，多くの場合 10000 ステップ以下でゴールに到達していることが見られる．それに対し図 14 は，上限 20000 ステップを越えている事が頻繁に見られる．また，カメラ有りの平均ステップ数は 2573 であるのに対し，カメラ無しの平均ステップ数は 7053 であることから，カメラ有りのほうがカメラ無しより少ないステップ数でゴールに到達できることがわかった．しかし，カメラの有無に関わらずステップ数の収束が見られなかったので学習がなされているとは言い難い．学習がなされなかった原因として，行き止まりの判別だけでは学習に対する有効な情報として不十分であることが考えられる．

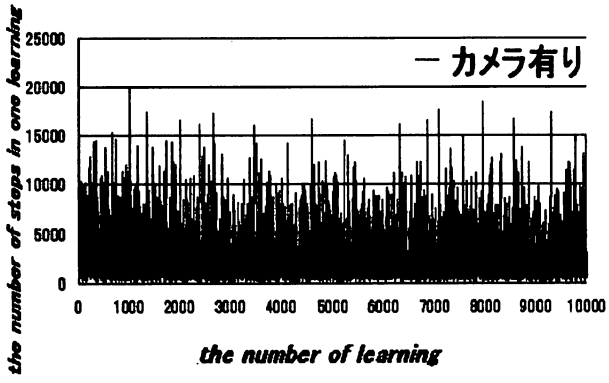

図 13 カメラ有りの実験結果

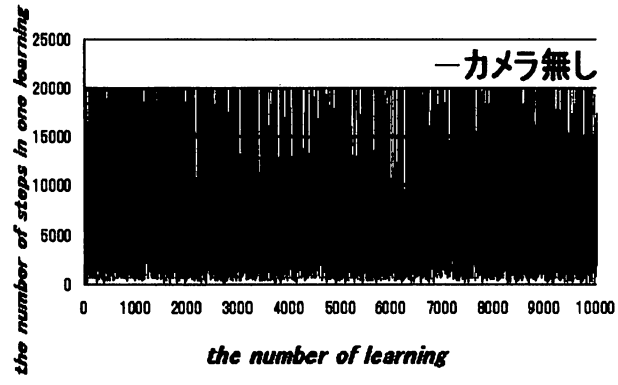

図 14 カメラ無しの実験結果

5. むすび

文献[1] では，目標類似状態が多数存在する迷路問題に対してエージェントが目標類似状態と目標状態を判別できず学習の進行が見られない問題があった．

本論文では，エージェントにカメラという視覚情報を与えることで目標類似状態と目標状態を判別できるのではないかと考えた．そして強化学習エージェントに対する視覚情報の有効性についての検証を行うため，目標類似状態が複数存在する迷路問題を対象としてエージェントに Profit Sharing を適用し，カメラ

を搭載する場合としない場合の比較実験を行った．その結果，エージェントにカメラを搭載した結果のほうがゴールに到達し易く，少ないステップ数でゴールに到達できることを示した．しかし，カメラの有無に関わらず学習がなされていないことがわかった．

本研究では，目標状態と目標類似状態の判別を行えるようにしたが，それ以外に迷路のT字路や十字路などの特徴的な場所を判別することが可能になれば，学習がなされるのではないかと考えられる．

参考文献

[1] 馬場安彦, 片山謙吾, 成久洋之, "Khepera ロボットを用いた強化学習手法の比較," 岡山理科大学紀要, 第 40 号, pp.129–136, 2004.

[2] Bellman, R. E., "A Markov decision process," Journal of Mathematical Mechanics, Vol.6, pp.679–684, 1957.

[3] Grefenstette, J. J., "Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, " Machine Learning, Vol.3, pp.225–245, 1988.

[4] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore, "Reinforcement Learning: A Survey," Journal of Artificial Intelligence Research, Vol.4, pp.237–285, 1996.

[5] 片上大輔, 山田誠二, "対話型分類子システムによる実環境ロボット学習～記述困難なプログラムを人間の教示から自動抽出する～" 第 1 回 MYCOM 資料, pp.50–53, 2000.

[6] 片上大輔, 山田誠二, "対話的進化ロボティクスの観測に基づく教示の設計," システム制御情報学会論文誌, Vol.16, No.6, pp.279–286, 2003.

[7] 木村元, 宮崎和光, 小林重信, "強化学習システムの設計指針," 計測自動制御学会, 計測と制御, Vol.38, No.10, pp.618–623, 1999.

[8] 北村新三, 片山修, "ニューラルネットとロボットの学習," 日本ロボット学会, Vol.13, No.1, pp.63–67, 1995.

[9] 宮崎和光, 山村雅幸, 小林 重信, "強化学習における報酬割当ての理論的考察," 人工知能誌, Vol.9, No.4, pp.580–587, 1994.

[10] 野田彰一, 浅田稔, 細田耕, "強化学習によるロボットの行動獲得のための状態空間の自律的構成," 日本ロボット学会誌, Vol.15, No.6, pp.886–892, 1997.

[11] Richard S. Sutton, Andrew G. Barto, [著] 三上貞芳, 皆川雅章共訳, "強化学習," 森北出版, 2000.

[12] 柴田克成, 西野哲生, 岡部洋一, "Actor-Q アーキテクチャに基づく能動認識学習システム," 信学論, Vol.J84-D-II, No.9, pp.2121–2130, 2001.

[13] 植村渉, 辰巳昭治, "Profit Sharing 法における強化学習に関する一考察," 人工知能論文誌, Vol.19, No.4A, pp.197–203, 2004.

[14] 畝見達夫, "実例に基づく強化学習法," 人工知能学会誌, Vol.7, No.4, pp.697–707, 1992.

[15] 畝見達夫, "強化学習法とロボットへの応用," 日本ロボット学会, Vol.13, No.1, pp.51–56, 1995.

[16] 内部英治, 浅田稔, 野田彰一, 細田耕, "視覚に基づく強化学習による移動ロボットの多重タスクの遂行のための協調行動の獲得," 日本ロボット学会, Vol.13, No.1, pp.68–74, 1995.

[17] Watkins, C. J. C. H., and Dayan, P., "Q-learning," Machine Learning, Vol.8, pp.279–292, 1992.

[18] 山田和明, 黒山和宏, 中村陽一郎, Mikhail Svinin, 上田完次, "実例に基づく強化学習の一手法 (Instance-Based Classifier Generator(IBCG) の連続空間への拡張)," 日本機械学, ロボティクス・メカトロニクス講演会'98, 講演論文集, No.98, pp.1BI1–7, 1998.

[19] 山田誠二, 斎藤淳也, "マルチロボットによる箱押しのための明示的通信を用いない適応的行為選択," 日本ロボット学会誌 Vol.17, No.6, pp.818–827, 1999.

[20] 山口智浩, 増渕元臣, 藤原一継, 谷内田正彦, "抽象化副報酬の自動生成による実ロボット強化学習の高速化," 人工知能学会誌, Vol.12, No.5, pp.60–71, 1997.

Effectiveness of Visual Information for Robot using Reinforcement Learning

Yasuhiko BABA, Kengo KATAYAMA* and Hiroyuki NARIHISA*

Graduate School of Engineering, Okayama University of Science
**Department of Information and Computer Engineering, Faculty of Engineering, Okayama University of Science*
1-1 Ridai-cho, Okayama, 700-0005, Japan

(Received September 30, 2005; accepted November 7, 2005)

Reinforcement learning is known to be a framework of the learning control by which an agent adapts himself to environment through trial and error. In our past research, we observed that the reinforcement learning agent with distance sensors could not learn on the maze problem having places where are very similar to the goal point. To overcome this, we add a camera as visual information to the reinforcement learning agent. To show the effectiveness of visual information, we compare two agents with camera and no camera on the difficult maze problem. The result shows that the number of steps the agent with camera reaches the goal is less than that of the agent without camera.