計測自動制御学会東北支部 40 周年記念学術講演会
2004 年 12 月 22 日，23 日　日本大学工学部

講演番号：C09

全方位視覚センサを用いた 移動ロボットナビゲーションのための強化学習

古村直樹　○釜谷博行（八戸高専）　阿部健一（東北大学）

Reinforcement Learning for Mobile Robot Navigation Using an Omni-Directional Vision Sensor

N. Komura, ∗H. Kamaya (Hachinohe National College of Technology)
and K. Abe (Tohoku University)

Abstract– Generally, mobile robot navigation systems use the position information based on dead-reckoning. However, the exact position cannot be detected because of the measurement error due to wheel slippage, rough floor, etc. On the other hand, men and animals do not use the information on the exact position. They use the information acquired from the external environment. In this research, we propose a new autonomous mobile robot navigation system, which uses the only information acquired from external sensor. This system consists of state classifier module and reinforcement learning module. The state classifier module classifies much information obtained by an external sensor to fewer states based on Kohonen's Self-Organizing Maps. Using classified states, the reinforcement learning module acquires action rules, such as avoiding obstacles and going toward a goal. This system is applied to an autonomous mobile robot equipped with an omni-directional vision sensor. And validity of this system is verified by extensive computer simulations.

Key Words: Reinforcement Learning, Mobile Robot, Navigation, Omni-Directional Vision Sensor

1　はじめに

移動ロボットナビゲーションとは，障害物を自律的に回避しながら目的地 (ゴール) に向けて移動ロボットを安全に誘導することである．このような技術を用いることで，病院内における巡回ロボットなど多岐にわたる応用が考えられる．

自律移動ロボットのナビゲーションにおいて，ロボットの位置情報を取得するには，デッドレコニング法や GPS などが用いられる．しかし，デッドレコニング法は車輪滑りや床面の凹凸などによる移動誤差のために正確な位置検出が難しい．また，GPS は屋内環境において精度が低下するという問題がある．一方，人間や動物などは正確な位置情報を用いなくとも，視覚など外界から得られる情報に基づいて柔軟に行動している．この点を踏まえて，本研究ではロボット自身の位置情報は用いずに，外界センサ情報のみから行動を決定するナビゲーションシステムを提案する．ナビゲーション能力を自律的に獲得するため,「何をすべきか」をエージェントに報酬という形で指示しておくだけで「どう実現するか」をエージェントが自動的に獲得する枠組み [1] である強化学習 [2, 3] を用いてシステムを実現する．これにより，設計時に想定していないような状況が生じても適切にタスクを実行することが可能となる．

本研究では，移動ロボットの外界センサとして全方位視覚センサ [4] を用いる．このとき，得られる全方位画像は多種多様であり，そのまま強化学習器への入力として用いると状態数が膨大なものとなり，学習が困難となる．そこで本システムでは，全方位画像から有限個の状態へパターン分類する状態分類器という考えを導入し，その分類手法として Kohonen の自己組織化マップ (Self-Organizing Maps: SOM)[5] を用いる．本稿では，提案システムを用いてシミュレーション実験を行い，その有効性を確認する．

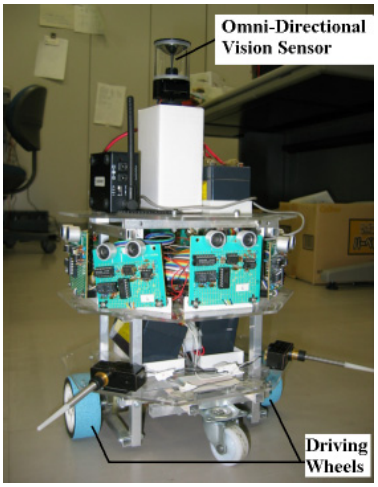

Fig. 1: Mobile robot

2　ハードウェア構成

Fig. 1 に本研究で用いる移動ロボットを示す．この移動ロボットは左右の車輪に取り付けられたモータを独立に駆動・制御することで直進・旋回ができる．また，外界センサとして，ロボット上部に全方位視覚センサを備えている．これは，鉛直下向きに設置した双曲面ミラーとその下に鉛直上向きに設置したカメラから構成される．これにより，Fig. 2 のようなロボットの周囲 360 °の全方位画像を得ることができる．

3　ナビゲーション学習システム

3.1　システム構成

Fig. 3 に提案するナビゲーション学習システムを示す．このシステムは，状態分類器と行動選択器から構成される．状態分類器は，全方位画像を有限個の状態へパターン分類するものである．行動選択器は状態に

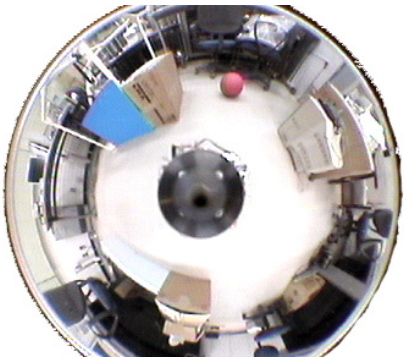

Fig. 2: Omni-directional image

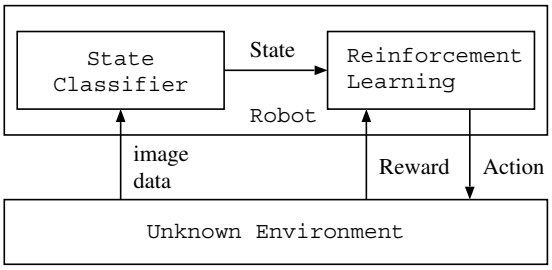

Fig. 3: Learning system

応じて行動を選択・実行し，その結果，環境から得られる報酬に基づき，行動選択の方策を改善する．このとき，ロボットは自身の位置情報を利用できないため，外界センサの時系列情報からゴールへ向かう行動を自律的に獲得しなければならない．ゴールへ向かうための正しい行動は各時点で与えられず，ゴールへ到達したときに与えられる報酬のみを学習の手がかりとする．そのため，行動選択の方策改善に，遅れ報酬を扱うことができる強化学習を用いる．

本システムではナビゲーション学習に先立って，状態分類学習を行う．状態分類学習は，状態分類器におけるパターン分類機能を自己組織化マップにより獲得する．

3.2 状態分類学習

状態分類学習では，環境中でロボットを移動させ，全方位画像から得られる移動可能領域ベクトルを収集する．その後，自己組織化マップを用いてオフラインでパターン分類学習を行う．

全方位画像から移動可能領域ベクトルを得るには，次の処理を行う．

1) 全方位画像から床面領域 (移動可能領域) を抽出する
2) 1) で得られた画像において，画像中心から 4.5[deg] 刻みで放射状に走査し，各方向の移動可能領域の長さを記録する
3) 全方位画像よりランドマークを抽出する
4) ランドマーク方向を基準にして，2) で得られたデータを規格化する
5) 22.5[deg] 毎の平均を取り，画像中心から 16 方向の移動可能領域ベクトルを求める

今回，床面は実験室の白い床を想定した．1) の処理において，移動可能領域を抽出するためには，床面と

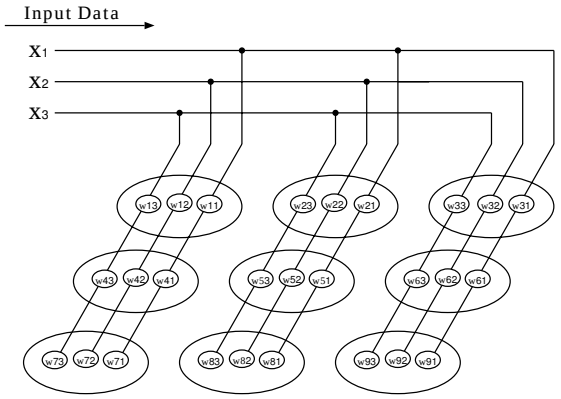

Fig. 4: Kohonen network

同じ白い色相の領域を抽出する必要がある．これは，以下の式で実現した．

$$f_{xy} = \begin{cases} 1 & \text{if } r_{xy} \geq 240 \ \& \ g_{xy} \geq 240 \ \& \ b_{xy} \geq 240 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

ここで，r_{xy} は対応する画素 (x, y) の赤色成分の値，g_{xy} は緑色成分の値，b_{xy} は青色成分の値を表す．$f_{xy} = 1$ となった画素が移動可能領域として抽出される．

3) のランドマークの検出には，まず色成分の演算を行う．ランドマークとして今回は赤いものを想定した．この場合の色成分の演算は次式で行う．

$$r'_{xy} \quad = \quad r_{xy} - g_{xy}. \tag{2}$$

2 値化処理は次式で行う．

$$h_{xy} = \begin{cases} 1 & \text{if } r'_{xy} \geq 50 \text{ (赤色領域)} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

その後，抽出された赤色領域の重心座標を求める．これがランドマークの座標となる．

ランドマーク方向を用いて規格化された移動可能領域ベクトルを，ニューラルネットワークの一種である自己組織化マップへ入力し，パターン分類学習を行う．複数の入力をもつノードが 2 次元状に配置されたネットワークを考える．実数値である M 個の入力信号 $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^T$ がすべてのノードに同時に提示され，各要素 x_j はそれぞれノードの結合強度 w_{ij} につながれている．

Fig. 4 に $M = 3$，ノード数 3×3 のコホネンネットを示す．このようなネットワークにおいて以下の処理を実行する．

$$c = \arg\min_i \|\boldsymbol{x}(t) - \boldsymbol{w}_i(t)\|, \tag{4}$$

$$\boldsymbol{w}_i(t+1) = \begin{cases} \boldsymbol{w}_i(t) + \beta(t)(\boldsymbol{x}(t) - \boldsymbol{w}_i(t)) & \text{for } i \in N_c(t) \\ \boldsymbol{w}_i(t) & \text{otherwise.} \end{cases} \tag{5}$$

(4) 式では，入力ベクトル $\boldsymbol{x}$ とのノルムが最小となる，つまり入力と一番よく似ている結合強度ベクトル $\boldsymbol{w}_i$ を持つノード c を見つける．(5) 式では，c を中心と

した近傍領域 $N_c(t)$ に含まれるノードの結合強度ベクトル $\boldsymbol{w}_i$ を入力 $\boldsymbol{x}$ に近づけるように更新し，それ以外のノードは更新しない．ここで，$\beta(t)$ $(0 < \beta(t) < 1)$ はステップサイズ・パラメータである．このアルゴリズムを用いることで，似たような特徴を持つ移動可能領域ベクトルは同じノードに対応付けられる．

3.3 ナビゲーション学習

ナビゲーション学習はオンラインで実行される．全方位画像を状態分類器へ入力すると，移動可能領域ベクトルの抽出後，ノルムが最小となるノードの検索が行われ，その番号が出力される．これが，強化学習器の状態入力となる．強化学習では，各状態において，障害物を回避しながらゴールへ向かうためにはどのような行動をとれば良いのかという行動決定戦略を学習する．

強化学習とは，試行錯誤を通じてエージェントが環境に適応する行動決定戦略を自律的に獲得する学習制御の枠組みである．未知環境中におかれたエージェントは，環境の状態を観測し，行動を選択・実行する．その結果，環境からエージェントに報酬が与えられる．エージェントは試行錯誤を繰り返しながら，より多くの報酬を得るように学習を行う．

ナビゲーション学習を行うための強化学習アルゴリズムとして Sarsa(λ)[2] を用いる．行動選択法として Max-Boltzmann 法 [6] を使用する．これは，確率 P_{max} で最大の評価値をもつ行動を選択し，確率 $1 - P_{max}$ で Boltzmann 分布に基づいて行動を選択するものである．Boltzmann 分布では，状態 s_t においてある行動 a_i を選択する確率は次式で与えられる．

$$\Pr(a_i|s_t) = \frac{e^{\frac{Q(s_t,a_i)}{\tau}}}{\sum_k e^{\frac{Q(s_t,a_k)}{\tau}}} \tag{6}$$

ここで，$Q(s,a)$ は状態–行動対の評価値を表し，Sarsa(λ) により逐次推定される．また，τ は温度係数と呼ばれるパラメータで，τ が大きくなるにつれて行動を選択する際のランダム性が大きくなる．通常，学習初期は τ を大きめに設定し，学習が進むにつれて τ を小さくしていく方法がとられる．

4 シミュレーション実験

4.1 実験設定

Fig. 5 に実験環境を示す．黒い部分は壁を，Start はロボットのスタート地点を，Goal はゴール領域を表す．このような環境において，状態分類学習およびナビゲーション学習を行う．

状態分類学習では，まず，ロボットを環境中で適宜移動させ，移動可能領域ベクトルのサンプルを収集した．つぎに，それらをもとに自己組織化マップによりパターン分類学習を行った．

ナビゲーション学習において，ロボットの行動はランドマーク方向を基準とした 45[deg] きざみの 8 方向へ移動するものとした．また，1 行動あたりの移動距離は 50[cm] とした．移動の際には車輪滑りを想定して，1 行動あたり移動距離に最大± 1[cm]，回転角度に最大± 2.5[deg] のノイズを確率的に加えた．Sarsa(λ) の各パラメータは学習率 $\alpha = 0.1$，割引率 $\gamma = 0.97$，減衰

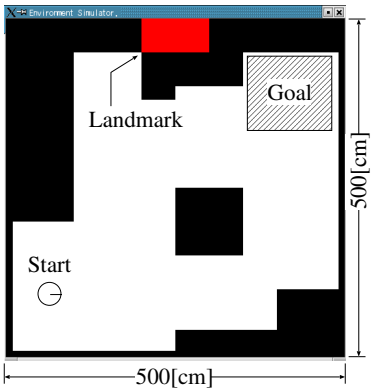

Fig. 5: Environment

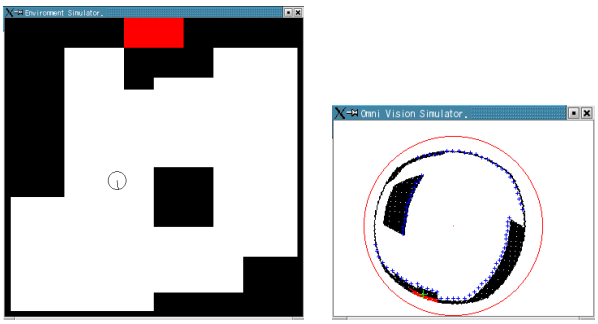

Fig. 6: Mobile robot simulator

率 $\lambda = 0.9$ とした．Max-Boltzmann 行動選択の温度係数 τ は次式にしたがって変化させた．

$$\tau = \frac{\tau_{init}}{\sqrt{trials}} \tag{7}$$

ここで $trials$ は試行回数，τ_{init} は τ の初期値を表し，$\tau_{init} = 0.9$ とした．また，P_{max} は試行回数とともに 0.9 から 1 まで直線的に増加させた．ロボットがゴール領域に入った場合報酬 10 を与え，スタート地点へ戻して次の試行を開始させた．また，ロボットが障害物に接近しすぎた場合に罰として報酬-1 を与え，ロボットを後退させて試行を継続した．その他は報酬 0 とした．

4.2 移動ロボットシミュレータ

実験のため，Fig. 6 に示す移動ロボットシミュレータを開発した．このシミュレータは行動環境シミュレータ (左) と全方位視覚シミュレータ (右) から構成される．

行動環境シミュレータは，ロボットや障害物の位置情報を管理し，ロボットの移動や障害物への衝突などの環境ダイナミクスをシミュレートする．

全方位視覚シミュレータは，行動環境シミュレータ上のロボットが得る全方位画像をシミュレートする．Fig. 7 に示した 3 次元環境中の点 $P(X,Y,Z)$ に対応する全方位画像上の点 $p(x,y)$ は次式で求められる [4]．

$$x = X \times f \times \frac{(b^2 - c^2)}{(b^2 + c^2)Z - 2bc\sqrt{X^2 + Y^2 + Z^2}} \tag{8}$$

$$y = Y \times f \times \frac{(b^2 - c^2)}{(b^2 + c^2)Z - 2bc\sqrt{X^2 + Y^2 + Z^2}} \tag{9}$$

ここで，定数 b, c は双曲線ミラー固有の，定数 f は CCD カメラ固有のパラメータである．これらを実験的

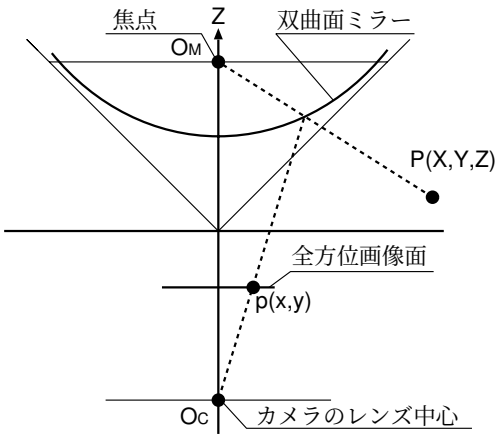

Fig. 7: Omni-directional vision sensor

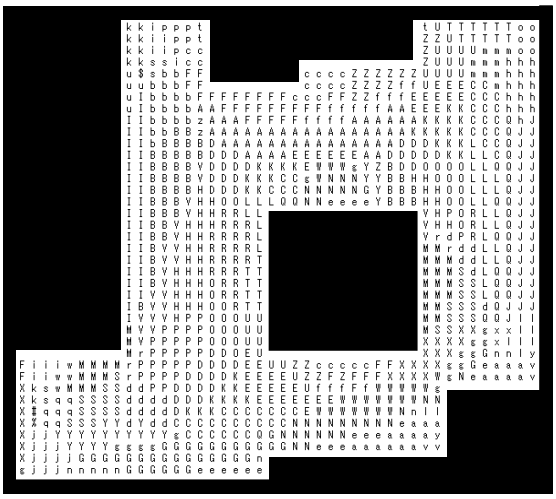

Fig. 8: Result of state classification

に求め，$b = 1.215$，$c = 1.635$，$f = 72.576$ とした．本稿では計算量の削減のため，床平面 (2 次元) のみを全方位視覚として再現した．

4.3 状態分類学習結果

環境中の各位置で得られた全方位画像が状態分類学習によって，どのような状態に分類されたかを Fig. 8 に示す．自己組織化マップにおけるノード数は 8×8 とした．広い分布を持つ状態の順に「A〜Z，a〜z，#，$，%」のラベルを付けた．これより，障害物までの距離に応じて規則的なパターンが表れていることや，似たような全方位画像が得られる距離的に近い場所では，同じ状態に分類されていることがわかる．

4.4 ナビゲーション学習結果

1 試行の最大ステップ数は 200，試行回数は 50 として実験を行った．最大ステップ数に達してもゴールに到達できない場合はその試行を中止し，ロボットをスタート地点に戻し，次の試行を開始した．乱数の初期値を変えた 30 回のシミュレーションを行い，その平均をとったものを実験結果とした．

Fig. 9 および Fig. 10 に，試行回数に対するゴールまでの平均ステップ数および試行回数に対する障害物への平均接近回数を示す．結果より試行回数の増加とともにゴールまでの平均ステップ数および障害物への接近回数が減少しており，約 15 試行という少ない試行回数で，障害物を回避しながらゴールへ向かう行動を学習することができた．全方位画像をそのまま状態として用いた場合には学習が不可能であることから考えると，提案したシステムは良好な性能を示したといえる．

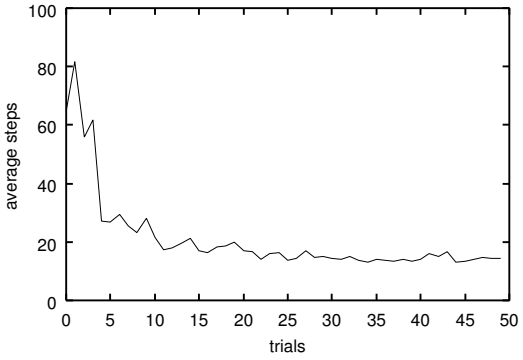

Fig. 9: Average steps to the goal

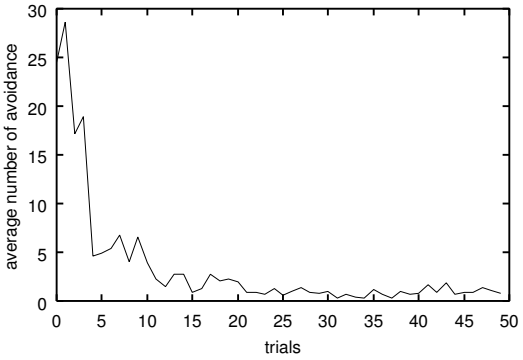

Fig. 10: Average number of avoidance

5 おわりに

本稿では，全方位視覚センサで得られた画像情報のみを用いて，障害物を回避しながらゴールへ向かう行動を強化学習により獲得する自律移動ロボットナビゲーションシステムを提案し，その有効性についてシミュレーション実験を通して確認した．

今後の課題として，環境中のランドマークの自動的な設定および認識，教示を用いた学習の高速化，部分観測性を持つ複雑な環境での検討などが挙げられる．

参考文献

[1] 木村　元，宮崎和光，小林重信: 強化学習システムの設計指針，計測と制御，Vol.38，No.10，pp. 618-623，1999.

[2] R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction., MIT Press，1998.

[3] J. Peng and R. J. Williams: Incremental Multi-Step Q-Learning, Proceedings of the 11th International Conference on Machine Learning, pp. 226-232, Morgan Kaufmann, San Francisco, 1996.

[4] 山澤一誠，八木康史，谷内田正彦: 移動ロボットナビゲーションのための全方位視覚系 HyperOmni Vision の提案，電子情報通信学会論文誌 (D-II)，Vol.J79-D-II，No.5，pp. 698-707，1996.

[5] T. Kohonen: Self-Organizing Maps 3rd ed., Springer-Verlag，2001.

[6] M. Wiering and J. Schmidhuber: HQ-Learning, Adaptive Behavior, Vol.6, No.2, pp. 219-246, 1997.