ロボットの行動獲得のための状態空間の自律的構成

Action-Based State Space Construction for Robot Learning

○浅田 稔，野田 彰一，細田 耕
Minoru ASADA, Shoichi NODA, Koh HOSODA
大阪大学工学部電子制御機械工学科
Dept. of Mech. Eng. for Computer-Controlled Machinery, Osaka University

This paper proposes an efficient method of robot learning by which a set of pairs of a state and an action are constructed to achieve a goal. Basic ideas of our method are as follows: i) Since autonomous construction of state and action spaces is generally a very difficult problem, we construct a state space so that a group of situations in which an action command to achieve the goal is the same can be merged into one state even if these situations appear to be different from each other. An action is defined as a sequence of the same action command in such a state. ii) Following the LEM (Learning from Easy Missions) paradigm [1], we first find a set of states (in terms of action) closest to the goal state, and then find a set of states closest to the set found previously. iii) In order to reduce an enormous number of trials to find such states, we place a robot so that it can observe objects which the state space consists of (in our case, a ball and a goal). iv) During the above process, the optimal action to achieve the goal is found in every state. This means that a robot can take an adequate action to achieve the goal from every state. We show the experimental results using a real robot system.

1 はじめに

動的な実世界でタスクを遂行することを学ぶ自律ロボットを実現することは，AIとロボティクスの中心課題の一つである．近年，環境とエージェントの相互作用を通して学習する手法として強化学習が注目されている [2]．強化学習に代表されるロボット学習法を実世界のロボットのタスクに適用するためには，二つの大きな問題を解決しなければならない．一つは，多大な数の試行を必要とし，学習時間が長いこと，もう一つは，ロボットが正しく学習できる統一的な状態と行動の空間が必要であることである．

前者に対して，タスク分割 [3] や外部からの批評者による導き法 (LEC：learning from external critic) [4] などが考えられている．これらの手法が，正しく働くためには，タスクに関する正確な知識が必要と考えられる．たとえば，タスク分割では，個々のサブタスクが独立で干渉しないこと，またLECでは，学習の収束を保証するために，常に正しいアドバイスを提供しなければならない．

二番目の問題は，**「状態と行動のずれ問題」**[1] と呼ばれ，物理的なセンサやアクチュエータを反映した状態空間や行動空間を構成する時に生じる．例えば，視覚センサの場合，観測者の近傍の変化は画面上で大きく写るが，遠方では同じ変化も小さな変化としてしか捉えられない．この問題に対して，Asada et al.[5] は，状態空間を先に固定し，**「ずれ」**が生じないように，行動空間を再構築した．この問題は，状態空間をどう構成するかという「状態の一般化問題」[6],「構造的信用割り当て問題」[7] などと関連し，ロボット学習の大きな課題である．これに対し，行動空間を固定して，状態空間を構成する手法など [6, 8] も提案されているが，最適な状態空間は，最適な行動空間により構成され，最適な行動空間は最適な状態空間によって構成されるべきという，「鶏と卵」の関係にあると考えられる．

これらの問題に対処するために，本稿では効率的なロボット学習法を提案する．基本的な考え方は以下の通りである．

1. 状態及び行動に関する最適な空間を構成することは，非常に困難な問題なので，ここでは，見掛けが異なっても，ゴールに到達する動作が同じ状況を同一の状態とする．行動は，このような状態で繰り返される同一の動作系列として定義する．

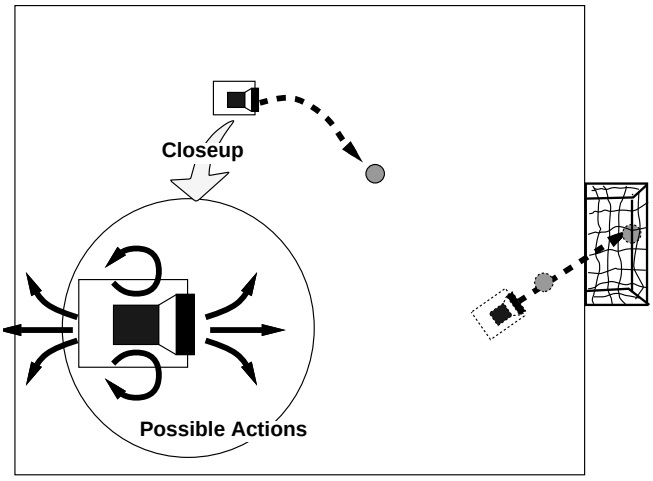
2. 効率的探索を実現するために，最初に，ゴール状態に最も近い行動に関する状態の集合を発見し，次に発見された状態集合に近い状態集合を発見する．試行回数を低減するために，状態空間を構成するオブジェクトが観測できる位置・姿勢を初期位置として生成する．

3. 上記の過程を経て，状態空間と行動空間が生成されると同時に，全ての状態において，ゴールに到達する最適行動が得られる．

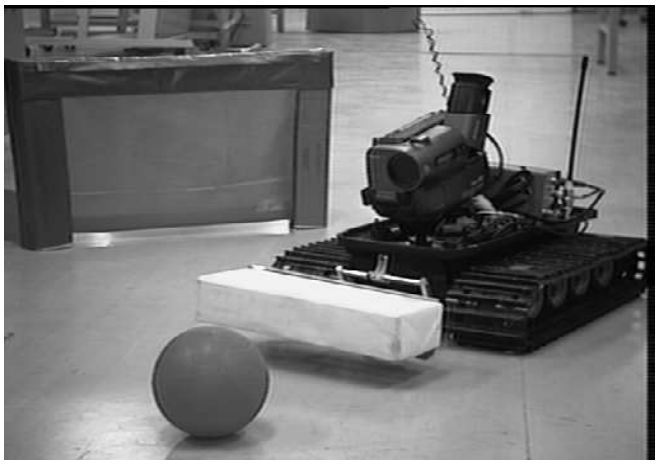
以下，本手法を適用するタスクを示し，次に本手法を説明する．最後に，実ロボットを用いた実験結果と討論を示す．

〒565 吹田市山田丘2-1 TEL(06)879-7347 FAX (06)879-7348 e-mail: asada@robotics.ccm.eng.osaka-u.ac.jp

2　タスクと仮定



(a) タスク



(b) 実験に用いたロボット

図 1: タスクとロボット

実ロボットの例として，ボールをゴールにシュートするサッカーロボットを考える．サッカーロボットとその環境を表したものを図1(a) に示す．環境内にはロボットの他に，ボールとゴールしか存在しないものとする．ロボットが得られる情報は，ロボットに搭載されたカメラからのボールとゴールについての画像情報のみである．ボールやゴールの大きさや距離などの三次元情報，カメラパラメータ，ロボット自身の動特性などの先験的な知識は一切与えられていない．

ロボットは，左右の車輪が二つのモータにより独立に駆動される PWS (Power Wheeled Steering) システムを持っている．左右輪がそれぞれ，前進，停止，後進の3段階の速度を出すので，合計9通りの行動要素が選択できる．ただし，この内，停止行動は状態の変化をもたらさないので選択せず，残りの8通りの中から行動要素を選択する．Fig. 1(b) に，実際に用いた移動ロボット，ボール，ゴールを示す．

3　状態・行動空間の自律的構成

ロボットが識別する状態空間を，人間が適当に分割しても，それがロボットの行動空間に対応するとは限らず，タスクにとって最適な分割になっている保障はない．このような状態空間では，

- 状態遷移のばらつき
- 不必要に分割された状態

が存在する恐れがある．同じ状態で同じ行動をとっても状態遷移にばらつきが生じる状態空間では，ゴール状態への状態遷移が不確実であるので，タスクの達成に障害となる．また，不必要な分割がなされている場合は，状態数が多くなり，学習に時間がかかる．そこで，ロボットが自らの経験を通して状態空間を構成することにより，上記の問題の解決を図る．

状態空間は，与えられたタスクに応じて，状態遷移が行動とできる限り1対1に対応するように構成すべきである．そのために，図2に示すような，ゴール状態と行動に基づいた状態空間を考える．ゴールに到達するための動作が同一の状況をまとめて一つの状態とし，そのような状態で繰り返される同一動作系列を行動と定義する．ゴール状態に一つの行動で到達できる状態をその行動の種類ごとに，$s_{1,k}, \quad k = 1, 2, 3, ..., n \leq |\mathbf{A}|$ とし，それらの集合を $\mathbf{S_1}$ とする．さらに，$\mathbf{S_1}$ に一つの行動で到達できる状態集合を $\mathbf{S_2}$ とする．同様にして，ゴール状態に最低 m 個の行動で到達できる状態集合は $\mathbf{S_m}$ となり，ゴールに到達可能なものは，いずれかの状態集合に含まれることになる．このような状態空間が構成されていれば，政策行動と状態遷移が1対1に対応することになり，パフォーマンスの向上が期待される．

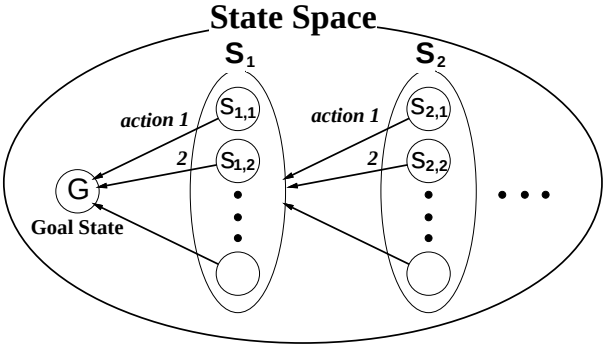


図 2: 行動・状態空間の構成

アルゴリズム

1. ゴール状態を目標状態とする．
2. ランダムに行動し，各行動をとった時，目標状態に到達可能な状態変数$\boldsymbol{x}$を蓄える．ただし，すでに区分された領域内にあるものは蓄えない．
3. 蓄えられた状態変数ベクトルを各行動ごとに状態として領域に区分する．状態空間がm次元の時，状態の分布はm次元の楕円体内で一様な分布とする．状態変数ベクトル$\boldsymbol{x}$の平均ベクトルを$\boldsymbol{\mu}$，分散共分散行列を$\boldsymbol{\Sigma}$とすると，楕円体の境界は，

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=m+2$$

で表される [9].
4. 各行動で区分された領域の論理和をとった領域を次の目標状態とする．重なる領域に対しては，分散で正規化した距離 (マハラノビス距離)

$$\Delta=(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

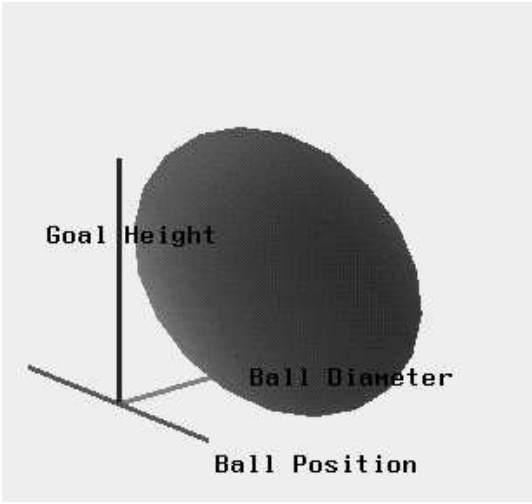
の近い方をとる．
5. 目標状態に到達可能な状態変数が無くなれば終了，さもなくば2に戻る．

4 実験結果

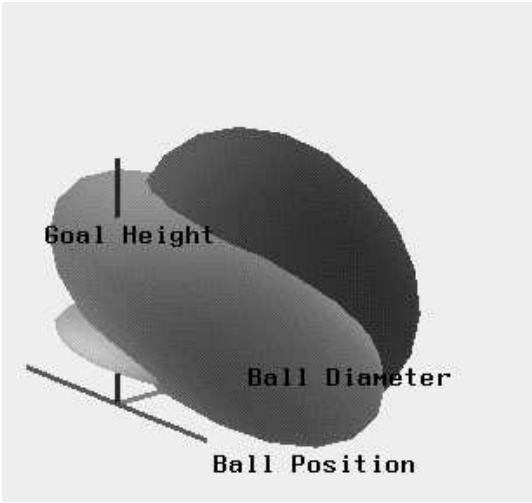
まず，シミュレーションにおける環境はFig.1(a) に示すような，3.0×3.0[m] の正方形のフィールドで，上辺の中央に幅0.9[m]，高さ0.23[m] のゴールがあり，全長0.45[m]，幅0.31[m] のロボットが直径0.09[m] のボールを蹴る．カメラはロボットの中央部についており，画角は36度である．ロボットの最大速度は1.1[m/s]，最大回転角速度は4.8[rad/s] である．ロボットの質量はボールと比べて十分大きいものとし，はねかえり係数は0.5とした．また，ボールの転がり速度は床との摩擦を考えて，各ステップ毎に0.8を掛けて減衰させている．また，画像処理による33[ms] の遅れおよび，モータの立上りの遅れ時間100[ms] を考慮している．

サッカーロボットの状態空間は，ボールの位置，大きさ，ゴールの位置，大きさ，向きの5次元のパラメータから構成される．画像のサイズは512×480である．ボール，ゴールのいずれかが観測されない状態は，これらのパラメータがわからないので，状態空間の自律的構成を行なう対象はボール，ゴールの両方ともが観測されている場合だけとする．

図3に，ゴールの位置，向きがいずれも0(真正面) での断面をとり，ボール位置，ボールとゴールの大きさの3次元で表現したものを示す．(a) では，大きな楕円体 ($\mathbf{S_1}$) が一つだけ得られ，直進運動に対応している．(b) では，二つの楕円体 ($\mathbf{S_2}$) が追加され，それぞれ直進と後退運動に対応している．著者らの以前の研究 [5] では，人間が直接状態空間を分割しており，それらは，この図で表せば，各軸に平行な直方体として各状態が表され，本手法とは大きく異なる．楕円体の集合でで覆われない残りの部分は,「ゴールが大きく，ボールが小さい」などの物理的に意味の無い状態などを表しており，以前の研究ではこのような意味のない状態も含まれていた．



(a) 第一段階



(b) 第二段階

図 3: 状態空間の構成過程

表1に比較結果を示す．探索時間は，以前の研究ではQ学習時間を示し，1/30秒を単位とするステップ数である (M=10^6)．状態数で約1/8，探索時間で約1/12と大きく改善されている．問題は，一つの状態のサイズが大きいことで，誤った状態への統合の確立が低くないことで，このことが，成功率が90%より低い原因の一つと考えられる．

表 1: 本手法と以前の手法との比較

	# of States	Search Time	Success Rate (%)
Previous work	243	500M*	77.4
Proposed method	33	41M	83.3

* はQ学習の探索時間を示す．.

図4に実ロボットの適用した際の結果を示す．実ロボットを用いた実験では，i)実ロボットを動かしてその経験をサンプリングする．ii) WS上でサンプルしたデータに基づいて提案する手法で行動・状態空間を構成する．iii)得られた政策を用いて実ロボットを動かす．ボールを発見し，ゴールにシュートしている様子が分かる．

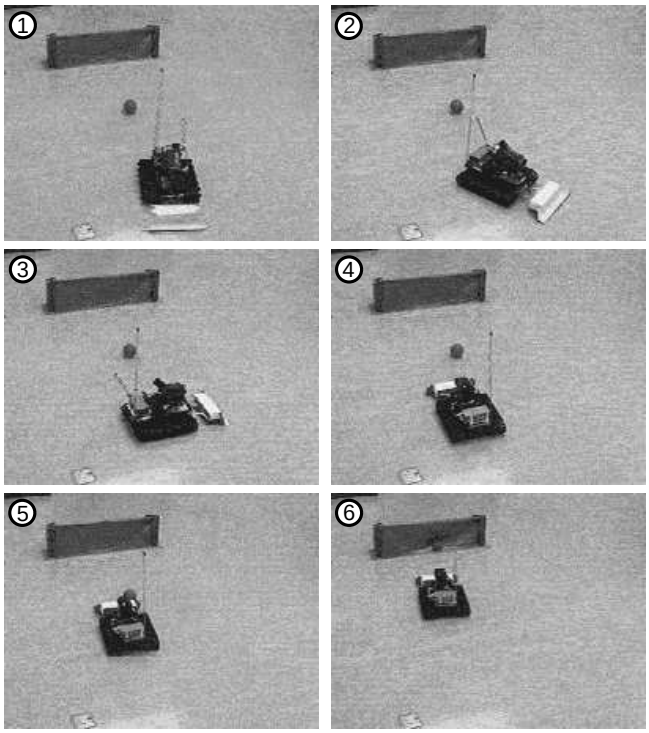


図 4: 実ロボットの実験

5 おわりに

本稿では，経験に基づく行動・状態空間構成法による効率的にロボット学習法を提案し，実ロボットを用いた実験でその有効性を示した．探索時間を早めるために，状態空間の滑らかさと均一性を仮定し，それに基づいて集中楕円体で状態を近似し，効率的な初期配置を実現した．

状態空間の構成時に問題となるのが，楕円体モデルの誤差である．モデル中に含まれるべき空間が含まれなかったり，含まれるべきでない空間が含まれることは理想的には避けたいが，実際のロボットには様々な不確定要因が存在するので完全に誤差がなくなることはない．たとえ誤差の少ないモデルであっても，複雑な計算や多くのメモリを必要とするものは望ましくない．このようなトレードオフを考慮して最適なモデルを選ぶ必要がある．

謝辞

本研究は，平成7年度文部省科学研究費重点領域研究(「創発システム」課題番号07243214)の補助を受けた．

参考文献

[1] M. Asada, S. Noda, S. Tawaratsumida, and K. Hosoda. Vision-based reinforcement learning for purposive behavior acquisition. In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pages 146–153, 1995.

[2] J. H. Connel and S. Mahadevan, editors. *Robot Learning*. Kluwer Academic Publishers, 1993.

[3] J. H. Connel and S. Mahadevan. “Rapid task learning for real robot”. In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 5. Kluwer Academic Publishers, 1993.

[4] S. D. Whitehead. “A complexity analysis of cooperative mechanisms in reinforcement learning”. In *Proc. AAAI-91*, pages 607–613, 1991.

[5] 浅田, 野田, 俵積田, and 細田. “視覚に基づく強化学習によるロボットの行動獲得”. **日本ロボット学会誌**, 13:1:68–74, 1995.

[6] D. Chapman and L. P. Kaelbling. “Input generalization in delayed reinforcement learning: An alogorithm and performance comparisons”. In *Proc. of IJCAI-91*, pages 726–731, 1991.

[7] J. H. Connel and S. Mahadevan. “Introduction to robot learning”. In J. H. Connel and S. Mahadevan, editors, *Robot Learning*, chapter 1. Kluwer Academic Publishers, 1993.

[8] A. Dubrawski and P. Reingnier. Learning to categorize perceptual space of a mobile robot using fuzzy-art neural network. In *Proc. of IEEE/RSJ/GI International Conference on Intelligent Robots and Systems 1994 (IROS ’94)*, pages 1272–1277, 1994.

[9] H. Cramër. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ, 1951.