様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２５　日現在

機関番号：　３２６５７
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１２４９３

研究課題名（和文）認知心理学におけるプライミング効果を基にした強化学習ロボットによる知識選択

研究課題名（英文）A policy selection method based on the priming effect in the cognitive psychology for reinforcement learning agent

研究代表者

鈴木　剛（SUZUKI, Tsuyoshi）

東京電機大学・工学部・教授

研究者番号：００３４９７８９

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：　本研究では，転移学習において複数方策から有用な知識を選択するために，人間の記憶や知識の思い出しや再認識を行うメカニズムである活性化拡散モデルを用いた転移学習手法を提案した．本手法は，1)複数方策をカテゴリに分類してネットワークを構築，2)カテゴリから方策を想起，3)方策を選択，4)活性値の拡散（活性化拡散），5)活性値の減衰，という処理をロボットエージェントの行動毎に反復実行し，方策の活性値を調整しながら，転移する方策を選択する．計算機シミュレーションにより，方策を用いない強化学習，単一方策のみを用いた転移学習，提案手法を用いた転移学習の学習効率を比較し，提案手法の有用性を確認した．

研究成果の概要（英文）：This research proposes a policy transfer method of a reinforcement learning agent for suitable learning in unknown or dynamic environments based on a spreading activation model in the cognitive psychology. The agent saves policies learned in various environments and learns flexibly by partially using suitable policy according to the environment. In the proposed method, an undirected graph is created between policies, and the network is constructed by them. The agent updates the activate value that policy has according to the environment while repeating processes of recall, activation, spreading, attenuation and learns based on the network. Agent uses this network in transfer learning. Experimental simulations comparing the proposed method with several existing methods are conducted to confirm the usefulness of the proposed method. Simulation results show that the agent achieves the task by selecting the optimal one from policies with the proposed method.

研究分野：　ロボット工学、情報通信工学

キーワード：　知識選択　活性化拡散モデル　転移学習　強化学習

１．研究開始当初の背景

ロボットシステムに実装される学習アルゴリズムとして，Q 学習やニューラルネットワークなどの様々な手法が提案されている．これらの手法では，通常，学習した知識はタスクや環境に対して 1 つだけ生成される．近年では様々なタスクや環境で学習したそれぞれの知識を個別に保存して，さらにそれら複数の知識をネットワーク構造として保存・再利用し，環境適応性を高める研究が行われている．しかし，保存された知識間のネットワークを記述する効果的な手法は確立されていない．

一方，人間では概念や意味がネットワーク構造として記憶されているという仮説が様々な心理学実験により確認され，プライミング効果の発現がそれを支持している．プライミング効果とは，先行的に得られる刺激が，後続の処理に無意識的に影響を及ぼす認知心理学の知見である．このプライミング効果が発現するように，ロボットシステム内に保存される獲得知識間のネットワーク構造を記述すれば，知識の検索と選択の正確性と効率を促進できると考える．そこで本研究では，強化学習で獲得した複数の知識をネットワーク構造で記述・保存し，効率的な知識選択のために知識ごとに活性レベルを採用する．また，活性拡散モデルに基づいた知識間ネットワークの記述手法と，プライミング効果の発現による知識選択の正確性の向上を目的とする．

２．研究の目的

本研究は，強化学習するロボット（以下，学習可能なロボットをエージェントと呼称）が獲得・保存した複数の知識の再利用における知識選択手法の確立を目的とする．強化学習アルゴリズムに，認知心理学で探求されてきた成果，特にプライミング効果の発現を目的として，強化学習アルゴリズムによる知識の保存方法，保存された知識を活性化・想起させる手法，知識間の関係性を記述する手法の開発などを検討し，活性拡散モデルに基づく知識間接続関係の記述手法と，それを用いた知識の効率的な再利用を実現する．

３．研究の方法

活性化拡散モデルとは，人間が獲得した概念同士が脳内でネットワーク構造として保存されていることを前提とし，ある概念が想起(思い出し，再認識等) されることで，関連する概念も活性化され，概念の利用が促進されるモデルである[①]．活性化拡散モデルには，関連性の強さに応じて関連している概念間の距離を変動させて配置する意味的距離が存在する．概念の活性化は，関連性によって構築されたネットワークを通じて行われ，各概念間に意味的な関連性の表現が存在する．活性化拡散モデルの例を図 1 に示す．図 1 では，活性化された概念から伸びる距離を経由して活性値と呼ばれる値が拡散，伝播している様子を表している．

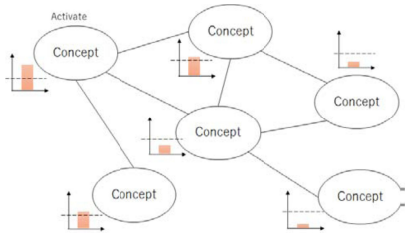


図 1　活性化拡散モデルの例

本研究では，活性化拡散モデルを強化学習へ応用し，エージェントが学習可能なアルゴリズムとして実装するために，方策間の関連性に基づきネットワークを構築して，そのネットワークを用いて転移学習を行う．

(1)　強化学習により獲得した知識の保存方法の開発

本研究では，エージェントが獲得した知識を再利用可能な形式で保存する必要がある．これは，1 つの知識（強化学習の方策）で全ての学習結果を記録することは現実的でなく，タスクや環境ごとに知識を分割し，選択させる形で知識の再利用を行うからである．強化学習の知識の保存方法に関しては，例えば，基本的な Look-uptable として保存する方法やニューラルネットワークとして記述する関数近似を用いた方法などがある．しかし，関数近似を用いた手法では近似誤差により獲得知識の品質が劣化する可能性がある．そのため，本項目では Look-up table 型の知識保存方法を採用し，エージェントが再利用しやすい形式を開発する．さらに，活性値などの直接知識内には記述できないパラメータの保存方法も開発する．

(2)　保存された知識を活性化・想起させる手法の開発

エージェントにより複数の知識が獲得され，それらを再利用するとき各知識に記述された活性値により知識の選択を行う．知識を再利用する際，各知識は次項で述べるネットワーク構造でお互いの知識の接続関係を持たせる．このとき，環境の先行入力（エージェントのセンサによる環境認識情報など）を用いて活性値が高い再利用知識の候補を上げる．さらに，それらの活性値を比較し任意の閾値（活性レベル）を超えた知識を再利用する．さらに，再利用された知識は活性化を受け，知識と共に保存されている活性値を上昇させる．これらの繰り返しにより，環境に対する適切な知識を選択していく．しかし，先行入力による知識選択が必ずしも環境に対して正しい知識であるとは限らない．そのため，本項の「知識の活性化・想起」手法では想起が正しかったのかどうか判断する評価関数や想起プロセスにまで検討を深める．また，知識選択後の行動選択部分においても先行入力による活性値を活用した手法を検討する．

(3) 知識間の関係性を記述する手法の開発

エージェントが獲得した知識は，システム的に個別に保存されているだけでは再利用するためのラベルがなく，また，ラベルを付加する場合でも知識獲得毎にラベル定義作業が必要である．そのため，知識間での関係性を知識獲得時に自動設定するようにしておき，知識選択時の想起頻度（選択頻度）に応じて知識間の関連度合いを高めるアルゴリズムが必要である．そこで，本項目では活性拡散モデルや意味ネットワークなどのモデルを参考に，エージェント内に知識間の関係性を記述する手法を開発する．具体的には，知識間の接続関係を記述する手法と想起頻度に応じて知識間距離を調整するアルゴリズムを開発する．

以上，3 つの検討を統合し，エージェントが獲得した知識を環境やタスクごとに保存・蓄積し，先行入力やモデルを用いて，適切な知識の選択・再利用を実現する手法を確立する．

４．研究成果

(1) 提案手法

① 前提条件

本研究では強化学習するエージェントが獲得した方策を選択しながら転移学習を行う手法を提案する．本節では，提案手法を述べるにあたり，関係用語や前提を述べる．強化学習には，Q 学習を用いる．時刻 t での状態 s における行動 a の価値である行動価値 $Q_s^{\pi_i}(s_t, a_t)$ (以下，Q 値と呼称) の更新式を次に示す．

$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha\{r_{t+1} + \gamma max Q(s_{t+1}, a_{t+1}\}$$

s_t, a_tは現在の状態及び行動，s_tは行動a_tを行った後の状態，αを学習率，γを割引率，r_{t+1}を報酬とする．

初期位置から目的地までの最短の経路を学習する最短経路問題を Q 学習のタスクとして設定する．学習した Q 値及び学習した環境における環境情報（障害物などの情報）は，Q-table と呼ばれる Look-up table と共に記述しておき，予め再利用可能な状態で保存する．方策再利用に関する転移学習の更新式を次に示す．

$$Q_c(s_t, a_t) \leftarrow \tau Q_s^{\pi_i}(s_t, a_t) + Q_t(s_t, a_t)\mathbf{d}$$

転移学習では，学習予定のタスク(以下，Target-task と呼称)に対して，予め学習したタスク(以下，Source-task と呼称)で獲得した方策を転移する．$Q_s^{\pi_i}(s_t, a_t)$は Source-task 方策(以下，i は方策の識別番号)，$Q_t(s_t, a_t)$は Target-task で学習中の方策，$Q_c(s_t, a_t)$は統合した方策を示している．は利用方策の Q 値を調整するパラメータ(以下，転移率と呼称) である．

② 提案手法の流れ

エージェントは，タスクの達成のため自身の周囲の環境情報を観測し，複数方策を用いて構築したネットワークと観測した環境情報を基に，方策を選択して転移する．この転移学習の目的は，$Q_s^{\pi_i}(s_t, a_t)$を手掛かりに$Q_t(s_t, a_t)$を獲得することである．提案手法では，複数方策を用いてネットワークを構築し，想起，方策選択，活性化拡散，減衰という処理をエージェントの行動毎に反復して行い，方策の活性値を調整しながら転移する方策を選択する．

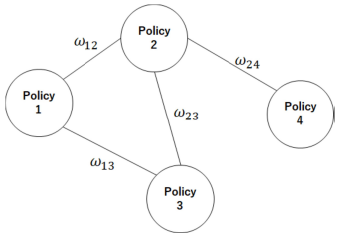


図 2　SAP-Net の例

③ 方策ネットワークの構築と方策のカテゴライズ

本研究では、複数方策をカテゴリに分類し，それを基にネットワークを構築する．本研究のカテゴリとは，複数の方策に関連性を見出し，関連性のある方策同士を集めたものを指す．方策同士をある観点で比較し，それが同じカテゴリに属するかを判定する．この観点のことをプロトタイプと呼称する．このプロトタイプは，予め学習した方策に含まれるデータで利用可能なものを用いる．また，分類したカテゴリ内で，方策間の関連性を記述するための方策間距離d_{ij}を生成する．方策間距離は，カテゴリ内ですべての方策接続パターンを網羅するように方策同士を全結合する．活性化拡散に基づく方策ネットワーク(以下，SAP-Net: Spreading Activation Policy Network と呼称) は無向グラフとし，結合する方策の集合Π，方策間の接続関係を示すE，距離の持つ重みωを用いて次式のように定義する．

$$\mathbb{G} = \Pi, E, \omega$$

無向グラフの頂点となる方策の集合の元は，$\pi_i \in \Pi$，$e \in E$であり，Eは最大で$E \subseteq \Pi \times \Pi$である．グラフの表現方法には，Tutte 行列を用いて表現する．Tutte 行列は$n \times n$の正方行列Mで表され，次式のように行列内の要素を定義する．

$$M_{ij} = \begin{cases} 0 & \left(\{\pi_i, \pi_j\} \notin E\right) \\ \omega_{ij} = 1 & \left(\{\pi_i, \pi_j\} = e \in E\right) \end{cases}$$

図 2 に SAP-Net の例を示す．図 2 を Tutte 行列で表現すると，次式のように表せる．

$$M = \begin{pmatrix} 0 & \omega_{12} & \omega_{13} & 0 \\ \omega_{12} & 0 & \omega_{23} & \omega_{24} \\ \omega_{13} & \omega_{23} & 0 & \omega_{34} \\ 0 & \omega_{24} & \omega_{34} & 0 \end{pmatrix}$$

SAP-Net を構築する際に，分類したカテゴリから生成したプロトタイプ行列によって，次式を用いて各要素の重みを調整する．プロトタイプ行列も SAP-Net と同様に Tutte 行列で表され，式(4)ようにカテゴリ内の方策間距離の接続から重みを生成する．

$$M = M + \delta\{\mathbb{P}_1 + \cdots + \mathbb{P}_n\}$$

n は距離の重複したカテゴリの数を表し，$\delta(-1.0 < \delta < 0)$はプロトタイプ行列の重みを調節する係数である．

④　方策の想起

想起では，観測した情報に基づいて，あるカテゴリを選択し，そのカテゴリに分類される方策の中から選択候補を求める処理が行われる．それぞれの方策には，活性値$\mathbb{A}_i$というパラメータを与えておき，エージェントは観測情報に基づいてカテゴリを選択することで，カテゴリ内の全方策の活性値を更新する．更新式を次に示す．

$$\mathbb{A}_i \leftarrow \mathbb{A}_i + A_{recall}$$

A_{recall}は，想起係数と呼び活性値の上昇を調節するための係数である．加えて，候補を求める際には，活性値を参照して候補に選択する方策の足切りを行う閾値関数$\mathbb{T}\ \mathbb{A}_i$ を定義する．次式に示す．H は任意の閾値を示す．

$$\mathbb{T}\ \mathbb{A}_i = \begin{matrix} 0 & \mathbb{A}_i < H \\ 1 & \mathbb{A}_i \geq H \end{matrix}$$

⑤　活性化拡散モデルに基づく方策選択手法

方策の選択には，エージェントが観測した情報と構築した SAP-Net を用いる．観測情報を用いた想起で得られた方策の選択候補から確率的に方策を選択する．活性値の大きさで選択確率P_iを算出する．算出式は次式に示す．j は候補方策の識別番号を示している．

$$P_i = \frac{\mathbb{A}_i}{\sum \mathbb{A}_j}$$

⑥　活性値の減衰

エージェントが行動する毎に，全方策の活性値を減少させていく．残った活性値をΦ_iとし，次に更新式を示す．λは減衰を調整する係数である．

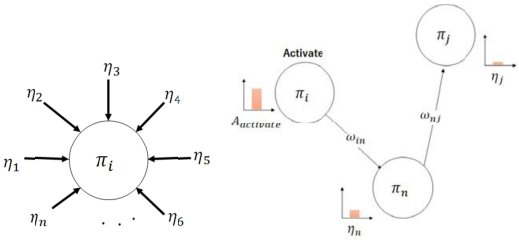
$$\Phi_i = \mathbb{A}_i e^{-\lambda}$$

⑦　活性化拡散

方策を選択してエージェントが行動した際には，そのフィードバックを活性値に与える活性化という処理を行う．統合方策に従った行動が Target-task の学習を助けるものになった場合(以下，正の転移)は使用した方策の活性値を増加させる．反対に，学習を妨げる行動につながった場合(以下，負の転移)は，活性値を減少させる．活性化は，残活性値に活性化係数$A_{activate}$を用いて次のように表される．

$$\mathbb{A}_i = \begin{cases} \Phi_i + A_{activate} & \text{(正の転移)} \\ \Phi_i - A_{activate} & \text{(負の転移)} \end{cases}$$

これに加えて，活性化した方策から伸びる方策間距離を経由して接続関係のある方策へと活性値の拡散処理を行う．拡散について定義するにあたり，SAP-Net 上に保存されている方策π_iへ拡散される活性値(以下，活性値入力) の様子を図 3 (a)に示す．ある方策π_iから活性値入力を，$\eta_k(k = 1,2,\cdots,n)$とし，拡散元



(a)　　(b)

図 3　活性値入力と拡散処理

π_iと拡散先π_jの二つの方策についての活性値の拡散を考えるとπ_jに拡散される活性値入力η_jはπ_jからの$A_{activate}$と複数の方策を経由した拡散も考慮し，経由した方策間距離の重みの和$\sum \omega$，経由した方策間距離の数 h を用いて次式で表す．

$$\eta_k = \begin{cases} 0 & \sum \omega \geq \omega_{threshold} \\ \dfrac{1}{h \sum \omega} A_{activate} & \sum \omega < \omega_{threshold} \end{cases}$$

活性値拡散のイメージを図 3(b)に示す．拡散する範囲の指定をしない限り永続的に拡散可能になるため，経由した方策間距離の重みの和で閾値$\omega_{threshold}$を定める．この計算は方策経由毎に再帰的に行い，経由した距離の重みの和が増えることによって拡散される活性値を減少させていく．最終的に，各方策の活性値入力の総和を求め，正の転移と負の転移で場合分けを行い，次式のように活性値を増減させる．

$$\mathbb{A}_i = \begin{cases} \Phi_i + \sum \eta & \text{(正の転移)} \\ \Phi_i - \sum \eta & \text{(負の転移)} \end{cases}$$

(2)　計算機実験および結果

複数方策を用いた転移学習の効果を検証するため，強化学習による学習，単一方策のみを転移した場合の転移学習，提案手法である複数方策を用いた転移学習を比較する．評価指標は，エージェントのタスク達成に必要となった行動回数(Step 数)とタスクの達成回数(Episode 数)で示される学習曲線と全体の学習で必要となった Step 数の総和(総 Step 数)を用いる．学習曲線と総 Step 数には，各学習結果を 5 回平均した値を用いる．

①　計算機実験の設定

エージェントは上下左右に停止を加えた 5 パターンの行動が可能であるとし，エージェントの行動選択関数には，ボルツマン分布を用いたソフトマックス手法を用いる．統合方策を参照したボルツマン分布を用いたソフトマックス手法を式(14) に示す．

$$p(a|s) = \frac{\exp\left(\frac{Q_c(s,a)}{T}\right)}{\sum_{b \in action} exp\left(\frac{Q_c(s,a)}{T}\right)}$$

実験で用いた強化学習及び転移学習，提案手

法のパラメータを表 1 に示す．Episode 数の上限は，1000 回に設定し，使用する方策も同様の回数に設定し学習する．提案手法の正の転移の判定には，エージェントの選択方策の履歴を使用する．目的地へ到達した際には，使用方策の履歴の新しいものが多く活性化されるよう設定する．次式を用いて使用した方策の活性化を行う．j は方策の識別番号，n は方策使用履歴の順序を降順に並べ替えた時の順序を示している．

$$A_{activate} = \frac{1}{n}$$

エージェントが統合方策に従った行動中に障害物に接触した場合を負の転移と設定する．提案手法で用いる方策数は 100 とする．方策同士が同一カテゴリであるかの判定には，以下のプロトタイプによって判定しカテゴリ毎にプロトタイプ行列を生成する．

- Source-Task 学習開始座標の周囲 1 グリッドの環境情報
- Source-Task の Start 座標から Goal 座標までの方向

エージェントは，行動毎に常に自身の周囲 1 グリッドを環境情報(障害物，通路)として観測しながら，自身の持つカテゴリと照合する．学習環境は，グリッドワールドによって構築する．学習に使用する Traget-task の環境を図 4 に示す．予め学習する Source-task の環境は，単一の方策を転移する場合も複数の場合もランダムに構築する．

② 実験結果

図 5 に強化学習，単一方策を用いた転移学習，複数方策を用いた提案手法による転移学習の学習曲線比較を示す．RL は強学習を，TL は転移学習を示している．これより，単一方策のみを転移した学習はタスクの達成により多くの Step 数を必要としていることが確認できる．ランダムに学習した方策を転移した場合は，必ずしもその方策が Target-task に適するとは限らないため，図 5 の学習曲線のように方策を使用しない場合よりも学習効率が低下する．この結果が負の転移である．そのため，環境に合う方策の転移の仕方が必要になる．強化学習と提案手法による転移学習を詳細に比較するため，図 6 に学習 episode 1-100 に注目した学習曲線を示す．それぞれの学習における総 Step 数を表 2 に示す．表 2 と図 6 より，提案手法において Step 数の減少が確認でき，かつ，総 Step 数が一番少ないことから，学習効率が最も良いことが確認された．ただし，転移学習の特徴である学習初期段階における Step 数の減少(Jump Start)が観測できなかったことから，Target-task に適する方策が少なかったことが推測される．今回の実験では，ランダムに生成した障害物を配置した環境の最短経路を学習した 100 の方策を使用していることから，Target-Task に適した知識が極めて少ない場合は，今回の結果よりも学習効率が悪くなる可能性も考えられる．

以上，活性化拡散モデルを基に，複数方策を用いて構築した SAP-Net と，エージェントの観測した情報に応じて知識を選択する手法を提案した．実験では，方策を用いない強化学習と単一方策を用いた転移学習，提案手法を用いた転移学習の学習効率の比較を行い，複数方策を用いた転移学習が有用である

表 1 実験における数値設定

Parameters	Symbol	Value
学習率	α	0.1
報酬	r	1
割引率	γ	0.9
転移率	τ	0.2
プロトタイプ係数	δ	−0.19
初期活性値	$\mathbb{A}_i$	0.1
想起係数	A_{recall}	0.3
想起の閾値	H	0.5
減衰係数	λ	0.09
活性化係数	$A_{activate}$	0.05
重みの閾値	$\omega_{threshold}$	0.75
温度定数	T	0.01

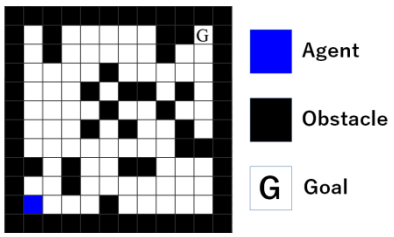


図 4 Target-task

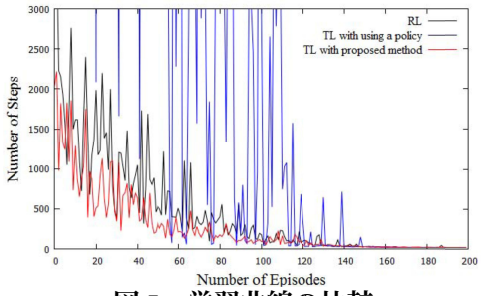


図 5 学習曲線の比較

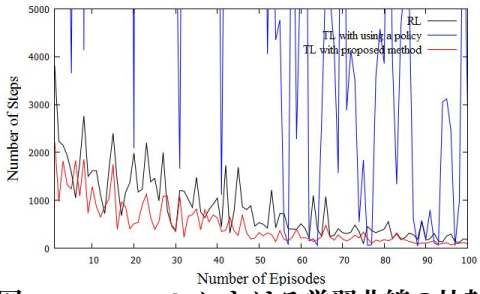


図 6 100episode における学習曲線の比較

表 2 総ステップ数の比較

学習方法	総 Step 数
Q 学習（転移学習なし）	110,908.4
転移学習（単一方策）	1,097,578.4
転移学習（提案手法）	72,131.2

ことを示した. 実験結果で得られた Jump Start の観測が可能な新たな手法や SAP-Net の学習中の最適化の検討や，獲得した方策数の学習に対する影響の検証等を行う．また，現在，図 7 に示す実機自律型全方向移動ロボットに提案手法を適用し，カメラ画像を用いた主に色認識に基づく環境情報を先行入力として用い，単純な環境で経路を探索する評価実験を行う準備を進めている．

＜引用文献＞

① A M. collins, E. F. Loftus, "A Spreading-Activation Theory of Semantic Processing", Psychological Review, Vol.82, No.6, pp.407-428, 1975.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

① H. Kono, T. Suzuki, A. Kamimura, K. Tomita, Y. Tamura, A. Yamashita and H. Asama, “Automatic Convergence Estimation by Utilizing Fractal Dimensional Analysis for Reinforcement Learning”, The Journal of Instrumentation, Automation and Systems, vol. 3, Issue 3, pp.58-70, 2016 (available online in 2017). DOI:http://dx.doi.org/10.21535/jias.v3i3.934, 2017.09.26, 査読有

〔学会発表〕（計 5 件）

① 河野 仁, 三浦昇三, 温 文, 鈴木 剛,「強化学習における方策転移度合い決定のための転移曲面の検討」，第 24 回画像センシングシンポジウム(SSII2018)，IS2-28，横浜(2018.06)

② 高桑優作，河野 仁，温 文，神村明哉，富田康治，鈴木 剛,「活性化拡散モデルに基づく強化学習エージェントの方策選択手法」,第 18 回計測自動制御学会システムインテグレーション部門講演会(SI2017)，3E1-04(DVD)，仙台(2017.12)

③ 河野 仁，三浦昇三，温 文，鈴木 剛,「強化学習における方策再利用評価のための転移曲面の検討」,第 18 回計測自動制御学会システムインテグレーション部門講演会(SI2017)，1D4-10(DVD)，仙台(2017.12)

④ 高桑優作，河野 仁，温 文，神村明哉，富田康治，鈴木 剛,「活性化拡散モデルに基づく強化学習エージェントの方策選択手法」，日本機械学会ロボティクス・メカトロニクス講演会 2017(Robomech2017), 2P2-E04 (DVD)，福島(2017.05)

⑤ 河野 仁, 伊藤祐希, 郡司拓朗, 神村明哉，富田康治，鈴木 剛,「強化学習の方策再利用時におけるステップ単位の方策忘却手法」，日本機械学会ロボティクス・メカトロニクス講演会 2017(Robomech2017), 2P1-F06 (DVD)，福島 (2017.05)

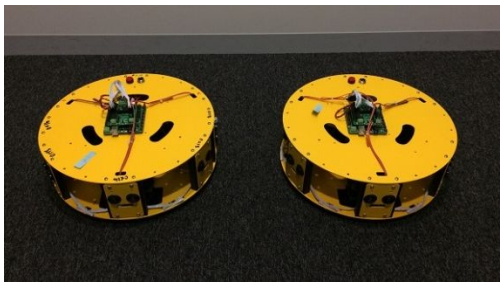


図 7　全方向移動ロボット

６．研究組織

(1) 研究代表者

鈴木　剛（SUZUKI, Tsuyoshi）
東京電機大学・工学部・教授
研究者番号：００３４９７８９

(2) 研究分担者

温　文（WEN, Wen）
東京大学・大学院工学系研究科（工学部）・特別研究員
研究者番号：５０６４６６０１

河野　仁（KONO, Hitoshi）
東京工芸大学・工学部・助教
研究者番号：７０７５８３６７