［研究助成（B）］

# 人間とロボットの持つ世界モデルの共有を通じた調和的行動生成の実現

Cooperative action generation by sharing world models between humans and robots

2211904

研究代表者　京都大学　大学院情報学研究科　教　授　森　本　　淳

## ［研究の目的］

今後，高齢化社会を迎えるにあたってロボットが人間の日々の生活をサポートすることが広く期待されている。しかし，これまでのロボットは，人間の生活圏とは離れた工場内での利用に特化した制御システムを用いてきたため，人間との調和を考慮した機能が十分に備わっていなかった。人間との調和的行動の生成のためには，ロボットの周りの人間の動作戦略や環境をいかに認識するかという，周りの世界の状況をロボットが理解する仕組みが必要となる。

本研究では，人間とロボットが調和的な行動生成を実現するための方法論の開発をおこなった。具体的には，ロボットが周りの世界を認識することで，その状況に応じて制御システムを適応させる手法の開発を目指した。

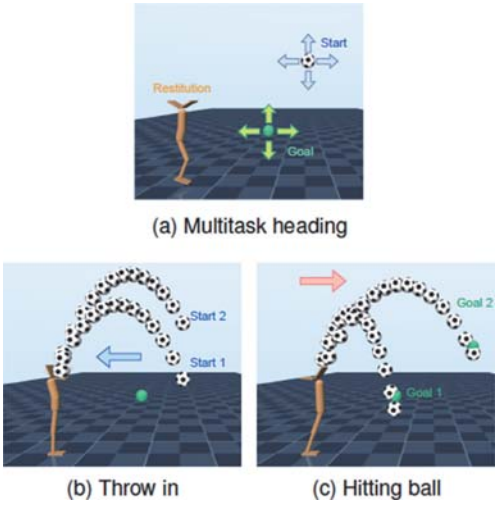
## ［研究の内容］

ロボットの意思決定問題はマルコフ決定過程として記述することが一般的だが，本研究では人間を含めたロボットの周りの世界が動的に変化することを扱う必要がある。そこで，ロボットの周りの状況を引数として考慮する文脈付きマルコフ決定過程によって動作課題を記述することを考えた。一方で，人間の行動戦略のように陽にはロボットに認識できないような問題に対応可能とするために，ロボット自身の経験からその文脈を推定するための方法論の検討を進めた。

具体的な実現方法として，次の３つのフェーズによって構成される方法論を導出した：

1）方策生成フェーズ：異なる文脈においてタスクを学習することにより，文脈にあわせた方策を生成することを可能にする。
2）方策選択フェーズ：生成された複数の方策を学習時のタスクパフォーマンスに応じて順位付けを行い，新しい文脈に適応するための基礎方策を選択する。
3）方策適応フェーズ：選択した基礎方策において，文脈を認識するエンコーダ入力を推定することにより，新しい文脈への適応を可能とする。

具体的な提案手法の評価のために運動学習課題への適用をおこなった。ここで，生成する動作としてサッカーのヘディング動作を考える。同じヘディング動作においても，ボールの目標位置は連続的に変化し得る。また，ボールの跳ね返り方もボールによって異なる。そこで，本研究では，図１に示すように１脚ロボットが人間のヘディング動作のようにバランスを維持しながら，跳ね返り係数が異なるボールを異なるゴール位置に向けて打ち返すタスクを考えた。ロボットはボールを打ち返したときのゴール位置との近さに応じて与えられる報酬のみの情報から，文脈に適応した動作を獲得できるかどうかを評価した。




（上） 異なるボールの初期位置およびゴール位置。
（左下）異なる初期位置からのスローイン。
（右下）異なるゴール位置へのヘディング動作。

図１　１脚ロボットのヘディング動作学習

まず方策生成フェーズにおいて，異なる文脈，つまりゴール位置やボールの反発係数などロボットの周りの世界の状況を乱択した場合のヘディング動作の学習をおこなう。ここで，動作生成を行うニューラルネットワークモデルとしてロボットの状態 s を入力とし目標関節角度 a を出力とする方策ネットワークを学習することを考える。この方策ネットワークは文脈 c を入力とし，潜在変数 z を出力とするエンコーダネットワークと接続される。また，価値関数を学習するネットワークを同時に用いる。これら３つのニューラルネットワークは図２のように構成した。

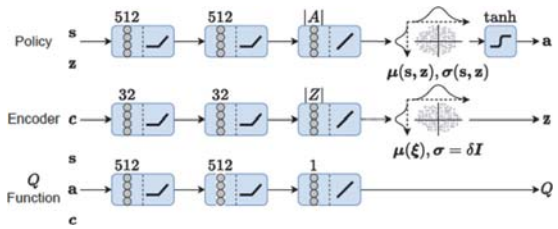



図２　方策，エンコーダ，価値関数を学習する
３つのニューラルネットワーク

方策選択フェーズにおいては，方策生成フェーズにおいて得られた方策を，各方策がタスク中に獲得した報酬の和に基づいて順位付けを行い，上位のものを方策適応フェーズにおいて活用するための選択を行う。３つの学習フェーズについて，図３および表１から３に各情報処理とそれらの接続について示した。

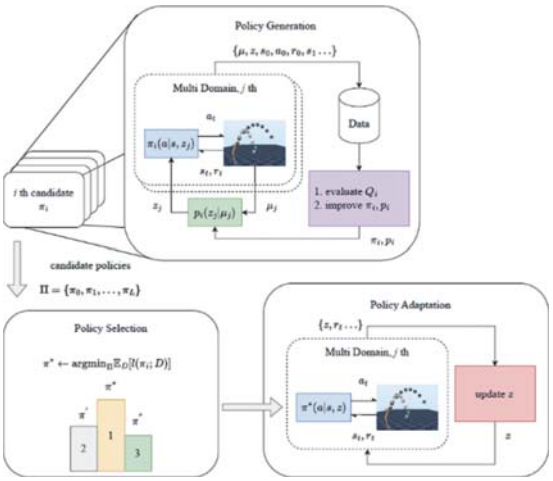



図３　（上）方策生成フェーズ：異なる文脈に対応する方策を生成。（左下）：方策選択フェーズ：生成された方策を順位付け，新たな文脈に適応させる方策を選択。（右下）：方策適応フェーズ：選択された方策を基礎として文脈を推定し，新しい状況に適応。

表１　方策生成フェーズ

**Algorithm 1** Generation Phase

**Input:** $\phi_{\{1,2\}}, \psi, \eta$
$\phi^{\text{target}}_{\{1,2\}} \leftarrow \phi_{\{1,2\}}$
$\mathcal{D} \leftarrow \emptyset$
**for** each epoch **do**
  sample domain and encoded parameter:
    $z \sim q(z|c;\eta), c \sim p(c)$
  **for** collect state-action steps **do**
    $a_t \sim \pi(a|s,z;\psi)$
    $s_{t+1} \sim p(s_{t+1}|s_t,a_t,c)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t,a_t,r(s_t,a_t,c),s_{t+1},c)\}$
  **end for**
  **for** each gradient step **do**
    update policy $\psi$
    update encoder $\eta$
    update $Q$-function $\phi_{1,2}$
    update temperature $\alpha$
    update target $Q$-function $\phi^{\text{target}}_{1,2}$
  **end for**
**end for**
**Output** $\phi_{\{1,2\}}, \psi, \eta$

表２　方策選択フェーズ

**Algorithm 2** Selection Phase

Reset candidate policies $\Pi = \phi$
Setup randomized domain with $\text{num}_c$ contexts $c$
**for** iteration $k = 1, \ldots, K$ **do**
  Train $\pi_k$ with contexts $\{c\}$ ▷ Generation Phase
  Rollout $\pi_k$ and collect $R(c) := \sum_t \gamma^t (r_t - \alpha \log \pi_t)$
  $l_k = \frac{1}{\text{num}_c} \sum_c \left[ \beta \log \frac{q(z|c)}{\rho(z)} - R(c) \right]$ in Eq. (10)
  $\Pi \leftarrow \Pi \cup \{\pi_k, l_k\}$
**end for**
Select $\pi_s$ based on sorted $\Pi$ by $\{l_k\}$
Deploy $\pi_s$ to domain with new context $c'$

表3　方策適応フェーズ

**Algorithm 3** Adaptation Phase

$\pi \leftarrow$ trained policy in Policy Search phase
$\mathcal{D} \leftarrow \emptyset$
$z^* = \mathbf{0}$
TPE model initialization
**for** iteration $k = 1, \ldots, k_{\max}$ **do**
  $z_k \leftarrow$ TPE suggests
  $J(z_k) \leftarrow$ Rollout policy $\pi(a|s, z_k)$ in Eq. (14)
  Store $(z_k, J(z_k))$ in $D$
  TPE updates model $p(z|J)$
**end for**
Select $z_k$ based on the sorted $\mathcal{D}$ by indices $J_k$

状態 s として，重心位置，頭部位置，足部位置，各関節角および角速度，ボール位置および速度などを観測した（表4参照）。

表4　観測値

| | Dimensionality | Type |
|---|---|---|
| Center of mass | 2 | float |
| Head position | 2 | float |
| Foot position | 2 | float |
| Joint angle | 5 | float |
| Joint angular velocity | 5 | float |
| Ball position | 2 | float |
| Ball velocity | 2 | float |
| One-step before action | 5 | float |
| Total dimensionality | 25 | |

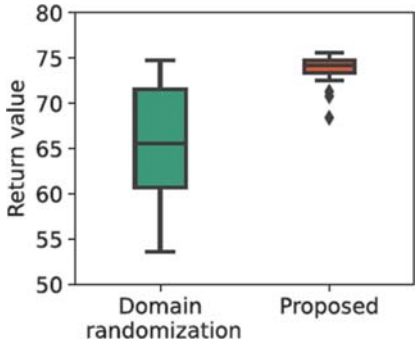
異なる環境およびタスクとしては，表5に示すようにボールの反発係数，スローイン位置，ゴール位置を変化させて方策生成をおこなった。

表5　異なる状況の設定

| | Type | Settings |
|---|---|---|
| Coefficient of restitution | float | {small, large} |
| Horizontal goal position | float | {near, far} |
| Vertical goal position | float | {low, high} |
| Horizontal throw-in position | float | {near, far} |
| Vertical throw-in position vertical-axis | float | {low, hight} |

## [研究の成果]

新しい文脈への方策の適応能力についての評価をおこなった。具体的には，1脚ロボットのヘディングタスクにおいて，32種の異なる環境（ボールの反発係数），異なるタスク設定（スローイン位置，ゴール位置）に対して，どの程度運動課題を達成できるかを評価した。ベースラインとして，一般に異なる状況に対応する方法論として用いられる環境乱択化の手法




（左）一般的な環境乱択化手法を用いた場合。適応能力にばらつきがある。
（右）提案手法。一貫して高い報酬が得られており，適応能力の高さを示している。

図4　新たな状況への適応能力の比較

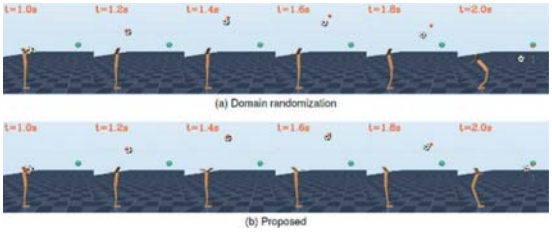



図5　（上）環境乱択化，（下）提案手法

を用いた場合と比較した。その結果を図4および5に示した。環境乱択化を用いる場合と比べて，提案手法では文脈によらず高い適応能力を示していることがわかる。

さらに，図6には適応前と適応後のタスク達成能力を定量的に示した。適応前と比べ，ロボットの周りの世界の状況の変化に対して，適応後においては得られる報酬の総和が高くなっていることがわかる。

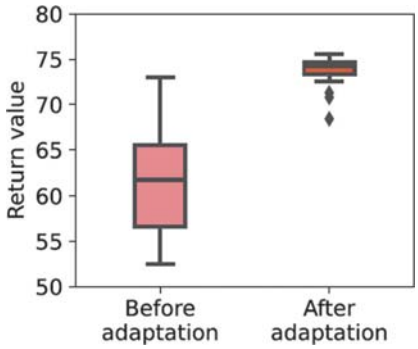



図6　（左）適応前，（右）適応後

実際にどのように，ロボットの周りの世界の状況変化に適応したかを図7に示した。適応前

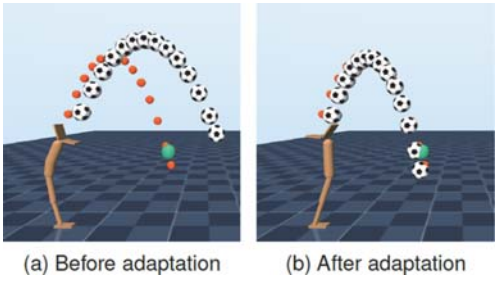



図７（左）適応前，（右）適応後

は緑点で示されたゴールにボールを打ち返すことができていないが，適応後には正確にボールを操ることができていることがわかる。

この新しい文脈変化に対しては，ロボットは提案手法を用いて，陽にその変化を観測することなく，得られる報酬の情報から新たな文脈を推定することに成功している。この推定は，方策生成フェーズにおいてエンコーダネットワークを用いて文脈を表現する適切な潜在空間を同定することで可能となっている。エンコーダネットワークの学習においては，表５に示した設定において，文脈ｃを陽に入力として与えることで潜在空間の同定をおこなっていた。そこで，実際にはあり得ないことであるが，仮に新たなロボットの周りの世界の状況ｃが陽に観測可能であるとして，エンコーダ入力に真の文脈を入力として与えた場合と提案手法の結果を比較した。

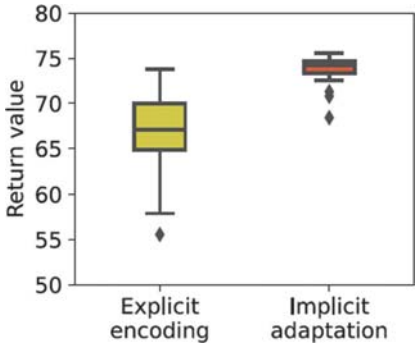



図８（左）文脈を与えた場合，（右）提案手法

図８に示した結果の通り，提案手法を用いて新しい文脈を推定した場合の方が良好なタスク達成結果が得られている。

これはエンコータネットワーク学習時には経験していない新たな文脈に対して適応する必要があるため，たとえ真の文脈情報が得られたとしても方策は必ずしも高い適応能力を示すわけではないことを示している。一方で，提案手法を用いた場合においては，得られる報酬から文脈推定を行うため，高い適応能力を示すことが確認された。これはロボットの周りの状況を表現する潜在空間が適切に同定されつつも，新たな状況に対しては，報酬の情報から文脈推定を行うことの有効性を示している。

## ［今後の研究の方向，課題］

近年，本研究で開発したような，文脈をエンコータネットワークで認識し，それに応じた制御をおこなう枠組みが有望なアプローチとして考えられるようになってきた。本研究においては現在，図９のようなロボット実験環境においてエンコーダネットワークへの入力として自然言語を用いる方法論の開発を進めている。

図９　ボトルに向かうように自然言語で指示した場合のロボット動作

## ［成果の発表，論文等］

・Satoshi Yamamori and Jun Morimoto, Foundational Policy Acquisition via Multitask Learning for Motor Skill Generation, arXiv: 2308. 16471, 2023
・特許出願済