ヒトの教示やインタラクションを活用したロボット学習

濵屋 政志

近年の目覚ましい学習技術の発展により、ロボットが複雑な作業を自律的に達成できる可能性を示してきた。しかし、その背景には、ロボットが適切に動作するまでにパラメータや制御目標（報酬関数）の設計をはじめとする、無視できない手間や労力がある。さらに、たとえ制御器（方策）を適切に学習できても、外乱など環境の変化に関して脆弱であることが知られている。この問題に対して、ヒトの知識を活用し、ロボット学習をより容易にさせる研究が行われている。ヒトがロボットに教示を与える、あるいは物理的に誘導することによって学習を加速できる。我々は、失敗の教示や、物理的な外乱など、「矛盾した」情報をあえてヒトが与えることによって、さらに学習の性能を向上させる手法を開発した。本稿では、我々の最新の成果である「矛盾からのロボット学習手法」に関して解説する。

Robotic Learning from Human Demonstrations and Interactions

HAMAYA Masashi

Thanks to recent learning techniques, robots can obtain skills for complex tasks automatically. Meanwhile, engineering efforts of tuning hyperparameters and designing appropriate reward functions are not trivial. Some researchers have leveraged human domain knowledge such as demonstrations and interactions to facilitate robotic learning. In this article, we introduce our recent works using contradictional information provided by humans. Concretely, we present learning from[1)] successful and failed demonstrations and[2)] advisory and adversarial interactions.

1. まえがき

近年の機械学習・強化学習の目覚ましい発展に伴い、ロボットに学習技術を適用し、自律的に作業戦略を獲得する研究が盛んになってきている。特に、深層学習による優れた汎化性や表現能力により、画像からロボットの行動を直接決定することができるなど、多感覚統合を実現することができる。また、解析的な表現が困難であるダイナミクスを持つシステムの制御に関しても学習技術は極めて有効である。

しかしながら、ロボットが適切に動作する背景には無視できない労力がある。たとえば、適切なハイパーパラメータや、報酬関数の設定などが挙げられる。これらの設定を実ロボットで試行錯誤的に検証することは、データを取得するための労力が大きい。特に、組立作業など対象物体とロボット間の接触を多く含む作業では、学習初期においてロボットが予期せぬ動作によって、部品の破損やロボット緊急停止が生じる可能性がある。さらに、たとえロボットが無事に作業を学習できたとしても、学習された制御方策は、外力、ノイズ、モデル化誤差などの外乱に対して脆弱であることが知られている。

この問題を解決するためには、いくつかの例が挙げられる。一つ目は、本誌の別章で紹介した「サンプル効率の良い学習手法」を適用することである。たとえば、モデルベース強化学習などのアプローチは、複数の試行回数で所望のスクを学習することができるため、設定したパラメータや報酬関数が理にかなっているかどうかを早い段階で評価できる。次に、シミュレータを使用し、学習した方策あるいはモデルを実世界に転移させることも考えられる。シミュレータにより、実世界におけるデータ収集の手間を大幅に削減できることが期待される。そして、これらに加えて、ヒトのドメイン知識を活用する方法が考えられる。たとえば、ヒトが教示を与える、動作の良し悪しを評価し、

Contact : *HAMAYA Masashi* masashi.hamaya@sinicx.com

ロボットにフィードバックを与える、あるいは、ロボットと物理的にインタラクションしながら、適切な動作を学習させる、などの方法が挙げられる。

このような作業を成功に導くヒトの教示やインタラクションに加えて、本稿では失敗の教示や敵対的なインタラクションをあえて与える「矛盾からの学習」を活用した我々の最新の研究成果を紹介する。さらに、物理的柔軟さを持つロボットを使用することで、教示やインタラクションが容易になることも併せて紹介する。第 2 節では、成功と失敗の教示を与えることで、失敗を回避し作業の成功率を向上させる学習手法[1] について紹介する。第 3 節では、ロボットの学習中に、作業を成功に導く外力を与えることで、学習を促進させる誘導インタラクションと、作業を阻害する外力を与えることで、方策を頑健にさせる敵対インタラクションによる学習手法[2] について紹介する。

2. 成功・失敗教示からの学習

本節では、図 1 に示すように成功と失敗の教示からロボットを学習させる手法について紹介する。ロボットが作業を学習する際、制御目標（報酬関数）の設計は極めて重要である。適切でない報酬関数を設計した場合、ロボットの学習性能が著しく低下する可能性がある。そこで、ヒトからの教示を活用して報酬関数を学習することを考える。

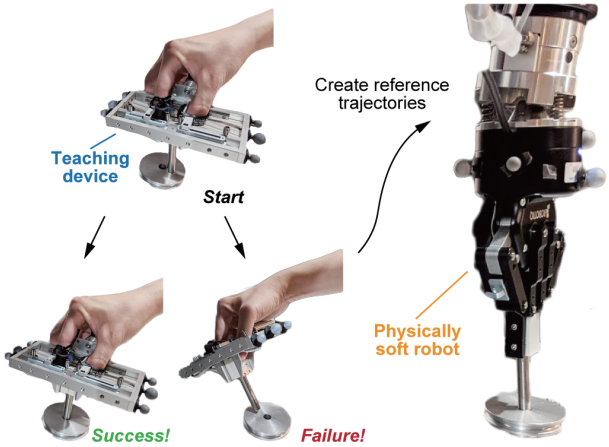


図 1　成功・失敗教示からの学習（文献[1] から引用）

教示からの学習は、Learning from Demonstration と呼ばれており、2000 年代から研究が盛んに行われている[3]。教示の手法は、主にロボットに直接触れて教示させるダイレクトティーチングや、特殊なデバイスで遠隔で操作する手法が挙げられる。しかし、ダイレクトティーチングは、ロボットの可動域を考慮する必要があるため、精密な教示は困難である。また、ゲームパッドなどのデバイスを使用してロボットと同期的に教示を与えるためには、教示者の熟練が必要になる。

そこで、ヒトの動作を直接計測して、ロボットに与えることができれば、より直感的な教示が可能となるが、ロボットとヒトの身体（キネマティクス・ダイナミクスの）差が問題となる。この問題点に対して、我々はハードウェアからのアプローチを考えた。我々は、直感的に教示ができ、かつロボットとの身体差を軽減させることができる新しい教示デバイスを開発した[1]（図 2）。このデバイスの動作部はロボットのグリッパと形状が近いため、身体の違いを軽減させることができるのが特徴である。

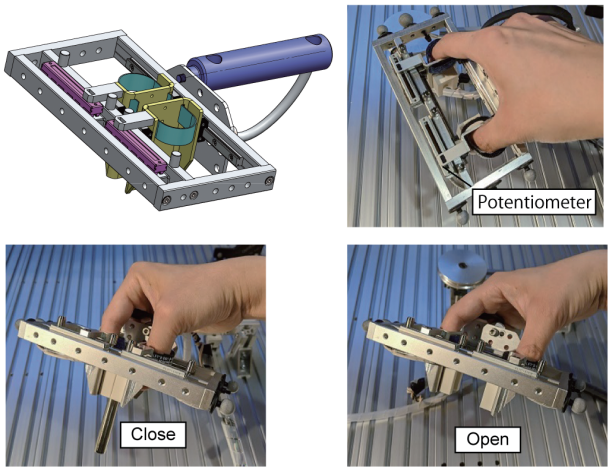


図 2　教示デバイス（文献[1] から引用）

さらに本研究では、柔軟手首を持つロボットを使用する。柔軟要素は環境との安全な接触を許容するので、教示デバイスから位置・姿勢の情報のみを与えるだけで、組立動作など接触を多く含む作業の教示が可能となる。つまり、教示デバイスと対象物体の接触力を計測する必要がなく、デバイスの軽量化・簡素化が期待できる。

提案した教示デバイスと柔軟なロボットにより、ヒトがより直感的に教示を行うことができるため、成功だけでなく、失敗の教示を与えることを考えた。ロボットの学習においては、過去の失敗から遠ざかるように拘束をかける手法[4] や、失敗や不完全な教示から学習する手法が提案されている[5]。これらの手法に対し、我々は、成功・失敗の両方の教示を与えて、失敗教示軌道をなるべく避けるように、成功軌道を生成する手法を提案する。

2.1 目標軌道生成・ロボット動作学習

図 3 に本提案手法の手順を示す。本手法は、ロボットの強化学習問題と考え、報酬関数となる目標軌道の生成を、失敗・成功軌道の教示によって生成する。

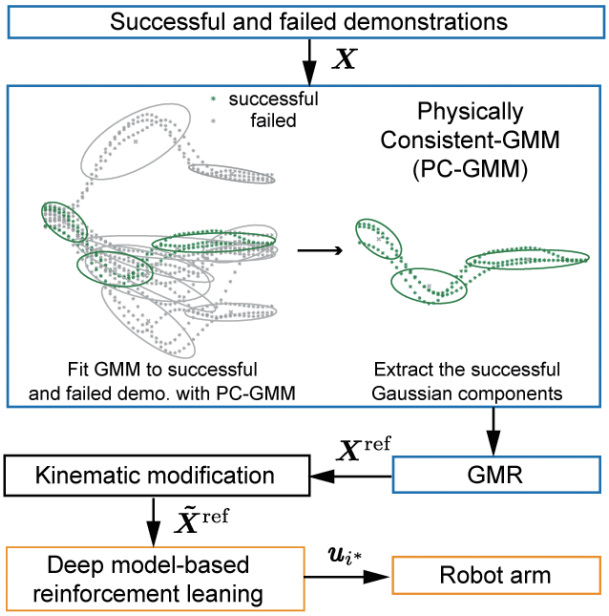


図 3　提案手法の手順（文献[1] から引用）

まず、教示デバイスで、成功・失敗の教示データを収集し、これらのデータから目標軌道を生成する。目標軌道を生成するために、Physically-consisted Gaussian Mixture Model（PC-GMM）[6] を適用する。GMM はガウス分布の線形重ね合わせで表現されるモデルであり、PC-GMM は、データにガウス分布を適合させるときに、データ同士の類似度を計算し、類似度が近いデータに同じガウス分布を割り当てるようにする。また、最適なガウス分布の数は、ノンパラメトリックベイズによって自動的に決定される。本研究では、失敗と成功データの類似度計算を行うときに、類似度を下げる項を導入し、成功・失敗データがそれぞれ離れるようにガウス分布が割り当てられるようにする。結果、失敗軌道を避けた状態で、成功軌道にガウス分布が適合される。そして適合された GMM から、連続軌道を生成するため、Gaussian Mixture Regression（GMR）を適用する。GMR によって、時間依存の目標の位置・姿勢の軌道を生成する。

次に、得られた軌道とロボットの動作軌道との誤差を報酬関数として、強化学習を行う。本研究では、ニューラルネットワークを使用したモデルベース強化学習[7] を適用する。この手法は、ニューラルネットワークで学習された複数の順モデルを同時に学習し、モデルをランダムに選択しながら、動作予測することで、不確実性に対処でき、結果としてサンプル効率が向上する。また、行動選択には、Cross Entropy Method（CEM）を使用する。CEM は反復的に最適行動を計算する手法である。ランダムに生成した行動をモデルに入力し、数ステップ間の期待収益が最も高くなった行動上位数種類からガウス分布を生成し、次の反復からはそのガウス分布から行動の候補をサンプルする。CEM は一様乱数的に行動の候補をサンプルするより効率的である。

2.2 実機実験

提案手法の有効性を検証するために、実機実験を行った。本実験の目的は、成功・失敗教示を両方使用することで、より高い作業成功率を示すことである。本研究では、ペグ挿入作業を課題とした。教示軌道は、成功軌道と 4 種類の失敗軌道をそれぞれ3回ずつ与えた（図4）。ロボットには柔軟手首[8] を搭載した。教示デバイスと、ロボットの手先の位置姿勢を計測するために、モーションキャプチャシステムを使用した。比較手法は、成功・失敗教示を与えた場合（success and failure）、成功教示のみ与えた場合（success only）、教示を与えず、挿入穴の位置のみを与えた場合（no demo）とする。

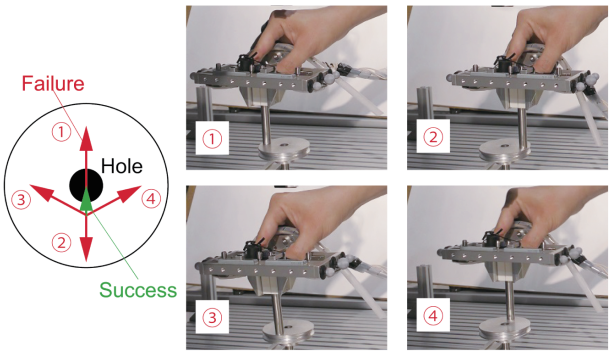


図 4　成功・失敗軌道の教示例（文献[1] から引用）

図 5 に、学習時における、穴の位置との誤差を示す。左は xyz 方向に、右図は z 方向（穴の挿入方向）における絶対平均誤差を示す。エラーバーは 10 回の学習実験における標準偏差を示す。この結果から、提案手法が最も小さい誤差を示した。また、学習後の作業成功率も、提案手法 8/10、成功教示のみ 5/10、教示無し 0/10 となり、提案手法が最も高い成功率を示した。成功軌道教示のみの場合は、ペグを穴の縁に当てる動作の際に、勢いが余り頻繁に穴を通り過ぎた。これに対し、提案手法は、この失敗を避けるように軌道が生成され、結果として高い成功率が示されたことが考えられる。

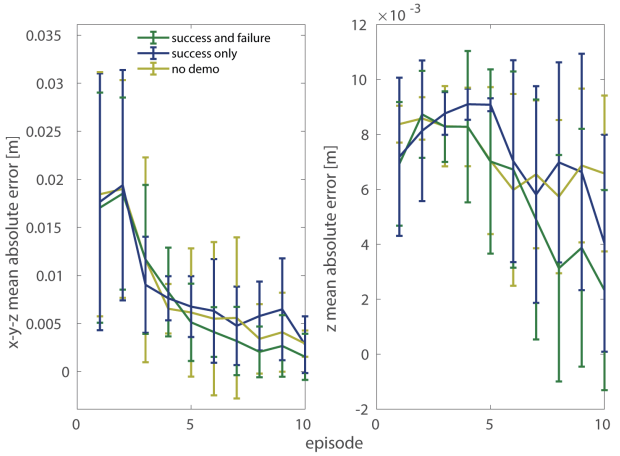


図 5　学習時における位置誤差（文献[1] から引用）

以上により、成功と失敗の教示を活用することで、ペグ挿入作業において高い成功率を示すことが確認できた。今後は、提案手法の理論保証や、どのように失敗教示を与えるかという課題について取り組む予定である。

3. 物理インタラクションからの学習

本節では、誘導・敵対インタラクションからロボットを学習させる手法について紹介する。前節の教示からの学習で、失敗・成功教を与えて学習させた手法は、教示を与えない場合より高い成功率を示した。しかし、適切な教示を与えるためには、熟練が必要となる。

熟練者でない場合においても、ヒトの知識を活用してロボット学習を促進させる方法はあるだろうか？一つの例として、ヒトのフィードバックを活用する方法が挙げられる。ロボットがある作業を遂行した後に、ヒトがロボットの作業の良し悪しを与える。ロボットは与えられたフィードバックをもとに、作業を改善するように学習する。方策を直接改善する手法[9]、あるいは報酬関数を改善する手法[10] が提案されている。しかしながら、作業後のフィードバックは、作業中のどの行動が良いか悪いかを与えることができないため、フィードバックの質が課題となる。

一方、たとえロボットが適切に作業を達成できたとしても、学習された方策・モデルは外乱、モデル化誤差に関して極めて脆弱であることが知られている[11]。この問題に対して、学習中に敵対的な外力を加えて、より頑健な方策を学習する手法が提案されている。実ロボットに適用する際には、作業を行うロボットと外乱を加えるロボットを同時に学習させる[12] 手法があるが、複数台のロボットのセットアップのために大きな労力が生じる可能性がある。

これらの二つの問題を同時に解決するために、本研究では、ヒトの物理インタラクションを活用する手法を提案する。ヒトとロボットが常に接触している状態であるため、任意のタイミングでヒトがフィードバックとしてのインタラクションを発生させることができる。そして、本研究では、誘導（Advisory）・敵対（Adversarial）インタラクションを提案する（図 6）。誘導インタラクションは、学習中の目標誤差を減らす役割を持ち、敵対インタラクションは学習方策を頑健にさせる役割を持つ。また、ヒトとロボットの安全な接触を保障するため、本研究においても柔軟要素を持つロボットを使用することを提案する。

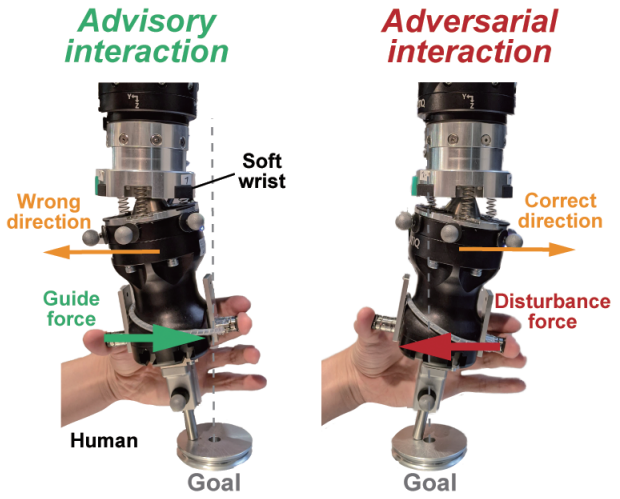


図 6　物理インタラクションからの学習（文献[2] から引用）

3.1 誘導・敵対インタラクションからの学習

本研究では、誘導・敵対インタラクションから学習する問題をモデルベース強化学習として定式化する。ヒトとロボットのインタラクションを考慮したダイナミクスモデルは $\boldsymbol{s}_{t+1} = f(\boldsymbol{s}_t, \boldsymbol{a}_t)$ で示される。ここで、$\boldsymbol{s} = [\boldsymbol{s}^{R\top}, \boldsymbol{s}^{H\top}]^\top$ は状態、$\boldsymbol{s}^R$ はロボットの状態（関節角度や手先位置など）、$\boldsymbol{s}^H$ はヒトの状態（ヒトがロボットに与える外力）、$\boldsymbol{a}_t$ はロボットの行動（関節トルクや、手先位置の目標速度指令など）である。本研究では、ヒトからのインタラクションをロボットに与えながらモデルを学習させる。報酬は $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$ で与えられ、累積期待収益を最大化させる行動を決定することがこの学習の目的となる。

誘導インタラクション：ロボットが作業中に、目標から離れるような間違った方向に動作したとき、ヒトが正しい方向に外力を加える。誘導インタラクションにおける報酬関数は、$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = r^R(\boldsymbol{s}^R, \boldsymbol{a}^R) - \alpha \left\| \boldsymbol{s}^H \right\|^2$ で与えられる。ここで、r^R は、ロボットにおける報酬であり目標の位置誤差であり、α は重み係数である。ヒトからの誘導外力が大きければ大きいほど、罰則が与えられるため、ロボットはヒトからの外力をなるべく受けないように動作することが期待される。つまり、ヒトから誘導インタラクションは、作業に使用される報酬の補助的な情報となっており、報酬の表現力が向上されると考えられる。その結果、学習中の目標状態との誤差を軽減できると期待される。

敵対インタラクション：誘導インタラクションによって

ロボットがある程度学習された後に、敵対インタラクションを適用する。ここでは、ロボットが目標に正しい方向に動作した場合、その方向とは逆にヒトが外力を加える。報酬は $r(s_t, a_t) = r^R(s^R, a^R) - \beta \left\| s^H \right\|^2$ で与えられる。ここで、β は重み係数である。大きな敵対外力を受けるほど、報酬も大きくなり、結果外乱に対して頑健な方策が学習される。

モデルベース強化学習は前節と同じアルゴリズムを使用する[7]。ヒトとロボットのインタラクションを含む複雑なモデルに対しても、少ない試行回数で方策を学習できると期待される。

3.2 シミュレーション実機実験

提案手法の有効性を検証するため、シミュレーションを行った。シミュレーションの目的は、提案手法がインタラクションを用いない場合と比較して、学習中の誤差を減少させ、頑健な方策を獲得できるかを確認することである。本研究においても検証対象としてペグ挿入課題を扱う。

本実験には 6 人の被験者が参加した。被験者はモニターに表示されるロボットに、ゲームパッドで外力を与えた(図 7)。ロボットシミュレータは、PyBullet という動力学シミュレータを使用して設計された。ロボットには、柔軟手首[8] が搭載されており、柔軟要素は 6 自由度の PID 制御によって再現されている。

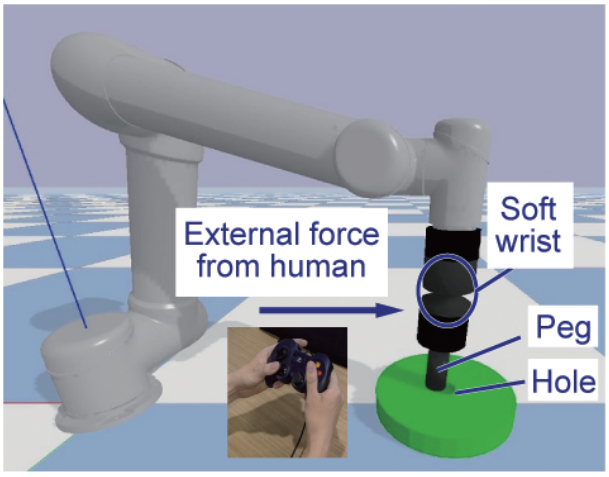


図 7　シミュレーションの概要（文献[2] から引用）

本シミュレーションでは、1 回の学習実験で、7 回の誘導インタラクション、7 回の敵対インタラクション、計 14 回行った。評価として、インタラクションなし（no interaction）、ランダム外力からインタラクション（random interaction）、経験則で設計した誘導・敵対インタラクション（heuristic interaction）、提案手法の 4 種類の状況で性能を比較した。

図 8 に学習中の目標位置との平均誤差を示す。橙色の線はインタラクションなし、黄色はランダムインタラクション、緑は経験則インタラクション、青は提案手法である。提案手法を使用した結果、誘導インタラクションにおいて、学習中の誤差を減少させることができた。また、敵対インタラクションにおいても、提案手法の誤差が一度増加するが、最終試行においては、インタラクションを与えないあるいはランダムなインタラクションを与えた場合よりも小さい誤差を示した。

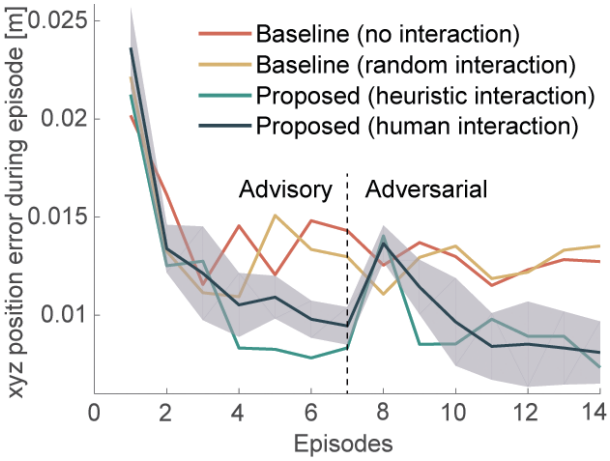


図 8　目標位置の誤差（文献[2] から引用）

次に、学習された方策の頑健性を検証するため、ロボットの質量や、穴の摩擦係数を変更してテストを行った。図 9 に未知環境におけるペグ挿入作業の成功回数を示す。この結果から、提案手法はすべての場合において、最も高い成功率を示した。以上の実験から、提案手法の有効性が示された。

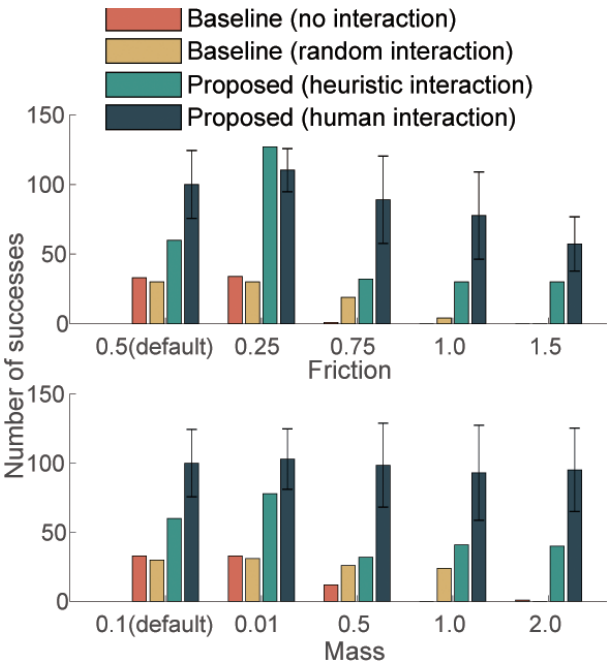


図 9　未知環境での作業成功回数（文献[2] から引用）

3.3 実機実験

最後に、実機を使用した実験を行った。提案手法が実世界においても適用可能かを検証するのが目的であった。

実機実験においても柔軟手首[8]を使用した。ロボット及び学習のセットアップは前節と同じであるが、ヒトの外力を計測するために、触覚センサを搭載した。ロボットを様々な方向から学習させた場合と、異なる素材のペグを適用したときの作業成功率を、インタラクションを与えない場合と比較した。

図 10 に、穴の 3 方向からペグ挿入作業を行った場合と、ある 1 方向から、様々な素材のペグで作業を行った時の成功回数を示す。橙色の棒グラフはインタラクションを与えない場合で、青は提案手法である。この結果より、提案手法はすべての場合において高い成功回数を示した。以上より、実機実験においても提案手法の有効性を示した。

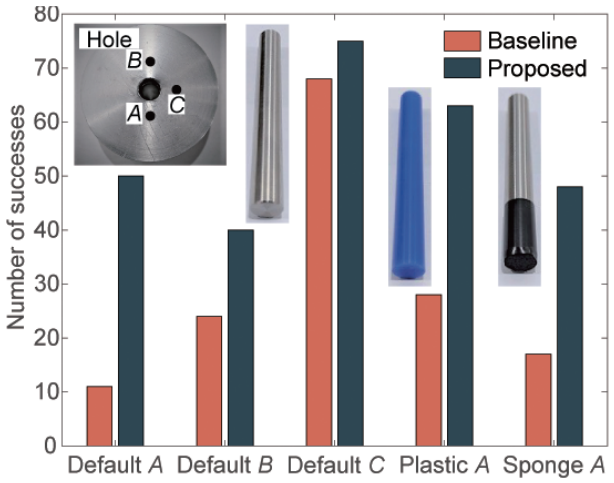


図 10　実機実験における作業成功回数（文献[2]から引用）

今後の課題として、更なる性能向上のために、修正・敵対インタラクションのより適切な順序やユーザビリティを向上させる手法を検討する予定である。

4. むすび

本稿では、ヒトの知識を活用して、ロボット学習をより容易にする我々の最新の研究成果について紹介した。核心となるアイデアは、「矛盾した」情報を与えることで、ロボット学習をより促進させるという点である。1 つ目の研究事例として、ヒトからの教示を利用する際に、成功軌道だけでなく、失敗軌道も陽に活用することで失敗を避けるような目標軌道生成手法を紹介した。2 つ目の事例として、ヒトがロボットに直接触れて学習する際、誘導インタラクションを活用することで、学習中の誤差を減少させ、学習後の方策を頑健にすることを紹介した。また、これらの作業を行う際に、対象物体やヒトとの安全な接触のために、柔らかさを持ったロボットが重要な役割を持つことも注目されたい。

本稿を読み、ヒトの知識を活用して学習を促進させる研究に興味を持っていただければ幸いである。

参考文献

1）Hamaya, M.; von Drigalski, F.; Matsubara, T.; Tanaka, K.; Lee, R.; Nakashima, C.; Shibata, Y.; Ijiri, Y. “Learning Soft Robotic Assembly Strategies from Successful and Failed Demonstration”. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020, p. 8309–8315.

2）Hamaya, M.; Tanaka, K.; Shibata, Y.; von Drigalski, F.; Nakashima, C.; Ijiri, Y. Robotic Learning From Advisory and Adversarial Interactions Using a Soft Wrist. IEEE Robotics and Automation Letters. 2021, Vol.6, No.2, p. 3878–3885.

3）Billard, A.; Calinon, S.; Dillmann, R.; Schaal, S. Survey: Robot programming by demonstration, Handbook of Robotics. Springer, 2008, Chapter 59, p.1371–1394.

4）Esteban, D.; Rozo, L.; Caldwell, D. G. “Learning deep robot controllers by exploiting successful and failed executions”. IEEE RAS International Conference on Humanoid Robots, 2018, p. 1–9.

5）Gao, Y.; Xu, H.; Lin, J.; Yu, F.; Levine, S.; Darrell, T. Reinforcement learning from imperfect demonstrations. International Conference on Learning and Representation. arXiv preprint arXiv: 1802.05313.

6）Figueroa, N.; Billard, A. “A physically-consistent bayesian nonparametric mixture model for dynamical system learning”. Conference on Robot Learning, 2018, p. 927–946.

7）Chua, K.; Calandra, R.; McAllister, R.; Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. Advances in Neural Information Processing Systems. 2018, Vol.31, p. 4754–4765

8）Tanaka, K.; von Drigalski, F.; Hamaya, M.; Lee, R.; Nakashima, C. Shibata, Y.; Ijiri, Y. “A Compact, Cable-driven, Activatable Soft Wrist with Six Degrees of Freedom for Assembly Tasks”. IEEE/RSJ International Conference on Intelligent Robots and Systems. 2020, p. 8752–8757.

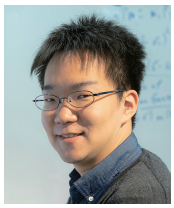
9）Celemin, C. E.; Maeda, G.; Ruiz-del Solar, J.; Peters, J.; Kober, J. Reinforcement learning of motor skills using policy search and human corrective advice. International Journal of Robotics Research. 2019, Vol.38, No.14, p. 1560–1580.

10）Bıyık, E.; Palan, M.; Landolfi, N. C.; Losey, D. P.; Sadigh, D. “Asking easy questions: A user-friendly approach to active reward learning”. Conference on Robot Learning, 2020, p. 1177–1190.

11）Ilahi, I.; Usama, M.; Qadir, J.; Janjua, M. U.; Al-Fuqaha, A.; Hoang, D. T.; Niyato, D.Challenges and countermeasures for adversarial attacks on deep reinforcement learning. arXiv preprint-arXiv:2001.09684, 2020.

12）Pinto, L.; Davidson, J.; Gupta, A. “Supervision via competition: Robot adversaries for learning tasks”. IEEE International Conference on Robotics and Automation, 2017, p. 1601–1608.

執筆者紹介



濵屋 政志　HAMAYA Masashi
オムロン サイニックエックス株式会社
リサーチアドミニストレイティブディビジョン
専門：ロボティクス、機械学習、強化学習
所属学会：日本ロボット学会、IEEE
博士（工学）

本文に掲載の商品の名称は、各社が商標としている場合があります。