2B1-1

ロボットマニピュレーションにおける画像内物体の深層学習による運動予測

Motion Prediction of Object in Image by Deep Learning for Robot Manipulation

室岡 雅樹 Masaki Murooka　二井谷 勇佑 Yusuke Niitani　和田 健太郎 Kentaro Wada
野沢 峻一 Shunichi Nozawa　垣内 洋平 Yohei Kakiuchi　岡田 慧 Kei Okada　稲葉 雅幸 Masayuki Inaba

東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

Predicting the motion of the manipulated object is useful for planning robot manipulation autonomously. In this paper, we propose the deep learning based approach to predict the 3D object motion from depth image and manipulation force. By generating the object motion dataset automatically with dynamics simulator and learning the deep learning model, which has Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), the complex object motion is predicted. We integrate the proposed object motion prediction with the robot manipulation system and show the effectiveness through the experiment with a real humanoid robot.

1. はじめに

ロボットによるマニピュレーションは計算機上での認識，探索，計画等の処理結果を現実世界に反映させる手段として重要な意義をもつ．ロボットは日常生活環境や災害現場のような多様な環境において活動することが期待されるが，このような環境は未知物体が複雑に配置され状態が時間を経て変化することから事前の完全なモデル化が困難であり，ロボットが適応的なマニピュレーションを自律的に行うための知能システムの構築が課題となる．筆者らは物体の幾何・物理特性モデルが既知であることを前提として，安定な物体運動遷移を表すグラフを用いたマニピュレーション計画法を提案している (Fig1)．

本論文では，モデル未知の物体に対して物体運動遷移に基づいたマニピュレーション計画を自動的に行うことを目的として，ロボットが物体にどのような操作を加えたときに物体がどのような運動をするかを深層学習により予測する手法を提案する．ロボットによるマニピュレーションを，視覚情報・操作情報から三次元物体運動への変換としてとらえ，Convolutional Neural Network (CNN) と Recurrent Neural Network (RNN) を統合したネットワークモデルによりこの変換を学習して再現する．ネットワークの学習に必要となるデータは動力学シミュレータを用いることで自動的に生成する．提案する物体運動予測手法によりヒューマノイドロボットが目的の物体操作を実現する実験を通して，本手法の有用性を示す．

1.1 物体運動予測の関連研究

人間は高度な感覚運動系を有しており，認知科学の分野では脳内に自身の身体や外界の環境を表す内部モデル [Wolpert 95] が構築されることでこれが実現されているとの仮説が立てられている．現実世界の物体は Newton の運動法則等の物理法則に従って運動を行うが，人間はこれを知らずとも物体の運動を予測することができる [Fischer 16]．本論文で扱う物体運動予測はこの内部モデルに相当するものとして位置づけることができ，学習モデルに運動法則を明示的に与えることなくこれを獲得することを目指す．

連絡先: 室岡 雅樹, 東京大学大学院 情報理工学系研究科 知能機械情報学専攻, 東京都文京区本郷 7-3-1 工学部二号館 73B2, murooka@jsk.imi.i.u-tokyo.ac.jp

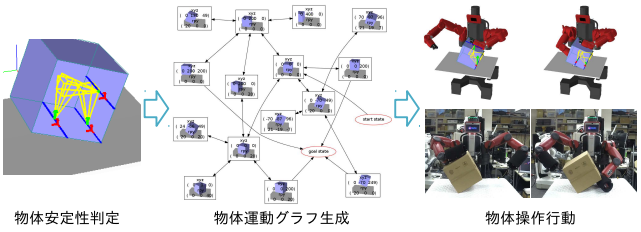

図 1: ロボットマニピュレーション計画 [Murooka 17]
幾何・物理特性モデルに基づいて安定な物体運動遷移を表すグラフを生成し経路を探索することで，ロボットのマニピュレーション行動を計画する．

コンピュータビジョンの分野においては，画像内の物理的意味理解 [Zheng 15] の延長として物体運動予測が研究されている [Wu 15]．これらの研究の多くはシーン画像内に含まれる情報から物体の運動を予測する Passive な予測である．ロボットはシーンに応じて能動的に操作を決定し実行することができるため，本論文ではシーン画像にこれとは独立な操作情報を合わせた情報を入力とする Active な予測を扱う．

三次元空間における物体運動は並進・回転変位の 6 自由度で表すことができる．先行研究の多くでは，回転運動を考慮せず並進運動のみを扱う [Mottaghi 16] ように自由度を限定した運動の予測を扱っており，本論文と同様に 6 自由度全ての運動を予測している研究は少ない [Byravan 16]．

一般にロボットのマニピュレーションは認識，計画，制御等の機能を統合することで実現されるが，最新の深層学習研究によってカメラ画像からロボットの動作を直接生成することが可能になりつつある [Levine 16]．これらの研究においては，対象タスクごとに多数のデータを作成しモデルの再学習を行う必要があることが汎用性向上の際の障壁であると考えられる．一方，本論文で扱う物体運動予測は，十分な汎化が達成されれば物体運動の伴う様々なタスクに適用され得る機能である．

1.2 ロボットマニピュレーションシステム

本論文で提案するロボットマニピュレーションシステムの構成を Fig2 に示す．ロボットが自律的にマニピュレーションを

実現するためには，視覚センサで得られた操作対象物の情報から，ロボットが動作するための物体操作行動指令を生成する必要がある．まず，視覚センサ情報から物体セグメンテーション処理によって物体の幾何的情報が抽出される．また，物体操作計画器が実行する候補となる操作を選択し物体操作力情報として出力する．これらの処理で得られた Depth-Mask 画像と物体操作力情報が深層学習モデルに与えられ，予測された物体運動が出力される．物体操作計画器はこの予測結果に基づいて新たな操作候補の選択や実機への操作行動の指令を行う．

本論文では，物体セグメンテーション処理や物体操作計画器については簡易な既存手法を用いるものとして，深層学習で物体運動を予測する手法について次章以降で詳しく扱う．

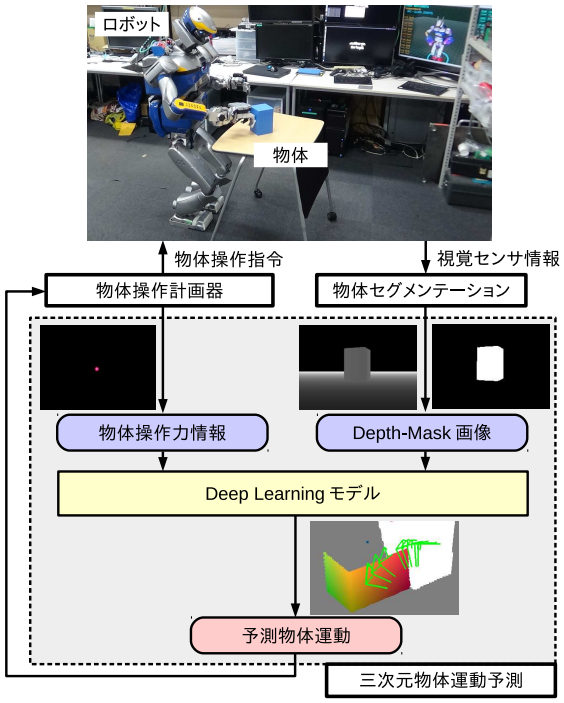

図 2: 物体運動予測を統合したロボットマニピュレーションシステム
ロボットの視覚情報から抽出された物体情報と物体操作計画器が出力した物体操作力情報を深層学習モデルに与えることで操作実行時の物体運動が予測される．物体操作計画器は予測された物体運動に基づいて目的のタスクを実現する物体操作行動を決定し実機へ指令する．

2. 深層学習による物体運動予測

2.1 物体運動予測の問題設定

物体操作は (a) 物体と環境の特性· 状態と (b) 物体に加える操作から (c) 物体の運動が生じる現象である．現実世界ではこの現象は Newton の運動法則や幾何干渉拘束などの物理法則に従って生じるため，(a) 物体環境情報や (b) 操作情報が完全に既知である場合には計算によって物理法則を再現することで (c) 物体運動を得ることができ，物理シミュレータとしてこれが実装されている．一般に人間やロボットが物体を操作する場合にはこれらの情報を事前に完全に得ることは難しい．特に (a) 物体環境情報に関しては，視覚センサにより幾何的な情報が観測が可能であるが，オクルージョンやセンサ分解能の限界により完全な観測は不可能である．また，物体の質量や摩擦係数といった物理特性を操作前に取得することも困難である．人間はこのような条件下でも経験に基づきおおよその物体運動を予測することが可能であり，本研究ではこの能力をロボットにおいて実現することを目指す．

本論文では，上記の (a),(b) を入力として (c) を出力とする深層学習モデルを提案する．(a) 物体環境情報としてはロボットに搭載された Depth カメラから得られた Depth 画像と物体領域を表す Mask 画像を用いる．これらの画像は物体と環境の三次元形状を再現するために十分な情報を含んでいる．質量や摩擦係数等の物体物理特性は取得が困難であることから入力として扱わない．(b) 操作情報としては，ロボットが物体に加える操作力の作用点と力の向き· 大きさを用いる．(c) 物体運動は，6 次元並進· 回転変位の時系列データとして表現する．

2.2 深層学習ネットワークモデル

本研究で提案する深層学習ネットワークモデルの概要を Fig3 に示す．物体· 環境の形状情報を表す Depth-Mask 画像とマニピュレーションにおける操作力を表す力画像が，それぞれ CNN に入力され特徴抽出された後に結合される．この特徴を RNN に与えることで，任意の時系列長の物体運動を出力として得る．

Depth-Mask 画像は，ロボットに搭載された Depth カメラから得られる Depth 画像と画像中の物体領域を表す Mask 画像を合わせた 2 channel 画像である．力画像は，Depth-Mask 画像と同じサイズで力の作用点に対応するピクセルが力の向きと大きさを表すような 3 channel 画像である [Mottaghi 16]．各 channel はカメラ相対の座標系において力の x,y,z 成分を表す．汎化性能の向上を期待して Depth-Mask 画像，力画像には平滑化フィルタ処理を施した後，ガウシアンノイズを加える．

物体運動については，任意の実数を出力する 6 つの出力ユニットから，並進変位の x,y,z 成分と，回転変位の x,y,z 軸周り成分が得られる．物体運動における変位は，Depth 画像を三次元復元することで得られた物体に属するポイントクラウドの重心の運動をカメラ座標系で表現したものである．

2.3 物体運動データセットの自動生成

提案した深層学習モデルの学習には入出力データのセットが数万個規模で必要になるが，本論文では動力学シミュレータを活用することでデータセット生成を自動化し負担を低減する．Fig4 に動力学シミュレータとその上でシミュレートされた視覚センサ情報を示す．本論文では視覚センサのシミュレーション機能を備えた動力学シミュレータとして Gazebo[*1] を，操作適用からデータ記録までの一連のシステムの構築に ROS[*2] を用いた．

動力学シミュレータ上では RGB カメラ，Depth カメラを任意の位置に配置しシミュレートすることが可能である．指定色を配色された物体に色フィルタ処理を適用して Mask 画像を生成し，Mask 画像からランダムに選ばれた点を操作力の作用点として，ランダムな向き· 大きさの操作力を物体に加える．運動中の物体姿勢は動力学シミュレータの機能により記録され，運動後の物体は初期位置に自動で再配置される．このようにランダムな操作を繰り返し加えることで，大規模なデータセットを自動的に生成することが可能である．

本手法によって 10 時間で 1 万個程度のデータセットが自動的に生成された．本論文では箱型の物体のみを扱っているが，Web 上に公開されている物体形状モデルデータセットを用いることで多様な物体を扱うことも可能である．

*1 http://gazebosim.org/
*2 http://www.ros.org/

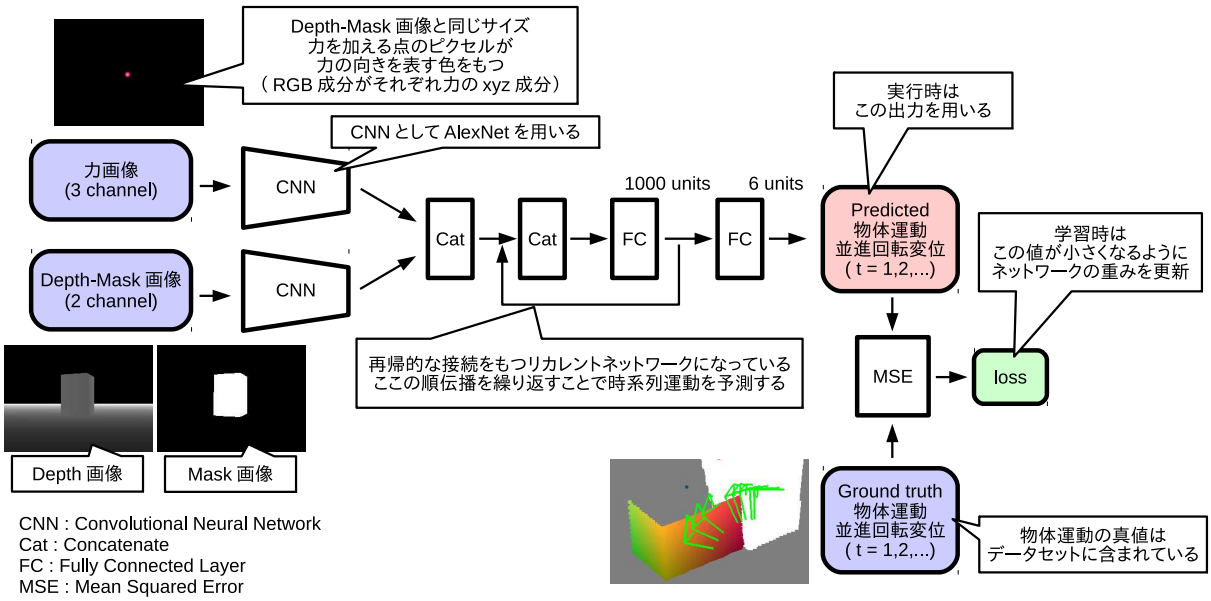

図 3: 物体運動予測のための深層学習ネットワークモデル
物体環境情報を表す Depth-Mask 画像と操作情報を表す力画像を入力として時系列物体運動を出力する．CNN で特徴を抽出した後に RNN によって時系列情報を生成することで物体運動が予測される．図中の青色で示される要素は学習データセットに含まれる情報を表す．

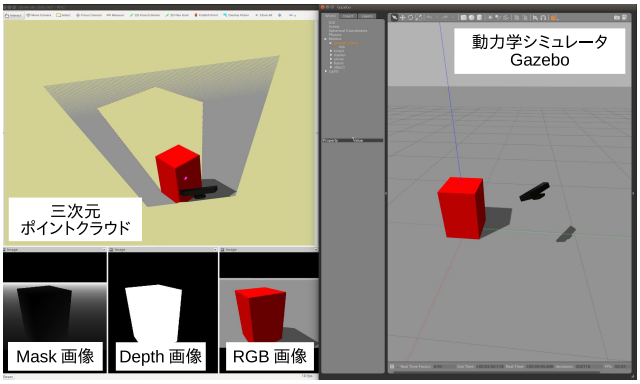

図 4: 動力学シミュレータを用いたデータセットの自動生成
ランダムな操作を動力学シミュレータ上で物体に加えその結果生じた運動を記録することで，多数のデータセットが自動的に生成される．動力学シミュレータ上では RGB カメラ，Depth カメラのシミュレートや物体運動の計測が可能である．

3. 物体運動の予測実験

3.1 物体運動モデルの学習

提案した深層学習モデルを自動生成したデータセットにより学習した．本論文では，物体は $100 \times 100 \times 150[\mathrm{mm}^3]$ の直方体形状物体に限定し，ランダムな操作を加えて生じた物体運動をランダムな視点から観測した 1 万通りのデータを生成した．各運動データは力を加えた直後から 250[msec] おきに 8 回取得された物体姿勢列を表す．ネットワークモデルの CNN には AlexNet[Krizhevsky 12] を用い，RNN では物体運動の長さに合わせて 8 回リカレント結合を順伝播する．物体運動の予測値と真値の自乗誤差平均を損失関数として (Fig3)，確率的勾配降下法を改良した Adam (Adaptive Moment Estimation) によってモデルの重みパラメータを更新した．深層学習フレームワーク Chainer[*3] でモデルを実装し学習により得られた学習曲線を Fig5 に示す．初期状態に比べると学習が進んではいるものの，訓練データ誤差に比べると検証データ誤差は減少しておらず，ネットワークモデルや学習手法の改良が今後の課題となる．

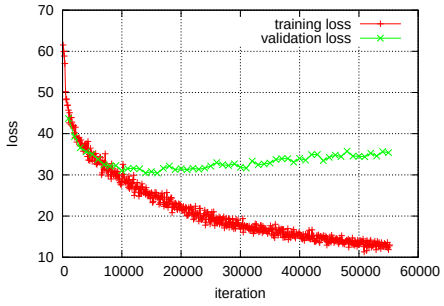

図 5: 物体運動予測モデルの学習曲線
1 万個のデータからなるデータセットを 8:2 で訓練用と検証用に分割して用いた．

3.2 学習済みモデルによる物体運動予測

学習済みモデルを用いて，カメラ位置や操作力が訓練データに含まれない設定での物体運動の予測を行った．物体は操作力の作用点や向き・大きさに応じて，滑りや転倒といった複雑な運動を行う．Fig6 に示すように，提案した深層学習モデルによってこれらの複雑な運動が概ね正しく予測されていることがわかる．複数の試行において物体の滑り・転倒の誤りや滑る向き・距離の誤差などが生じており，精度の評価や向上が今後の課題である．

3.3 ロボットによるマニピュレーションへの適用

物体予測を統合したロボットマニピュレーションシステムを構築し，等身大ヒューマノイドロボット HRP2-JSK[Okada 05] に適用して物体操作実験を行った様子を Fig7 に示す．学習時と同じ寸法の物体を対象として，ロボットが物体を滑らせる，倒すために物体を押す位置を物体運動予測結果に基づいて決定した．物体表面にサンプリングされた操作力作用点を入力として物体運動を予測し，目的の物体運動に近い運動を実現する作用点を探索した．これによって物体操作の結果から入力を求め

*3 http://chainer.org/

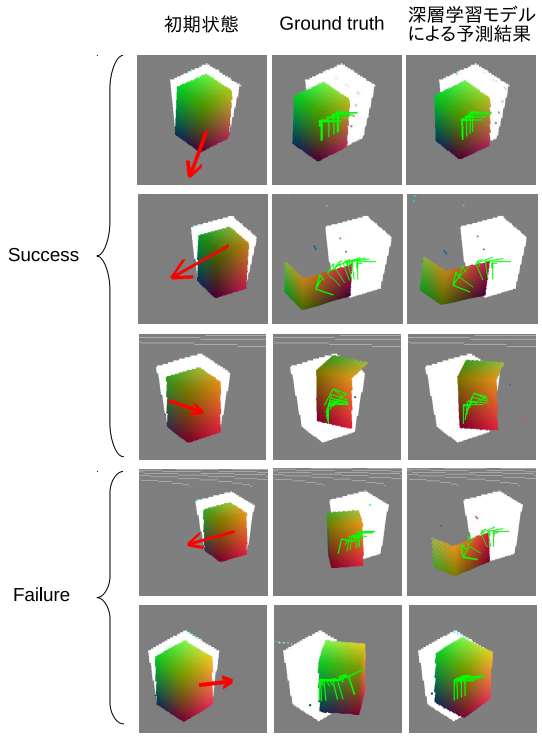

図 6: 深層学習モデルによる物体運動の予測結果
Depth 画像を三次元復元して得られたポイントクラウドに予測運動を反映した様子を示す．Mask 画像における Mask 内 (物体) の点はグラデーション色，Mask 外 (環境) の点は灰色で示されている．赤色矢印は操作力，緑色座標列は物体重心の時系列運動を表す．

ることが可能となっている．予測された物体運動では幾何的な干渉が生じているものの，滑り，転倒を判別するために十分な精度で物体運動が予測され，ロボットによる自律的なマニピュレーションが実現された．

4. 結論

本論文では視覚情報と操作情報から深層学習モデルにより三次元物体運動を予測する手法を提案した．視覚情報と操作情報は画像形式でネットワークモデルに入力され，CNN と RNN を経て時系列物体運動が出力される．提案手法をロボットマニピュレーションシステムに統合することで，物体運動予測によりロボットの自律的な物体操作が実現された．

参考文献

[Byravan 16] Byravan, A. and Fox, D.: SE3-Nets: Learning Rigid Body Motion using Deep Neural Networks, *CoRR*, Vol. abs/1606.02378, (2016)

[Fischer 16] Fischer, J., Mikhael, J. G., Tenenbaum, J. B., and Kanwisher, N.: Functional neuroanatomy of intuitive physical inference, *Proceedings of the National Academy of Sciences*, Vol. 113, No. 34, pp. E5072–E5081 (2016)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. eds., *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, Curran Associates, Inc. (2012)

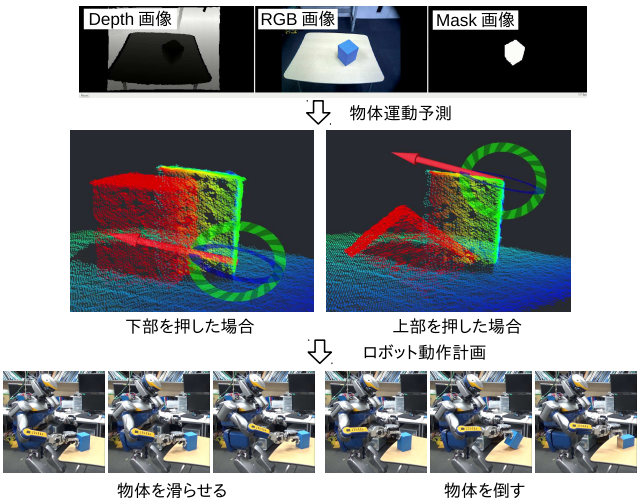

図 7: 物体運動予測によるロボットマニピュレーション実験
ロボット頭部の RGB-D カメラ画像から物体運動を予測した．中段の図はグラデーション色で示されたポイントクラウドは現在の状態，赤色で示されたポイントクラウドは予測運動後の物体の状態を表す．

[Levine 16] Levine, S., Finn, C., Darrell, T., and Abbeel, P.: End-to-end Training of Deep Visuomotor Policies, *J. Mach. Learn. Res.*, Vol. 17, No. 1, pp. 1334–1373 (2016)

[Mottaghi 16] Mottaghi, R., Rastegari, M., Gupta, A., and Farhadi, A.: *"What Happens If..." Learning to Predict the Effect of Forces in Images*, pp. 269–285, Springer International Publishing, Cham (2016)

[Murooka 17] Murooka, M., Ueda, R., Nozawa, S., Kakiuchi, Y., Okada, K., and Inaba, M.: Global planning of whole-body manipulation by humanoid robot based on transition graph of object motion and contact switching, *Advanced Robotics*, Vol. 31, No. 6 (in press) (2017)

[Okada 05] Okada, K., Ogura, T., Haneda, A., Fujimoto, J., Gravot, F., and Inaba, M.: Humanoid motion generation system on HRP2-JSK for daily life environment, in *Proceedings of The 2005 IEEE International Conference on Mechatronics and Automation*, Vol. 4, pp. 1772–1777 (2005)

[Wolpert 95] Wolpert, D. M., Ghahramani, Z., and Jordan, M. I.: An internal model for sensorimotor integration., *Science*, Vol. 269, No. 5232, pp. 1880–2. (1995)

[Wu 15] Wu, J., Yildirim, I., Lim, J. J., Freeman, B., and Tenenbaum, J.: Galileo: Perceiving Physical Object Properties by Integrating a Physics Engine with Deep Learning, in Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. eds., *Advances in Neural Information Processing Systems 28*, pp. 127–135, Curran Associates, Inc. (2015)

[Zheng 15] Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., and Zhu, S.-C.: Scene Understanding by Reasoning Stability and Safety, *International Journal of Computer Vision*, Vol. 112, No. 2, pp. 221–238 (2015)