家庭用品操作時の多視点画像に基づく VAE–GAN + LSTM によるホームロボットの動作生成

Motion Generation for Home Robot by Using VAE-GAN + LSTM Based on Multi-Viewpoint Images During Operation of Household Items

山内 翔太（都市大）　○学　松島 駿介（都市大）　徳永 夏帆（都市大）
正　佐藤 大祐（都市大）　正　金宮 好和（都市大）

Shota YAMAUCHI, Tokyo City University, yamauchi@rls.mse.tcu.ac.jp
Shunsuke MATSUSHIMA, Tokyo City University, matsushima@rls.mse.tcu.ac.jp
Natsuho TOKUNAGA, Tokyo City University, tokunaga@rls.mse.tcu.ac.jp
Daisuke SATO, Tokyo City University, sato@rls.mse.tcu.ac.jp
Yoshikazu KANAMIYA, Tokyo City University, nenchev@tcu.ac.jp

In this paper, we aim to apply motion generation method using machine learning based on human demonstration data to generation of task motion of home robot. A generation model by using VAE-GAN based on image data at task from multiple viewpoints is created, and a network model that can recognize the positional relationship between the robot hand and the object is also created. In addition, LSTM is learned by inputting latent variables generated by the image network model and joint angle data of the robot at each time, and network model capable of outputting the joint angle command value necessary for the task is created. Pushing task of the chair by the home robot is executed on the dynamics simulator by the learned network model and its usefulness is discussed.

Key Words: Home robot, Deep learning, Motion generation

1　緒言

近年，少子高齢化や女性の社会進出などからロボットによる生活支援や家事代行が望まれており，我々は，ユーザごとに求める家庭内作業の内容に応じてロボットのシステム構成を選択することが可能な MMM コンセプトに基づいたモジュラーホームロボットを開発している [1]．この MMM コンセプトの一つはユーザが求める複数作業（Multitask）の実現であるが，ここ数年，ロボットの動作生成の研究分野では，機械学習を用いた作業動作生成の研究が非常に活発になっている．

Google の Levine ら [2] や Robobarista の Jaeyong ら [3] は，産業用ロボットに適用されてきたあらかじめ決められた単一の作業を教示する従来の動作生成手法ではなく，ロボットに対して深層学習や強化学習を利用して動作を獲得させる研究を行っている．また，Rouhollah ら [4] は，ユーザがオペレータとなってロボットを操作して家庭内作業の動作を実演し，その際の作業の様子とロボットの関節角度情報を学習モデルの入力データとして，生成モデルである Variational Approaches for Auto-Encoding Generative Adversarial Networks（VAE-GAN）と Long Short-Term Memory（LSTM）を統合した深層学習を用いることで，ロボットに家庭内作業を獲得させる研究を行っている．この研究では，ロボットを感覚的に操作できることから，動作生成の際に専門的な知識が不要であることが大きな利点である．

このように，特別でも最適化もされていないが家庭内で人間が頻繁に実行する動作で，従来手法による動作生成のコストに見合わない作業動作，例えば，さまざまな日用品を把持することや，ものを運ぶ・片付ける動作などを実現するための研究は，ホームロボットにおいては非常に重要であり，我々もこのような新しい動作生成手法の研究課題に取り組む．

本研究では，7 自由度モバイルマニピュレータのホームロボットに対して作業時の動作画像を多視点から記録可能な形に拡張した文献 [4] の手法を用い，作業動作の学習を向上させること狙い，具体例として，いすを押す動作を実現することに取り組む．

2　動作生成のためのネットワーク構成

本研究での深層学習のネットワークモデルは，VAE-GAN と LSTM を統合したものである．VAE-GAN とは，Convolutional Neural Network（CNN）を使用した手法であり，作業実演時の状況を画像データとして記録し，ロボット手先と操作対象物の位置関係を認識するために使用する．また LSTM とは，連続した時系列データを学習データとして持つことが可能な Recurrent Neural Network（RNN）の一つであり，ロボットの関節角度データを生成するために利用する．

2.1　VAE-GAN

VAE-GAN を使用して画像を学習させる流れを Fig. 1 に示す．VAE-GAN は，深層学習の CNN を使用した生成モデルの VAE と GAN を組み合わせた生成モデルの手法である [5]．

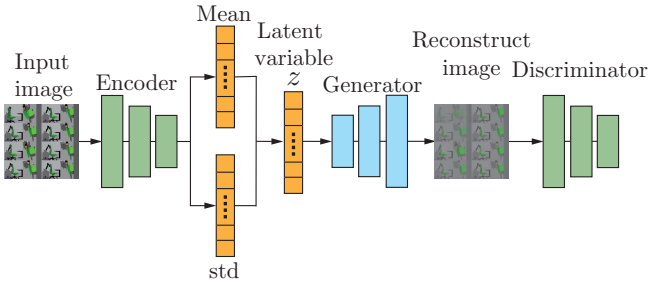


Fig.1 Network for learning using VAE-GAN.

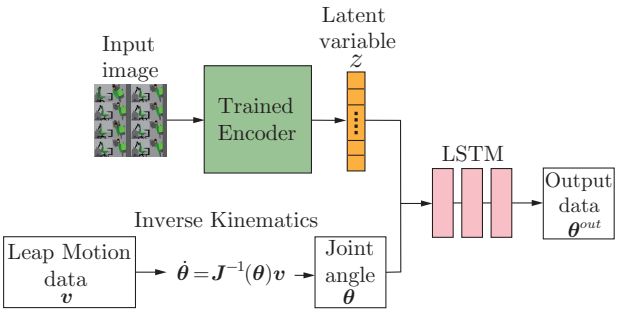


Fig.2 Network for learning using LSTM.

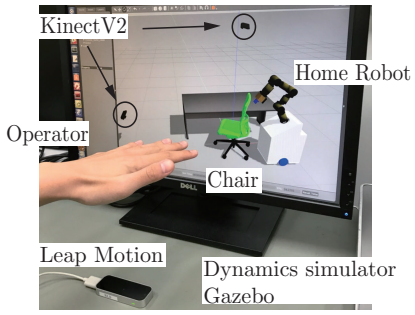


Fig.3 Demonstration of chair push task for creating training data using Leap Motion.

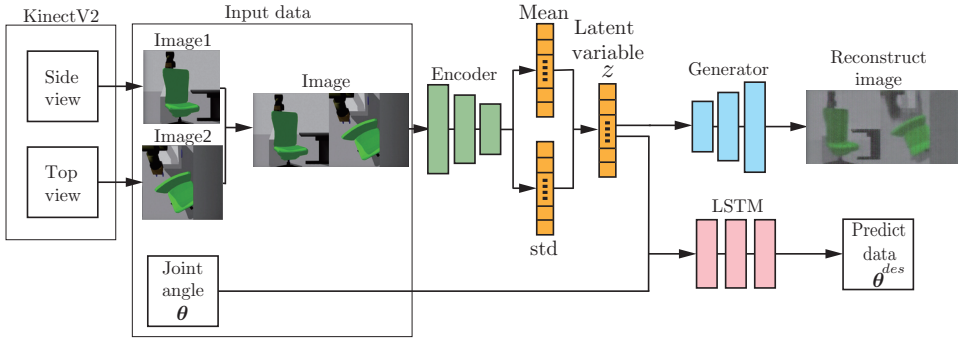


Fig.4 Network model composed of networks created for motion generation.

CNN は，畳み込みとプーリングを行い，画像データに対して高い認識精度を実現したニューラルネットワークであり，CNN を使用した生成モデルの VAE は，データ次元を圧縮する Auto Encoder に確率分布である潜在変数を作成したものである．Encoder によって作成した平均と標準偏差から潜在変数を作成する．GAN は，Generator と Discriminator の二つのネットワークがあり，Generator によって生成されたデータと入力データを Discriminator によって識別させ学習する生成モデルである．

VAE-GAN は，VAE の Encoder から作成した潜在変数から GAN の Generator によってデータが生成され，Discriminator によって識別される．この VAE と GAN を組み合せることで安定な学習ができ，鮮明な画像が生成される．

2.2 LSTM

RNN は再帰型の構造を持ち，時系列データの前後関係から次の時間ステップのデータを予測できるニューラルネットワークであり，過去の時間ステップの隠れ状態を多く保持するために LSTM を使用する [6]．LSTM は RNN の隠れ層のニューロンを置き換えたものであり，RNN の学習時に生じる誤差の発散や消失などの問題を解決する．LSTM の中に入力ゲート，出力ゲート，忘却ゲートの三つのゲートが存在し，ネットワーク内に保持している過去のデータを調整しながら学習することができる．

2.3 動作生成のための学習の流れ

動作生成のためには，作業動作の画像データとロボットの関節角度データを入力し，学習する必要がある．まず，画像データを入力し VAE-GAN を使用する．次に，VAE-GAN で学習された Encoder によってできた潜在変数と関節角度データを一つの時系列データに統合し，LSTM の入力として使用する．この流れを Fig. 2 に示す．Fig. 2 では，入力画像を VAE の Encoder によって次元を圧縮し，確率分布の潜在変数とすることで，関節角度と同じように時系列データとして扱えるようにし，それを学習させることで動作生成のためのネットワークモデルが作成される．

3 多視点画像

文献 [4] の手法のままでは，操作対象物がマニピュレータの裏に隠れてしまいカメラの死角になってしまうことや対象物を三次元的に操作する場合など，単一視点では対応が難しい動作が存在するため，カメラを増やし多視点にすることで問題に対応する．

多視点画像の学習方法は，複数の画像に対してその数に対応する分の VAE-GAN を用意する方法や，多視点画像のすべてを一つの VAE-GAN に入力する方法などが考えられるが，今回は多方向からカメラで撮影した画像を記録し，それらの画像を一枚に組み合わせる処理を行い，一枚の画像として学習させる方法を選択した．この方法では，多視点画像データを一枚の画像として処理するため，画像データの時間的な関係性を保持できる．

4 家庭用品操作のデモンストレーション

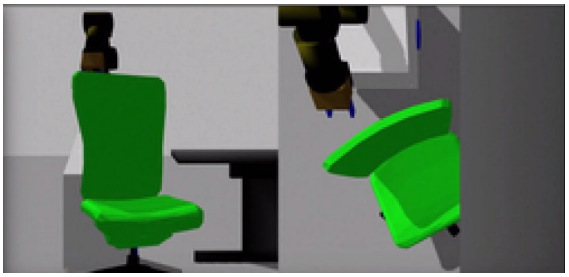
深層学習を使用して動作生成させるためには，生成させる動作を教示する必要がある．今回はいすを操作する動作実演を行った．

4.1 シミュレーション環境

デモンストレーションする環境には，Robot Operating System（ROS）を用いた動力学シミュレーションである Gazebo を使用し，7 自由度のマニピュレータを有するホームロボットを動作させた．デモンストレーションするロボットの操作方法としては，小型 USB カメラである Leap Motion 社の Leap Motion から得たユーザの手の位置情報からホームロボットのマニピュレータの手先に対応するように逆運動学から算術した関節角度を指令値として与えることで遠隔操作した．また，シミュレーション環境内には，操作対象物である椅子とホームロボット，机，および椅子とロボットのマニピュレータを撮影できるよう仮想 KinectV2 の二つを Fig. 3 に示すように設置した．

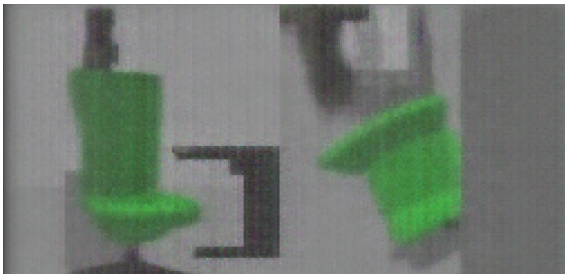
4.2 デモンストレーション動作

動作生成学習用の訓練データをデモンストレーションにより作成した．その様子を Fig. 3 に示す．行ったデモンストレーションの動作は，机の前にある椅子の背面をロボットマニピュレータの手先で押し，机に対し椅子を正面の向きに戻し，椅子を机に入



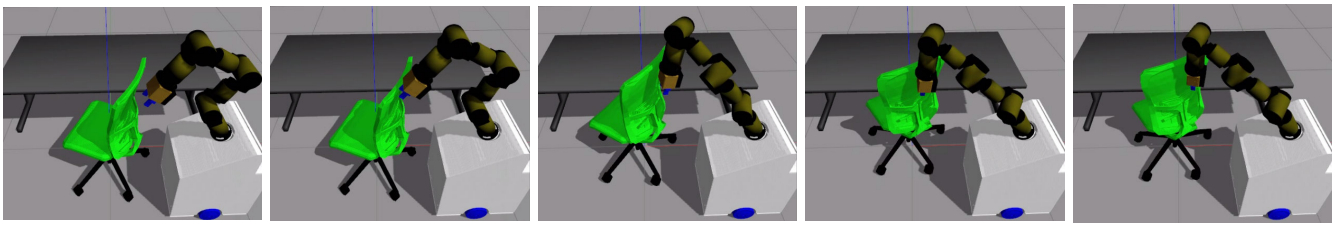
(a) (b)

Fig.5 Input image recorded by KinectV2, (a)Image taken at the side viewpoint, (b)Image taken at the top viewpoint.



(a) (b)

Fig.6 Image reconstructed by network, (a) Side image, (b) Top image.



(a) (b) (c) (d) (e)

Fig.7 Motion generated by outputting joint angle datas from the network. (a) (b) Press the back of the chair, (c) (d) align the chair with the table, and (e) put the chair on the desk.

れた．この動作のデモンストレーションを 5 回記録した．このとき，椅子押し作業のデモンストレーションを椅子に対し上方向と横方向に設置した二つの KinectV2 によって，それぞれ 128 × 128 pixel の RGB 画像と七つの関節角度およびグリッパ開閉のデータを時系列データとして 33 Hz で記録した．また，二つの KinectV2 から記録された二枚の画像を横に組み合わせ，128 × 256 pixel の RGB 画像として保存した．

5　ネットワークの学習

デモンストレーションで記録したデータを Fig. 1，Fig. 2 のネットワークに入力して動作生成するためのネットワークを作成した．2 枚の画像を組み合わせ保存された画像データを Fig. 1 に示す Input image の入力し，Encoder と Generator および Discriminator のネットワークを作成した．その後，Fig. 2 に示すネットワークに画像データと関節角度データを入力して LSTM のネットワークを作成した．VAE-GAN を用いた学習を 100 epoch 行い，LSTM を用いた学習を 10000 epoch 行った．

6　学習されたネットワークによる動作生成

Fig. 1，Fig. 2 の学習で作成されたネットワークによって構成された動作生成のためのシステム全体のネットワークを Fig. 4 に示す．Fig. 4 の Input data に最初の関節角度を入力することで，学習されたネットワークから次の関節角度が予測することができる．また，二つの KinectV2 から取得した画像を学習させるときと同様に組み合わせ一つの画像とし入力して，Generator によって生成したデータを出力させた．生成された関節角度を Gazebo 内のシミュレーションモデルに送信することで，動作生成した．

動作生成中に KinectV2 によって撮影した画像を Fig. 5 に，その画像を入力し Generator によって生成された画像を Fig. 6 に示す．Fig. 6 より，入力画像の特徴を捉えて画像を生成できていることがわかる．また，この動作生成手法によって生成された動作の様子を Fig. 7 に示す．図から机から出ている椅子に対して，椅子の背面を押して机に入れる動作が生成できている．

7　結言

動力学シミュレータ上で視点を多視点にした動作生成手法を用いてホームロボットによる椅子の押し作業を実現した．この手法を用いることで単一の視点のときロボットの死角になっていた位置でも多視点にすることで認識できるようになり，操作対象物を三次元的に操作することも期待できる．

今後の課題として，単一の視点画像のときと多視点画像との動作，および違う方向からの視点画像を組み合わせたときの成功率を比較し，有効性を検証する必要がある．また，実際のロボットに適応するために，拘束条件やトルクの条件を考慮する必要があると考えられる．

参考文献

[1] T. Tsuchiya, Y. Shiraki, S. Sekido, A. Yamamoto, D. Sato, and D. N. Nenchev, “Modular home robot system based on the MMM concept-design instance with detachable symmetric arm module,” in *IEEE/ASME Int. Conf. Adv. Intell. Mechatron.*, 2013, pp. 280–285.

[2] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen, “Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection,” in *Proc. Int. Symp. Exp. Robot.*, 2016.

[3] S. Jaeyong, L. Ian and, S. Ashutosh, “Deep Multimodal Embedding: Manipulating Novel Objects with Point-clouds, Language and Trajectories,” in *IEEE Int. Conf. on Robot. and Automat. (ICRA)*, 2017.

[4] R. Rouhollah, A. Pooya, and B. Ladislau, “Vision-Based Multi-Task Manipulation for Inexpensive Robots Using End-To-End Learning from Demonstration,” *CoRR*, vol. abs/1707.02920, 2017. [Online]. Available: http://arxiv.org/abs/1707.02920

[5] A. B. L. Larsen, S. K. Snderby, H. Larochelle, and O. Winther, “Autoencoding beyond pixels using a learned similarity metric,” in *Int. ernational Conf. on Machine Learning,* vol.48, 2016, pp. 1558-1566.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol.9, no.8, pp. 1735–1780, 1997.