J-004

コミュニケーションロボットのための
畳み込みニューラルネットワークを用いた物体認識手法の検討

Object recognition using a convolutional neural network for communication robots

川﨑 邦将† 長田 茂美†
Kunimasa Kawasaki Shigemi Nagata

1. はじめに

近年、人に対してサービスの提供や人とコミュニケーションを行うロボットの研究開発が活発化している。このようなロボットを実現するためには、人とロボットとの円滑なインタラクションを支援する技術が不可欠である。人とロボットとのインタラクションの中では、コミュニケーション相手となる人に対して、ロボットが人からのインタラクションに応じて自らの行動を適切に選択できることが要求される。ロボットが人から受けた教示や行動を認識、記憶し、人に返すインタラクションの内容を適応的に変化させることで、人とロボットとのより深いコミュニケーションが実現されると考えられる。

一般に、コミュニケーションロボットは人の生活環境下で機能させるべきものであり、事前にその環境を明確に定義することはできない。そのため、コミュニケーションロボットはさまざまな環境下で人から受けた教示や行動を正確に認識できなければ、人とのコミュケーションは成立し得ないことから、環境変化に強いコミュニケーションロボットのための認識手法が必要となる。

本研究では、人とコミュニケーションロボットとのインタラクションの中で、人からロボットに物体の教示が行われる場面を想定し、環境変化に頑健な物体認識を実現するための学習データの前処理の方法と畳み込みニューラルネットワークを用いた物体認識手法について検討する。

2. 畳み込みニューラルネットワークとは

2.1 概要

畳み込みニューラルネットワーク（Convolutional Neural Network、以下 CNN）は、主に画像認識に応用されている順伝播型のニューラルネットワークの一つである。生物の脳の視覚野の神経細胞の受容野の局所性と単純型細胞、複雑型細胞の生物学的知見に基づいている[1]。入力パターンに対して位置変化に敏感な単純型細胞と、位置変化に鈍感な複雑型細胞の処理を組み合わせることで選択性を持ったニューラルネットワークといえる。

2.2 物体認識での CNN の利点と欠点

本研究で、教示された物体の認識を行うために CNN を適用する上での利点と欠点を挙げる。

「利点」

CNN は入力画像から画像特徴量を得ることができるため、学習時の環境に合わせた画像特徴量を得ることが可能である。これにより、物体の教示が行われた場所に合った物体認識を行える。

また、CNN は画像上の物体位置の変化に強く、認識の際に学習時と異なる画像上の物体位置に対してロバストな物体認識が可能である。

「欠点」

CNN は学習に時間を要するため、リアルタイム学習を行う場合は、入力画像サイズを非常に小さくする必要がある。

加えて、ハイパーパラメータが多くあり認識精度を向上させるためのパラメータ調整が難しい。

3. 関連研究

CNN を用いた物体認識の研究として、大規模教師付き画像データセット ImageNet[2]の一部を用いたコンペティション型ワークショップの ImageNet Large-scale Visual Recognition Challenge（ILSVRC）で トロント大学の Hinton らが 8 層の CNN を用いて 1,000 クラス識別のエラー率を SIFT 特徴を用いた手法よりも 10%近く下げることに成功している[3]。

ロボットに適用した事例として、Sergey Levine らのアームロボットの把持動作を画像から学習するネットワークで、画像から把持する物体の特徴量を得るために CNN が用いられており、複数の物体がある中でも目標の物体の画像特徴を獲得し、最終的な把持動作を行える実験結果が示されている[4]。

4. 提案手法

4.1 提案手法の概要

本研究では、人の生活環境下でコミュニケーションロボットが人からの物体教示が行われる場面を想定して、CNN による物体認識手法を検討する。人からロボットに教示する物体を常に同じ（画像）位置で提示することは難しいため、画像上の物体位置の変化に頑健な CNN が有効であると考える。

また、環境変化に頑健な物体認識を実現するためには、教示物体とその他の物体や机や壁などの背景とを分離し、教示物体のみの特徴をできるだけ正確に抽出する必要がある。そのため、ロボットのカメラから入力される連続動画像から背景推定を行い、教示時の画像と推定した背景との差分をとり、教示物体のみを抽出する手法を検討する。

CNN を多クラス分類問題に適用する場合、学習していない未知のクラスのデータも、学習した既知のいずれかの

†金沢工業大学

クラスの分類されてしまうという問題がある。これでは、学習していない物体も学習した物体と誤認識する可能性が高まる。この問題に対しては、焦点の概念を導入し、焦点が合っている注視領域と焦点が合っていない非注視領域を考え、焦点が合っていない非注視領域はぼかし効果がかかっていると仮定して、非教示物体の未知の学習クラスとして利用することで、教示物体と非教示物体の分類を行う手法を提案する。また、注視領域を画像全体ではなく、局所領域とすることで、認識する画像領域を小さくし、学習時間の短縮を図っている。

4.2 提案手法の構成

図 1 に、提案手法の構成を示す。また、物体教示開始から教示した物体の認識までを以下の step1~5 で行う。

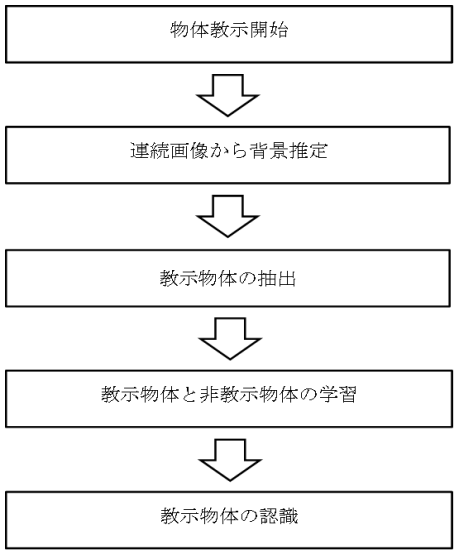

図 1　提案手法の構成

step1:　物体教示開始
　人がロボットに教示する物体とその物体のラベルを提示し、ロボットに物体教示の開始を指示する。

step2:　連続画像から背景推定
　教示物体と背景を分離するために、物体教示開始後と開始前の背景差分をとるだけでは、人の手の影などが動くことに起因する画像上の変化など、教示物体以外の領域を誤抽出してしまう可能性がある。この問題を軽減するために、物体教示開始までの間に得られている連続動画像から次時刻の画像を予測し、その予測画像を背景画像とする。具体的には、CNN と時系列データを扱うための Long short-term memory（LSTM）を組み合わせて学習器を構成し、現時刻の画像を入力データ、次時刻の画像を教師データとして学習させ、現時刻の画像から次時刻の画像を生成することによって、背景画像を推定している。このようにして、物体教示開始後から人の影などの動きを予測した背景画像を生成し、教示物体の誤抽出を軽減できる。

step3:　教示物体の抽出
　予測した背景画像と物体教示が行われている現時刻の画像との差分をとり、教示物体を抽出する。また、非常に小さい領域が抽出された場合には、ノイズとして除去する。

step4:　教示物体と非教示物体の学習
　抽出した教示物体以外の領域にぼかし効果を適用後、教示物体の領域内とそれ以外の領域を、予め設定した注視領域サイズで切り出して学習データを生成する。この学習データを用いて、教示物体領域とそれ以外の領域に分類できる CNN を構築する。

step5:　教示物体の認識
　学習した CNN を使用して、注視領域サイズで画像全体を探索し、教示物体と非教示物体の分類を行う。

4.3 提案手法の評価実験

提案手法を簡易的に実装して実験を行った。実験では、人の影などが画像に影響しないようにカメラを設置し、Step2 の背景推定を行わずに、物体教示開始直前の画像との背景差分で画像を抽出するように簡易化した。実験の結果、ぼかし画像の適用前よりも教示物体と非教示物体を有意に分類できることが確認できた。図 2 に、その分類結果を示す。

この結果に加えて、背景推定を行い人の影や周辺の環境変化により頑健な物体抽出を行い、注視領域を小さく切り出しても同様の効果が得られるか確認されれば、本手法が、物体教示時の環境変化に対して頑健な認識手法といえる。

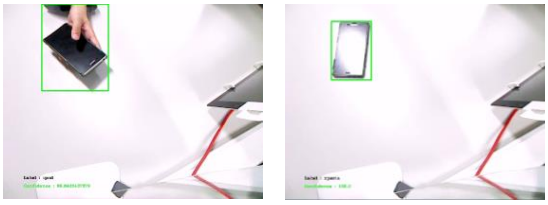

図 2　評価実験の分類結果

5. 今後の課題

本稿では、コミュニケーションロボットが人からの物体教示を受けた場合を想定し、環境変化に対してロバストな物体認識手法について検討した結果を述べた。

今回は、環境をある程度制限しての実験を行ったが、より自然なコミュニケーションが行えるように教示物体以外の周辺環境が画像に影響を及ぼす環境でも認識が可能かどうかの検討を行う予定である。

また、教示された物体が単体の場合のみを検討したが、複数になった場合にどのように学習を行うのか、リアルタイム性を向上させることも視野に入れながら、認識時の探索手法も検討する必要がある。

参考文献

[1] 福島 邦彦, “位置連れに影響されないパターン認識機構の神経回路のモデル-ネオコグニトロン-”, 電子通信学会論文誌 A, Vol.J62-A, No.10 (1979).
[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, “ImageNet: A large-scale hierarchical image database”,CVPR, (2009).
[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet classification with deep convolutional neural networks”, NIPS, (2012).
[4] Sergey Levine, Peter Pastor, Alex Krizhevsky, Deirdre Quillen, “Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection”.arXiv,(2016)