4L1-2

実環境におけるロボットの物体認識
Object Recognition for Robot in Actual Environment

原田 侑弥[*1] Yuya Harada　中村 郁仁[*1] Fumihito Nakamura　松村冬子[*1] Fuyuko Matsumura　原田 実[*1] Minoru Harada

[*1] 青山学院大学　理工学部　情報テクノロジー学科
Department of Integrated Information Technology, Faculty of Science and Engineering, Aoyama Gakuin University

In this paper, an object recognition method using Convolutional Neural Networks, and a three-dimensional object position estimation method were applied to a humanoid robot. The proposed system has two types of object recognition mode: a mode to detect something and a mode to search the specified object. As an experimental result, the latter mode of object recognition was almost succeeded whereas the average of success rate of the former mode was 75%. The average of success rates of position estimation in both modes was 69% and it tends to fail when the object is thin.

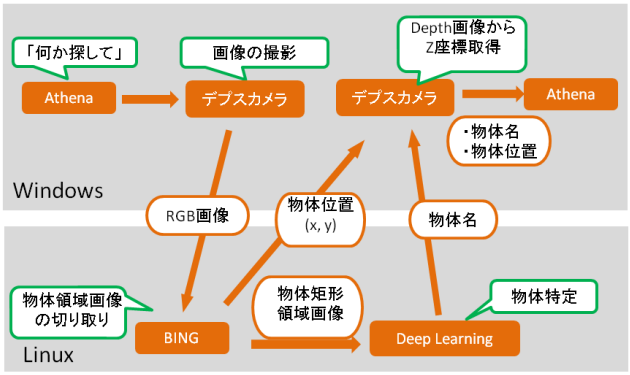
1. はじめに

近年，対話型ヒューマノイドロボットが台頭しており，人間とロボットのより親和的な対話を行うための研究が盛んに行われている．たとえば原田研究室では Aldebaran 社の Nao を用いた対話的動作指示システム Athena[1]の開発を行っているほか，Softbank 社では Pepper に感情 Map を利用した対話などを搭載している．

本稿では，Pepper を用いたロボットの対話システムのための物体認識，座標推定方法の検討をする．

2. 物体認識システム Horus

図 1. システム構成図



対話中では，ロボット自身が物を見つける場合と，発話者がロボットに目の前のものを見つけてほしい場合の 2 パターンあるため，検知モードと探索モードの二つのモードを用意した．検知モードは，ロボットが目の前の物を見つけるためのモードであり，キャプチャ画像から物体画像を 1 枚切り出し，1 つの物体を認識する．探索モードは，発話者がロボットに物を見つけさせるためのモードであり，物体画像を複数枚切り出し，指定された物体名と一致する認識結果の物体画像に対する矩形の情報を返す．Horus のシステム構成を図1に示す．

連絡先：原田侑弥，青山学院大学理工学部，神奈川県相模原市，e-mail: a5812089@alumni.it.aoyama.ac.jp

2.1 物体認識

本研究では，物体のクラス名認識に Deep Learning のフレームワークの一つである Caffe を用いた．Caffe は Deep Learning の一つの手法である CNN(Convolutional Neural Network)を使っている．CNN は，畳込み層と呼ばれる局所的な特徴を抽出する層とプーリング層と呼ばれるずれを吸収する層を交互に重ねることによって構成されており，画像認識に利用する際は，画像を入力すると分類結果であるクラス名が出力される．Caffe は，すでに学習済みのリファレンスモデルが配布されているため，これを用いる．リファレンスモデルを利用することにより，1000 種類の犬や虫，ボール，机などさまざまな物体を認識することができる．

2.2 物体矩形領域検出

ロボットが認識した物体を活用するためには画像中にある物体の座標取得が必要である．このため，画像中に写る物体の矩形領域を切り出し，矩形の中心を物体の x, y 座標であると推定することとした．この矩形領域検出は BING アルゴリズム[2],[3]を用いて検出を行った．この BING アルゴリズムは物体の輪郭を正確に切り出すことはできないが実行速度が速く，ロボットのようにリアルタイム性が要求される環境に適している．さらに，一枚の認識画像に複数の物体が存在する場合に，それぞれの物体に対して矩形領域を切り取ることにより，複数の物体を認識することができる．この物体領域検出は物体認識の前処理として行うことにより，物体以外の部分を除去できるため物体認識の精度の向上が期待できる．

2.3 座標取得

本研究では，Intel 社が提供する RealSenseCamera(F200)というデプスカメラを使用する．このカメラは，カラー画像の他，赤外線カメラの搭載により距離データの取得が可能であり，距離データの有効値として 20～1200mm とヒューマノイドロボットの動作可能範囲内である．また，このカメラで取得した x, y 座標から RealSenseSDK の座標変換機能を用いることにより奥行きとなる z 座標を取得する．z 座標について，領域検出で出力した物体画像全体について座標変換を行うと処理にとても時間がかかるため物体画像の中心点1点のみに対して座標変換を行った．

3. 評価実験

検知モードと探索モード，それぞれのモードに関して物体認識と座標取得の成功率を測る．公開されているリファレンスモデルを使用することにより 1000 種類の一般物体認識を行うことができるが，本実験では，物体認識に使用する物体はオレンジ，マウス，木製スプーンの 3 種類とした．物体は机の上に設置し，真上から撮影した．設置位置は撮影範囲を 10 分割し，設置する回数が均等になるように 1 回ごとに規則的に換えて設置した。それぞれの物体に対して試行を 20 回ずつ行った．図 2 はマウスでの実験の画像である．赤枠が BING で認識した物体矩形領域であり，その中心点を赤の点で示した．また，認識したクラス名と推定した座標はタイトルバーに示した．

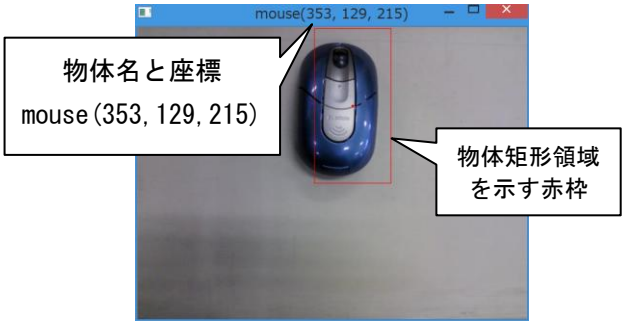


図 2. マウスでの実験

3.1 実験1-検知モード-

ロボットが目の前の物を見つけるためのモードである検知モードでは，認識結果の物体名が設置された物体を示せていたことを示す物体検出と，物体検出を行った後に 3 次元座標の全てが取得できていることを示す座標取得の二つの精度を測定することとした．x, y 座標に関しては，画像に検出した物体の矩形領域の中心点をプロットし，その中心点が画像上の物体の上にあるとき成功，また z 座標に関しては，実際にメジャーで距離を測定し，2cm 以上の差がないとき成功とした．物体検出と座標取得の成功率を表 1 に，また，誤認識したときの物体名の上位 3 つを表 2 に示す．

表 1. 検知モードの実験結果

設置した物体	物体検出	座標取得
オレンジ	75%	100%
マウス	65%	84%
木製スプーン	85%	35%
合計	75%	73%

表 2. 実験 1 の誤認識結果の一部

設置した物体	誤認識結果
オレンジ	ピンポン球, マダニ, ザクロ
マウス	携帯電話, サンドバッグ, ライター
木製スプーン	絆創膏, ヘラ

実験 2-探索モード-

発話者がロボットに物を見つけさせるためのモードである探索モードでは，与えられた物体名と一致する認識結果の物体が検出されたことを示す物体検出と，物体検出を行った後に 3 次元座標の全てが取得できていることを示す座標取得の二つの精度を測定することとした．先ほどと同様に，x, y 座標に関しては，画像に検出した物体の矩形領域の中心点をプロットし，その中心点が画像上の物体の上にあるとき成功，また z 座標に関しては，実際にメジャーで距離を測定し，2cm 以上の差がないとき成功とした．物体検出と座標取得の成功率を表 3 に示す．

表 3. 探索モードの実験結果

設置した物体	物体検出	座標取得
オレンジ	100%	75%
マウス	95%	78%
木製スプーン	95%	47%
合計	96%	66%

4. 考察

検知モードでは，マウスの物体検出の精度がややほかと比べ低かった．表 2 にあるように誤認識物体には，携帯電話やサンドバックといった似た形状のものが出力されており，人工物であるマウスはその他の人工物とも似てしまうことが一つの原因だと考えられる． 探索モードでは，物体検出，座標取得ともにオレンジ，マウスの結果が高く，実用できる範囲であると考える． 検知モード，探索モードに共通して，木製スプーンの座標取得の精度がほかと比べ低かった．木製スプーンの座標取得で失敗しているものに関しては，検出した物体の矩形領域の中心が物体の上にないものの近い位置を指していることが多かった．このことから，木製スプーンは他の 2 種よりも細いため，矩形領域の中心が物体をとらえ辛かったことが一つの原因だと考える．

また，マウスの側面に中心点があったとき，z 座標の値にエラー値が返ってきていた．これは，デプスカメラで使用されている赤外線の性質により，反射あるいは吸収されたためその点の値が取得できなかったと考えられる．これらの問題を解決するため，今後は中心点の 1 点だけでなく，中心付近のいくらかの範囲で物体との距離を測るといった対策が必要であることがわかった．

5. おわりに

Pepper を用いたロボットの対話システムのための物体認識，座標推定方法の検討を行った．物体認識に関しては，Deep Learning を用いることにより，多種の物体を精度よく認識することが可能となった．座標推定に関しては，BING アルゴリズムとデプスカメラを組み合わせることにより，物体の 3 次元座標の推定が可能となったが，精度においては更なる向上が求められる．

参考文献

[1] 田村 優樹, 長崎 達也, 中野 雅広, 原田 実: 意味解析に基づくロボット指示システム Athena2011, 情報処理学会研究報告, Vol.2012-NL-206, No.10, pp.1-8, 2012

[2] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, Philip Torr: BING: Binarized Normed Gradients for Objectness Estimation at 300fps, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3286-3293, 2014

[3] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari: Measuring the objectness of image windows, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 34, No. 11, pp. 2189-2202, 2012