マルチモーダル学習を用いた力覚センサによる把持物体の識別

○山崎雅幸 山下隆義 山内悠嗣 藤吉弘亘 (中部大学)
堂前幸康 (産業技術総合研究所) 白土浩司 (三菱電機)

1. はじめに

ロボットには多数のセンサが搭載されておりそれぞれタスクに合わせた役割を持っている．しかし，人間は視覚，聴覚，触覚などの複数の感覚を用いて外部環境の知覚を行っている．ロボットと人間の知覚を同等にするためにマルチモーダル学習が必要とされている [1]．マルチモーダル学習において動画を音声認識に併用することで音声認識の認識精度が向上している [2]．また，ロボットへの応用として画像，音声，モータ角を用いてモーターの経路予測が行われている [3]．

従来の物体識別は，Deep Convolutional Neural Network[4] 等の機械学習を用い，ビジョンセンサにより取得した画像より，識別することが一般的である．しかし，画像からでは箱の内包物の種類や量などの取得できない情報がある．これらのシーンに対して通常，対象を持ち上げた時の重さや，振ったときの振動を感じとることで中身を判別することが一般的である．これらの情報は画像情報から取得できないため，別にセンサを用いることが必要である．

ロボットに触覚の機能を付随するセンサとして力覚センサがある [5]．力覚センサは主にロボットのハンドに取り付けられ，ロボットが接触を用いて精密な動作をするために活用されている [6]，[7]．力覚センサをハンドにつけることで接触以外にも，物体を把持した時にハンドにかかる力を測ることができる．

本研究は把持に使用する画像情報と，ロボットハンドに装着された力覚センサを用いた把持物体識別を対象する．力覚センサにおいて，物体を把持した時の物体ごとに異なるモーメントの変化に注目する．また，力覚センサはハンドにかかる力のみを測定するため，ロボットの姿勢を加味することができない．そこでロボット関節の電流フィードバックを用いることでロボットの関節にかかった負荷を見ることができる．力覚データを測定するシーンはアームを動作させ，停止した後のデータを用いる．停止後を用いることで慣性による影響や，非剛体の形状変化が生じるため，重量を測るよりも有効な特徴を得ることができる．

力覚センサから得られるデータは連続的なデータであり，系列を加味した学習が必要となる．Recurrent Neural Network(RNN)[9] の一種である Long Short Term Memory(LSTM)[8] は動画，音声などの系列データを扱うときに用いられ，系列性を加味した学習を行うことができる．RNN は 1 時刻前の中間層の出力をサンプルとともに入力する．学習時のパラメータ更新は，Back Propagation Through Time(BPTT)[10] と呼ばれる方法を用い中間層から時系列をさかのぼりながら行うため連鎖的に過去の情報を得ることができる．LSTM は通常の RNN で起こる勾配消失問題を解決したもので，長時系列を扱うことができる．

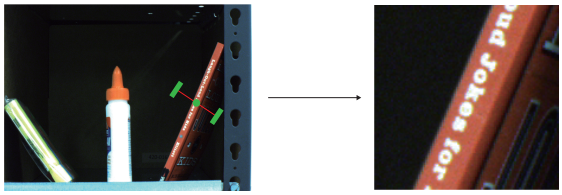

図 1　把持点画像の切り出し

本研究では画像情報，力覚センサと，電流フィードバック (FB) より LSTM を用いて把持物体の識別方法を検討する．

2. 画像情報

本章では識別に使用する画像情報について述べる．画像情報を用いた識別に CNN ベースのアイテム認識を導入する．CNN はあらゆる視点から撮影されたアイテム画像を認識する必要があるが，計算コストが高くなってしまう．そこで，検出された把持位置の周辺画像を用いることで効率的にアイテム識別を可能にする．ビジョンセンサより得られた計測データから Fast Graspability Evaluation[11] を用いて把持点を検出する．図 1 に示すように，検出された把持位置の周辺画像を切り出して学習に使用するデータとする．

3. 力覚センサと電流フィードバック

本章では識別に使用される力覚センサ，電流 FB について述べる．力覚センサはロボットに触覚を付与させる目的で，様々なロボットに装備されている．特に産業ロボットの微細な部品はめ込み等に利用されている．力覚センサを用いることで接触を利用した動作を可能とさせている．力覚センサは図 2 に示すようにセンサ上部と下部との力の変位を計測し，静電気容量の変化から変位を計測し力とモーメントを求める．取得できるデータは力ベクトル (Fx, Fy, Fz) と回転ベクトル (Mx, My, Mz) を合わせた 6 軸のデータである．

電流 FB は，アームが動作する際に関節のモータの電流を測定した値である．取得されるデータは 6 軸関節に流れる電流データ $(J_1, J_3, J_3, J_4, J_5, J_6)$ である．

3.1 把持物体ごとの力覚センサと電流 FB の変化

図 3 は 2 つのアイテム (book，mailer) を把持し，ロボットアームを左右に動作させ停止した後の力覚値 Fy の遷移で示す．縦軸はアームの動作停止後に y 軸方向へかかったモーメントであり単位は N，横軸は取得開始時からのスレッド (1 スレッド約 $7m/s$) である．550 や 560 スレッドにおいて力覚センサデータの波形のピー

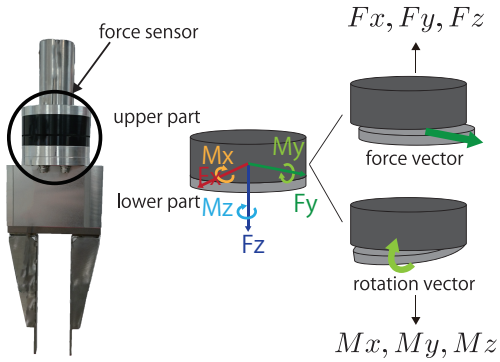

図 2　力覚データの測定

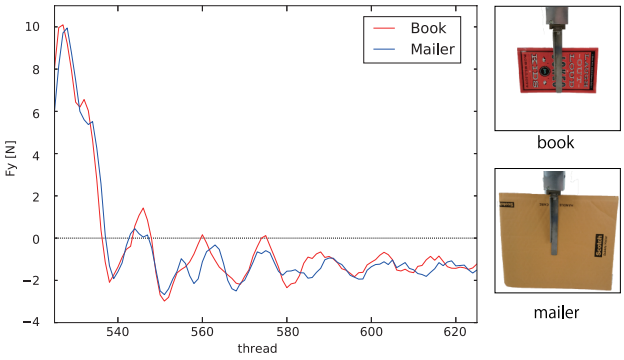

図 3　アイテムごとの力覚データの比較

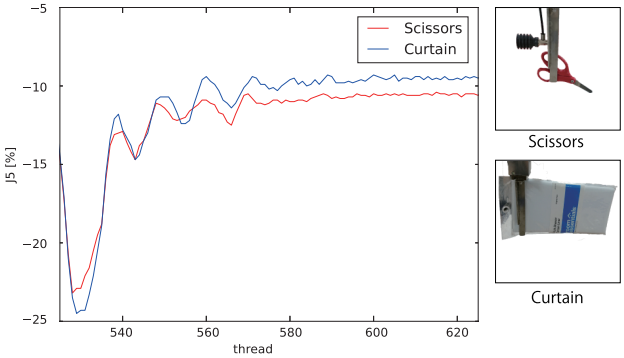

図 4　アイテムごとの電流 FB の比較

クの位置が異なっている．mailer が book に比べ細かい振動をしている．これにより，物体形状が似ていてもセンサ値の変動が異なることが分かる．

電流 FB において変化量の多い電流 J_5 を例に比較を行う．図 4 は 2 つのアイテム (scissors，curtain) を把持し，ロボットアームを左右に動作させ停止した後の電流 FB の遷移である．縦軸が停止後の J5 軸に流れた電流値であり単位は%，横軸が取得開始時からのスレッド (1 スレッド約 $7m/s$) である．停止動作を実現するために，電流 FB の収束が把持物により変化する．curtain の方が Scissors に比べて重量があるため制御に使用する力が大きいためである．これにより，把持する物体により関節にかかる電流が変化することが分かる．

3.2　把持位置による力覚データの変化

図 5 は 1 つのアイテム (book) を 2 種類の把持位置で把持し，ロボットアームを左右に動作させ停止した後の力覚値 Fy の遷移である．同一アイテムであるが

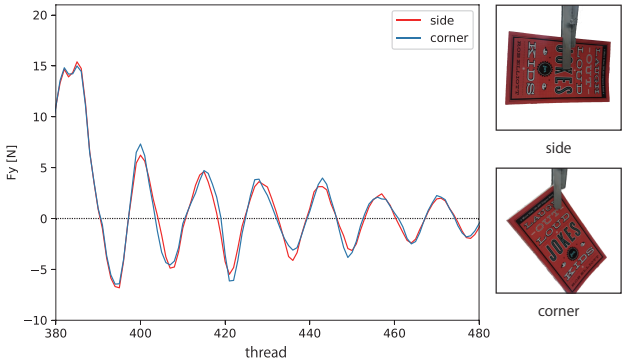

図 5　把持点による力覚データの比較

把持位置に波形のピークの位置が異なっている．把持点が違うことによりハンドにかかるモーメントの方向が少し変化する．これにより同一アイテムでも把持位置によりセンサ値の値が変化するため，様々な把持パターンを学習する必要がある．

4.　提案手法

本章では，提案手法の流れを述べる．本研究では，産業用ロボットに装着されたビジョンセンサを用いて物体を把持し，把持点の画像と把持力覚センサ，電流フィードバック (FB) より得られたデータを用いて LSTM により学習する．力覚データを取得するシーンはアイテムを把持した後のハンドを動作させた時のデータを用いる．把持点画像は CNN を用いて特徴抽出を行う．取得された系列データを正規化し，把持点画像の特徴と結合した後 LSTM に入力する．フレームごとに出力される LSTM の出力を一つの結果に統合し，識別を行う．

4.1　取得シーン

本研究では，あらかじめ棚に複数のアイテムを入れて置き，画像情報を取得し把持点を検出する．画像情報から把持点周囲の画像を切り出し入力データに用いる．把持したアイテムは図 6 のようにロボット座標系で x，y，z 方向にそれぞれ 1 往復させ，力覚データを取得する．x，y 軸は地面に水平方向に動作させ，z 軸は地面に垂直に動作している．学習時には 3 方向を混在させたものを使用する．取得した力覚データには移動時のデータ停止，制動までのデータが含まれており，1 回往復した長い系列データを扱うことになる．この中でも本実験では動作停止後 100 フレームを学習データとして使用する．図 7 に 1 往復データから抽出する学習箇所を示す．縦軸が Y 軸にかかるモーメントで横軸が取得時からのスレッドになる．青色の波形が取得した全体の波形で赤色の部分が識別に使用する部分になる．

4.2　力覚データの正規化

力覚データはベクトルごとの関係を持たせるためベクトルごとに正規化を行う．正規化は力ベクトル，回転ベクトル，電流 FB ごとの最大値をもとに求める．力ベクトルの最大値を $F_{\max}$，回転ベクトルの最大値を $M_{\max}$，電流 FB の最大値を $J_{\max}$ としたとき，i フレーム目の入力データ $\boldsymbol{x}_i$ を求める式 (1) を示す．

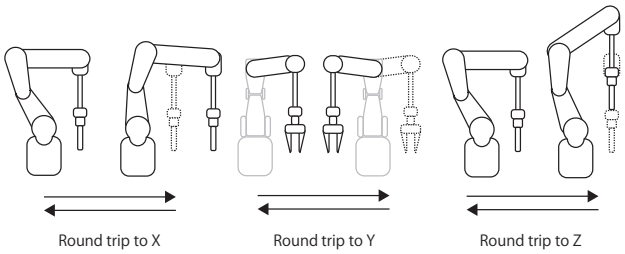

図 6　取得するロボットの動作

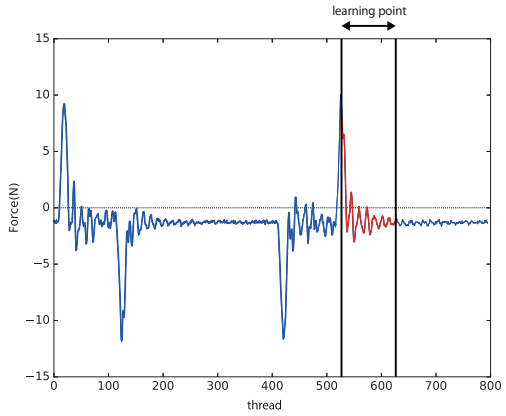

図 7　1 往復より抽出する Fy 軸データ範囲

$$\boldsymbol{x}_i = \Big[\frac{Fx_i}{F_{\max}}, \frac{Fy_i}{F_{\max}}, \frac{Fz_i}{F_{\max}}, \frac{Mx_i}{M_{\max}}, \frac{My_i}{M_{\max}}, \frac{Mz_i}{M_{\max}},$$

$$\frac{J_{1i}}{J_{\max}}, \frac{J_{2i}}{J_{\max}}, \frac{J_{3i}}{J_{\max}}, \frac{J_{4i}}{J_{\max}}, \frac{J_{5i}}{J_{\max}}, \frac{J_{6i}}{J_{\max}}\Big] \qquad (1)$$

4.3　ネットワーク構成

本研究で使用するネットワーク構造を図 9 に示す．棚に並べられたアイテムから把持点を検出する．検出した把持点をもとにアイテムをピッキングし力覚データを取得する．センシングした画像の把持点周囲を切り出し 156 × 156 ピクセルのパッチ画像を生成する．Alexnet[12] ベースの CNN に画像を入力し特徴ベクトルを取得する．取得された特徴ベクトルと力覚データを結合し LSTM にで識別を行う．ロボットを動作させた時の力覚データを蓄えておき，フレームごとにネットワークに入力する．学習係数は 0.01，入力層には力覚データ，電流 FB のを 12 ユニット与え，出力層は 21 クラスに対応した 21 ユニット，中間層は 250 ユニットの LSTM を 2 層を使用する．

4.4　識別結果の統合

RNN は i フレーム毎にクラス c の確率 $P_i(c|\boldsymbol{x}_i)$ を出力するため，1 つの系列データの結果を統合する．RNN は通常出力値を入力とする場合や系列データを入力し，最終フレームの出力を用いる場合が多い．しかし，本研究で使用するデータは後半になるにつれて収束していくため特徴が失われていくことが考えられる．そこで今回は，入力に使用したフレーム全てを加味した識別結果を出力するように統合を行う．式 (2) により 100

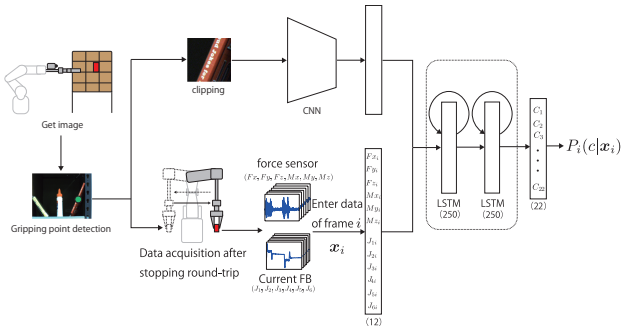

図 8　ネットワーク構成

フレームの確率を統合し，最終的な識別結果 $\hat{C}$ を式 (3) より求める．

$$P(c) = \frac{1}{100}\left(\sum_{i=1}^{100} P_i(c|\boldsymbol{x}_i)\right) \qquad (2)$$

$$\hat{C} = \arg\max_c P(c) \qquad (3)$$

5.　評価実験

力覚センサと画像を用いた把持物体の識別の有用性を示すために評価実験を行う．本実験では，三菱電機社製ロボットである MELFA(RV-7F)，力覚センサ (1F-FS001-W200)，ロボット用小型三次元ビジョンセンサ (4 F-3DVS2-PKG1) を用いる．実験対象は，Amazon Picking Challenge 2016 データセットの 21 アイテムを対象とする．

取得する動作シーンはロボット座標系で X，Y，Z の 3 軸方向へ往復させたデータを，学習用データ 3,326 セット，評価用データは 375 セットを使用する．学習用データは X が 1124，Y が 1124，Z が 1078 セットの力覚データが用意されている．評価用データはそれぞれ 125 セットの力覚データが入っている．

本実験では，画像のみ，力覚データのみ，マルチモーダルに学習した場合の識別結果を比較する．

5.1　実験結果

画像のみ，力覚データのみ，同時に用いた場合の Confusion matrix を図 9 に示す．画像のみで学習した場合識別率は 80.7%，力覚データのみでは 79.7%となり，画像と力覚データを組み合わせた場合識別率が 87.5%となった．画像のみで識別した場合よりも 6.8%向上している．図 10 のように画像はテクスチャが似ているアイテム同士で誤識別をしてしまう場合がある．また，力覚データは軽いアイテムにを誤認識する傾向にある．画像と力覚データを用いることでアイテムのテクスチャとモーメントから識別に有効な特徴を得られたと考えられる．

また，剛体のアイテムはシャツなどの非剛体のアイテムに比べ，高い識別率を達成した．非剛体のアイテムは不規則に振動するため，学習誤差が収束せず，その結果誤識別の要因となっている．非剛体を 1 クラスとした全 21 クラスを LSTM で学習すると，平均識別率は 90.1%を得た．力覚データを用いた学習は剛体，非剛体のカテゴリ分類に有効な手段だと考えられる．

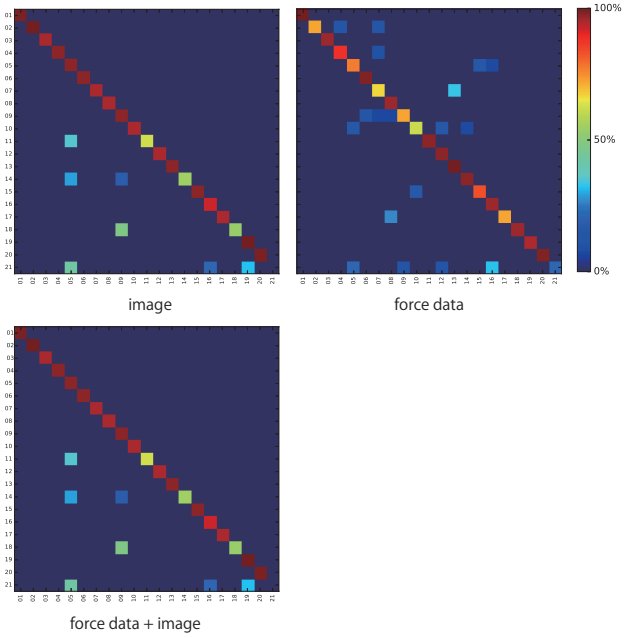

図 9　画像と力覚データを用いた場合の Confusion matrix

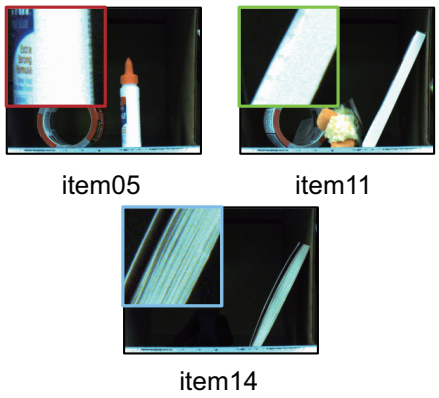

図 10　画像識別が難しい物体

6.　おわりに

本研究では，力覚センサと画像情報を用いたマルチモーダル学習による把持物体の識別法を提案し, その有効性を示した．画像と力覚データを組み合わせた場合識別率が 87.5%となり，それぞれのデータのみに比べ，識別率が向上した．今後の課題として複数パターンの動作による比較を行う．

参　考　文　献

[1] K. Noda, H. Arie, Y. Suga and T. Ogata“Multimodal integration learning of robot behavior using neural networks,”Robotics and Autonomous Systems, vol.62, issue 6, pp.721-736, 2014.

[2] T. Kawabe, W Roseboom, S Nishida“The sense of agency is action-effect causality perception based on cross-modal grouping”, Proc. R. Soc. B: Biological Sciences, vol.280, issue 1763, (1763) (2013), 2013.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng,“Multimodal deep learning”, International Conference on Machine Learning, Bellevue (28): 689-696, 2011.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner,“Gradient-Based Learning Applied to Document Recognition”, Proceedings of the IEEE, pp. 2278-2324, 1998.

[5] T. Yoshikawa, T. Miyazaki,“Development of a six-axis force sensor”, Proc. Japan-USA Symp. Flexible Automation, pp. 531-538, 1988.

[6] K. Hirai:“Current and Future Perspective of Honda Humanoid Robot,”Proc. IEEE/RSJ Int. Conference on Intelligent Robots and Systems, pp.500-508, 1997.

[7] M. Raibert, J. Craig,“Hybrid position/force control of manipulators”, ASME J. Dynamic Syst. Meas. Contr., 981.

[8] S.Hochreiter,“Long Short-Term Memory”, Neural Computation, 9(8): 1735-1780, 1997.

[9] J. Elman,“Finding Structure in Time”, Cognitive Science 14 (2): 179-211, 1990.

[10] P. Werbos,“Generalization of backpropagation with application to a recurrent gas market model”, Neural Networks 1 (4): 339-356, 1988.

[11] Y. Domae, H. Okuda, Y. Taguchi, K. Sumi, and T. Hirai,“Fast graspability evaluation on single depth maps for bin picking with general grippers”, International Conference on Robotics and Automation, pp.1997-2004, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton,“ImageNet classification with deep convolutional neural networks”, Neural Information Processing Systems, pp. 1106-1114, 2012.