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に関する研究

TP17030 **山崎 雅幸**　　**指導教授：山下 隆義**

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，それぞれのセンサに適したタスクがある．人間も同様に視覚，聴覚，触覚などのセンサのような複数の感覚を持ち，それらを用いて外部環境を知覚している．ロボットが特定のタスクにおいて高い識別精度を達成するためには，人間と同様に複数のセンサを利用するマルチモーダル学習 [1] が重要となる．従来の物体識別は，Deep Convolutional Neural Network(DCNN) 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を単一のモーダルとして利用する．しかし，画像だけでは把持対象物体における内包物の種類や量などの取得できない情報がある．本研究では，このような物体に対して力覚センサを用いて物体にかかるモーメントを用いることで，画像のみの単一モーダルを用いる場合よりも精度を向上させることを目的とする．本研究では画像情報，力覚センサ，電流フィードバック (FB) をマルチモーダル情報とし，Recurrent Neural Network の一種である Long Short Term Memory(LSTM)[2] を用いて把持物体の識別を行う手法を提案する．また，識別に有効な力覚センサのための動作を考察する．

2. 把持物体のマルチモーダル情報

マルチモーダル学習は，画像，音声，関節データなど複数の要素を用いた学習である．マルチモーダル学習を行うことで単一モーダルとして学習した場合よりも高精度な認識が可能となる．本研究では，画像及び力覚センサと電流 FB を利用する．

画像情報

画像情報はロボットの視覚に相当し，様々なタスクに使用されている．産業ロボットでは，物体の把持やキャリブレーションに用いられ，さらに通常の作業だけでなく，様々な作業タスクの効率化に用いられている．画像情報を用いるためには，画像から特徴抽出を行う必要がある．DCNN を用いることで識別に適した特徴を獲得することができる．

力覚センサ

力覚センサはロボットの触覚に相当し，様々なロボットに装備されている．力覚センサを用いることで接触を判定できるため，特に産業ロボットの微細な部品のはめ込み等に利用されている．力覚センサは，図 1 に示すようにセンサの上部と下部にかかる力の変位を静電気容量の変化から計測し，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) である．

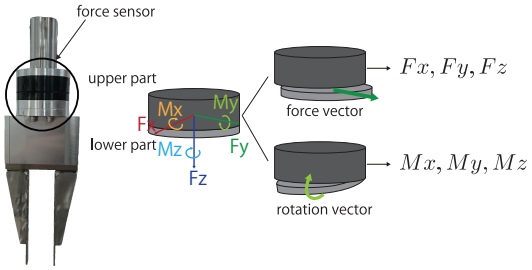

図 1：力覚データの測定

電流フィードバック

電流 FB は，アームが動作する際に関節のモータにかかる電流を測定した値である．関節に流れる電流値を示しており，アームにかかる力により増減する．

3. 提案手法

本研究における提案手法の流れを図 2 に示す．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，把持点周囲の画像，把持力覚センサ，電流 FB より得られたデータを取得する．取得し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．

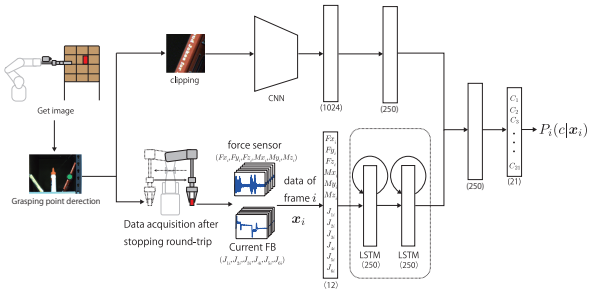

図 2：マルチモーダル学習による把持物体識別の流れ

3.1 画像情報の取得

本研究では，複数のアイテムがランダムに入っている箱や棚を想定している．物体検出を用いる場合，様々な向きに対応しなければならず，検出の計算コストが高くなってしまう．そこで，物体の種類に依存しない Fast Graspability Evaluation(FGE)[3] を用いて把持可能領域を検出し，把持点の周辺画像を切り出してから物体識別に用いる．アイテム画像の把持点周囲を 156 × 156 ピクセルで切り出し，224 × 224 ピクセルに拡大し，パッチ画像を作成する．局所的な画像を作成することで，隣接したアイテムの影響を減らし，アイテム識別の高精度化が期待できる．

3.2 力覚データと電流 FB の取得

把持したアイテムを一定方向にそれぞれ 1 往復させ，力覚データを取得する．取得した 1 往復中の力覚データには，移動時のデータ停止，制動までのデータが含まれており，長い系列データを扱うことになる．特に動作中の力覚データは，物体固有の特徴が得られない．そこで，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．

まず，取得したデータの正規化を行う．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行う．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．これにより，物体固有の振動を得ることができる．

$$\boldsymbol{x}_i = \Big[\frac{Fx_i}{F_{\max}}, \frac{Fy_i}{F_{\max}}, \frac{Fz_i}{F_{\max}}, \frac{Mx_i}{M_{\max}}, \frac{My_i}{M_{\max}}, \frac{Mz_i}{M_{\max}}, \frac{J_{1i}}{J_{\max}}, \frac{J_{2i}}{J_{\max}}, \frac{J_{3i}}{J_{\max}}, \frac{J_{4i}}{J_{\max}}, \frac{J_{5i}}{J_{\max}}, \frac{J_{6i}}{J_{\max}}\Big] \tag{1}$$

3.3 識別結果の統合

本ネットワークは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．系列データを扱う場合，最終フレームを入力した時の出力を結果として用いる場合が多い．しかし，動作データの場合，振動が収束していくため特徴が失われていく．そこで，全フレームの出力を考慮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 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．

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

$$\hat{C} = \arg\max_{c} P(c) \tag{3}$$

表 1：動作パターンごとの識別率 [%]

	画像のみ	動作パターン									
		1	2	3	4	5	6	7	8	9	10
力覚センサのみ	-	32.2	64.6	65.2	30.6	60.6	64.2	20.6	34.2	30.4	40.6
AlexNet	69.4	59.8	73.2	72.6	52.4	72.6	70.4	52.2	54.2	55.8	56.4
VGG-16	72.4	60.0	74.8	73.4	54.2	71.8	71.2	55.2	55.6	56.8	57.8
ResNet50	73.2	60.0	**75.2**	72.2	53.8	73.2	72.0	58.0	59.8	59.8	60.2

4. 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．力覚センサ，画像情報をそれぞれ単一モーダルとして用いた識別，マルチモーダル学習として力覚センサと電流 FB，画像情報を利用した識別精度を比較する．

4.1 実験概要

複数の動作パターンを比較することで識別に有効な動作方法を調べる．また，識別に有効な画像識別器の検討を行う．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 (4F-3DVS2-PKG1) を用いる．実験対象は，Amazon Picking Challenge データセットの 21 アイテムとする．取得する動作パターンを図 3 のように 10 種類比較する．動作パターン 1〜6 はハンドを地面に垂直に向け，Fx, Fy, Fz, Mx, My, Mz 方向に動作させたものである．動作パターン 7〜10 はハンドを地面に平行に向け，Fx, Fy, Fz, Mz 方向に動作させたものである．学習用データ 5000 セット，評価用データは 500 セットを使用する．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．

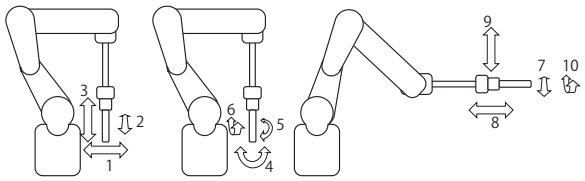

図 3：力覚データを取得するロボットの動作

4.2 実験結果

各動作パターンを用いた時の識別結果を表 1 に示す．動作パターン 1，4，7，8，9，10 の力覚データを用いた場合の識別精度が，画像のみの識別に比べ低下している．動作パターン 1，4 の動作についてはグリッパに対して垂直な方向に動作させているため，有効な把持物体の振動が力覚センサに伝わらなかったと考えられる．動作パターン 7，8，9，10 の動作については，動作姿勢が地面と平行になり動作時に生じた不要な振動がノイズとなり，識別精度が低下した．動作パターン 2，3，5，6 の場合，力覚データのみが最も識別精度が低い．一方，マルチモーダル学習として画像情報と力覚センサの両方を利用することで，画像のみの識別率を上回ることができた．ネットワーク構造に ResNet50 を用いて，動作パターンを 2 とした場合，75.2% と最も高い識別率を達成した．画像のみを用いた識別では，テクスチャが似ているアイテムの識別率に低下がみられた．また，力覚データを用いた識別では，多くのアイテムで誤識別が見られた．これには，力覚データ取得環境の差が起因すると考える．一方，マルチモーダルに学習することで，画像識別における誤識別をしたテクスチャの似たアイテムを正答することができた．

力覚センサのみ，画像のみ，マルチモーダルに用いた識別の confusion matrix を図 4 に示す．図 4(b) に示すアイテム 17〜20 は，ぬいぐるみや衣類といった非剛体の物体であり，力覚データを用いる場合，識別率が低下した．しかし，画像と共に用いることで，非剛体の物体に対して識別率が向上し，画像のみに比べても全体的に識別率が向上した．これより，画像と力覚データを用いることで，識別に有効な見え情報と動き情報の特徴を得られたと考えられる．

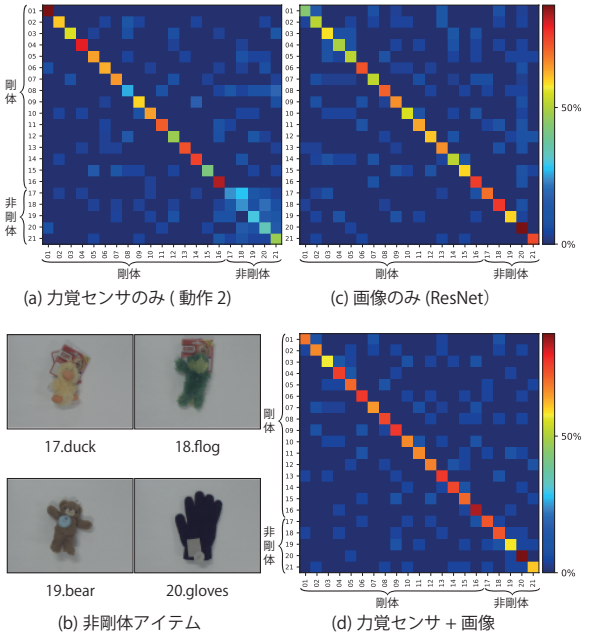

図 4：confusion matrix による比較

5. 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手法を提案し，その有効性を示した．また，識別に有効な力覚データ取得動作の比較を行った．動作する姿勢，ハンドにかかる力を考慮した動作をすることにより，識別に有効な力覚データを得ることができた．画像と力覚データを組み合わせた場合，識別率が 75.2%となり，各単一モーダルのみに比べ，識別率が向上した．今後の課題として学習データ数による精度，汎化性の向上などが挙げられる．

参考文献

[1] K. Noda, *et al.*“Multimodal integration learning of robot behavior using neural networks,” Robot. Auton Syst, Vol. 62, No. 6, pp.721–736, 2014.

[2] S. Hochreiter, “Long Short–Term Memory”, Neural Computation, Vol. 9, No. 8, pp.1735–1780, 1997.

[3] Y. Domae, *et al.*“Fast graspability evaluation on single depth maps for bin picking with general grippers”,ICRA, 2014.

研究業績

[1] 山崎雅幸 等, “LSTM による力覚データを用いた把持物体の識別”, 画像センシングシンポジウム, 2017.

[2] 山崎雅幸 等, “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”, 日本ロボット学会 学術講演会, 2018.

(他 学会発表 2 件)