B-06

マルチモーダル情報を用いたロボットによる物体概念獲得のシミュレーション

Simulation of the object concept acquisition with the robot using the modal information

笹野　仁†　　吉野　幸一郎†　　中村　哲†
Jin Sasano　　Koichiro Yoshino　　Satoshi Nakamura

1. はじめに

人間は、視覚や聴覚のような五感から得られる様々な情報を用いて物体の概念を獲得している。このような物体概念の獲得を、身体性を持つロボットが得る様々な情報から教師なしで学習する試みが行われている [1]。こうした学習の枠組みでは、人間が行うようにその物体に関する様々な情報を得るための試行を繰り返し、その結果の類似から試行に用いた物体のクラスタリングを行う。このクラス一つ一つを、ロボットが獲得した物体概念として扱う。ロボットにより学習された物体概念には、対話システムや生活支援システムへの応用など、実世界の物体概念を扱う必要があるシステムでの利用が考えられている。

しかし、人間が普段の生活で扱う物体の数は膨大で、これらすべての物体概念を実際のロボットを用いて獲得することは難しい。これは、大量の学習対象の入手やロボットの運用などの物理的な制約が大きなコストとなるためである。そこで、ロボットの物体概念の獲得を仮想環境で行うことができれば、こうしたコストを低減することができる。具体的には、ロボット実機で行った概念獲得の結果と仮想環境でシミュレートした概念獲得の結果が一致すれば、仮想環境での学習結果を実機に転用することができる。

本研究では、まず先行研究 [1] で行われているように、ロボット実機での聴覚と視覚を利用した概念獲得を再現する。また、同様の物体クラスに対して、聴覚・視覚を再現した仮想環境上での概念獲得を行う（シミュレーション）。最後に、実機とシミュレーションでの実験の比較を行い、仮想環境上での概念獲得結果がどの程度実機での概念獲得結果と一致するかを確認する。

2．ロボットによる概念獲得

人間は、五感から得られる情報を用いて物体を認知し、その繰り返しによって同じ種類のものが同じクラスに属するという概念を獲得すると考えられている [2]。例えばコーヒーカップを認知する場合は、コーヒーカップを触った時の感触や見た目など、複数の感覚器の情報を統合して物体の認知・学習を行っている。こうした情報は単一の種類の情報ではなく、多様な種類の情報（マルチモーダルな情報）が用いられている。マルチモーダルな情報により、同じ種類のものを同じクラスのものとして分類する、物体概念という物体の認知が形成される。

このような物体概念の獲得を、マルチモーダル情報を用いてロボットにより獲得する試みがすでに研究されている [1]。ロボットにより獲得された物体概念は、様々な応用が考えられる。 例えば、生活支援システムの研究 [3, 4] では、システムは人間の身体性に基づいて物体の概念を理解する必要があるとされている。また対話システムの研究においては言葉だけでなく表情や身振り手振り、語調といった複数のチャンネルを使用したマルチモーダルなコミュニケーションに関する議論が盛んに行われている [5, 6, 8]。

これらの研究では、人間の身体的な感覚を理解するうえでマルチモーダルな概念獲得が有効であるとされている。しかし、物体概念の学習・獲得のために実際のロボットを用いることはコストが大きい。例えば、学習対象が消耗品である場合や高価なものである場合などである。また、実ロボットを運用する費用もコストを増大させる要因である。加えて、実ロボットには物理的な制約があり、学習に長い時間を要する。これらに対して、仮想環境で物体概念の獲得を行うことができれば、こうしたコストの問題を解決することが可能である [7]。

そこで本研究では、ロボットによる物体概念の獲得に関して仮想環境上でシミュレーションを行い、学習のコストを低減する方法を検討する。

3. ロボット実機による概念獲得

まず先行研究同様、実機によって聴覚・視覚を用いた概念獲得を行う。実験に用いるロボットとして Aldebaran Robotics の NAO[9] を使用した。ロボットは４種類の物体（図 1）を握り、握った物体をカメラに近づけ撮影する行動と、物体を振りそれにより生じる音を録音する行動を各物体に対して行う（図 2）。ロボットは頭部にカメラ (CMOS 640 x 480 camera) ２台とマイク４台を搭載しており、これらから得られる情報をそれぞれ視覚情報、聴覚情報として利用する。

(1) 模様あり、鈴あり　(2) 模様あり、鈴なし　(3) 模様なし、鈴あり　(4) 模様なし、鈴なし

図 1　ロボットが認識する 4 種類の物体

これらの行動によって、得られた視覚情報・聴覚情報を、LDA[10] によって教師なしで分類する。LDA で扱う特徴量には、Bag of features モデルを用いる。

†奈良先端科学技術大学院大学(NAIST)

(a) 振って音を聞く（聴覚） (b) 近づけて見る（視覚）

図 2　ロボットが物体に行う 2 種類の動作

3.1　Bag of features モデル

Bag of features 特徴量は各離散特徴量の出現頻度ヒストグラムをベクトルにしたもので、ベクトルの成分が各特徴量の出現回数となる [11]。この特徴量を作成するため、あらかじめ与えられたサンプル画像の特徴量ベクトルをクラスタリングし、サンプル画像を代表する特徴量（コードブック）を計算する。分析対象の画像ベクトルをこのコードブックを用いてベクトル量子化することで、各コードブックにおける特徴量の出現頻度ヒストグラムを計算することができる。

先行研究 [1] ではこの Bag of features モデルを視覚情報だけでなく、聴覚情報と触覚情報にも拡張したモデルであるマルチモーダル Bag of features モデルを提案している。本研究ではこれにならい、視覚情報と聴覚情報を以下のように処理する。

3.2　視覚情報の処理

ロボットのカメラで撮影した画像は、SIFT 特徴量 [12] を用いて 1 枚ごとに 128 次元の特徴ベクトル 300〜400 個に変換する。変換された特徴ベクトルは、学習とは関係のない画像 10 枚から計算された 100 次元の代表ベクトルを用いてベクトル量子化する。つまり視覚情報は、100 次元のヒストグラムに変換される。このヒストグラムのインデックスが LDA で扱う特徴量となる。

3.3　聴覚情報の処理

ロボットがマイクでとらえた音声は、MFCC(Mel-frequency cepstrum)[13] を用いて 13 次元の特徴量ベクトルに変換する。この特徴ベクトルは白色雑音や複数の音楽、人間の音声を用いて計算した 30 の代表ベクトルによりベクトル量子化する。視覚特徴量は 30 次元のヒストグラムに変換され、このヒストグラムのインデックスが LDA で扱う特徴量となる。

3.4　LDA による教師なし分類

ベクトル量子化された視覚情報と聴覚情報を結合し、一つのベクトルとみなして LDA による教師なしの分類を行う。クラス数は 4 としてあらかじめ与えた。今回の実験では、4 種類の物体に対しそれぞれ 10 回ずつ計 40 回物体の視覚情報と聴覚情報を取得し、それらの情報を入力として LDA でラベルを割り当てた。この LDA によるラベルの割り当て結果と、実際の物体番号の一致率を表 1 に示す。評価には正解ラベルに対する適合率、再現率およびその調和平均（F 値）を用いる。support は各正解ラベルの個数である。全ラベルをあわせた Accuracy は 78%であった。

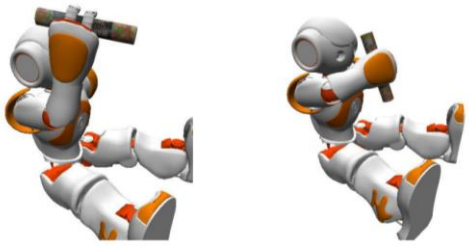
表 1　実機による実験の結果

クラス	適合率	再現率	F 値	support
1	0.58	0.70	0.64	10
2	1.00	0.40	0.57	10
3	0.77	1.00	0.87	10
4	0.91	1.00	0.95	10

この結果から、先行研究と同様に、物体の教師なし分類にある程度成功しており、概念獲得が行われていることがわかる。

4. 仮想環境における概念獲得のシミュレーション

次に、実機と同様の状況を仮想環境で再現し、概念獲得のシミュレーションを行う。仮想環境には Unity[14] を使用した。Unity 上で Aldebaran Robotics が提供する 3DCG モデルを用いて実機での実験を再現した。図 3 に実験の様子を示す。



(a) 振って音を聞く（聴覚） (b) 近づけて見る（視覚）

図 3　仮想環境上でロボットが物体に行う 2 種類の動作

シミュレーションに使用する物体は blender[15] を使用し作成した（図 4）

(1) 模様あり、鈴あり　(2) 模様あり、鈴なし　(3) 模様なし、鈴あり　(4) 模様なし、鈴なし

図 4　仮想環境上でロボットが認識する 4 種類の物体

視覚のシミュレーションとしては、Unity 上でロボットのカメラの位置から捉えられる画像情報を用いた。また聴覚のシミュレーションとしては、物体の中の鈴のシミュレーションを行い、その衝突に合わせてあらかじめ録音した鈴の音を鳴らすことで、実機のように音声データを得ることができるようにした。これらのシミュレーション環境で、実機と同様の試行（4 種類の物体をそ

れぞれ 10 試行）を行い、獲得された特徴量に対する LDA を用いたクラスタリングを行った。この結果を表 2 に示す。

表 2　シミュレーションによる実験の結果

クラス	適合率	再現率	F 値	support
1	1.00	0.50	0.67	10
2	0.67	1.00	0.80	10
3	1.00	0.70	0.82	10
4	0.77	1.00	0.87	10

この結果から、仮想環境上であっても物体の教師なし分類がある程度成功しており、シミュレーションを用いた物体の概念獲得が行われていることがわかる。すべてのラベルをあわせた Accuracy は 80%で、実機での結果と大きく変わらないことがわかる。

5. シミュレーションと実機の比較

4 節では、シミュレーションにおいても実機で行う場合と同程度の精度で物体に対する概念獲得が可能であることが示された。しかし、これを実機で利用する場合の概念獲得として転写する場合、この結果がどの程度実機での結果と一致しているかが重要となる。そこで、実機とシミュレーションで得られた 40 試行のラベルの割り当ての一致率を確認した。これを表 3 に示す。ここでは実機で獲得されたラベルを正解とし、シミュレーションで獲得されたラベルの精度を調べたものを、これまでの実験同様適合率、再現率とその調和平均で表す。

表 3　実機とシミュレーションの実験の比較

クラス	適合率	再現率	F 値	support
1	1.00	0.42	0.59	12
2	0.57	1.00	0.73	4
3	0.87	1.00	0.93	13
4	0.85	1.00	0.92	11

全体の Accuracy としては 83 %で、実機と仮想想環境で獲得されたラベルにはある程度の一致が見られるものの、完全には一致しなかった。しかし今回は、実機・仮想環境双方の概念獲得結果も実際のラベルと一致していない部分が一部あり、素性の改善や触覚などの新しいセンサー情報の付与により、これらの結果が近づいていく可能性がある。

6. まとめ

本研究では視覚と聴覚を通して得られる 2 種類の情報を利用してロボットによる教師なし学習での物体概念獲得を行った。また、同様の実験を行うことができる仮想環境を構築し、仮想環境上での概念獲得のシミュレーションを行った。加えてシミュレーションと実機の両方の実験結果を比較した。この比較の結果、実機とシミュレーションの両方について結果がある程度一致することが示された。

今後の課題として、特徴量の選択による教師なし学習でのラベルの割り当ての精度の向上、より多くの物体を用いた実験などがあげられる。また、視覚と聴覚以外に触覚のモダリティを使用した実験を行う予定である。

参　考　文　献

[1] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai and Naoto Iwahashi: “Grounding of Word Meanings in LDA-Based Multimodal Concepts”, Journal of Intelligent and Robotic Systems, pp.1-18, Jul.2011

[2] Erdogan, G., Yildirim, I., & Jacobs, R. A. “From sensory signals to modality-independent conceptual representations: A probabilistic language of thought approach.”, PLoS Computational Biology, 11, e1004610.,2015

[3] 伊藤麻里, 矢入健久: “部屋の片付けのための情報支援システムの提案”, 人工知能学会全国大会論文集 28, 1-3, 2014

[4] 井手麻友美, 荒木雅弘: “マルチモーダル対話記述における関数による解釈表現の検討”, 人工知能学会全国大会論文集, 1K3-OS-17a-3, 2013

[5] 角 康之: “マルチモーダルデータに基づいた多人数会話の構造理解”, 情報処理学会研究報告, ヒューマンコンピュータインタラクション研究会報告, 2011-HCI-145(7), p1,2011

[6] 高橋裕己, 中野幹生, 岩橋直人, 左祥, 船越孝太郎, 岡夏樹,菅野重樹 “マルチモーダル情報を利用した未知語を含む発話のドメイン選択精度の向上”, 第 76 回情報処理学会全国大会講演論文集 2014(1), 443-444, 2014-03-11

[7] MIT Technology Review “To Get Truly Smart, AIMightNeedtoPlayMoreVideoGames”,https://www.technologyreview.com/s/601009/to-get-truly-smart-ai-might-need-to-play-more-videogames/

[8] T. Tagniguchi, K. Hamahata, and N. Iwahashi, “Unsupervised segmentation of human motion data using sticky hdp-hmm and mdl-based chunking method for imitation learning”, Advanced Robotics, vol. 25, no. 17, pp. 2143-2172, 2011.

[9] “Aldebaran Robotics NAO”, https://www.ald.softbankrobotics.com/en/cool-robots/nao

[10] Blei, D. M.,　Ng, A.Y. and Jordan, M.I.: “Latent Dirichlet Allocation”, Journal of Machine Learning Research 3, pp.993-1022, 2003.

[11] Csurka, G., Dance, C.R., Fan, L., Willamowski, J. and Bray, C: “Visual categorization with bags of keypoints”, ECCV International Workshop on Statistical Learning in Computer Vision (2004).

[12] D. G. Lowe.: “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, 60(2):91-110, 2004.

[13] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄.: “音声認識システム”, オーム社, 2001.

[14] “Unity4.0”, https://unity3d.com/jp/unity/whats-new/unity-4.0/

[15] “blender”, https://www.blender.org/