LI-006

物体のマルチモーダルカテゴリゼーション

Multimodal Categorization of Objects

長井隆行 [1]　岩橋直人 [2,3]
Takayuki Nagai　Naoto Iwahashi

1　まえがき

近年, 大量の画像を用いた物体の教師なし学習が盛んに研究されている [1][2]. これは, 様々な環境において柔軟に物体認識を行うために重要であり, タスクの設定や正解ラベルを用いた学習を必要とする従来の物体認識の枠組みを大きく変えることができる. このような物体の教師なし学習は, 物体の視覚的特徴を統計的手法を用いてクラスタリングすることで実現可能である. つまり, 画像中の特徴量の共起性がカテゴリを形成する手がかりとなる. しかし物体のカテゴリは, 常に視覚的な情報のみで決定できるわけではなく, 人間が物体をカテゴリに分類する際は, より高次の情報を用いていると思われる. そこで, ロボットが画像情報だけでなく音声や触覚の情報を利用することで, より人間の感覚に即した物体のカテゴリ分けを自動的に行うことを考える. ロボットがカテゴリ分類を教師なしで自律的に行うことができれば, 物体の認識や, 機能の推定などが可能となるため非常に有用である.

本稿ではロボットの身体性を利用することで, 物体を見るだけでなく, 実際に掴み様々な視点から観察することで得られるマルチモーダル情報を利用した物体のカテゴリゼーションを提案する. 提案法は, 統計的手法である pLSA(probabilistic Latent Semantic Analysis) をベースとしており, 物体のカテゴリゼーションを行うとともに, 物体の機能を確率的に推測する枠組みを提供するものである. 実際のロボットによる実験を通して, 提案手法の有効性を示す.

2　マルチモーダル Bag of words モデル

2.1　Bag of words モデルによるトピックモデル

自然言語処理では, 与えられた大量のテキストデータから意味のまとまりであるトピックを教師なしで見つけ出す手法が研究されている. 一つの重要な考え方は, トピックとは単語の出現頻度のパターンで定義されるというもので,Bag of words モデルと呼ばれる. 従ってトピックは, 単語の出現位置や順序に関係なく, その頻度を基にモデル化される.

一方, 画像において同様の手法を適用することで, 物体のカテゴリゼーションが可能となる [1]. この際トピックモデルにおける文書 d が画像 (シーン) に, 単語 w が局所的な特徴量 (ベクトル量子化したもの) に, トピック z がカテゴリに対応する. pLSA のモデルでは, 文書 d, 単語 w, トピック z の同時確率を,

$$p(d, w, z) = p(d)p(z|d)p(w|z) = p(z)p(d|z)p(w|z) \quad (1)$$

と書くことができる. 但し右辺は, 文書と単語に関して対称なモデルとなっていることに注意が必要である [5].

[1] 電気通信大学大学院電気通信学研究科電子工学専攻, UEC
[2] 独立行政法人 情報通信研究機構, NICT
[3] ATR 音声言語コミュニケーション研究所, ATR

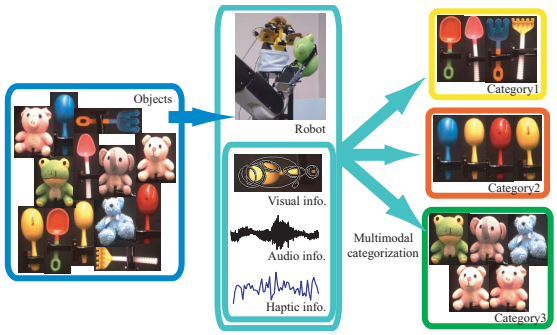

図 1　システムの概要

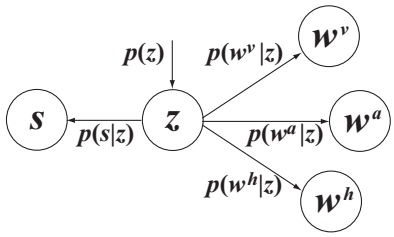

図 2　マルチモーダルカテゴリゼーションのグラフィカルモデル

本稿では以降これを, 視聴覚及び触覚を用いたマルチモーダルなカテゴリゼーションに拡張する.

2.2　提案手法の概要

ロボットは, 物体を掴み, 様々な角度からこれを観察することが可能であり, その間, 同一の物体を観測しているという情報を積極的に利用することができる. ここでは, 同一の物体を観測している間に得られる視覚情報, 聴覚情報 (物体を振ったときに生じる音), 触覚情報 (硬さ) を位相情報を考慮することなく生起回数の情報として利用する. これは, 各特徴量を「単語」と考えれば, 前述の Bag of words モデルであり, 本稿ではこれをマルチモーダル情報に適用することで物体のカテゴリゼーションを行う. 図 1 にシステムの概要を示す. ロボットはカメラ, マイク, アーム, ハンド, 感圧センサを備えており, 様々な物体を実際に掴むことで観察する. その間に得られる, 画像情報, 音情報, 触覚情報を用い, 物体の性質の類似性から物体を分類する. この際, 物体の性質とは, 物体の見た目や振った際の音, 物体の硬さを意味している. 分類は確率モデルの学習として実現されるため, その後新たに観測される未知物体に対して, その物体のカテゴリや性質を確率的に推論することが可能となる.

2.3　マルチモーダルカテゴリゼーションのためのグラフィカルモデル

提案するマルチモーダルカテゴリゼーションでは, 図 2 のグラフィカルモデルを用いる. 図において,w^v,w^a,w^h はそれぞれ視覚, 聴覚, 触覚情報を示している. また Z は

(a) (b)

図 3　ロボットによる物体の観測と視覚的特徴

カテゴリを,S はシーンを表している.

図から分かるように, このモデルでは各センサ情報は独立に出力される. つまりカテゴリ Z が決まった場合, 各センサ情報は他の情報とは無関係に決まることになる. 実際には, 各センサ情報同士には何らかの関係性があると考えられる. 例えば, ある視覚情報には, ある特定の音や硬さが関係している可能性がある. 但しこれらの関係性を正しく捉えるためには, 非常に精細なセンサ情報を得る必要があると考えられる. またグラフィカルモデルが若干複雑になるため, 学習や推論のための計算が複雑になる. 従って, 図 2 の独立なモデルは, センサや計算の面において現実的であると考えられ, ここではこのモデルを用いることとする.

2.4　画像情報

各画像から抽出する特徴量として, 文献 [3] の局所的重要領域を用いる. さらに各抽出領域を SIFT 記述子 [4] を用いて表現することで特徴ベクトルとする. これによって得られる特徴量は, アフィン変換に対する不変性をもち, 物体を様々な視点から観測する際の特徴量として優れている. また, 位相情報を用いないため, オクルージョンの問題を回避することができる. 特にここではロボットが自ら物体を掴んで観測するため, 自分の手によるオクルージョンが常に起こる. 自らの手に対する特徴量はあらかじめ手のみの画像を取得することで計算しておき, 物体画像の特徴量のうち, 手によって生じた特徴量は取り除くことが可能である. 図 3 にロボットがぬいぐるみを掴み様々な角度から観測している画像と, その際に抽出された特徴領域を示した.

特徴ベクトルは, 学習画像とは全く関係のない背景画像（室内シーンの画像 100 枚）から計算した 600 の代表ベクトル（コードブック）を用いてベクトル量子化する. 従って, 画像特徴量 w^v は実際にはコードブックのインデックスを表すことになり,$w^v \in \{1, 2, \cdots, 600\}$ である.

2.5　音声情報

音声情報も画像情報同様に, ひとつの物体を観測している間に得られる音声信号をフレームに分割し, それぞれを Bag of words モデルにおける単語として扱う. 特徴量としては 13 次元の MFCC を用いる. 後に述べる実験では, 男女それぞれ 3 名の音声と 3 種類の雑音から計算した 100 の代表ベクトルを用いてベクトル量子化した.

2.6　触覚情報

触覚情報としては, ロボットハンドに組み込んだ感圧センサによって得られる電圧値を用いる. これは物体の硬さを表現していると考えられるが, 物体は位置によって硬さが異なる可能性がある. 今回の実験では, 人間がロボットに物体を手渡すことである程度掴む位置が一定となるようにした. また, ロボットはひとつの物体を何度か掴むこととする. 最終的に, センサの電圧値は 4 段階に量子化する.

2.7　pLSA による学習

学習には,pLSA[5] を用いる. 既に述べたように, 提案するマルチモーダルカテゴリゼーションは, 図 2 のグラフィカルモデルで表現することができる. 従って, 観測シーン s, カテゴリ z, 画像特徴量 w^v, 音声特徴量 w^a, 触覚特徴量 w^h の同時確率は,

$$p(z, s, w^v, w^a, w^h) = p(z)p(s|z)p(w^v|z)p(w^a|z)p(w^h|z) \quad (2)$$

と書くことができる. 隠れ変数 z を含んでいるため, 各パラメータは EM アルゴリズムを用いて推定する. ここで Q 関数は次のように書くことができる.

$$Q(\theta|\hat{\theta}) = \left\langle \log p(z, w^v, w^a, w^h, s|\theta) \right\rangle_{p(z|s,w^v,w^a,w^h,\hat{\theta})} \quad (3)$$

但し,$\theta, \hat{\theta}$ はパラメータを表している.

最終的に EM アルゴリズムは次のようになる.

[**E ステップ**]

$$p(z|s, w^v, w^a, w^h) = \frac{p(z)p(s|z)p(w^v|z)p(w^a|z)p(w^h|z)}{\sum_z p(z)p(s|z)p(w^v|z)p(w^a|z)p(w^h|z)} = p(z|\boldsymbol{D})$$

[**M ステップ**]

$$p(w^v|z) \propto \sum_s \sum_{w^a} \sum_{w^h} n(s, w^v, w^a, w^h)p(z|\boldsymbol{D})$$

$$p(w^a|z) \propto \sum_s \sum_{w^v} \sum_{w^h} n(s, w^v, w^a, w^h)p(z|\boldsymbol{D})$$

$$p(w^h|z) \propto \sum_s \sum_{w^v} \sum_{w^a} n(s, w^v, w^a, w^h)p(z|\boldsymbol{D})$$

$$p(s|z) \propto \sum_{w^v} \sum_{w^a} \sum_{w^h} n(s, w^v, w^a, w^h)p(z|\boldsymbol{D})$$

$$p(z) \propto \sum_s \sum_{w^v} \sum_{w^a} \sum_{w^h} n(s, w^v, w^a, w^h)p(z|\boldsymbol{D})$$

$$n(s, w^v, w^a, w^h) = n(s, w^v)n(s, w^a)n(s, w^h)$$

但し,$n(s, w^*)$ はシーン s における特徴量 w^* の生起回数を表している. ランダムな初期値からスタートし E ステップと M ステップを繰り返すことで, 各パラメータを決定することができる. 但し, 得られるパラメータが最適解であることは保証されない. 後に示す実験では, ランダムな初期値を複数回選び直し, 得られた結果の中で Q 関数が最大となるものを選択した.

3　カテゴリゼーションに基づく物体の認識

本章では, 学習した確率モデルを用いた未知物体のカテゴリ推定について述べる. また, 視覚情報から聴覚情報を予測するなど, モダリティ間の推論を確率的に行うことが可能であり, その手法についても述べる.

3.1　カテゴリ認識

物体 (シーン s) が与えられた場合, そのカテゴリは $p(z|s)$ を最大とするカテゴリ z とすれば良いことになる.

従って, 学習したパラメータより $p(z)p(s|z)$ を計算し, これを最大とする z が推定されたカテゴリである. しかし, パラメータ $p(s|z)$ は与えられたシーン s ごとに点推定しているため, 未知のシーン $\bar{s}$ に対してはそのまま適用することができない. そこで, fold in ヒューリスティックを用いる [5]. これは, 新たな $\bar{s}$ に対して EM アルゴリズムによってパラメータを決め直すものである. 但し, パラメータ $p(w^v|z)$,$p(w^a|z)$,$p(w^h|z)$ を固定し,$p(\bar{s}|z)$,$p(z)$ を変化させる. 最終的に未知物体のカテゴリは,

$$z = \underset{z}{\operatorname{argmax}}\, p(z|\bar{s}) = \underset{z}{\operatorname{argmax}}\, p(z)p(\bar{s}|z) \qquad (4)$$

によって決めることができる.

3.2 機能の推定

提案するマルチモーダルカテゴリゼーションの有効性は, 物体のカテゴリ分類だけでなく, あるセンサ情報を得ることによって他のセンサ情報を推測することができる点にもある. つまり, 物体を見ることによって, それが音を出すかどうか, またどのような音を出すか, 硬さなどの情報を推測することができる. これにより例えばロボットは, 音を鳴らすおもちゃが欲しい場合複数の物体から最も音を出す可能性の高い物体に手をのばすことが可能となる. 例として, 画像情報から音声情報を推定する場合を考える. 但し, 他の場合も同様に考えることができる.

$$w^a = \underset{w^a}{\operatorname{argmax}}\, p(w^a|w^v, \bar{s}) \qquad (5)$$

右辺はさらに次のように書くことができる.

$$p(w^a|w^v\bar{s}) = \frac{\sum_z p(w^a|z)p(w^v|z)p(\bar{s}|z)p(z)}{\sum_z p(w^v|z)p(\bar{s}|z)p(z)} \qquad (6)$$

ここで,$p(\bar{s}|z)$ は前節同様に fold in ヒューリスティックにより決定する. しかしこの場合は w^a,w^h に関する情報は得られていないため,EM アルゴリズムではそれらの情報に関して周辺化を行う.

一方,

$$w^a = \underset{w^a}{\operatorname{argmax}}\, p(w^a|w^v) \qquad (7)$$

とすることで, シーンを介さずに, 視覚的特徴から直接音声情報を推定することも可能である. しかし w^v は複数の物体で共有されるため（単語における多義語に相当する）, 推定される w^a は, 学習データに最も多く含まれる w^v と w^a の組み合わせによって決まると考えられる. 従って, ここでは式 (5) を用いることとする.

4 実験

提案手法の有効性を確かめるために, カテゴリゼーション, カテゴリ認識, 機能の予測に関する 3 つの実験を行った. 実験に用いるロボットは, 図 1 に示したもので, 頭部に 2 つのカメラと 2 つのマイクロフォン, 手には感圧センサを装備している. これらのセンサにより得られるマルチモーダル情報を, 前述の方法により処理した. オブジェクトとしては, 50 個のおもちゃ(ぬいぐるみ, がらがら, タンバリン, ゴム人形, マラカス, ボール, 砂場道具など) を用いた.

4.1 カテゴリゼーション

まず, 50 個のオブジェクト全てを 8 人の被験者に手動でカテゴリ分類してもらった. この際, 分類の基準は自由

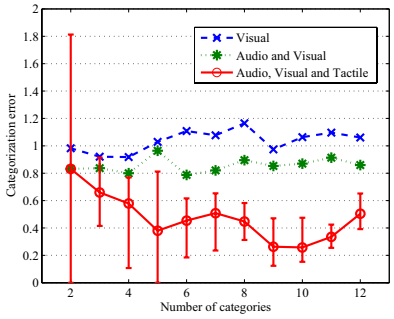

図 4　カテゴリ数 vs. カテゴリ分類エラー

であり, 感覚的に最も納得できる分類をするように注意を与えた. また, カテゴリ数は 12 から 2 まで順次変化させた. これによって得られる人手による分類結果と, 画像情報のみでの分類及び, 提案するマルチモーダル情報を用いた分類を比較することで, カテゴリ分類の性能を評価した. カテゴリ分類の類似度を測るために, 次のようなカテゴリマトリクス C と, カテゴリ分類エラー E を定義する.

$$E = \frac{1}{\sum C(i,j)} \sum_i \sum_j |C(i,j) - \hat{C}(i,j)| \qquad (8)$$

但し $C(i,j)$ は, 物体 i と j が同じカテゴリの場合 1, それ以外では 0 となる. 2 つのカテゴリ分類の結果からそれぞれカテゴリマトリクス C,$\hat{C}$ を生成し, それらの差を合計することで, 分類の類似度を測る. E の値は, 小さいほど 2 つのカテゴリゼーションが似通っていることを示している.

図 4 に結果を示す. 図において, 破線が画像情報のみ, 点線が画像と音声を用いたもの, 実線が画像, 音声, 触覚を用いた結果である. これは, 被験者 8 人の分類とのカテゴリ分類エラーの平均をプロットしたものであり, 3 つの情報を使った場合が最も値が小さいことが分かる. また, 画像, 音声, 触覚を用いた結果には, 8 人中最も値が小さかったものと最も値が大きかったものをエラーバーとして示した. これより, 人手による分類にも個人差があり必ずしも同じ分類をしていないことが分かる. これは, 特にカテゴリ数が極端に少ないときに顕著であり, 例えば 2 つに分類する場合, 複数の分類基準から 1 つだけ選択する必要があり, どの基準を選択するかによって分類結果が大きく異なる. 図 4 のカテゴリ数が 2 の場合, カテゴリエラーの最小値はゼロであり, 人手と全く同じ分類をしている. この際分類の基準として用いられているのは, 物体が音を出すかどうかであり, 実際に同じ分類をした人のカテゴリ分類の基準は,「音が鳴るもの」と「音が鳴らないもの」であった. また, 音声と画像を用いた分類でもカテゴリ数 2 の場合は, 同じ結果となった. 一方, カテゴリエラーが大きい分類は, 異なる基準を用いたもので, 例えば「タオル地のもの」と「それ以外」などである.

カテゴリ数が適切な場合, 分類の個人差にほとんど影響を受けないカテゴリも存在する. 被験者 8 人が 5〜12 のカテゴリに分類した際に全員に共通に現れるカテゴリのみを抽出したところ, 42 オブジェクト, 9 カテゴリとなった. そこで分類の個人性を除くために, この 42 オブジェクトを新たな実験セットして以降の実験を行う.　分類の結果を図 5 に示す. この図において, 縦がオブジェクトのインデックス, 横がカテゴリのインデックスであり,

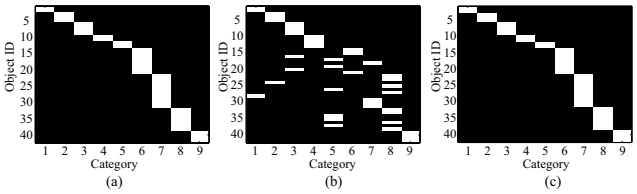

図 5　カテゴリゼーションの結果 (a) 人手による分類 (b) 画像情報のみを用いた分類 (c) 画像, 音声, 触覚情報を用いた分類

図 6　各カテゴリのオブジェクト

白いところがオブジェクトのカテゴリを表している. 図 5(a) は人手による正しい結果であり,(b) は画像のみ,(c) は 3 つのモダリティーを用いた結果である. また, 図 6 はカテゴリごとのオブジェクトを示しており, カテゴリの番号は, それぞれ図 5 のカテゴリインデックスに対応している. この結果より, マルチモーダルな情報を利用することで, 画像のみに比べより人手に近いカテゴリゼーションを行うことができていることが分かる.

図 7 は, 各カテゴリにおいて上位の $p(w^v|z)$(つまりカテゴリを表現する視覚的特徴) を楕円で示したものである. 例えばガラガラでは, 手に持つための輪の部分が特徴的であり, タンバリンでは音を出す円形の金属部分が特徴となっている. 一方,(e)(f)(g) はマラカスとシャベルが同じ特徴を共有している例であり, 画像だけではこれらが同じカテゴリとして分類されてしまう. これらが異なるカテゴリとして分類されるためには, 音声情報が必要である. (h)(i) はゴム人形とぬいぐるみが同じ特徴を共有している例であり, これらも画像情報だけでは同じカテゴリに分類されてしまう. これらをそれぞれ別のカテゴリとして分類するためには, 触覚の情報が必要である.

4.2　未学習オブジェクトのカテゴリ認識

未学習オブジェクトのカテゴリ認識性能を調べるために, 前述の 42 オブジェクトを用いて leave-one-out 法により評価を行った. 既に述べたように, 未学習の 1 つのオブジェクトについては $p(s|z)$ を $p(w|z)$ を固定した EM アルゴリズムによって適応させることになる.

$\max p(z|s)$ によって認識した結果,42 のオブジェクトは全て正しいカテゴリとして認識された.

4.3　機能の推定

学習によって得られるモデルを用いることで, 複数の物体から音のするおもちゃを選択したり, 物体の硬さを見た目や音から推測することが可能である. 図 8 は実際に画像情報から, 音声情報を推測した結果である. 図 8(a) は, いくつかの物体が重なるように配置されたシーンであり, 図 8(b) 上の楕円は全ての視覚的特徴, 図 8(c) は視覚的特徴の内, 音を出す確率が高いところ (確率 0.7 以上) を示している. このように重なりあって配置されている場合でも, タンバリンに音を出す可能性の高い局所的視覚特徴が多く配置されており, ロボットはタンバリンに手を伸ばすことが可能となる.

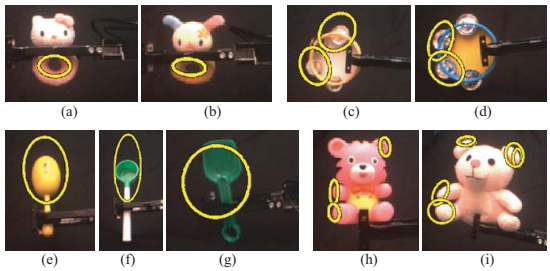

図 7　カテゴリを表現する画像特徴の例

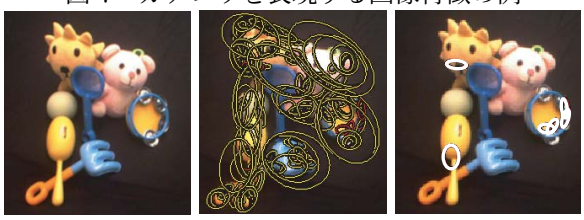

図 8　視覚的特徴による機能予測

5　まとめ

ロボットが自ら物体を観測し, マルチモーダル情報を利用することで物体をカテゴリに分類する手法を提案した. 50 個の子供用おもちゃを 10 程度のカテゴリに分類する実験を行い, 視覚的な特徴のみを用いるよりも, 聴覚や触覚 (硬さ) の情報を併用することで, より人間が行う分類に近づくことを示した. しかし人間の分類は, ただ単にボトムアップな特徴の類似性だけに頼っているわけではなく, むしろオブジェクトのもつ意味や機能が本質的に重要である. また, 手触りや材質といったより詳細な情報を用いていることも確かであり, 今後こうした点を考慮する必要がある. さらに, モデル選択によるカテゴリ数の自動決定や階層的な分類, オンラインアルゴリズムも今後の課題である.

謝辞

本研究は, 国立情報学研究所共同研究「高次元環境知覚データにおける情報構造の発見的認識に関する研究」による研究助成を受け実施したものである.

参考文献

[1] J.Sivic, B.C.Russell, A.A.Efros, A.Zisserman and W.T.Freeman, “Discovering Object Categories in Image Collections”, *AI Memo*, 2005-005, pp.1-12, Feb.2005

[2] R.Fergus, P.Perona and A.Zisserman, “Object Class Recognition by Unsupervised Scale-invariant Learning”, *in Proc. CVPR*, 2003

[3] K.Mikolajczyk and C.Schmid, “Scale & Affine Invariant Interest Point Detectors”, *Int. Journal of Computer Vision*, 60(1), pp.63-86, 2004

[4] D.G.Lowe, “Distinctive image features from scale-invariant keypoints”, *Int. Journal of Computer Vision*, 60(2), pp. 91-110, 2004

[5] T.Hofmann, “Probablistic Latent Semantic Analysis”, *in Proc. of UAI'99*, 1999