2G5-OS-19b-2

マルチモーダルカテゴリ推定のための最適な行動選択

Optimal Action Selection for Multimodal Object Category Recognition

福田 一[*1] 竹下 卓哉[*2] 谷口 忠大[*2]
Fukuda Hajime Takeshita Takuya Taniguchi Tadahiro

[*1]立命館大学大学院 [*2]立命館大学
Graduate school of Ritsumeikan University Ritsumeikan University

In this paper, we propose an optimal action selection technique for multimodal object category recognition. It is shown that a robot can obtain several multimodal categories through interaction with daily objects. However, when a robot observes a target object, the robot cannot obtain haptic and auditory information without action to the object. The robot has to decide which action it should take to recognize the object. We propose to use expected KL-divergence to select next action. We give an experiment and show its effectiveness.

1. はじめに

ロボットが物体のカテゴリを認識するには，事前に物体カテゴリがマルチモーダル情報から形成されていることが必要である．ロボットによる自律的なマルチモーダル物体カテゴリの形成を行った研究として Nakamura ら [Nakamura 11] の研究が挙げられる．Nakamura らは，Hierachical Dirichlet Process(HDP)[Teh 06] をマルチモーダル情報に拡張し，視覚，聴覚，触覚のマルチモーダル情報を用いたカテゴリ分類を行うマルチモーダル HDP というモデルを提案した．そして，ロボットが多様な日用品について，人の分類結果に近い物体のカテゴリ分類を行えることを示した．ただし，Nakamura らの研究では一部のマルチモーダル情報によって認識を行わせることができるものの，どのマルチモーダル情報を取得すべきかということの判断については扱っておらず，外部から与えられるものとしていた．

人間が得る情報には視覚情報に代表される比較的受動的で遠隔的なものから，触覚情報に代表されるような能動的で近接的なものまで幅広い．後者については，実際にその情報を取得しようと物体に働きかけないことには，その情報を得ることは出来ないため対象を認識するために，最小限の行動を選択することが重要になる．このような研究は能動学習の文脈で研究されている．Sugiura らは対話ロボットの学習においてロボットの質問生成に能動学習を用いることで，効率的な学習を行う手法を提案している [Sugiura 11]．本研究では，未観測のマルチモーダル情報を取得し，未知物体のカテゴリ推定を素早く行うための最適な行動選択法を提案する．

2. マルチモーダルカテゴリゼーション

本研究では，最適な行動選択の事前準備としてロボットに物体の視覚，聴覚，触覚情報からマルチモーダルカテゴリゼーションを行わせる．物体カテゴリを獲得するためのマルチモーダル情報は視覚情報・振ったときの聴覚情報・叩いたときの聴覚情報・触覚情報の 4 種類を用いる．マルチモーダルカテゴリゼーションには Nakamura らのマルチモーダル HDP を用いる．

連絡先: 福田一，立命館大学大学院情報理工学系研究科，
fukuda@em.ci.ritsumei.ac.jp

図 1: 実験に用いたロボット

2.1 マルチモーダル情報

2.1.1 視覚情報

ロボットは一定速度で回転する回転台の上に置かれた物体を，ロボット頭部に搭載されたカメラで固定点から撮影することにより 1 周分の画像を取得する．取得した各画像から Scale-Invariant Feature Transform(SIFT)[Lowe 04] を用いて特徴量を抽出する．その結果，1 つの画像から 128 次元の特徴ベクトルを複数得ることが出来る．各画像から得られた特徴ベクトルの数は異なっており特徴を比較する際に扱い辛いため，すべての特徴ベクトルを k-means 法でクラスタリングすることにより Bag-of-Features(BoF) に変換し，視覚特徴量とする．

2.1.2 聴覚情報

物体を叩いたときの音，振ったときの音をそれぞれ取得する．ただし，叩いたときの聴覚特徴量と振ったときの聴覚特徴量は別々の特徴量として扱う．それぞれの取得した聴覚情報はフレームに分割し，各フレームを 13 次元の MFCC(Mel-frequency cepstral coefficients) に変換する．この変換したデータを特徴ベクトルとし，視覚情報の時と同様に，k-means 法でクラスタリングすることにより BoF に変換し，聴覚特徴とする．

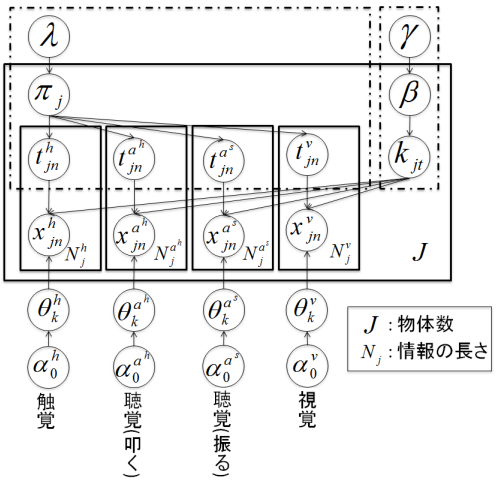


図 2: マルチモーダル HDP のグラフィカルモデル

2.1.3 触覚情報

物体を把持することで触覚情報を取得する．ロボットハンドの前に物体を置き，ロボットは徐々にハンドを閉じ，物体から一定の応力が返ってくるまでハンドを閉じ続ける．ここでは，ロボットがハンドを閉じ始めてからハンドを停止させるまでを把持とする．物体から取得する触覚情報は，関節角度と変化量を使用する．関節角度は，ロボットハンドと物体が接触した際の関節角度と，ロボットハンドが停止した際の関節角度を使用する．また，変化量は，ロボットハンドを停止した際の関節角度と接触した際の関節角度の差を用いる．この 3 つの角度を触覚情報として使用する．この試行を 10 回行い，取得した各観測値を 3 次元の特徴ベクトルとし，k-means 法でクラスタリングすることにより BoF に変換し，触覚特徴量とする．

2.2 マルチモーダル HDP によるカテゴリ分類

本研究では，Nakamura らの提案したマルチモーダル HDP を用いて，物体のカテゴリ分類を行う．Nakamura らの使用していたマルチモーダル HDP はマルチモーダル情報が視覚・聴覚・触覚の 3 種類であったが，本研究では視覚・聴覚 2 種・触覚の 4 種類とする．図 2 は，本研究で使用するマルチモーダル HDP のグラフィカルモデルである．図 2 中の v，a^s，a^h，h は視覚情報，聴覚情報 (振る，叩く)，触覚情報の各モダリティ m を表している．x^v_{jn}，$x^{a^s}_{jn}$，$x^{a^h}_{jn}$，x^h_{jn} は j 番目の物体の各モダリティを持つ n 番目の特徴を表している．各モダリティの観測値は多項分布 θ^m_k のパラメータから生成される．また，θ^m_k はディリクレ事前分布 α^m_0 のパラメータによって生成される．Nakamura らはマルチモーダル HDP の生成過程を Chinese Restaurant Franchise(CRF)[Teh 06] により表現した．学習および物体認識については Gibbs Sampling を用いることで，テーブルの割り当て（マルチモーダル情報のトピック）t_{jn}，料理の割り当て（物体カテゴリ）k_{jt} を推定することが出来る [Nakamura 11]．詳しくは Nakamura らの文献を参考されたい．

3. 最適な行動選択

本章では，物体を認識するために視覚情報の次に何の情報を取得すべきかを決定する手法を提案する．

Nakamura らは，既観測の視覚情報 $\bar{x}^v$ から触覚情報を推定するなど，最初に取得した情報から未観測情報の推定を行う手法も提案した．視覚情報から未観測の触覚情報 x^h の推定は以下の式で求められる．

$$P(x^h|\bar{x}^v) = \sum_k P(k|\bar{x}^v)P(x^h|k) \tag{1}$$

この式では視覚情報より推定されるカテゴリ k を経由して，触覚情報を生成し，それをカテゴリ k により周辺化している．モダリティは入れ替え可能で触覚情報から視覚情報を推定することも可能である．

本研究ではこの未観測情報の推定を応用することで未知物体の識別を行う上で最適な行動選択を行わせる．視覚 (v) により物体を確認した後に，ロボットは，(1) 振って音を確認するか (a^s)，(2) 叩いて音を確認するか (s^h)，(3) 掴んで触覚で確認するか (h)，の選択を求められる．もともとのマルチモーダルカテゴリ自体が 4 つのモダリティ全ての情報からボトムアップに形成されているために，全てを得られれば最もよい結果が得られると仮定する．この場合，カテゴリ分類の視点から考えて，視覚の情報と最も重複しない，情報利得の大きなモダリティ情報を得るべきだと考えられる．

そこで，そのモダリティ情報を得た時に生じるカテゴリ認識の変化について，カルバックライブラー情報量 (KL 情報量) を尺度とし，その期待値が最も大きくなるようなモダリティ情報を得る行動を選択する．

行動選択を行うため，それぞれのモダリティ情報を取得する行動 m に対する評価式を以下に示す．

$$E_m = \int \mathrm{KL}(P(k|x^m, \bar{x}^v), P(x^m|\bar{x}^v))dx^m \tag{2}$$

$$\approx \frac{1}{I}\sum_i \sum_k P(k|x^m_i, \bar{x}^v) \log \frac{P(k|x^m_i, \bar{x}^v)}{P(k|\bar{x}^v)} \tag{3}$$

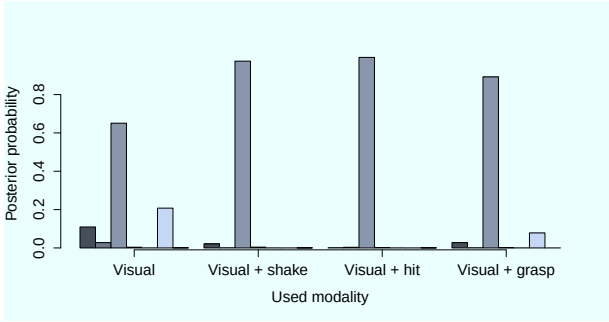
ここで x^m とは各モダリティを持った情報であり，x^v，x^{a^h}，x^{a^s}，x^h が入る．式 2 の積分を解析的に求めることは困難である．そこで，モンテカルロ法によりこれを近似する．$P(x^m|\bar{x}^v)$ から未観測情報の擬似的な情報としてサンプリング値 x^m_i を得て，それらに基づく KL 情報量の平均を求める（式 3）．各モダリティについて，その動作を行った時に得られる KL 情報量の期待値を求め，より大きな KL 情報量を得る動作を選択する．

4. 実験

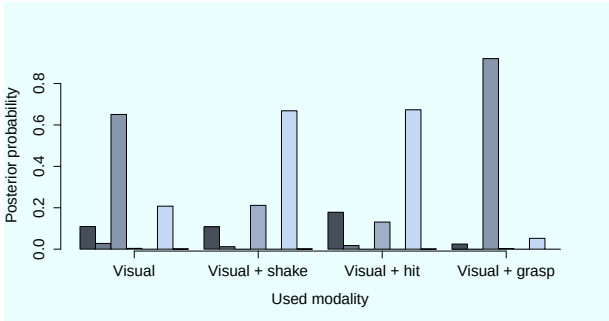
4.1 実験条件

まず，最適な行動選択を行う上で重要となるカテゴリ分類の学習データセットをマルチモーダル HDP を用いて作成する．本研究ではアールティ社製の上半身ロボット RIC-Torso を使用した (図 1)．視覚情報は頭部に搭載されている Xtion PRO LIVE を使用し画像を取得した．聴覚情報については RODE 社の NTG-2 を使用し取得した．また，ハンドにタッチエンス社のショッカクキューブを搭載し触覚情報を取得した．実験時に用いる物体は，図 3 に示す日用品 17 物体を用意した．各物体には図 3 に示すように番号が与えられている．今回の実験では視覚・聴覚 2 種・触覚の BoF はそれぞれ 25 次元，25 次元，25 次元，5 次元とした．これは，k-means 法のクラスタ数に対応する．まず，マルチモーダルカテゴリゼーションによるカテゴリ分類の結果を図 3 に示す．ボールは硬さの違い，コップ

図 3: 実験に用いた日用品 17 物体とカテゴリ分類の結果



(a) カテゴリ 3



(b) カテゴリ 6

図 4: 物体 14 の視覚情報と各カテゴリから各モダリティの擬似観測情報を生成して推定したカテゴリ確率分布．各棒は物体認識におけるカテゴリ 1〜7 の事後確率を左から順に表す．

は叩いた時の音の違い，ペットボトルは振った時の音の違いなどマルチモーダル情報が影響しカテゴリ形成されたと考えられ，視覚だけでは正しい認識が行えないデータセットになっている．式 3 のモンテカルロ法の定数 I は 10 とした．

4.2 実験結果

4.2.1 擬似観測情報と KL 情報量の期待値

図 4 は物体 14 について視覚情報のみで推定されたカテゴリ分布と，それらにカテゴリ 3,6 から，擬似的な観測情報を各モダリティ毎に生成し，その情報を付加した上で再度認識を行った際の，カテゴリ分布を示している．カテゴリ 6 を前提とした擬似観測情報を付加した場合，振った音，叩いた音ならば認識が変化するが，掴んだ触覚では変化しないことがわかる．カテゴリ 3,6 の違いはペットボトルの中に入っている鈴であり，このカテゴリ分布の違いは妥当と考えられる．次に図 5 に KL 情報量の期待値 E_m を示す．各物体によって，選ぶモダリティによって得られる KL 情報量の期待値が異なっている様子がわかる．

4.2.2 最適な行動選択の認識への影響

ロボットに視覚のみの情報で認識させた結果 (Visual) と，KL 情報量の期待値が最大化される行動 (max KL) と，KL 情報量の期待値が最小化される行動 (min KL) と，残り 3 つともものモダリティ情報を得る行動 (All) で得られる情報を付加して認識させた結果の 4 条件についていくつかの視点から比較を行った．

まず，図 6(a) に真のカテゴリの事後確率値の全物体に関する平均を示す．学習セットとテストセットは分け，学習セットに含まれないデータを再度取得し，ロボットに与えている．ここでは，Visual + max KL 条件と All 条件がほぼ変わらない事後確率を得ている事がわかる．次に，図 6(b) に正答率を示す．これについては min KL と max KL であまり差がなかった．しかしながら，提案手法では，音の違いによるペットボトルの混同については，視覚のみでは間違うところを，適切に行動選択を行った結果，全て正解することができていた．一方で，視覚のみで正解できていたコップ（プラスチック）の認識において間違うことがあった．典型的な事例を，図 7(a) と図 7(b) に示す．視覚のみの情報で正しい推定ができている際には，それを崩すような効果が本手法にはあることがわかる．

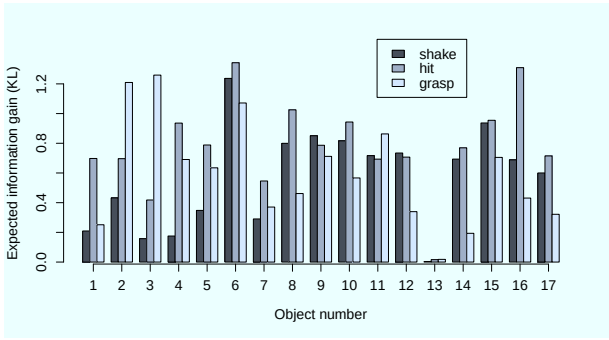
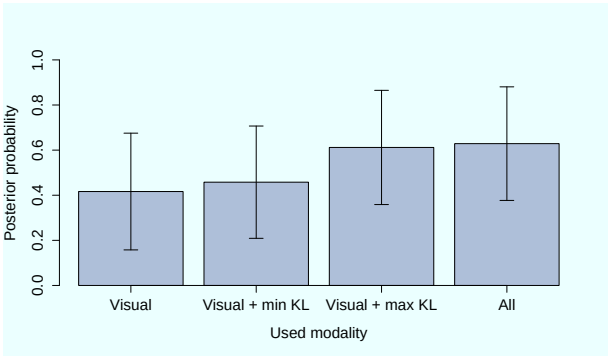
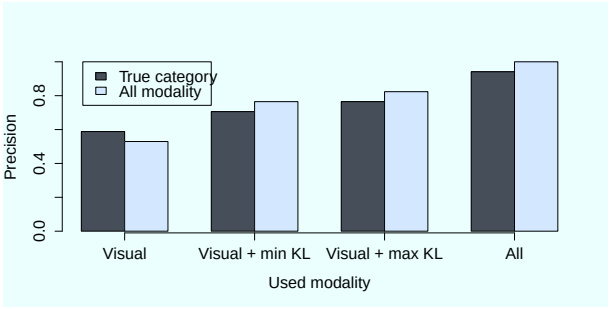


図 5: 各行動に対して得られる KL 情報量の期待値 E_m

手法の提案動機の上で，重要なのは，全てのモダリティ情報を取得した結果と同等のカテゴリ分布に早期に至ることである．その評価のため，各条件の結果えられるカテゴリ分布と全てのモダリティ情報を用いた場合のカテゴリ分布との間の対称

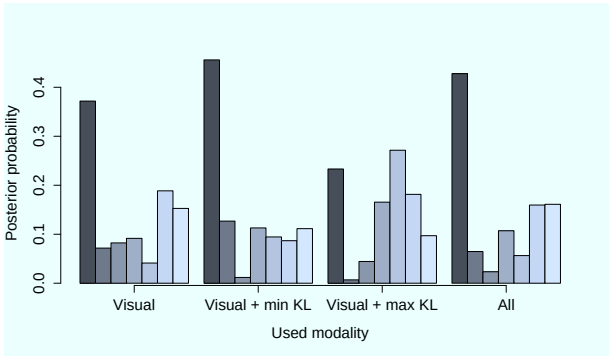


(a) 真のカテゴリの事後確率

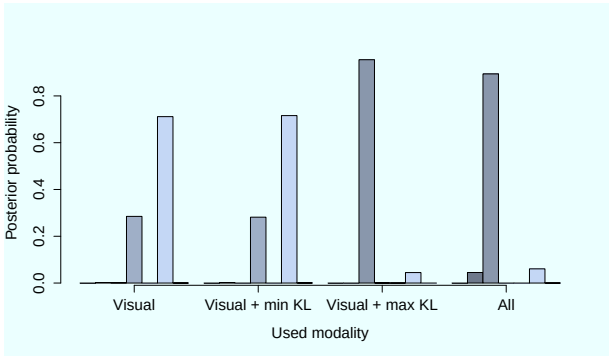


(b) 正答率

図 6: 各条件に対する（上）真のカテゴリの事後確率,（下）正答率



(a) 物体 7



(b) 物体 16

図 7: 各条件でのカテゴリ事後分布（上）物体 7,（下）物体 16

化 KL 情報量*1を計算し比較した．全物体についての平均値と標準偏差を図 8 に示す．提案手法により，迅速に全てのモダリティを得た後の認識結果に近いカテゴリ分布に至れている事がわかる．

5. まとめ

未知物体に遭遇した際に，最適な行動選択を行うことで素早くカテゴリを推定し物体を適切に扱うことを目指し，マルチモーダルカテゴリ推定のための最適な行動選択手法を提案した．KL 情報量の期待値を最大化させるような行動を取ることによって，効率的にマルチモーダルカテゴリ推定を行うことが示された．

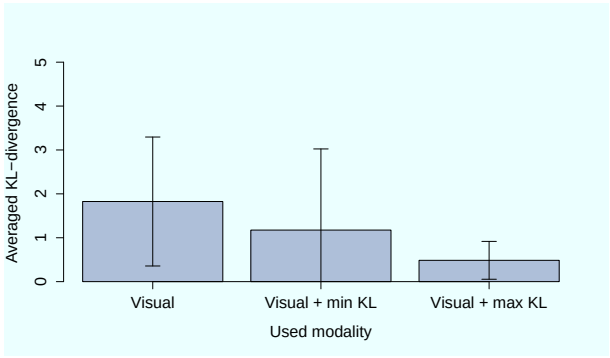


図 8: 全てのモダリティ情報を用いた場合のカテゴリ分布と各条件でのカテゴリ分布の対称化 KL 情報量

参考文献

[Nakamura 11] Nakamura Tomoaki, Nagai Takayuki, Iwahashi Naoto: “Multimodal categorization by hierarchical dirichlet process”, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1520-1525, California, Sep.2011

[Teh 06] Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: “Hierarchical Dirichlet Processes”, Journal of the American Statistical Association, Vol. 101, No. 476, pp. 1566-1581 (2006)

[Sugiura 11] K. Sugiura, N. Iwahashi, H. Kawai, S. Nakamura: “Situated Spoken Dialogue with Robots Using Active Learning”, Advanced Robotics, Vol.25, No.17, pp. 2207-2232, 2011

[Lowe 04] Lowe, D.G.: “Distinctive image features from scale-invariant keypoints”, International journal of computer vision, 60(2), pp. 91-110 (2004)

∗1 両方向の KL 情報量を計算し平均したもの