2G5-OS-19b-1

階層マルチモーダルカテゴリゼーションによる多様な概念と語意の学習

Hierarchical Multimodal Categorization for Learning of Various Kinds of Concepts and Words

安藤義記 Yoshiki Ando　中村友昭 Tomoaki Nakamura　荒木孝弥 Takaya Araki　長井隆行 Takayuki Nagai

電気通信大学大学院情報理工学研究科
Faculty of Informatics and Engineering, The University of Electro-Communications

In recent studies, it has been revealed that robots can form concepts and understand the meanings of words through inference. The key idea underlying these studies is "multimodal categorization" of a robot's experience. However, previous studies considered only nonhierarchical categorization methods, which led to nonhierarchical concept structures. Our concepts have a hierarchical structure, thus ensuring that the resulting inferences are more efficient and accurate. In this paper, we propose a novel hierarchical categorization method. The method involves extending multimodal latent Dirichlet allocation (MLDA) to hierarchical MLDA using the nested Chinese restaurant process, which makes it possible for robots to acquire concepts in a hierarchical structure. We show that a robot can form a hierarchical concept structure based on self-obtained multimodal information.

1. はじめに

事物のカテゴリ分類は，人間の認知機能において重要な役割を果たしていることが指摘されている．人間はカテゴリを形成することで，経験した物事を全て参照することなく，必要最小限の認知的処理によってより多くの情報を得ることができる [Rosch 99]．さらに，カテゴリ分類の重要性は，経験を通して形成したカテゴリを利用した予測が可能な点にある．人間は，未知の物事に対しても様々な予測を行い，柔軟に対応している．すなわちロボットにおいても，このような経験をカテゴリ分類する能力を持つことは非常に重要であると考えられる．

これまで著者らは，自然言語処理の分野で盛んに研究されてきた統計モデルの一つである latent Dirichlet allocation (LDA) [Blei 03] をベースに，物体カテゴリを教師なしで形成する手法を提案してきた [Nakamura 08, Nakamura 09]．これらの研究では，物体の視覚や聴覚，触覚などのマルチモーダル情報を LDA によりカテゴリ分類することで，ロボットがマラカスやタンバリン，ぬいぐるみといった人間の感覚に即した物体のカテゴリ（概念）を形成できることを示した．しかし，これらの研究は物体のカテゴリの形成にのみ注目しており，その階層的な関係性を捉えていないという点で，人間の概念を十分にモデル化していないと言える．実際人間が形成する各カテゴリは独立しているわけではなく，それぞれが関係し合い階層的な構造をなしていることは直観的にも明らかであろう．例えば，マラカスやタンバリンといったカテゴリはパーカッションというカテゴリの一部であり，さらには他の多くのカテゴリとともに楽器というカテゴリを形成している．このように，物体概念は階層的な構造をしており，各階層においてカテゴリのメンバは共通した特徴を有している．この共通した特徴に注目することで，物体概念だけでなく，その他の様々な概念をロボットは獲得することが可能となる．

本稿では，LDA に nested Chinese restaurant process (nCRP) を導入した hierarchical latent Dirichlet allocation (hLDA) [Blei 10] をマルチモーダル入力に拡張し，階層的なカテゴリ分類を行う．文献 [Blei 10] では，hLDA を文書の分類に適用しており，文書をトピック毎に分類し，トピックが 1 つのパスによって表現されている木構造を構成している．このモデルでは，単語は各ノードからその抽象度に応じて生成される．すなわち，上位のノードで生成される単語は複数のトピックで共有されており，より抽象的な単語となっている．文書を物体，トピックをカテゴリ，単語を物体から発生する特徴として考えることで，このモデルを物体の階層的な分類へと応用する．さらに本稿では，単一の知覚情報と選択し，単語情報と組み合わせてカテゴリ分類を行うことにより，選択した知覚情報に即したカテゴリを形成できることを示す．これにより，物体概念 (名詞) だけでなく色や硬さといった形容詞に相当する概念を獲得することが可能となる．

連絡先: 安藤 義記，電気通信大学大学院情報理工学研究科，東京都調布市調布ヶ丘 1-5-1，y4422@apple.ee.uec.ac.jp

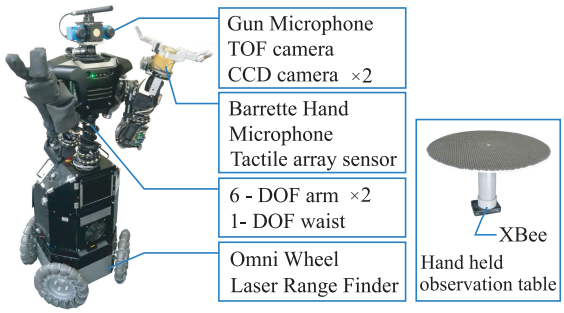

図 1: ロボットプラットフォーム

2. 提案手法

2.1 マルチモーダル情報の処理

本稿では，図 1 のロボットを用いることを想定する．ロボットは，物体を発見し自律的にマルチモーダル情報を取得する [Araki 12]．ここでは取得するマルチモーダル情報と，その処理に関して述べる．

視覚情報

まず観測した物体の画像を複数枚取得する (後述する実験では，各物体に対して 36 枚の画像を取得した)．本稿では特徴量として 128 次元の DSIFT を用い，これにより 1 枚の画像から多数の特徴ベクトルを得ることができる [Vedaldi 10]．これらの特徴ベクトルを，学習画像とは関係のない背景画像から計算した 500 の代表ベクトルを用いてベクトル量子化することで得られる 500 次元のヒストグラムを視覚情報として取り扱う．

さらに，2 つ目視覚情報として，Lab 表色系の補色次元 a 及び b の 2 次元ヒストグラムを用いた．ビンの数はそれぞれ 5

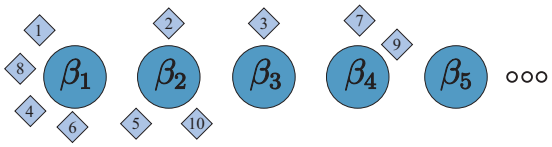

図 2: Chinese restaurant process.

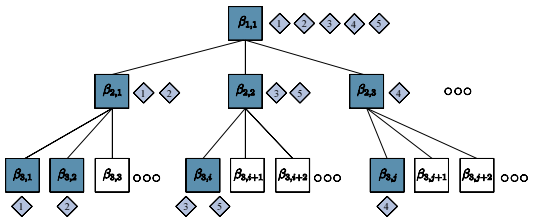

図 3: Nested CRP

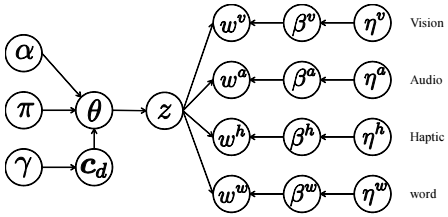

図 4: hierarchical multimodal LDA グラフィカルモデル

とし，合計 25 次元のヒストグラムとした．なお，後述する複数のモダリティを用いたカテゴリ分類実験においては SIFT による視覚情報，視覚・単語情報を用いたカテゴリ分類実験においては Lab による視覚情報を用いている．

聴覚情報

取得した音情報は 0.2[sec] 毎のフレームに分割し，フレーム毎の特徴量に変換する．特徴量としては，音声認識でよく利用されている MFCC を用い，各フレームは 13 次元の特徴ベクトルとなる．これにより，物体から発生した音から複数の特徴ベクトルを得ることができる．この特徴ベクトルを，あらかじめ計算した 50 個の代表ベクトルによりベクトル量子化を行い，各代表ベクトルの発生頻度を表すヒストグラムを聴覚情報として使用する．

触覚情報

触覚情報には，162 個のセンサから構成された触覚センサにより取得した時系列データを用いる．取得したデータは近似を行い，近似パラメータを各センサの特徴ベクトルとして扱う [中村 10]．さらに k 平均法により予め計算した 15 の代表ベクトルを用いてベクトル量子化を行い，15 次元ヒストグラムを触覚情報として用いる．

単語情報

ロボットは，物体を観察中に人から発せられた教示発話を単語情報として利用する．人の教示発話は，音声認識され，形態素解析をすることで単語へと分割される．最終的に，単語の発生頻度ヒストグラムを単語情報として用いる．

2.2 Chinese Restaurant Process [Aldous 10]

Chinese restaurant process (CRP) は，無限次元の多項分布を生成するディリクレ過程の 1 つである．CRP では，無限にテーブルが存在する中華料理店を考えることで多項分布を生成する．今，$n-1$ 人の客が既に K 個のテーブルを使用しているとすると，n 番目の客は以下の確率に従い着席するテーブル z_n を選択する．

$$P(z_n = k|\gamma) = \begin{cases} \frac{N_k}{\gamma+n-1} & (k = 1, \cdots, K) \\ \frac{\gamma}{\gamma+n-1} & (k = K+1) \end{cases} \quad (1)$$

ここで，N_k は現在 k 番目のテーブルに座っている客の人数であり，γ は客が新しいテーブルを選択する確率を制御するパラメータである．図 2 に，10 人の客がいる中華料理店の例を示す．新たな客は，既に各テーブルに座っている客の人数に応じて，テーブルを選択することとなる．

2.3 Nested CRP [Blei 10]

nCRP は CRP の拡張であり，市内に無限個のテーブルを有する中華料理店が無限に存在すると考え，次のようなプロセスにより階層的な構造を確率的に考えることができる．無限個の中華料理店のうち 1 軒を大元となる店舗とし，その中華料理店のテーブルには別の中華料理店の名前が書かれたカードが置いてあるとする．また，それら名前の書かれた中華料理店のテーブルにも同様に他の中華料理店の名前が書かれているカードが置いてあるとし，この構造は無限に繰り返される．なお，各中華料理店は 1 度しか参照されないとする．ある観光客がこの街を訪れたとし，1 日目の夜，彼は大元となる店舗を訪れ，式 (1) を用いてテーブルを選ぶ．2 日目の夜，1 日目のテーブルで指定された中華料理店を訪れ，式 (1) を用いてテーブルを選ぶ．このプロセスを L 日繰り返すことは，無限の木構造上で大元となる中華料理店から L 個の階層を持った中華料理店へのパスが構成されたことと同じ意味を持つ．図 3 に $L = 3$，観光客を 5 人とした場合のパスの例を示す．なお，各ボックスは中華料理店を表しており，hLDA において各中華料理店はデータを生成するパラメータを $\beta_{\ell,i}$ とする確率分布を持っている．

2.4 Hierarchical Multimodal Latent Dirichlet Allocation

文書のトピックモデルに nCRP を導入することで，階層的な文書分類を可能としたモデルが hLDA である．この hLDA を，マルチモーダル情報の階層的分類が可能なモデルである hMLDA へ拡張する．提案する hMLDA のグラフィカルモデルを，Fig.4 に示す．この図において，$\boldsymbol{c}_d$ は γ をパラメータとする nCRP によって生成される木構造上のパスである．さらに，z は物体のカテゴリであり，α と π をパラメータとする Stick breaking process により生成される．w^v, w^a, w^h, w^w は，それぞれ視覚・聴覚・触覚・単語情報であり，β^* をパラメータとする多項分布から生成される．また，β^* は，η^* をパラメータとするディリクレ事前分布によって決定される．

2.5 階層的カテゴリ分類

物体の分類は，図 4 のモデルのパラメータを物体の情報を用いて学習することに相当する．本論文では，モデルのパラメータ学習に Gibbs Sampling を用いる．hMLDA では，物体ごとのパス $\mathbf{c}_d$ と，これらパス内の物体 d のモダリティ m の n 番目の特徴量へ割り当てられるカテゴリ $z^m_{d,n}$ を事後分布からサンプリングすることで，パラメータの推定を行う．

2.5.1 カテゴリのサンプリング

全物体のパスが割り当てられているとして，d 番目の物体の m 番目のモダリティの n 番目の特徴量 $w^m_{d,n}$ に割り当てられるカテゴリ $z^m_{d,n}$ を以下の式からサンプリングを行う．

$$p(z^m_{d,n}|\mathbf{z}^m_{-(d,n)}, \mathbf{c}, \mathbf{w}^m, \alpha, \pi, \eta^m) \propto p(z^m_{d,n}|\mathbf{z}^m_{d,-n}, \alpha, \pi) p(w^m_{d,n}|\mathbf{z}, \mathbf{c}, \mathbf{w}^m_{-(d,n)}, \eta^m) \quad (2)$$

ここで $\mathbf{c}$ と $\mathbf{w}^m$ は，それぞれ全物体に割り当てられたパスの集合と，物体の特徴量の集合を表している．また，負の添字は，その情報を含まないことを意味し，$\mathbf{w}^m_{-(d,n)}$ と $\mathbf{z}^m_{-(d,n)}$ はそれ

ぞれ，$w^m_{d,n}$ と $z^m_{d,n}$ を除いた全物体の特徴量と，特徴量に割り当てられたカテゴリの集合を表している．また，$\mathbf{z}^m_{d,-n}$ は，物体 d のモダリティ m の特徴に割り当てられたカテゴリの集合 $\mathbf{z}^m_d$ から，n 番目の特徴量に割り当てられたカテゴリ $z^m_{d,n}$ を除いた残りである．

式 (2) の 1 つ目の項は，Stick breaking process によって生成される多項分布である．d 番目の物体のモダリティ m の n 番目の特徴量のカテゴリが k となる確率は以下のようになる．

$$
\begin{aligned}
&p(z^m_{d,n}=k|\mathbf{z}^m_{d,-n},\alpha,\pi)\\
&=E\left[V_k\prod_{j=1}^{k-1}(1-V_j)|\mathbf{z}^m_{d,-n}\alpha,\pi\right]\\
&=E\left[V_k|\mathbf{z}^m_{d,-n},\alpha,\pi\right]\prod_{j=1}^{k-1}E\left[1-V_j|\mathbf{z}^m_{d,-n},\alpha,\pi\right]\\
&=\frac{(1-\alpha)\pi+\#[\mathbf{z}^m_{d,-n}=k]}{\pi+\#[\mathbf{z}^m_{d,-n}\geq k]}\prod_{j=1}^{k-1}\frac{\alpha\pi+\#[\mathbf{z}^m_{d,-n}>j]}{\pi+\#[\mathbf{z}^m_{d,-n}\geq j]} \qquad (3)
\end{aligned}
$$

ここで，$\#[\cdot]$ は与えられた条件を満たす配列の要素数である．また，V_j は，Stick breaking process において枝を折る割合を決めるパラメータである．

式 (2) の 2 つ目の項は，パス $\mathbf{c}_d$，カテゴリ $z^m_{d,n}$ から特徴量が生成される確率である．特徴量を生成する多項分布のパラメータが，ハイパーパラメータを η^m とするディリクレ分布から生成されるという仮定から，以下の式を得ることができる．

$$
\begin{aligned}
&p(w^m_{d,n}|\mathbf{z},\mathbf{c},\mathbf{w}^m_{-(d,n)},\eta^m)\propto\\
&\qquad\#[\mathbf{z}^m_{-(d,n)}=z^m_{d,n},\mathbf{c}^m_{z^m_{d,n}}=c_{d,z^m_{d,n}},\mathbf{w}^m_{-(d,n)}=w^m_{d,n}]+\eta^m \qquad (4)
\end{aligned}
$$

これはパス $\mathbf{c}_d$ において，特徴量 $w^m_{d,n}$ にカテゴリ $z^m_{d,n}$ が割り当てられた回数を表す．

2.5.2 パスのサンプリング

次に，物体の全ての特徴量に対してカテゴリが割り当てられているとして，次式よりパスのサンプリングを行う．

$$
\begin{aligned}
&p(\mathbf{c}_d|\mathbf{w}^v,\mathbf{w}^a,\mathbf{w}^h,\mathbf{w}^w,\mathbf{c}_{-d},\mathbf{z},\eta^v,\eta^a,\eta^h,\eta^w,\gamma)\\
&\quad\propto p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)\\
&\qquad\times p(\mathbf{w}^v_d|\mathbf{c},\mathbf{w}^v_{-d},\mathbf{z}^v,\eta^v)p(\mathbf{w}^a_d|\mathbf{c},\mathbf{w}^a_{-d},\mathbf{z}^a,\eta^a)\\
&\qquad\times p(\mathbf{w}^h_d|\mathbf{c},\mathbf{w}^h_{-d},\mathbf{z}^h,\eta^h)p(\mathbf{w}^w_d|\mathbf{c},\mathbf{w}^w_{-d},\mathbf{z}^w,\eta^w) \qquad (5)
\end{aligned}
$$

$\mathbf{c}_{-d}$ は $\mathbf{c}$ から $\mathbf{c}_d$ を除いた残りを表している．$p(\mathbf{w}^m_d|\mathbf{c},\mathbf{w}^m_{-d},\mathbf{z},\eta^m)$ は特定のパスからモダリティ m の特徴量が生成される確率であり，$p(\mathbf{c}_d|\mathbf{c}_{-d},\gamma)$ は nCRP によって作られる事前確率である．各モダリティ m において，特徴量が生成される確率は，多項分布のパラメータを周辺化することで以下のようになる．

$$
\begin{aligned}
&p(\mathbf{w}^m_d|\mathbf{c},\mathbf{w}^m_{-d},\mathbf{z}^m,\eta^m)\\
&=\prod_{\ell=1}\frac{\Gamma(\sum_w\#[\mathbf{z}^m_{-d}=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}^m_{-d}=w]+V^m\eta^m)}{\prod_w\Gamma(\#[\mathbf{z}^m_{-d}=\ell,\mathbf{c}_{-d,\ell}=c_{d,\ell},\mathbf{w}^m_{-d}=w]+\eta^m)}\\
&\qquad\times\frac{\prod_w\Gamma(\#[\mathbf{z}^m=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}^m=w]+\eta^m)}{\Gamma(\sum_w\#[\mathbf{z}^m=\ell,\mathbf{c}_\ell=c_{d,\ell},\mathbf{w}^m=w]+V^m\eta^m)} \qquad (6)
\end{aligned}
$$

2.5.3 ギブスサンプリングによる学習

$\mathbf{c}_1$〜$\mathbf{c}_D$ 及び $\mathbf{z}_1$〜$\mathbf{z}_D$ に対してランダムな初期値が与えられているとして，以下の手順を収束するまで繰り返し計算することにより，パラメータを推定することができる．

1. 各物体 $d\in\{1,\dots D\}$ に対して以下の処理を繰り返す

(a) パスをサンプリング

$$\mathbf{c}_d\sim p(\mathbf{c}_d|\mathbf{w}^v,\mathbf{w}^a,\mathbf{w}^h,\mathbf{w}^w,\mathbf{c}_{-d},\mathbf{z},\eta^v,\eta^a,\eta^h,\eta^w,\gamma) \qquad (7)$$

(b) 物体の各モダリティ m の n 番目の特徴量に対してカテゴリをサンプリング

$$z^m_{d,n}\sim p(z^m_{d,n}|\mathbf{z}^m_{-(d,n)},\mathbf{c},\mathbf{w}^m,\alpha,\pi,\eta^m) \qquad (8)$$

最終的に，このアルゴリズムを繰り返すことで，全物体のパスとカテゴリは，それぞれ $\hat{\mathbf{c}}$, $\hat{\mathbf{z}}$ へと収束する．

図 5: 実験に使用した 67 物体

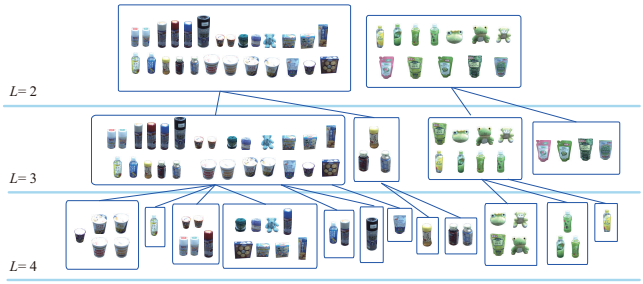

図 7: 視覚 (Lab)・単語情報を用いたカテゴリ分類

3. 実験

図 1 に示すロボットにより，取得した視覚 (SIFT)・視覚 (Lab)・聴覚・触覚・単語情報を用いて実験を行った．単語情報として，5 人の被験者が物体の特徴を音声にて教示し，その音声認識結果を用いた．なお、実験には図 5 に示す 67 個の物体を使用し，カテゴリ分類実験を行った．

まず，視覚 (SIFT)・触覚・聴覚・単語情報を用いて 67 個の物体を hMLDA により階層的カテゴリ分類を行った．その結果を図 6 に示す．図 6 の $L=2$ の階層において，カテゴリ 1 はスプレー缶のみで構成されており，その下の $L=3$ の階層ではスプレー缶の大きさごとにカテゴリが形成された．同様に，$L=2$ の階層のカテゴリ 5 も $L=2$ の階層においてはカップ麺のみのカテゴリが形成され，その下の階層では種類ごとに分類できている．また，$L=2$ の階層において，カテゴリ 2 はペットボトル・ビン・シャンプー・フローリングワイパー・クッキーから構成されており，1 つ下の $L=3$ の階層ではフローリングワイパー及びクッキーが正しく個々のカテゴリとして形成された．さらに，$L=4$ の階層ではペットボトル・シャンプー・ビンがそれぞれ正しく分類できた．$L=2$ の階層において，カテゴリ 4 はぬいぐるみ・ガラガラから形成されており，$L=3$ の階層ではぬいぐるみとガラガラが正しく分類できた．以上のように，カテゴリ 5 のようにスナックとクッ

図 6: 視覚 (SIFT)・聴覚・触覚・単語情報を用いたカテゴリ分類

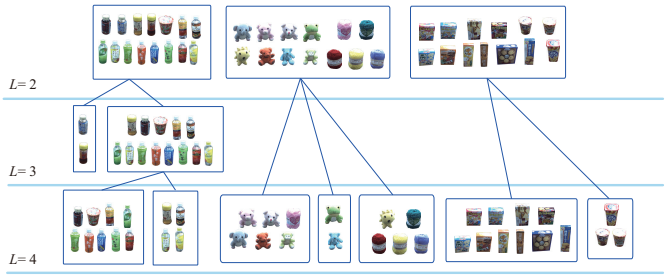

図 8: 聴覚・単語情報を用いたカテゴリ分類

キーが混ざったカテゴリなど一部誤ったカテゴリが存在してはいるが，hMLDA を使用することで，ロボットは視覚・聴覚・触覚・単語情報の類似性を手掛かりに，広範なカテゴリから具体的なカテゴリまでを教師なしで自動的に形成することができたと言える．

次に，視覚 (Lab)・単語情報のみを用いて，同様に hMLDA により階層的カテゴリ分類を行った．結果を図 7 に示す．視覚 (SIFT)・触覚・聴覚・単語情報を用いて分類した場合，シャンプー・ガラガラ・ぬいぐるみ・ペットボトルが 1 つのカテゴリに分類されることはなかったが，図 7 の $L = 2$ の階層において，それら緑色をした物体は 1 つのカテゴリとして分類されている．また $L = 4$ の階層において，青色をしたぬいぐるみ・毛糸・クッキー・スプレー缶が 1 つのカテゴリとして分類されている．これより，視覚 (Lab)・単語情報を用いた分類によって色に依存した形容詞的な概念が形成されることが分かる．

最後に，聴覚・単語情報のみを用いて，hMLDA により階層的カテゴリ分類を行った．結果を図 8 に示す．図 8 の $L = 2$ の階層において，ぬいぐるみ・毛糸が 1 つのカテゴリとして分類されている．これは，ぬいぐるみ・毛糸がともに音が鳴らないため，1 つのカテゴリとして分類されたと考えられる．また，スナック・クッキーが 1 つのカテゴリとして分類されている．これはスナック・クッキーはともに箱の中に固形の物体が入っており，振った際に似た音が鳴るためである．さらに，中に液体の入っているペットボトル・ビンが 1 つのカテゴリとして分類されている．以上のように，聴覚・単語情報を用いてカテゴリ分類を行うことにより，音の種類ごとにカテゴリが形成できることが分かった．

4. まとめ

本稿では，ロボットが取得した視覚・聴覚・触覚・単語情報を用い，hMLDA により階層的なカテゴリ分類を行う手法を提案した．これにより，抽象的なカテゴリから物体カテゴリまで階層的なカテゴリ分類が可能となることを実験を通して明らかにした．また，ある特定のモダリティのみに注目することで，「赤い」などの形容詞に関する概念の形成が可能であることを示した．

今後さらに実験を進めることで，モダリティを選択した各モデルの各階層にどのような単語が結びついているのを明らかにすると共に，各モデルの予測性能を評価する必要がある．また，モダリティの選択は，各モダリティに対する重みの決定問題として一般化することができるため，重みを単語や予測性能といった規範で自動的に決定する枠組みを検討したいと考えている．学習のオンライン化も今後の重要な課題である．

参考文献

[Rosch 99] Rosch,E.: “Principles of categorization,” Concepts: core readings, pp.189–206, 1999.

[Blei 03] Blei,D.M. *et al.*: “Latent dirichlet allocation,” Journal of Machine Learning Research, vol.3, pp.993–1022, 2003.

[Araki 12] Araki,T.*et al.*: “Online object categorization using multimodal information autonomously acquired by a mobile robot,” Advanced Robotics, Vol.26, Issue 17, pp.1995–2020, 2012.

[Nakamura 08] 中村ほか: “ロボットによる物体のマルチモーダルカテゴリゼーション,” 電子情報通信学会論文誌 D, vol.91, pp.2507–2518, 2008.

[Nakamura 09] Nakamura,T.*et al.*: “Grounding of word meanings in multimodal concepts using LDA,” in Proc. of IROS, pp.3943–3948, 2009.

[Blei 10] Blei, D.*et al.*: “The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies,” Journal of the ACM, vol.57, no.2, p.7, 2010.

[Aldous 10] Aldous, D. : “Exchangeability and related topics,” École d’Été de Probabilités de Saint-Flour XIII-1983, pp.1–198, 1985.

[Vedaldi 10] Vedaldi, A. *et al.*: “VLFeat: An open and portable library of computer vision algorithms,” ACM International Conference on Multimedia, pp.1469–1472, 2010.

[中村 10] 中村ほか: “把持動作による物体カテゴリの形成と認識”, 情報処理学会全国大会 2010, 5V-3, 2010