2G5-OS-19b-7

マルチモーダルLDAと教師なし形態素解析による認識誤りを含む文章からの概念・語意獲得

Multimodal LDA and Unsupervised Word Segmentation for Learning Word Meanings Using Speech with Phoneme Recognition Errors

中村 友昭 *1 Tomoaki NAKAMURA
荒木 孝弥 *1 Takaya ARAKI
長井 隆行 *1 Takayuki NAGAI
長坂 翔吾 *2 Syogo NAGASAKA
谷口 忠大 *2 Tadahiro TANIGUCHI
岩橋 直人 *3 Naoto IWAHASHI

*1電気通信大学　知能機械工学専攻
Dept. of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications

*2立命館大学　情報理工学部
Dept. of Human and Computer Intelligence, Ritsumeikan University

*3(独) 情報通信研究機構
National Institute of Information and Communications Technology

In this study, we propose a method for concept formation and word acquisition for robots. The proposed method is based on multimodal latent Dirichlet allocation (MLDA) and the nested Pitman-Yor language model (NPYLM). A robot obtains haptic, visual, and auditory information by grasping, observing, and shaking an object. At the same time, a user teaches object features to the robot through speech, which is recognized using only acoustic models and transformed into phoneme sequences. As the robot is supposed to have no language model in advance, the recognized phoneme sequences include many phoneme recognition errors. Moreover, the recognized phoneme sequences with errors are segmented into words in an unsupervised manner; however, not all words are necessarily segmented correctly. The words including these errors have a negative effect on the learning of word meanings. To overcome this problem, we propose a method to improve unsupervised word segmentation and to reduce phoneme recognition errors by using multimodal object concepts. In the proposed method, object concepts are used to enhance the accuracy of word segmentation, reduce phoneme recognition errors, and correct words so as to improve the categorization accuracy. We experimentally demonstrate that the proposed method can improve the accuracy of word segmentation and reduce the phoneme recognition error and that the obtained words enhance the categorization accuracy.

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．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．さらに，このようなカテゴリが概念を形成しており，概念と単語が結びつくことで，我々は単語の意味を理解することができる．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．

そこで著者らは，これまで LDA (Latent Dirichlet Allocation)[Blei 03] を拡張したマルチモーダルカテゴリゼーションを提案し，複数のモダリティを用いることにより，より人間の感覚に近いカテゴリを形成することが可能となることを示し，さらに単語と形成されたカテゴリを確率的に結びつけることで語意の理解が可能となることを示した [Araki 11]．提案手法は確率モデルに基づいており，学習したグラフィカルモデルを用いることで，未学習物体のカテゴリ認識が可能である．さらに，学習したモデルを用いた未観測の情報である単語の推定を可能とした．すなわち，ロボットは得られたセンサー情報を単語で表現することが可能となった．

しかし，これらの研究では，語彙はあらかじめ持っているものとし，人の発話は音声認識により認識し，また単語を切り出す際には形態素解析器を用いてきた．すなわち，語彙に含まれない単語には全く対応できないといった問題があった．人は言語を獲得する過程において，音素列を単語に分節化することで語彙を構築している．このような教師なしで音素列を単語へ分節化する能力は，柔軟な言語獲得において非常に重要である．そこで，さらに著者らは人の発話を音素認識器で認識し，

連絡先: 中村友昭，電気通信大学，東京都調布市調布ケ丘 1-5-1，042-443-5203，naka_t@apple.ee.uec.ac.jp

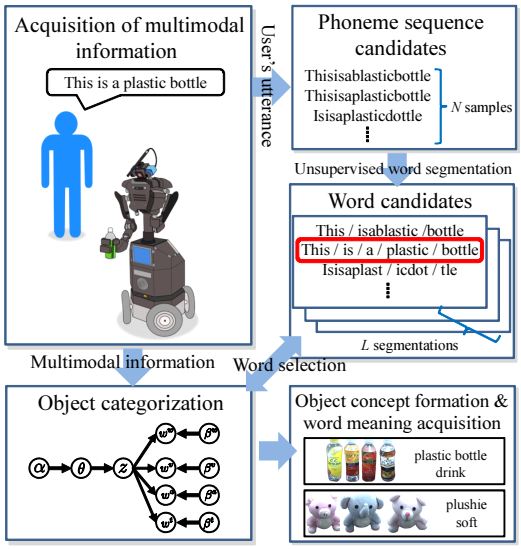

図 1: Overview of proposed method

Nested Pitman-Yor Language Model(NPYLM)[持橋 09] を用い，教師なしで音素列の分節化を行い，切り出された単語を概念と結びつけることで，その語意の学習を行った [Araki 12]．しかし，言語モデルを持たない音素認識器では正しく認識することは困難である．さらに NPYML において正しい単語の分節化を行うためには音素誤りのない大量の学習文章が必要となる．今回のようなロボットと人が物体に関する対話を行うシナリオでは十分な文章を得ることは困難となる．

そこで，本稿では物体概念を用いることで，音素認識とその分節化の精度を向上する手法を提案する．図 1 が提案手法の概

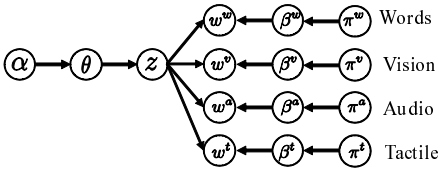

図 2: Graphical model of multimodal LDA

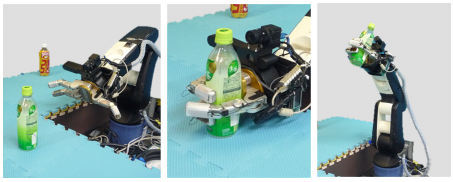

図 3: (a) Capture of visual information, (b) Capture of haptic information, and (c) Capture of auditory information

要である．まず，人の発話は音素認識され，N-best の N 個の認識結果を音素列の候補として利用する．さらに，このように取得した N 個の音素列それぞれを NPYLM により単語へと分割を行う．NPYLM はサンプリングによるパラメータ推定をしているため，初期値によって異なる結果となる．そこで，各音素列に対して L 回 NPYLM を適用することで，$N \times L$ 個の単語の候補を計算する．最終的に，マルチモーダル情報によって形成された物体概念を表現する単語を，候補の中から選択する．ここで重要な事は，単語はロボットが形成した概念を意味しており，単語と概念が独立しているのではなく相互に関係している点である．すなわち，同じカテゴリに含まれる物体には，同一の単語が与えられる可能性が高く，また逆に同じ単語が与えられた物体は，共通する特長を有している可能性が高いと言える．ロボットは概念形成と教示発話の分節化を行う際に，このような情報を利用することにより分節化の精度と，分類精度の両方を高める事が可能となる．

関連研究として，視覚情報を用いた物体カテゴリの学習 [Sivic 05, Fergus 03, Fei-Fei 05, Wang 09] や，ロボットが物体に触れた際の音を用いた研究等が行なわれている [Sinapov 11]．しかし，人間がカテゴリ分類する際には単一のモダリティだけではなく，複数のモダリティを用いていると思われるため，より人間の感覚に近い分類を実現するためにはマルチモーダルな情報が必要である．

2. マルチモーダルカテゴリゼーション

ロボットは実際に物体を観察して得られるマルチモーダル情報をカテゴリ分類することで，概念の形成を行う．図 2 がマルチモーダル LDA のグラフィカルモデルである．w^v，w^a，w^t，w^w はそれぞれ視覚情報，聴覚情報，触覚情報，単語情報であり，z が物体カテゴリである．また，β^*，θ は，多項分布のパラメータであり，それぞれ π^*，α をパラメータとするディリクレ事前分布によって生成される．物体の分類は，これらのモデルのパラメータを学習データから推定することに相当する．

2.1 マルチモーダル情報

マルチモーダル情報は，図 3 に示したロボットにより取得した．ここでは，ロボットが取得するマルチモーダル情報（視覚・聴覚・触覚・単語情報）の詳細について述べる．

視覚情報　ロボット (図 3) はアームの先に CCD カメラと TOF カメラを搭載しており，観察することで得られる画像を視覚情報として利用する (図 3(a))．各画像から抽出する特徴量として，Dense Scale Invariant Feature Transform (DSIFT) [Vedaldi 10] を用いる．最終的に，これらの特徴ベクトルは，500 の代表ベクトルによりベクトル量子化することで，500 次元のヒストグラムとする．

触覚情報　触覚情報の取得には，アームに取り付けられたバレットハンドと，そのハンドに取り付けられた触覚アレイセンサーを用いる．図 3(b) のように，ロボットが実際に物体を把持することで得られるセンサーの時系列データの近似を行い，その近似パラメータを各センサーの特徴ベクトルとして扱う [中村 10]．さらに，この特徴ベクトルをベクトル量子化することで，15 次元のヒストグラムを触覚情報として用いる．

聴覚情報　図 3(c) のように，ロボットが物体を把持し，振ることで発生する音をロボットのハンドに取り付けられたマイクにより取得し，聴覚情報として利用する．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，フレーム毎に 13 次元の MFCC(Mel-Frequency Cepstrum Coefficient) を計算する．これにより，各フレームは 13 次元の特徴ベクトルとなる．最終的にこの特徴ベクトルも，ベクトル量子化を行い，50 次元のヒストグラムとする．

単語情報　ロボットが物体を観察している間に，ユーザーが各物体の特徴を音声にて教示する．ロボットは認識された音素列を，教師なしで形態素解析を行い単語へと分割する．最終的に，単語の出現頻度を表すヒストグラムを，単語情報として用いる．

2.2 物体概念の学習

物体の分類は，図 2 のグラフィカルモデルのパラメータを，ロボットが取得したマルチモーダル情報を用いて学習することに相当する．パラメータの学習にはギブスサンプリングを用いた．ギブスサンプリングでは，j 番目の物体のモダリティ m の情報の i 番目に割り当てられるカテゴリ z_{mij} は，θ，β^m を周辺化した条件付確率

$$p(z_{mij} = k|\boldsymbol{z}^{-mij}, \boldsymbol{w}^m, \alpha, \pi^m) \propto (N_{kj}^{-mij} + \alpha)\frac{N_{mw^mk}^{-mij} + \pi^m}{N_{mk}^{-mij} + W^m\pi^m} \quad (1)$$

からサンプリングされる．ただし，W^m はモーダル情報の次元数である．N_{mw^mkj} は，j 番目の物体のモダリティ m の情報が w^m となり，かつカテゴリ k が割り当てられた回数を表している．ただし，$N_{mw^mk} = \sum_j N_{mw^mkj}$，$N_{kj} = \sum_{m,w^m} N_{mw^mkj}$，$N_{mk} = \sum_{w^m,j} N_{mw^mkj}$ となる．ギブスサンプリングでは，各物体 j のモダリティ m の i 番目の情報へのカテゴリの割り当てを，式 (1) に従いサンプリングを繰り返すことでパラメータの推定を行う．

2.3 未観測モダリティの予測

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，物体の硬さ，それが音を出すかどうか，またどのような音を出すか，さらにどのような単語で表現可能かなどの情報を推測することができる．例として，視覚情報から単語情報を推定する場合を考える．物体の視覚情報 $\boldsymbol{w}_{obs}^v$ から，ある単語情報 w^w が発生する確率を次のように書くことができる．

$$p(w^w|\boldsymbol{w}_{obs}^v) = \int \sum_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}_{obs}^v)d\theta \quad (2)$$

3. 教師なし形態素解析

これまで我々は，語彙は既知であるとし，形態素解析器を用いてきた [Nakamura 09]．しかし，これでは形態素解析器内の語彙に含まれていない語には全く対応できないといった問題が存在する．そこで，本稿では教師なしで形態素解析が可能なモデル NPYLM を用いることで，この問題を解決する．

3.1 Hierarchical Pitman-Yor Language Model

Hierarchical Pitman-Yor Language Model(HPYLM) は，階層 Pitman-Yor 過程を用いた，n-gram 言語モデルである．HPYLM では，文脈 h の後に単語 w が続く確率は以下のよう

になる．

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + \sum_w c(w|h)} + \frac{\theta + d \cdot \sum_w t_{hw}}{\theta + \sum_w c(w|h)} p(w|h') \quad (3)$$

ただし，h' は $(n-1)$-gram の文脈である．よって，$p(w|h')$ は h より一つ短い文脈での単語 w が続く確率であり，再帰的に計算される．また，$c(w|h)$ は文脈 h での単語 w の発生回数であり，t_{hw} は，$c(w|h)$ のうち，文脈 h' から w が発生した回数である．d と θ は，Pitman-Yor 過程のハイパーパラメータであり，ギブスサンプリングを用いてデータから推定される．

3.2 Nested Pitman-Yor Language Model

前節の HPYML では，単語ユニグラムの場合，式 (3) の $p(w|h')$ は辞書が与えられていれば，語彙数の逆数を設定すればよい．しかし，ここでは辞書はあらかじめ用意されていないため，教示発話内の全ての部分文字列が単語となる可能性があるため，計算することが困難である．そこで，単語ユニグラムの基底測度として，文字 HPYLM を使用する．これは，単語 HPYLM の基底測度に文字 HPYLM が埋め込まれているモデルであるため，Nested Pitman-Yor Language Model(NPYLM) と呼ばれている．この NPYLM では，ブロック化ギブスサンプリングと動的計画法により高速に単語の分割が可能である．

4. 語意獲得

NPYLM により，認識した音声を教師なしで形態素解析し，単語に分割することが可能となった．認識誤りのない十分な学習用の文章を得ることができれば，高い精度での単語の分割が可能となる．しかし，言語モデルを持たない音素認識器では，誤りのない単語を得ることは困難である．また，人がロボットに物体に関する特徴を表す語を教示する際に，学習に十分な文章を与えることは困難であるといえる．さらに，NPYLM の学習はサンプリングをベースとしているため，学習毎に単語の切り出しの結果が変わってしまうといった問題も存在する．そこで，音素認識により音素列の候補を，教師なし形態素解析により単語の候補を複数個計算し，マルチモーダルカテゴリゼーションによって，物体概念と確率的に結びつきの強い単語の選択を行うことで，この問題を解決する．

1. まず，j 番目の物体に対する i 番目のユーザーの教示発話は音素認識により認識され，N-best の音素列 $p_{jin}(1 \leq n \leq N)$ を得る．

2. 音素列 p_{jin} に対して，初期値を変え NPYLM を L 回適用することで単語分割を行い，L 個の単語の発生頻度ヒストグラム $\bar{\boldsymbol{w}}^w_{jinl}(1 \leq l \leq L)$ を計算する．すなわち，j 番目の物体に対する i 番目のユーザーの教示発話から $N \times L$ 個の単語ヒストグラムを得ることができる．ここで，単語ヒストグラムの候補の集合を $\bar{\boldsymbol{W}}^w = \{\bar{\boldsymbol{w}}^w_{jinl} | 1 \leq j \leq J, 1 \leq i \leq I_j, 1 \leq n \leq N, 1 \leq l \leq L\}$ とする．ただし，I_j は j 番目の物体に与えられた教示発話の総数である．

3. 以下，全物体 $j(=0, \cdots, J)$ に関して繰り返す．

   (i) j 番目の物体の情報を除いたマルチモーダル情報 $\boldsymbol{W}^v_{-j}, \boldsymbol{W}^a_{-j}, \boldsymbol{W}^t_{-j}, \bar{\boldsymbol{W}}^w_{-j}$ から物体概念の形成を行う．ただし，負の添字は，$\boldsymbol{W}^v, \boldsymbol{W}^a, \boldsymbol{W}^t, \bar{\boldsymbol{W}}^w$ から j 番目の物体の情報を除いた残りを表している．

   (ii) 単語ヒストグラムの候補 $\bar{\boldsymbol{w}}^w_{jinl}(1 \leq n \leq N, 1 \leq l \leq L)$ の中から，j 番目の物体に対する i 番目の発話の単語ヒストグラムとして，最も物体を表現しているものを選択する．

$$\boldsymbol{w}^w_{ji} = \underset{\boldsymbol{w}^w \in \bar{\boldsymbol{W}}_{ji}}{\operatorname{argmax}} \, p(\boldsymbol{w}^w | \boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j) \quad (4)$$

図 4: Objects used in the experiment

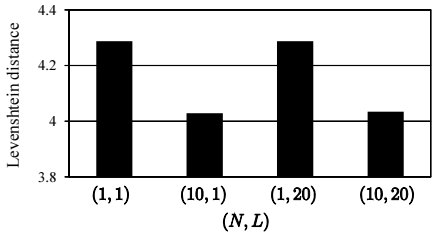

図 5: Average of Levenshtein distance between selected phoneme sequences and correct phoneme sequences

この式において，$p(\boldsymbol{w}^w | \boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j)$ は，j 番目の物体の情報 $\boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j$ から，単語 $\boldsymbol{w}^w$ が発生する確率を表しており，(i) において学習されたモデルと，式 (2) を用いて計算される．また，$\bar{\boldsymbol{W}}_{ji}$ は，j 番目の物体に対して与えられた i 番目の発話から計算された単語ヒストグラムの候補の集合である．

(iii) j 番目の物体に与えられたすべての文章の単語ヒストグラム $\boldsymbol{w}^w_{ji}$ の和をとることで，j 番目の物体の単語ヒストグラム $\boldsymbol{w}^w_j$ とする．

$$\boldsymbol{w}^w_j = \sum_i^{I_j} \boldsymbol{w}^w_{ji} \quad (5)$$

4. 最終的に，選択された単語ヒストグラム $\boldsymbol{W}^w = \{\boldsymbol{w}^w_1, \boldsymbol{w}^w_2, \cdots, \boldsymbol{w}^w_J\}$ と，物体から得られたマルチモーダル情報 $\boldsymbol{W}^v, \boldsymbol{W}^a, \boldsymbol{W}^t$ を用いて，MLDA により物体概念を形成する．

以上のように，音素認識・教師なし形態素解析の結果と，マルチモーダルカテゴリゼーションによる単語情報の発生確率を統合することにより，物体の特徴を表す単語を正しく獲得することが可能となる．

5. 実験

提案手法の有効性を検証するための実験を行った．実験では，図 4 に示したペットボトルやぬいぐるみなど 10 カテゴリ 50 個の物体を使用した．また，ユーザーが各物体に関する特徴をロボットへ音声により教示し，その発話を音素認識により音素列へと変換した．

5.1 単語の選択

まず，実際にロボットが取得したマルチモーダル情報と，ユーザーからの教示発話を用いて，提案手法により単語の選択を行った．この際，音素認識の候補として 1-bset の認識結果のみを用いる場合 $(N = 1)$ と，10-bset の認識結果を用いる場

表 1: The number of words in the selected phoneme sequences

(N, L)	$(1, 1)$	$(10, 1)$	$(1, 20)$	$(10, 20)$
# of words	255	204	238	208

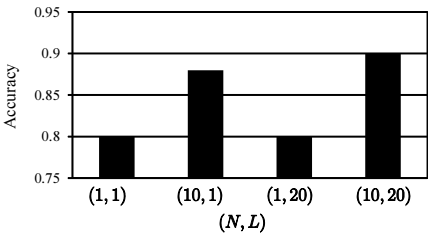

図 6: Classification accuracy by using a word histogram selected from (a)$(N, L) = (1, 1)$, (b)$(N, L) = (10, 1)$, (c)$(N, L) = (1, 20)$, and (d)$(N, L) = (10, 20)$

合 ($N = 10$) の 2 パターン行った．また，それぞれの認識候補に対して，NPYLM による音素列の分節化を 1 回のみサンプリングし 1 候補のみ用いる場合 ($L = 1$) と，初期値を変え 20 回サンプリングし 20 個の候補を用いる場合 ($L = 20$) の 2 パターン行った．すなわち，単語の候補として $(N, L) = (1, 1), (1, 20), (10, 1), (10, 20)$ の 4 パターンから単語の選択を行った場合を比較した．なお，$(N, L) = (1, 1)$ の場合が，文献 [Araki 12] で提案された手法となる．

まず，それぞれの場合で選択された音素列がどの程度正解の音素列に近いかを評価した．ここで，選択された音素列と正解との違いをレーベンシュタイン距離で評価した．これは，2 つの文字列がどの程度異なっているかを示す数値であり，文字の挿入や削除，置換によって，一つの文字列を別の文字列に変形するのに必要な手順の最小回数として与えられる．図 5 が提案手法により選択された音素列と正解となる音素列とのレーベンシュタイン距離の全教示発話の平均である．この図より，$N = 1$ の場合，すなわち音素認識の結果の最尤の物を用いる場合に比べ，10-best を用いた $N = 10$ の場合の平均距離は短くなっており，正解に近い音素列を候補の中から選択できていることが分かる．さらに，それぞれの場合において NPYLM によって切り出された単語数が表 1 である．この表より，選択候補が 1 つの場合 ($(N, L) = (1, 1)$) では，単語数が最も多くなっている．これは，同じ単語であっても，音素の誤認識や単語の切り出し位置が異なることで，違う単語となってしまっているためであり，例えば「ぬいぐるみ」の一部が誤認識され「ういぐるみ」となった場合は，異なる単語として扱われてしまうためである．一方，10 個の音素認識の候補から選択する場合 ($N = 10$) では，単語数は $N = 1$ の場合に比べ少なくなった．これは，単語が同じ意味を表す場合には，同じ音素列で表現した方が尤度が高くなり，例えば「ういぐるみ」の候補の中に「ぬいぐるみ」が存在すれば，この候補を選択できているためである．

以上のように，提案手法によりマルチモーダル情報と NPYLM を統合することで，より正解に近い単語が選択可能であることが分かる．

5.2 マルチモーダルカテゴリゼーション

次に，前節の実験によって選択された単語を用いて，物体概念の形成を行った．各条件における分類の精度が図 6 である．音素認識の 1-best のみを使用した場合 ($N = 1$) は，他の場合より精度が低いことが分かる．1-best のみでは音素の誤認識が多く含まれているため，NPYLM による単語の候補を複数計算したとしても ($(N, L) = (1, 20)$)，正しい単語を得ることができなかったことが原因として考えられる．一方，音素認識の 10-best から選択した ($(N, L) = (10, 1)$) の場合の分類精度は 88%であり，正しい単語を選択することができ，物体の分類の精度が向上した．さらに，音素認識の 10-best の認識結果に対して，NPYLM から複数の候補を計算した場合 ($(N, L) = (10, 20)$) の分類精度は 90%となり，最も高くなった．

以上の結果より，物体概念を用いることでより正解に近い単語を選択することができ，このようにして選択された単語を用いることで物体の分類精度の向上ができていることが分かる．

6. まとめ

本稿では，ロボットが取得可能なマルチモーダル情報と，人からの教示発話を用いてロボットによる概念・語意獲得を行った．ロボットは語彙を持たないことを想定し，人からの教示発話を音素認識し，認識された音素列に対して NPYLM を適用し単語へと分割した．しかし，音素認識では認識誤りを含み，さらに十分な文章数が得られないため，正しい単語を得ることは困難である．そこで，音素列の候補を複数計算し，それらの音素列に対して NPYLM を複数回適用することで，単語の候補を計算した．このようにして得られた単語の候補から，物体カテゴリから単語が発生する確率を考慮し単語の選択を行った．実験により，音素誤りの少ない単語が選択でき，さらに選択した単語を用いることでより正解に近い物体概念が形成できることが示された．今後，我々がこれまで行なってきたオンラインマルチモーダルカテゴリゼーション [Araki 12] へ，この手法を適用することでよりインタラクティブに学習が可能なシステムを構築する予定である．

参考文献

[Araki 11] Araki, T., Nakamura, T., Nagai, T., Funakoshi, K., Nakano, M., and Iwahashi, N.: Autonomous Acquisition of Multimodal Information for Online Object Concept Formation by a Robot, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1540–1547 (2011)

[Araki 12] Araki, T., Nakamura, T., Nagai, T., Nagasaka, S., Taniguchi, T., and Iwahashi, N.: Online Learning of Concepts and Words Using Multimodal LDA and Hierarchical Pitman-Yor Language Model, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1623–1630 (2012)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Fei-Fei 05] Fei-Fei, L.: A bayesian hierarchical model for learning natural scene categories, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 524–531 (2005)

[Fergus 03] Fergus, R., Perona, P., and Zisserman, A.: Object Class Recognition by Unsupervised Scale-Invariant Learning, in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 264–271 (2003)

[Nakamura 09] Nakamura, T., Nagai, T., and Iwahashi, N.: Grounding of word meanings in multimodal concepts using LDA, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3943–3948 (2009)

[Sinapov 11] Sinapov, J. and Stoytchev, A.: Object Category Recognition by a Humanoid Robot Using Behavior-Grounded Relational Learning, in *IEEE International Conference on Robotics and Automation*, pp. 184–190 (2011)

[Sivic 05] Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., and Freeman, W. T.: Discovering Object Categories in Image Collections, in *IEEE International Conference on Computer Vision*, pp. 17–20 (2005)

[Vedaldi 10] Vedaldi, A. and Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms, in *ACM International Conference on Multimedia*, pp. 1469–1472 (2010)

[Wang 09] Wang, C., Blei, D., and Fei-Fei, L.: Simultaneous image classification and annotation, in *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 0, pp. 1903–1910 (2009)

[持橋 09] 持橋 大地, 山田 武士, 上田 修功 : ベイズ階層言語モデルによる教師なし形態素解析 (言語モデル・ウェブ解析), 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2009, No. 36, p. 49 (2009)

[中村 10] 中村 友昭, 西田 匡志, 長井 隆行 : 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, 5V-3 (2010)