2B2-OS22a-6

移動ロボットによる マルチモーダル情報の自律的取得と概念・語意獲得

Autonomous Multimodal Information Acquisition by Mobile Robots for Concept Formation and Acquisition of Words

荒木 孝弥[*1] Takaya Araki　中村 友昭[*1] Tomoaki Nakamura　長井 隆行[*1] Takayuki Nagai　船越 孝太郎[*2] Kotaro Funakoshi　中野 幹生[*2] Mikio Nakano　岩橋 直人[*3] Naoto Iwahashi

[*1]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

[*2]ホンダ・リサーチインスティチュート・ジャパン
Honda Research Institute Japan

[*3]独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

This paper proposes a robot that acquires multimodal information, i.e. auditory, visual, and haptic information, fully autonomous way using its embodiment. We also propose an online algorithm of multimodal categorization based on the acquired multimodal information and words, which are partially given by human users. The proposed framework makes it possible for the robot to learn object concepts naturally in everyday operation in conjunction with a small amount of linguistic information from human users. In order to obtain multimodal information, the robot detects an object on a flat surface. Then the robot grasps and shakes it for gaining haptic and auditory information. For obtaining visual information, the robot uses a hand held small observation table, so that the robot can control the viewpoints for observing the object. As for the multimodal concept formation, the multimodal LDA using Gibbs sampling is extended to the online version in this paper. The proposed algorithms are implemented on a real robot and tested using real everyday objects in order to show validity of the proposed system.

1. はじめに

近年，人間との共存を目的としたロボットの研究が盛んに行われており，中でも重要な課題の一つとして，ロボットによる物体の扱いが挙げられる．実環境でロボットが扱うべき物体は数多く，同時に動作環境によって異なるため，事前に全ての物体をロボットに登録しておくことは現実的ではない．このことは，実環境においてロボットが物体の見た目や名前の学習を行う必要性を物語っていると言え，ロボットは自律的に物体の情報を取得し，人の手を煩わせることなく自動的に学習を進めていくことが望まれる．

著者らは，データマイニングや自然言語処理などに広く用いられてきた統計モデルを自律型ロボットに応用することで，物体のカテゴリ分類及び物体概念の形成を行う手法を提案している [中村 10]．物体概念を獲得することで，未知の物体に対しても，その特性や性質を予測することが可能となり，人間と同様に柔軟な対応をとることができる．この様なレベルでの物体学習を実際の環境で行うためには，ロボットが物体の視覚や聴覚，触覚などのマルチモーダルな情報を自律的に取得する必要がある．複数の感覚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類を自動的に行うことが可能になると考えられるためである．つまり，最終的に実現したいのは，部屋内を動き回り，未知の物体を発見した場合には，自律的かつ自動的にその物体のマルチモーダル情報を取得し，物体や概念の学習を行う図 1 のようなロボットである．一方，物体概念を形成する上で，言語情報も重要である．文献 [Nakamura 07] で我々は，マルチモーダル情報のカテゴリゼーションによって形成された概念と単語の結び付きを学習することで，ロボットが語意を獲得できることを示した．ここでは，さらに単語の情報とマルチモーダル情報を併用することで，語意を含む物体概念全体を学習することを考える．物体の概念を形成する上で，語意情報を用いることで，より人間の感覚に即した物体のカテゴリ分類が実現し，同時に未知の物体に対する言語的な予測を行うことも可能となる．単語情報は人間による教示が必要であるが，先述したように単語情報全てを人間の手によって与えることは現実的ではないため，断片的な単語情報を適切な感覚情報と結び付けた概念形成を行う能力が必要となる．ロボットが物体のマルチモーダル情報取得を全自動で行い，カテゴリ分類と語意獲得を同時にかつ自律的に行うことができれば，未知物体の認識や機能の推定，単語情報の想起などが可能となるため非常に有用である． そこで本稿では，ロボットに搭載された各種センサを用いて，ロボット自身が複数の情報を自律的に取得するシステムを提案する．これによりロボットは，搭載されたカメラからの視覚情報，物体を振った際の音情報，物体を把持した際の感圧センサによる触覚情報を取得できる．また，一部の単語情報を利用した学習を行い，Gibbs Sampling をベースとしたオンラインマルチモーダル LDA による，自律的カテ

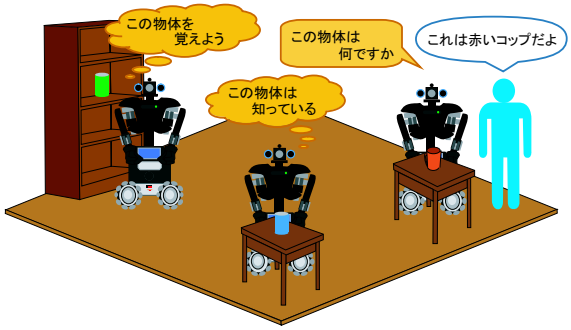

図 1: 自律的物体情報取得ロボット

連絡先: 荒木 孝弥，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，taraki@apple.ee.uec.ac.jp

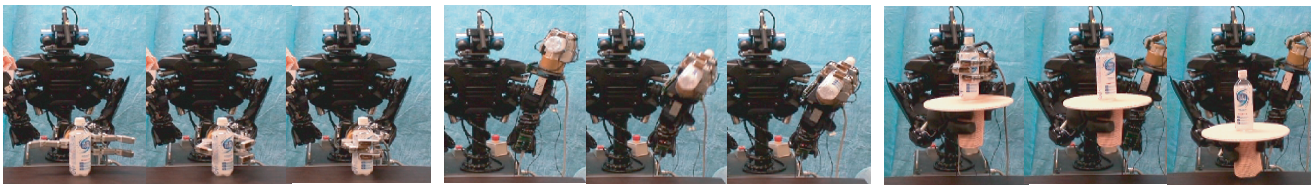

図 3: マルチモーダル情報の取得 (左から触覚情報，聴覚情報，視覚情報の取得)

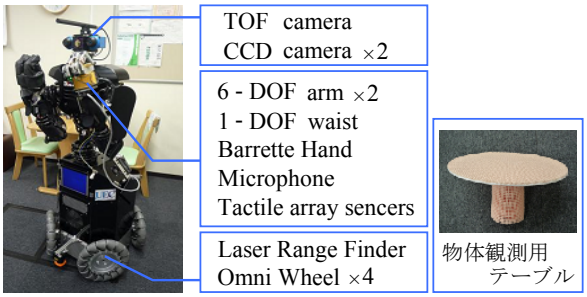

図 2: ロボットプラットフォーム

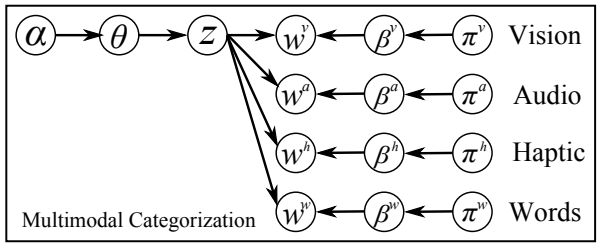

図 4: マルチモーダル LDA のグラフィカルモデル

ゴリゼーションと語意の獲得を同時に行う手法を提案する．本稿では，Bag-of-Features モデルに基づく LDA をマルチモーダルに拡張し，パラメータ推定に Gibbs Sampling の原理を適用する．さらにシステムをオンライン化することで，データを保持する必要なく，ロボットと人間によるインタラクティブな物体の学習と認識が可能となる．

2. マルチモーダル情報取得システム

2.1 マルチモーダル情報取得

本稿では図 2 に示すロボットを想定した．ロボットの下半身にはレーザーレンジファインダー及びオムニホイールの台車が搭載されており，予め室内の地図を保持しておくことで，障害物を回避しながら全方位への移動が可能である．ロボットは，物体の探索を行いながら室内を自由に移動し，未知の物体を発見した際には，視覚情報として，CCD カメラ及び赤外線 TOF カメラによる複数視点からの画像情報，3 次元情報，反射強度情報を，触覚情報として，ハンドに搭載された感圧センサによる物体を把持する際の圧力情報と指の角度情報を，聴覚情報として，ハンドに搭載されたマイクから取得される，物体を振動させた際の音情報を取得する．

ロボットが自律的に未知物体の情報を取得する際には，未知物体をどの様に発見するか，マルチモーダル情報をどのように観測するか，という問題が考えられる．本稿では，物体が机などの平面上にあることを想定して，平面検出を利用した物体検出手法 [Attamimi 10] を用いて物体の検出を行う．また，視覚情報取得時に把持した際に物体が変形したり，指で隠れてしまうなどの問題を解決するため，片方の手に簡易的な物体観測用テーブルを持たせておき，把持した未知物体をテーブル上に物体を載せ，そのテーブルをロボット自身が動かすことで様々な方向からの観測を実現する．観測用テーブル上の物体を再度平面検出により検出することで，物体の色・テクスチャ情報，距離情報，反射強度情報を取得する．触覚情報は，平面検出により未知物体を検出し，把持動作を 1 つの物体に対して 5 回行い，一定速度でハンドを閉じた際の触覚アレイセンサの出力を取得する．聴覚情報は，触覚情報取得時に使用したハンドに取り付けたマイクを用いて物体を振った際に発生する音を取得する．本稿では腕を振ることによるモータ等のノイズの影響を解決するために，何も持たずに腕を振った際の音を同様に取得しておき，特徴量レベルでのノイズ除去を行うこととする．実際にロボットが各情報を取得している様子を図 3 に示す．

2.2 マルチモーダル情報処理

2.2.1 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 10 枚の画像を取得した)．本稿では特徴量として 36 次元の PCA-SIFT 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 [Lowe 07]．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し，500 次元のヒストグラムとして視覚情報を取り扱う．

2.2.2 聴覚情報

音声信号を 0.2[s] 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，音声認識で最もよく使用されている MFCC を用いることとし，これにより各フレームは 13 次元の特徴ベクトルとなる．この特徴ベクトルを，予め計算した 50 の代表ベクトルを用いてベクトル量子化し，50 次元ヒストグラムとして聴覚情報を扱う．また音声取得時の雑音を取り除くため，何も持たずに腕を振った際の音を予めに取得しておくことで，特徴量のレベルでノイズ除去を行う．

2.2.3 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アレイ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，その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，最終的に得られる 15 次元ヒストグラムを触覚情報として用いる．

3. オンラインマルチモーダル LDA

3.1 概念のグラフィカルモデル

提案するロボットによる概念形成のグラフィカルモデルを図 4 に示す．ロボットは自律的にマルチモーダル情報を取得し，その際に部分的に得られた単語情報に基づいて，物体概念を形成する．先述したように，取得した全ての知覚情報は Bag-of-Features モデルとして扱い，多数・多次元の局所的な情報を，位相情報に依存しない生起回数の情報として利用する．実際に取得する視覚，聴覚，触覚，単語情報 w^v，w^a，w^h，w^w は，それぞれハイパーパラメータ π^v，π^a，π^h，π^w によって決まるディリクレ事前分布に従う，パラメータ β^v，β^a，β^h，β^w の多項分布によって発生する．また，z はカテゴリを示し，

カテゴリ z の出現確率分布を表す多項分布のパラメータを θ とする．このパラメータ θ は，ハイパーパラメータ α により決まるディリクレ事前分布に従う．

3.2 Gibbs Sampling に基づくオンライン MLDA

本稿における物体のカテゴリゼーションは，図 4 のモデルのパラメータを，物体から取得した情報と人間が付与した一部の単語情報を用いて学習することに相当する．これにより，マルチモーダル情報による物体のカテゴリ分類と同時に，各物体に対する語意の獲得が可能となる．本稿では，モデルのパラメータの学習に，近似式を用いず計算過程も簡易である Gibbs Sampling の原理を適用する．各モダリティ(視覚，聴覚，触覚，単語) のインデックスを m，モーダル情報集合を w^m とすると，対象物体における m 番目のモダリティ情報の i 次元目の情報に割り当てられるカテゴリ z_{mi} をサンプリングする式は以下のようになる．

$$P\left(z_{mi}=k|\boldsymbol{z}^{-mi},\boldsymbol{w}^m,\alpha,\pi^m\right)\propto (N_k^{-mi}+\alpha)\cdot\frac{N_{mw^mk}^{-mi}+\pi^m}{N_{mk}^{-mi}+W^m\pi^m} \quad (1)$$

但し，W^m は m 番目のモーダル情報の次元数，N_k は対象物体における全情報についてカテゴリ k が割り当てられた回数，N_{mw^mk} は対象物体におけるモダリティ m の情報 w^m についてカテゴリ k が割り当てられる回数，N_{mk} は対象物体におけるモダリティ m の情報についてカテゴリ k が割り当てられた回数を表す．式 (1) に従って，繰り返しサンプリングを行うことで，結果がある値 $\hat{N}_*$ へと収束する．収束結果から，K をカテゴリの総数とする時，最終的なパラメータの推定値 $\hat{\beta}^m_{w^mk}$，$\hat{\theta}_k$ は以下のようになる．

$$\hat{\beta}^m_{w^mk} = \frac{\hat{N}_{mw^mk}+\pi^m}{\hat{N}_{mk}+W^m\pi^m} \quad (2)$$

$$\hat{\theta}_k = \frac{\hat{N}_k+\alpha}{\sum_k\hat{N}_k+K\alpha} \quad (3)$$

従来手法では学習を行う全物体のデータに対し，式 (1) によりサンプリングを繰り返すことで，パラメータの推定を行っていた．そのため複数の物体を分類するためには分類する物体の全データを保持しておくことが前提となり，扱うデータに応じて大量のメモリを消費する問題がある．さらに，新たなデータが入力された際に，新たなデータを加えた全データを使用しバッチ学習を行う必要があった．バッチ学習は計算時間が長く，本稿のようなインタラクティブな学習には実用的でない．

本稿ではこの問題を解決するために，新たな入力データのみから学習したモデルのパラメータを逐次更新することで，パラメータをオンラインで学習する手法を用いる．しかし，事前に学習したモデルを次回の学習時に利用する場合，初期値や学習物体の順番によって，最終的に学習されるモデルが大きく変化することが予想される．そこで本稿では，忘却率 $\lambda(0<\lambda<1)$ を導入し，

$$\hat{N}^{(j+1)}_{mw^mk}=(1-\lambda)\hat{N}^{(j)}_{mw^mk} \quad (4)$$

のように，j 番目の物体の学習時に収束した値の一部を忘却し，$j+1$ 番目の学習時の初期値とするモデルの作成を行う．さらにこの時 λ を微小に変化させたモデルを数十から数百個作成し，そのモデル群の中から，最適なモデル選択を行うことでこの問題を解決する．ここで，本稿における単語情報は人間のアノテーションによって与える情報であるため，取得した感覚情報から予測される単語の正解率が高いほど，そのモデルは適切であると考えられる．そこで本稿では，モデル選択の指標として，物体の視覚，聴覚，触覚情報 $\boldsymbol{w}^v$，$\boldsymbol{w}^a$，$\boldsymbol{w}^h$ による単語情報の予測確率

$$P(w^w|\boldsymbol{w}^{v,a,h})=\int\sum_z P(w^w|z)P(z|\theta)P(\theta|\boldsymbol{w}^{v,a,h})d\theta \quad (5)$$

を用い，作成した各モデルの中で，単語の予測精度が最も高いモデルを次の学習モデルとして選択する．すなわち，新たな物体のデータが入力された際に，式 (4) によりそれまでの学習の一部を忘却したモデルを複数個作成し，単語の予測精度の高いモデルの選択を行い，式 (1) を収束するまで繰り返すことで，オンラインによる物体のカテゴリゼーションが可能となる．なお，単語の予測確率 $P(\theta|\boldsymbol{w}^{v,a,h})$ は，学習したパラメータ $\beta^{v,a,h}$ を固定し，前述の Gibbs Sampling を適用することで θ を再計算することにより求めることができる．

図 5: 実験に使用した物体

表 1: オブジェクトに付与した単語一覧

赤	鈴	液体	サル	ラップ	クッキー
青	頭	硬い	ナス	四角い	シャンプー
緑	黄色	長い	ピンク	取っ手	柔らかい
紫	黄緑	毛糸	カエル	オレンジ	ぬいぐるみ
黒	灰色	ゾウ	ヒヨコ	ライオン	プラスチック
白	茶色	クマ	アルミ	ビニール	ペットボトル
音	単色	ブタ	コップ	カラカラ	カラースプレー
布	動物				

4. 実験

図 2 に示すロボットに提案手法を実装し，物体からの自動情報取得実験，及び取得情報と人が付与した単語情報を用いたオンライン MLDA によるカテゴリゼーション実験を行った．実験には，図 5 に示す 8 つのカテゴリに分類される 48 個の物体を使用した．またこれらの各物体に対して，人間が表 1 に示す単語の教示を行い，未知物体に対する発生単語の予測の精度を検証した．

4.1 マルチモーダル情報の取得実験

図 6 に，実際にロボットが自動で取得した情報の一部を示す．提案手法により，ロボットは全ての物体から自律的に視覚，聴覚，触覚情報を取得し，また人間の教示によって単語情報を取得することができた．

4.2 オンライン MLDA とモデル選択の有用性の検証

本章では，提案するオンライン MLDA におけるモデル選択の有用性の検証のため，実験 4.1 で取得した情報によるオンライン MLDA の学習を行った．この際，忘却率をある一意に固定した場合の学習と，逐次最適なモデル選択を行った場合の学習，さらに従来手法であるバッチ学習での認識精度の比較を行った．なおオンライン学習では，λ を $0.0, 0.1, 1.0$ とし，全 48 物体の情報をランダムに並び替え逐次学習を行い，1 つの物体を学習する毎に，全物体の認識を行い認識精度を算出した．

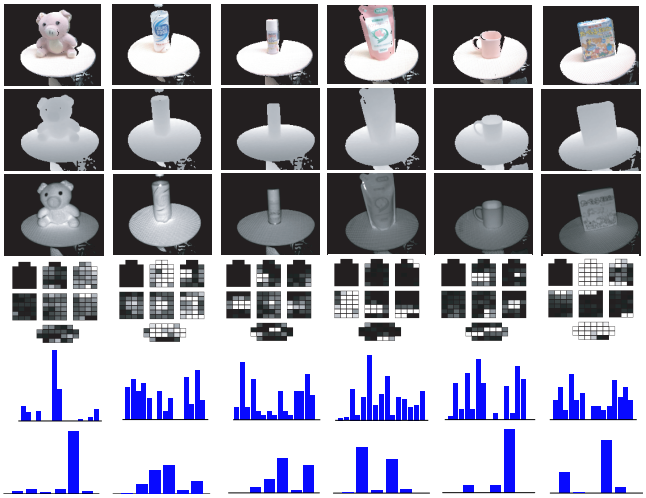

図 6: 取得した情報の例 (上から CCD 画像を 3 次元情報にマッピングした画像，距離画像，反射強度，物体を握りきった際の触覚センサ出力，触覚情報ヒストグラム，特徴的な 6 次元 MFCC ヒストグラム)

実験結果を図 7 に示す．なおグラフの横軸は学習物体数，縦軸は認識精度を表す．従来手法である Gibbs Sampling によるバッチ学習に比べて，いずれも低い値となったが，提案するオンライン学習により，学習する物体数の増加に伴って認識精度が高くなる結果が得られた．これは提案するオンライン学習アルゴリズムにより正しく分類ができており，全ての学習データを保持することなく，逐次学習が可能であることを示している．学習モデルを忘却せずに利用する ($\lambda = 0.0$) 場合，初期値や物体の順番による影響が大きくなり，$\lambda = 0.1$ の場合に比べて認識精度が低くなったと思われる．さらに忘却率を増加させた場合，認識精度は徐々に降下していき，$\lambda = 1.0$ とした時，全てを忘却するため学習は進まず，学習物体数によらずほぼ一定の精度となった．また，提案するモデル選択手法を導入した学習結果では，忘却率を固定した学習に比べて分類精度が向上しており，逐次最適なモデルの選択が実現したものと言える．

4.3 オンライン MLDA による認識と語意獲得

図 5 に示す物体を，学習用物体と認識用の未知物体にランダムに分割し，認識用物体に対して，提案手法により正しく単語の予測ができるか検証した．まず学習用物体の視覚，聴覚，触覚，単語情報を用いて，提案するオンライン MLDA により学習を行う．次に，認識用物体のマルチモーダル情報 $\bar{w}^v$，$\bar{w}^a$，$\bar{w}^h$ から単語 w^w が発生する確率 $P(w^w|\bar{w}^v, \bar{w}^a, \bar{w}^h)$ を求め，その上位 3 つの単語を評価する．さらに，学習時に付加する単語数と物体数を変化させ，付加単語数及び学習物体数に応じた単語の予測精度を検証した．

実験結果を図 8 に示す．グラフは，学習時に付加する単語数と学習物体数における，単語の予測精度である．学習時に 5 つの物体に対して単語を付与するだけでも，未知物体に対して，約 30〜50%の精度で正しい単語を予測することが可能となり，その後も付与単語数の増加に伴い精度は向上した．また学習物体数も同様に，物体数に比例して単語予測精度が増加し，最終的に 40 個の物体に単語を付与することで，77.8%の精度となった．以上の結果から，一部の単語情報を用いたオンライン学習によって，語意の獲得ができ，未知物体に対しても予測が可能であると言える．

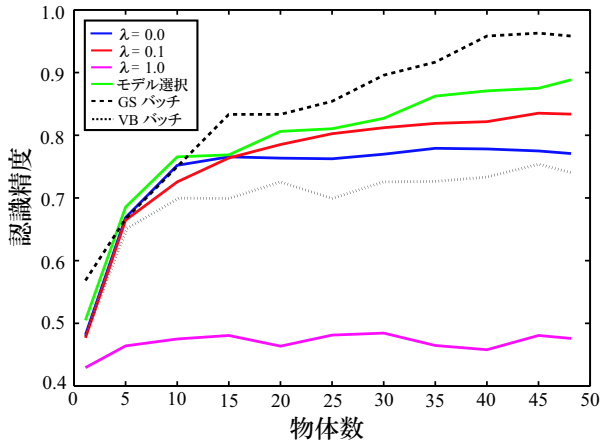

図 7: 物体の認識精度

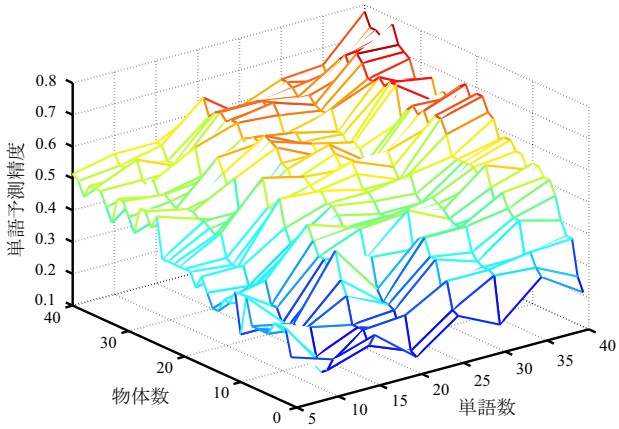

図 8: 単語予測精度

5. まとめ

本稿では，ロボットによる自律的なマルチモーダル情報取得システムとオンライン MLDA による概念と語意の獲得手法を提案した．情報取得システムにより，ロボットは物体の視覚，聴覚，触覚情報，さらに人間による単語情報の付与を行うことで，これらの情報を自律的に取得することが可能となった．またオンライン MLDA により，情報を保持する必要なく，インタラクティブな物体のカテゴリ分類が可能となり，同時に未知物体の認識と単語予測が実現した．今後の課題として，モデルの階層化やカテゴリ数の自動決定等が挙げられる．

参考文献

[中村 10] 中村ほか: “複数のマルチモーダル LDA を用いた抽象的概念の形成”, 人工知能学会全国大会 2010, 1J1-OS13-3, 2010

[Nakamura 07] Nakamura,T. *et al.*: “Grounding of word meanings in multimodal concepts using LDA”, in Proc. of IROS2009, pp.3943–3948, 2009

[Attamimi 10] Attamimi,M. *et al.*: “Learning Novel Objects Using Out-of-Vocabulary Word Segmentation and Object Extraction for Home Assistant Robots”, in Proc. of ICRA2010, pp.745–750, 2010

[Lowe 07] Lowe,D. : “Distinctive Image Features from Scale-invariant Keypoints”, Int. J. Comput. Vision, vol. 60, no. 2, pp.91–110, 2004

[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