2G4-OS-19a-3

多層マルチモーダルLDAを用いた人の動きと物体の統合的概念の形成

Integrated Concept of Objects and Human Motions based on Multi-layered Multimodal LDA

*1Muhammad Fadlil　池田圭佑*1 Keisuke Ikeda　阿部香澄*1 Kasumi Abe　中村友昭*1 Tomoaki Nakamura　長井隆行*1 Takayuki Nagai

*1電気通信大学
The University of Electro-Communication

The human understanding of things is based on prediction which is made through concepts formed by categorization of their experience. To mimic this mechanism in robots, multimodal categorization, which enables the robot to form concepts, has been studied. On the other hand, segmentation and categorization of human motions have also been studied to recognize and predict future motions. This paper addresses the issue of how these concepts are integrated to generate higher level concepts and, more importantly, how the higher level concepts affect each lower level concept formation. To this end, we propose multi-layered multimodal latent Dirichlet allocation (mMLDA) to learn and represent the hierarchical structure of concepts. We also examine a simple integration model and compare with the mMLDA. The experimental results reveal that the mMLDA leads to better inference performance and, indeed, forms higher level concepts integrating motions and objects that are necessary for real-world understanding.

1. はじめに

近年，知能ロボットの研究開発が盛んに進められている．そのようなロボットの要素技術として物体のカテゴリ分類・認識があり，未知の環境で柔軟に動作するためにも物体のカテゴリの認識ができることが重要であると考えられる．そのため現在まで，物体から取得可能な特徴量を用いた物体や画像のカテゴリ分類・認識に関する研究が行なわれている．著者らも，これまで pLSA (probabilistic Latent Semantic Analysis) や LDA (Latent Dirichlet Allocation) 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物体カテゴリを教師なしで形成できることを示した [Nakamura 07, Nakamura 09]．しかし，ロボットが物体を扱うためには，物体のカテゴリ認識だけでは不十分であり，物体と動作やその使い方など，物体概念と他の概念との関係を獲得する必要があると言える．

そこで本稿では，特に物体と動作の関係に着目し，これまでの研究で行なってきたマルチモーダルカテゴリゼーションにより形成された物体概念と，その物体を扱う人の動作の概念を統合することで，物体と動作の関係を学習可能なモデルを提案する．物体概念は，物体をロボットが観測し，得られるマルチモーダル情報をマルチモーダル LDA（MLDA）で分類することで形成する．また，動作概念はロボットに搭載された Kinect から，物体を扱っている人の関節角を取得し，これらの情報を MLDA 分類することで形成する．さらに，提案するモデルでは，これら物体と動作の 2 つの概念を統合する MLDA を上層に配置し，これらの関係を表す概念を形成する．すなわち，提案モデルは，多層の MLDA から構成されており，下層の MLDA では動作と物体の概念がそれぞれ形成され，上層の MLDA でこれらの概念を統合している．これにより，例えば図 1 に示すように，下位層では，ジュースという物体概念と，物を口に運ぶ動作概念が形成され，上位層でこれらの関係が学習され飲むという行動概念を形成することが可能となる．さらに，このように複数の概念を統合することで，未観測の情報の予測が可能となる．例えば，ジュースを見ることで飲む動作を予測することや，逆に動作からその動作に関連している物体が予測可能となる．

連絡先: Muhammad Fadlil, 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, mfadlil@apple.ee.uec.ac.jp

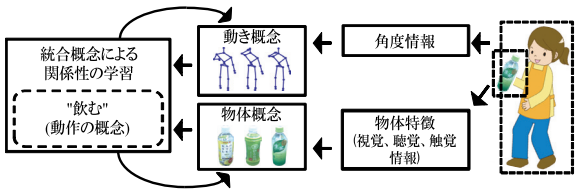

図 1: 統合概念形成の概念図

関連研究として，物体や画像のカテゴリ分類に関する研究があり，これにより柔軟な物体認識が可能となる．また，動作のモデル化に関する研究も近年盛んに行なわれている [谷口 10]．しかし，これらの研究では，物体の分類や動作のモデル化など単一の概念のみを考えており，それらの関係の獲得は考えられていない．また，Recurrent Neural Network with Parametric Bias(RNNPB) を用いて，物体の動きとその時に発生する音情報を関連付けて学習し，他方の情報から未観測の情報を予測する研究が行なわれている [Ogata 10]．RNNPB を用いることで，時系列の情報を保持したまま物体の動きや，音などを学習している．しかし，この研究の目的は，未観測の情報を予測することであり，物体のカテゴリ分類は考えられていない．また，RNNPB では学習データ数のスケーラビリティの問題がある可能性があり，どこまで複雑な情報が扱えるかは必ずしも明らかではない．

2. マルチモーダル LDA

ここでは，図 2 に示す MLDA の学習・認識について述べる．図 2 のモデルにおいて，モダリティ n の情報 x^n は，それぞれハイパーパラメータ ϕ^n 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^n の多項分布によって生成されるモデルである．また z はカテゴリを表し，ハイパーパラメータ α によって決まるディリクレ事前分布に従うパラメータ θ の多項分布により生成される．本稿におけるカテゴリ分類は，実際に取得した情報 x^n に基づき，モデルのパラメータ θ および β^n を推定することに相当し，パラメータ推定にはギブスサンプリングを用いる．ギブスサンプリングにおいて，j 番目の物体のモダリティ n の情報の i 番目に割り当てられるカテゴリ z_{nij} は，θ，β^n を周辺化した条件付確率からサンプリングされる．

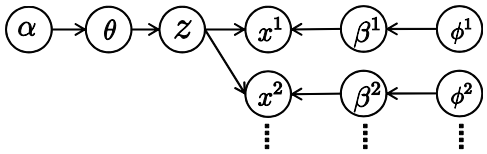

図 2: マルチモーダル LDA のグラフィカルモデル

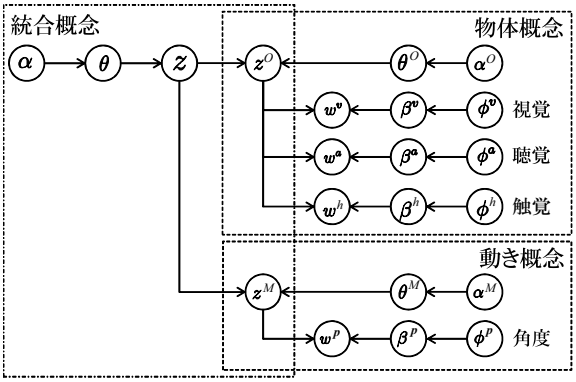

図 3: 多層マルチモーダル LDA のグラフィカルモデル

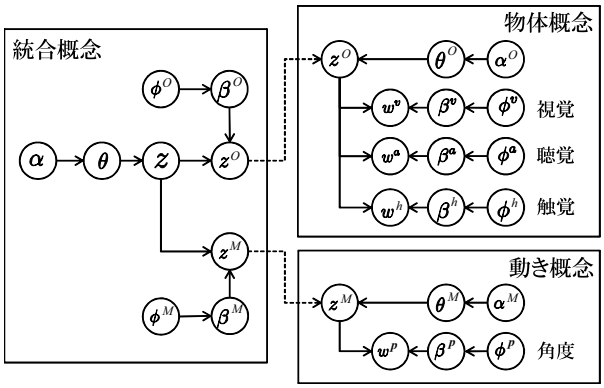

図 4: 統合概念形成 LDA の近似モデル

$$p(z_{nij}=k|\boldsymbol{z}^{-nij},\boldsymbol{x}^n,\alpha,\pi^n)\propto (N_{kj}^{-nij}+\alpha)\frac{N_{nx^nk}^{-nij}+\pi^n}{N_{nk}^{-nij}+W^n\pi^n}, \quad (1)$$

但し，W^n はモダリティ n の次元数を表す．サンプリングを繰り返すことで，N_* がある値 $\bar{N}_*$ へと収束し，パラメータの推定値 $\hat{\beta}^n_{x^nk}$，$\hat{\theta}_{kj}$ は以下のようになる．

$$\hat{\beta}^n_{x^nk} = \frac{\bar{N}_{nx^nk}+\phi^n}{\bar{N}_{nk}+W^n\phi^n} \quad (2)$$

$$\hat{\theta}_{kj} = \frac{\bar{N}_{kj}+\alpha}{N_j+K\alpha} \quad (3)$$

また，学習した確率モデルを用いて，未知物体のカテゴリを推定することが可能である．未知物体のマルチモーダル情報 $\boldsymbol{x}'^1,\boldsymbol{x}'^2\cdots$ が与えられた場合，そのカテゴリは $P(z|\boldsymbol{x}'^1,\boldsymbol{x}'^2,\cdots)$ を最大とするカテゴリ z から求められる．

$$\hat{z}=\underset{z}{\operatorname{argmax}}\sum_{\theta}P(z|\theta)P(\theta|\boldsymbol{x}'^1,\boldsymbol{x}'^2,\cdots) \quad (4)$$

3. 概念の統合モデル

本稿では，MLDA を用いて形成された物体と動きの概念を統合することで，より上位の概念を形成することができる．図 3 は提案する多層 MLDA (mMLDA) のグラフィカルモデルを示し，左側の z は統合概念，右側の z^O，z^M はそれぞれ物体と動きの概念を表す．一方，図 4 に示したように複数の独立した MLDA を結合することで，統合することも可能である．本稿ではこれを近似モデルと呼び，mMLDA と比較する．

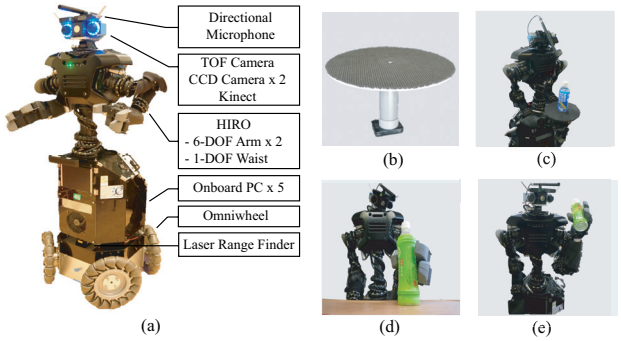

図 5: (a) ロボットプラットフォーム，及びマルチモーダル情報の取得：(b) 物体観察用回転テーブル，(c) 視覚情報の取得，(d) 触覚情報の取得，(e) 聴覚情報の取得

3.1 物体概念

物体概念はロボットによって実際に取得したマルチモーダル情報をカテゴリ分類することにより形成する．ここでは図 2 における x^1,x^2,x^3 をそれぞれ視覚・聴覚・触覚情報と考えることで，MLDA により物体概念を形成することが可能である．マルチモーダル情報は，図 5(a) に示した家庭用サービスロボット DiGORO により取得する．

視覚情報 図 5(b) に示した回転テーブルを用いて，物体を観測し，画像を 10 枚取得する (図 5(c))．本稿では特徴量として 36 次元の DSIFT[Vedaldi 10] 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．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．

触覚情報 触覚情報には，各物体を数回握り (図 5(d))，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．取得したデータは曲線近似を行い，その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．

聴覚情報 各物体を振った際に取得した音声信号 (図 5(e)) 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．

3.2 動き概念

前述の物体概念と同様に，図 2 における x^1 を人が物体に対して行なう動きの情報と考えることで，動き概念の形成を行なう．動き情報は，人の動作中の関節角度を Kinect を用いて取得した．取得した関節角は 20 箇所であり，動作開始から動作終了まで連続して取得した．本稿では，動きの情報は対象となる物体によって分節することができると仮定している．1 つの動作から複数の 20 次元の特徴ベクトルが得られ，それ

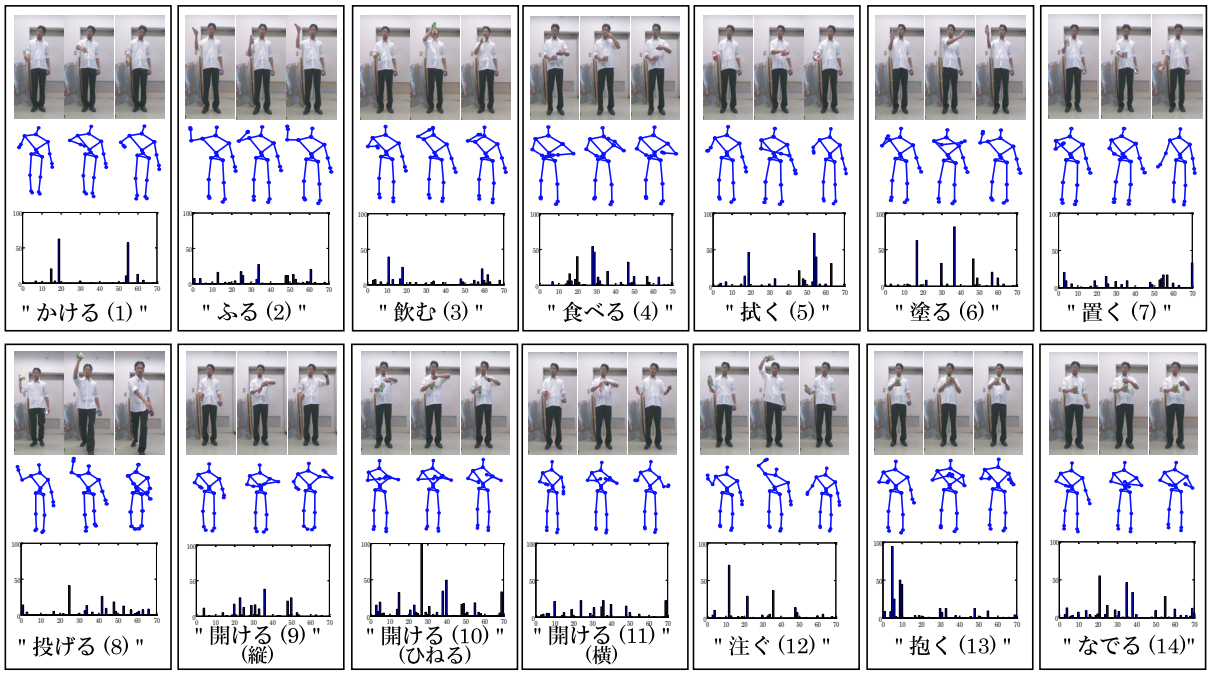

図 6: 各動きから取得した情報の例：（上から）実際の動き，Kinect から取得した情報，70 次元のヒストグラム

をあらかじめ計算した 70 の代表ベクトルによりベクトル量子化することで，70 次元のヒストグラムを動き情報として用いる．このような，特徴量を動作認識に使用することは，既に [Mangin 12] において提案されており，その有効性が示されている．

3.3 概念の統合

図 3 に示したモデルを用いて，物体と動きの関係性を学習することで，統合概念を形成することができる．

3.3.1 多層マルチモーダル LDA (mMLDA)

図 3 より，このモデルは二層構造となっていることがわかる．このモデルにおいて，各概念を表す z，z^O，z^M は直接観測できない隠れ変数であり，観測データ w^v，w^a，w^h，w^p から学習する．具体的には，事後確率から隠れ変数をサンプリングすることで，各パラメータを推定する．w^v，w^a，w^h，w^p はそれぞれ，ハイパーパラメータ ϕ^v，ϕ^a，ϕ^h，ϕ^p によって決まるディリクレ事前分布に従う β^v，β^a，β^h，β^p をパラメータとする多項分布によって生成される．またカテゴリ z,z^O,z^M は，それぞれハイパーパラメータ α，α^O,α^M によって決まるディリクレ事前分布に従うパラメータ θ，θ^O，θ^M をパラメータとする多項分布によって生成されるモデルである．各パラメータは，以下の式を用いて，Gibbs Sampling により推定する．

$$\begin{aligned} P(z, z^M, z^O, w^v, w^a, w^h, w^p | \boldsymbol{z}, \boldsymbol{z}^O, \boldsymbol{z}^M, \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h, \boldsymbol{w}^p) \\ = P(z)P(z^M|z)P(z^O|z)P(w^p|z^M)P(w^a|z^O) \\ P(w^v|z^O)P(w^h|z^O) \end{aligned} \quad (5)$$

$$P(z|\boldsymbol{z}) = \frac{\alpha + N_{jz}}{K\alpha + N_j} \quad (6)$$

$$P(z^*|z, \boldsymbol{z}, \boldsymbol{z}^*) = \frac{\alpha^* + N_{zz^*}}{K\alpha^* + N_z} \quad (7)$$

$$P(w^m|z^*, \boldsymbol{z}^*, \boldsymbol{w}^m) = \frac{\phi^m + N_{z^* w^m}}{W^m \phi^m + N_{z^*}} \quad (8)$$

但し，N_{jz} は物体 j の全モダリティに上位カテゴリ z が割り当てられた数であり，$N_{z^* w^m}$ はモダリティ m の特徴量 w^m に下位カテゴリ z^* が割り当てられた回数である．

学習過程は，図 3 の右側に示したような下位概念 z^* の形成から始まり，形成された下位概念を初期値として，mMLDA を用いて，モデル全体の学習により，統合概念 z の形成を行う．式 (5)〜(8) と観測データを用い，サンプリングを繰り返すことで N_* がある値へと収束する．K を上位カテゴリの総数とする時，最終的なパラメータの推定値 $\hat{\beta}^m_{w^m z^*}$，$\hat{\theta^*}_{zz^*}$，$\hat{\theta}_{jz}$ は以下のようになる．

$$\hat{\beta}^m_{w^m z^*} = \frac{N_{z^* w^m m} + \phi^m}{N_{z^* m} + W^m \phi^m}, \quad (9)$$

$$\hat{\theta^*}_{zz^*} = \frac{N_{zz^* m} + \alpha^*}{N_{zm} + K\alpha^*}, \quad (10)$$

$$\hat{\theta}_{jz} = \frac{N_{jz} + \alpha}{N_j + K\alpha}, \quad (11)$$

但し，W^m はモダリティ m の次元数を表し，$N_{z^* w^m m}$ はモダリティ m の w^m に下位カテゴリ z^* が割り当てられた回数を表す．

学習したモデルを用いた未観測情報の予測は，以下の式で実現できる．

$$\hat{z}^M = \underset{z^M}{\operatorname{argmax}} \sum_z \sum_{z^O} P(z)P(z^M, z^O|z)P(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h|z^O) \quad (12)$$

$$\hat{z}^O = \underset{z^O}{\operatorname{argmax}} \sum_z \sum_{z^M} P(z)P(z^O, z^M|z)P(\boldsymbol{w}^p|z^M) \quad (13)$$

3.3.2 近似モデル

上述した提案モデル以外にも，各概念を MLDA により独立的に形成し，フィードフォワード的に接続することで，簡易的に物体，動き，統合概念を形成できると考えられる．図 4 に示したように，図の右側の物体概念 z^O と動き概念 z^M を学習した後，統合概念 z を学習することになる．しかし，後に示す実験の結果からわかるように，各概念を独立的に学習することで，下位概念での学習誤りがそのまま上位概念の学習に影響を及ぼし，モデル全体の精度を下げてしまうことになる．

4. 実験

実験は，表 1 に示した，図 6 の 14 種類の動きと図 7 の 10 カテゴリに分類される 50 の物体を組み合わせたサンプルを取

図 7: 実験で使用した物体

表 1: 物体に対して行った動き（括弧内の数字はカテゴリ番号）

動き	物体	動き	物体
かける (1)	ドレッシング (3)	置く (7)	カップ麺 (4)
ふる (2)	スプレー缶 (1)		スナック (7)
	ペットボトル (2)		クッキー (8)
	ドレッシング (3)	投げる (8)	ぬいぐるみ (9)
	ガラガラ (10)	開ける (9)	スプレー缶 (1)
飲む (3)	ペットボトル (2)	開ける (10)	ペットボトル (2)
食べる (4)	カップ麺 (4)	開ける (11)	フローリング ワイパー (6)
	スナック (7)	注ぐ (12)	シャンプー (5)
	クッキー (8)	抱く (13)	ぬいぐるみ (9)
拭く (5)	フローリング ワイパー (6)	なでる (14)	ぬいぐるみ (9)
塗る (6)	スプレー缶 (1)		

得することで行った．赤い四角の 10 個の物体は未観測情報の予測実験，残り 40 個の物体は分類実験に使用した．

4.1 分類実験

まず，得られた物体のマルチモーダル情報を分類することで物体概念の形成を行った．この際，カテゴリ数は 10 とした．その結果が図 8 であり，縦軸が物体のカテゴリ番号，横軸がモデルによって分類されたカテゴリを表している．図 8(a) に示した人手による分類を正解とした時，図 8(b) に示した提案手法 (mMLDA) による分類結果の精度は 87.5%である．一方，図 8(c) に示した近似モデルの精度は 85.0%である．同様に，動きの概念形成の結果を図 9 に示した．正解の分類 (図 9(a)) と比較すると，mMLDA(図 9(b)) の分類精度は 72.5%となり，近似モデル (図 9(c)) の分類精度は 62.5%となる．

最終的に学習されたモデルの上位層で，物体と動きの関連性を表す MLDA について考察する．表 1 に示した各物体と動きの関係の学習サンプル数から同時確率 $p(z^O, z^M)$ を求め，図 10(a) に色の濃淡で表示している．縦軸と横軸は，それぞれ物体と動きのカテゴリ番号を表す．これを正解基準とした時に，両モデル mMLDA(図 10(b)) と近似モデル (図 10(c)) の学習結果を比較する．実際に近似モデルの結果と正解，及び mMLDA の結果と正解の KL 距離を求めると，それぞれ 50.25 と 46.50 であり，mMLDA の学習結果が正確に近いことが分かった．

4.2 未観測情報の予測実験

次に，未観測情報の予測性能を評価するための実験を行った．実験では，図 7 に示した赤い四角の 10 個の物体を認識用データとして用い，残りの 40 個を学習用のデータとした．動き概念 z^M は観測された物体のマルチモーダル情報 $(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ から予測を行った．同様に，物体概念 z^O は観測された動きの情報 $\boldsymbol{w}^p$ から予測を行った．

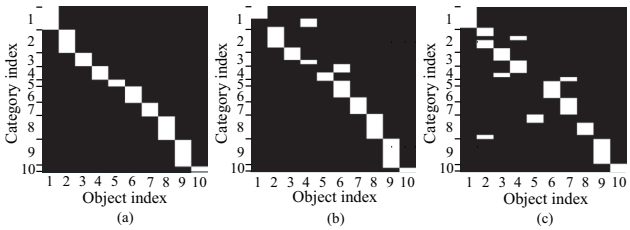

図 8: 物体の分類結果:(a) 正解, (b) mMLDA, (c) 近似モデル

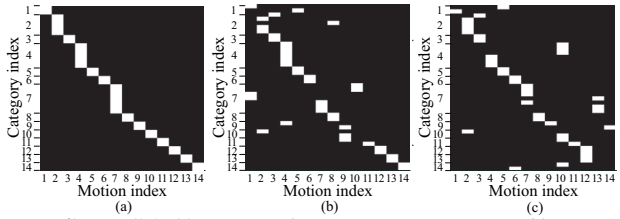

図 9: 動きの分類結果:(a) 正解, (b) mMLDA, (c) 近似モデル

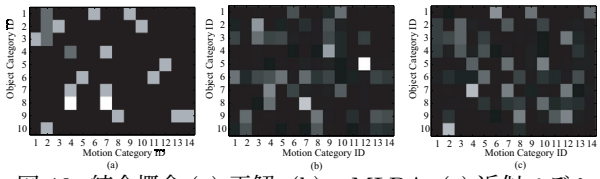

図 10: 統合概念:(a) 正解, (b) mMLDA, (c) 近似モデル

まず，提案した mMLDA を用いた動き概念 z^M の予測精度は 80.0%となり，近似モデルを用いた予測精度は 70.0%となった．同様に，観測された動きの情報から物体の概念 z^O を予測した時の予測精度は，mMLDA と近似モデルでそれぞれ，70.0%と 60.0%となった．

5. まとめ

本稿では，下位概念の関係性を表す上位概念を形成するための多層マルチモーダル LDA を提案した．実験結果より，提案した mMLDA が簡易的な近似モデルに比べ予測性能が高いことが明らかとなった．これは，上位・下位概念が相互に影響し合うことが，多層概念形成において重要であることを物語っている．今後，mMLDA を用いた様々な概念の統合を行い，提案した mMLDA の有効性を評価したいと考えている．

参考文献

[Nakamura 07] Nakamura,T. *et al.*: “Multimodal Object Categorization by a Robot”, in Proc. of IROS 2007, pp.2415–2420, 2007

[Nakamura 09] Nakamura,T. *et al.*: “Grounding of Word Meanings in Multimodal Concepts Using LDA”, in Proc. of IROS 2009, pp.3943–3948, 2009

[谷口 10] 濱畑慶太ほか: “ディリクレ過程と相互情報量による非分節対象物操作のからの動作抽出,” 人工知能学会全国大会, 1J1-OS13-11, 2010.

[Ogata 10] Ogata,T. *et al.*: “Inter-modality Mapping in Robot with Recurrent Neural Network”, Pattern Recognition Letters, vol.31, pp.1560–1569, 2010

[Vedaldi 10] Vedaldi,A. *et al.*: “Vlfeat: An open and portable library ofcomputer 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

[Mangin 12] Mangin,O. *et al.*: “Learning to Recognize Parallel Combinations of Human Motion Primitives with Linguistic Descriptions using Non-negative Matrix Factorization”, in Proc. of IROS 2012, pp.3268–3275, 2012