平成25年度修士論文

多層マルチモーダルLDAを用いた
複数概念の統合に関する研究

学 籍 番 号　1232061

氏　　名　MUHAMMAD FADLIL

知能機械工学専攻　先端ロボティクスコース

主指導教員　長井 隆行　准教授

副指導教員　金子 正秀　教授

提　出　日　平成26年 2月 5日

平成25年度修士論文

多層マルチモーダルLDAを用いた
複数概念の統合に関する研究

学 籍 番 号　　1232061

氏　　名　　MUHAMMAD FADLIL

知能機械工学専攻　　先端ロボティクスコース

主指導教員　　長井 隆行　准教授

副指導教員　　金子 正秀　教授

提　出　日　　平成26年 2月 5日

概要

知能ロボット開発において，ロボットが物体を扱うために，物体のカテゴリ分類だけではなく，物体と動きやその使い方など，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．さらに，ロボットによる真の理解を実現するために，場所や人物といった物事に対する概念の獲得も必要とする．

本研究では，多層マルチモーダル LDA(mMLDA) に基づく，ロボットによる多様な概念形成及び統合を実現する．mMLDA によって，概念の形成と統合が同時に獲得が可能であるため，各概念の形成が互いに影響しあって，より正しく形成できる．

さらに，我々が用いている言語もカテゴリに基づいており，ロボットもカテゴリ分類を通じて物体の概念を学習することで，未観測情報の予測や言語の理解が可能になると考えられる．言語理解のためのロボットによる語意の獲得問題についても，mMLDA を用いて実現することが可能である．本研究では，単語と概念間の相互情報量を用いることで，どの単語が本来どの概念に結びついているのかを自動的に推定する手法を提案する．また，単語と概念の結び付きを用いて，教示発話における概念の発火順を学習することが可能であり，これを学習することで，観測した情報を表現する文章を生成することができる．提案したこれらのモデルを実験によって，その有効性を示した．

目次

図一覧

表一覧

第 1 章

序論

1.1 はじめに

近年，知能ロボットの研究が盛んに進められている．そのような知能ロボットの要素技術として，物体のカテゴリ分類や認識があり，未知の環境でロボットが柔軟に動作するためにも物体のカテゴリが認識できることは重要である．現在まで，物体から取得可能な特徴量を用いた物体のカテゴリ分類・認識に関する研究が数多くなされている [2, 1, 3, 4, 5, 6].

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が提案され，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ロボットが教師なしで学習できることが示された [7, 8]．ここで重要なのは，学習された物体カテゴリを基盤とした未観測情報の予測であり，これがロボットによる理解につながる [9]．また，こうした物体カテゴリが教師なしで学習されることが重要であり，学習されたカテゴリを物体概念と呼ぶ [9].

しかし，ロボットが物体をより深く理解するためには，物体概念の学習だけでは不十分である．なぜなら，ほとんどの物体にはそれを使うという人の行為，使う人物の情報，場所の情報などが関連しており，物体とその物体に関連する様々な要素を学習する必要があるためである．つまり，物体概念，人の動きの概念，人物の概念と場所の概念を学習すると同時に，それらの関係性を獲得する必要があると言える．

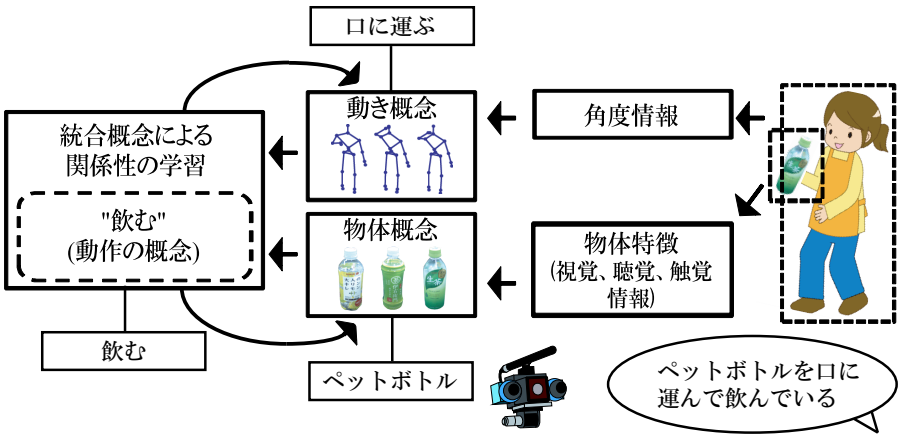

図 1.1: 統合概念形成の模式図

一方，人の動きの分節化と範疇化は，従来多くの研究がなされている [11, 12, 13]. 例えば高野らは，動きのシンボルツリーを生成するために人間の動きを分節化し，階層的に分類している．そして生成したシンボルツリーを用いて，人間の動きを予測することを提案した [11] ．また谷口らは，人間の動きの二重分節化を提案している [13]．これらの研究は，人間の動きと記号 (言語) を接続するための鍵を提供しているという意味において，非常に興味深く重要である．しかしこれらの研究では，物体を陽には表現しておらず，その点においては，本論文で提案する手法とは大きく方向性が異なる．人の動きの多くは，物体の使用に深く関連しているため，物体と動きの概念の関係性を表現する統合概念を学習することは重要であると考えられる．

また，場所の分類も同様に，従来一つの問題として研究されている．例えば Jianxin Wu らは，ロボットが取得したその場所の視覚情報を用いて，場所の分類化や認識を行なう．しかしながら，これらの研究の多くは，物体の概念や人の動きの情報が考慮されていません．さらに，人物の概念に関しては，人の顔認識や顔の分類化を行なう研究はあげられます．しかし，同様にこれらの研究は，物体や場所などの様々要素との関係性が考慮されていません．

本論文では，これらの多様な概念をそれぞれ獲得すると同時に，それらの関係性を表すより高いレベルの概念を形成することを考える．例えば，物体概念と動き概

念の場合を考え，図 1.1 は，本論文の目的を表現した模式図である．この図において，“飲み物”（物体概念）と “何かを口に運ぶ”（動き概念）という 2 つの下位概念が表現されている．さらにこれらの概念が統合されることで，より高いレベルの概念である “飲む”（動作概念）という概念が形成される．この図において重要なのは，様々なレベルでの推論が可能であるということである．上記の例においてロボットは，与えられたペットボトルの視覚的な情報から，“何か口に運ぶ” 動きを想起することができる．逆に，“何か口に運ぶ” 動きから，“ペットボトル” という物体を推論することも可能であり，これはいわゆるジェスチャーの理解と捉えることができる．また，上位概念の形成過程が下位概念の形成、つまりは物体や動きのカテゴリ分類に影響を及ぼすことは注目に値するであろう．例えば，全く異なるテクスチャをもちながらもボトルの形である物体は，飲み物とは別の物体カテゴリに分類される可能性があるが，この物体が “何かを口に運ぶ” 動きと共に使用される場合，統合概念である “飲む” が下位層の分類に影響することで，“ペットボトル”（物体概念）といった単一の物体概念を形成することに寄与する．一方で，物体が異なる動きに関係する場合，見た目の似た物体であっても異なるカテゴリに分類される可能性がある．

本論文ではこうした仕組みを実現するために，多層マルチモーダル LDA（multi-layered Multimodal LDA: mMLDA）を提案する．mMLDA は，下位層の物体概念と動き概念，および上位層でこれらを統合した統合概念で構成される．ロボットは学習プロセスにおいて，人の動きと使用される物体を観測する．物体概念は，ロボットが物体に関して取得したマルチモーダル情報，すなわち視覚，聴覚及び触覚情報をマルチモーダル LDA (MLDA) を用いることで形成する．同様に動き概念は，ロボットに搭載したキネクトから取得される人の関節角度情報に基づいて MLDA によって形成される．これら二つの MLDA は，上位の MLDA によって結合され，この上位層において下位概念間の関係性を表現するような上位概念，すなわち統合概念を形成することになる．ただし，全ての層の分類プロセスは相互に依存していることに注意が必要である．こうした相互依存的なモデルとは異なり，それぞれの概念を表現する複数の MLDA をフィードフォワードに接続させる簡易的な近似モデルを考えることも可能である．本論文ではこれをベースラインモデルと呼ぶことにするが，このベースラインモデルと mMLDA との比較を通して提案

する mMLDA の有効性を明らかにする．また，これらの 2 つの概念だけではなく，場所概念と人物概念を含めた概念の統合を行う．

さらに，我々が用いている言語もカテゴリに基づいており，ロボットもカテゴリ分類を通じて物体の概念を学習することで，未観測情報の予測や言語の理解が可能になると考えられる．言語理解のためのロボットによる語意の獲得問題についても，mMLDA を用いて実現することが可能である．これは，[8] などで提案されているように，単語を Bag of Words (BoW) 表現としモダリティーの一つと考えることで解決することができる．しかし階層のない MLDA と異なる点は，どの階層のどの概念にどの単語が結びつくかを考える必要がある点である．この情報は教示発話には明示的に含まれていないため，学習アルゴリズムが何らかの基準に従ってこの結びつきを見出す必要がある．本稿では，単語と概念間の相互情報量を用いることで，どの単語が本来どの概念に結びついているのかを自動的に推定する手法を提案する．さらに，単語と概念の結び付きを用いて，教示発話における概念の発火順を学習することが可能であり，これを学習することで，観測した情報を表現する文章を生成することができる．

1.2 関連研究

関連研究としては，センサ情報に基づいた物体のカテゴリ分類に関する研究が挙げられる [1]-[6]．また，人間の動きのモデル化についても多くの研究がなされている [11]-[13]．本論文では，知覚情報の分類が主眼であり，その点においては上記の研究とは同様の方向性であると言える．しかし，本論文で提案するモデルでは複数の概念 (特に動き概念と物体概念) とそれらの関係性を同時に学習することを目的としているという点で上記の研究とは大きく異なると言える．したがって提案モデルでは，概念間の推論が可能であるのに対し，これらの研究ではそうした点については考慮されていない．

一方で尾形らは，Parametric Bias を用いた Recurrent Neural Network（RNNPB）を用いることで，異なるモダリティ間の情報をマッピングすることのできる手法を提案している [14]．このシステムは，物体の動きによって生成される音を表現する運動をロボットが生成できるように学習することが可能である．従って論文の目的

は，ロボットが異なる種類のセンサからの信号間のマッピングを学習するモデルを，RNNPB によって構築することである．このモデルを用いることで，ロボットは音から関連する動作を生成することができるようになり，これは本論文の目的と非常に関連している．しかしながら，[14] ではカテゴリー（概念）とそれらの相互依存関係を明示的に扱っているわけではなく，複数の概念を統合することにより獲得される上位の概念といったことも考慮していない点で本論文の提案するモデルとは大きく異なると言える．また，RNNPB は，スケーラビリティに問題がある可能性がある．実際，文献 [14] では，5 つの物体で実験を行っているのみであり，物体数などが大幅に増えた際にモデルが実際に機能するかどうかは必ずしも明らかではない．

さらに文献 [15, 16] では，感覚運動マッピングとしてのアフォーダンス学習を提案している．著者らは，ベイジアンネットワークを用いて，物体・動作・効果の関係性をモデル化している．しかし，提案されているモデルの構造は非常にシンプルであるため，本論文で扱う複雑な概念構造を表現するのは困難である．また，扱う動作は固定されており，ロボットが新規な動きの概念を学ぶことができないという問題もある．これは，与えられた概念間の関係性のみを学習していることに相当していると言える．これに対して本論文では，センサ入力から動きや物体の概念を獲得すると同時にそれら概念間の関係性も同時に学習する枠組みとなっている点で大きく異なる．

コンピュータビジョンの分野では，human-object interaction (HOI) なる考え方が提案されている [17, 18] ．これは，人間の動作の認識には使用されている物体が何かということが手掛かりとなると同時に，物体を認識する際に人間の動作や姿勢が重要になるという考え方である．つまり，HOI をモデルに組み込むことで，物体検出およびヒトの姿勢推定の性能を大幅に向上させることができる．しかしこれらの研究は，教師あり学習であり，本論文で扱う教師なしの学習問題とは大きく方向性が異なる．

1.3 本論文の構成

本論文は，以下のように構成されている．まず第 2 章で基本的な原理を述べ，第 3 章では提案手法を述べる．第 4 章で提案モデルによる実験，および考察を述べ，第 5 章で本論文をまとめる．

第 2 章

理論

2.1 DSIFT

DSIFT(Dense Scale Invariant Feature Transform) は，スケールや照明・回転等に対して普遍な局所的な視覚特徴量で，画像に対して同サイズ，同方向に特徴量のサンプリングを行うことで，記述子の計算を高速に行う手法である．まず入力画像に対して，ビンの中心点（キーポイント）が画像境界における座標上で常に整数値を取ることを制約条件とし，

$$\begin{aligned} \{0,\ldots,W-1\} \ni T_x + m\sigma x_i &= T_x + m\sigma i - \frac{N_x - 1}{2} \\ &= \bar{T}_x + m\sigma i, \qquad (i = 0,\ldots,N_x - 1) \end{aligned} \tag{2.1}$$

を満たすキーポイントのサンプリングを行う．式 (2.1) において，T はビンの中心点，$\bar{T}$ は左上のビンの中心点を表しており，これが整数値となるため，式 (2.1) の制約条件は

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \leq \bar{T} = \begin{bmatrix} \bar{T}_x^{\min} + p\Delta_x \\ \bar{T}_y^{\min} + q\Delta_y \end{bmatrix} \leq \begin{bmatrix} W - 1 - m\sigma N_x \\ H - 1 - m\sigma N_y \end{bmatrix}, \ \bar{T} = \begin{bmatrix} T_x - \frac{N_x - 1}{2} \\ T_y - \frac{N_y - 1}{2} \end{bmatrix} \tag{2.2}$$

と置くことができる．これは左上のビン $\bar{T}$ を基準として，整数のグリッドを想定し，図 2.1 のように以降のビン T をサンプリングすることを意味している．ビンのサイズ，方向，サンプリングステップを任意に変化させることで，制約条件を満たしながらキーポイントのサンプリングを行うことが可能となる．

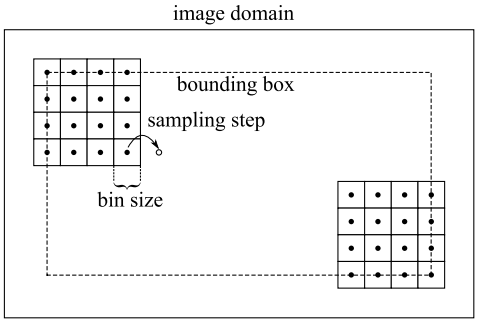

図 2.1: DSIFT 記述の幾何

次に，制約条件のもと抽出した各特徴点に対して，周辺の 4×4 ピクセルを使用して，SIFT 記述子の計算を行う．実際のキーポイントは

$$\mathbf{x} = m\sigma\hat{\mathbf{x}} + T \tag{2.3}$$

$$h(t,i,j) = m\sigma \int g_{\sigma_{sin}}(\mathbf{x}-T)\omega_{ang}(\angle J(\mathbf{x}) - \theta_t)\omega(\tfrac{x-T_x}{m\sigma} - \hat{x_i})\omega(\tfrac{y-T}{m\sigma} - \hat{y_i})\,|J(\mathbf{x})|d\mathbf{x} \tag{2.4}$$

に従ってサンプリングされており，窓関数を用いたビニング処理によって，以下のように近似を行うことができる．

$$T' = T + m\sigma \begin{bmatrix} x_i \\ y_i \end{bmatrix} \tag{2.5}$$

$$h(t,i,j) = m\sigma \int g_{\sigma_{sin}}(T' - \mathbf{x} - \mathbf{x}_{ij})\omega_{ang}(\angle J(\mathbf{x}) - \theta_t)\omega(\tfrac{T'_x - x}{m\sigma}\omega(\tfrac{T'_y - y}{m\sigma})\,|J(\mathbf{x})|d\mathbf{x} \tag{2.6}$$

ここで，

$$k_i(x) = \frac{1}{\sqrt{2\pi}\sigma_{win}} exp(-\frac{1}{2}\frac{(x-x_i)^2}{{\sigma_{win}}^2})\omega(\frac{x}{m\sigma}) \tag{2.7}$$

$$k_j(y) = \frac{1}{\sqrt{2\pi}\sigma_{win}} exp(-\frac{1}{2}\frac{(y-y_j)^2}{{\sigma_{win}}^2})\omega(\frac{y}{m\sigma}) \tag{2.8}$$

とおくと，式 (2.6) は

$$\bar{J_t}(\mathbf{x}) = \omega_{ang}(\angle J(\mathbf{x}) - \theta_t)\,|J(\mathbf{x})| \tag{2.9}$$

$$h\,(t,i,j) = \left(k_i k_j * \bar{J_t}\right)\left(T + m\sigma \begin{bmatrix} x_i \\ y_i \end{bmatrix}\right) \tag{2.10}$$

と記述できる．式 (2.10) に従って記述子の計算を行うことで，最終的に 128 次元の特徴量情報を得ることができる．

2.2 MFCC

MFCC(Mel-Frequency Cepstrum Coefficient) は，音声認識で一般的に用いられている音声特徴量である．提案手法では，物体を振った際の音を表す特徴量として使用した．

まず，音声から連続する数十 ms 程度の時間長の信号区間を切り出し，切り出された信号が定常確率過程に従うと仮定して，スペクトル解析を行う．すなわち，与えられた信号 $s(n)$ に長さ N の分析窓を掛けることで以下のように信号系列 $s_w(m;l)$ を取り出す．

$$s_w(m;l) = \sum_{m=0}^{N-1} w(m)s(l+m) \qquad (l = 0, T, 2T \dots) \tag{2.11}$$

ここで，添え字 l は，信号の切り出し位置に対応している．すなわち，l を一定間隔 T で増加させることで，定常と見なされる長さ N の音声信号系列 $s_w(n)(n = 0, \dots, N-T)$ が間隔 T で得られる．この処理はフレーム化処理と呼ばれ，N をフレーム長，T をフレーム間隔と呼ぶ．また，フレーム化処理を行う窓関数 $w(n)$ としては，ハミング窓 (式 (2.12)) やハニング窓 (式 (2.13)) がしばしば用いられている．

$$w(n) = 0.54 - 0.46cos(\frac{2n\pi}{N-1}) \qquad (n = 0, \dots, N-2) \tag{2.12}$$

$$w(n) = 0.5 - 0.5cos(\frac{2n\pi}{N-1}) \qquad (n = 0, \dots, N-2) \tag{2.13}$$

フレーム化処理によって得られた音声信号系列の短時間フーリエスペクトルは，離散時間フーリエ変換 (DTFT) により以下で与えられる．

$$S(e^{j\omega}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\omega n} \tag{2.14}$$

実際の信号処理過程では，離散フーリエ変換 (DFT) をその高速算法である FFT を用いて実行し，当該音声区間のスペクトル表現とすることが一般的である．すなわち

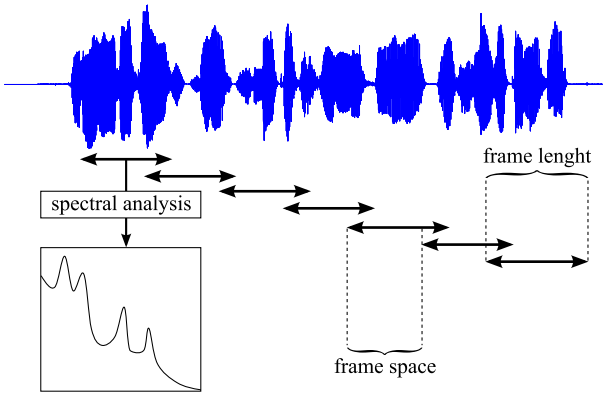

図 2.2: フレーム化処理

$$S'(k) = S(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\frac{2\pi}{N}kn} \qquad (k = 0, \ldots, N-1) \tag{2.15}$$

なる複素系列 $S'(k)$ が音声のスペクトル表現として最も一般的に用いられる．フーリエ変換の性質から，実数系列として与えられた $s_w(n)$ のスペクトル対称性 $s_w(n)$ のスペクトルには

$$\mathrm{Re}[S'\{(-k)_{modN}\}] = \mathrm{Re}[S'(k)] \tag{2.16}$$

$$\mathrm{Im}[S'\{(-k)_{modN}\}] = \mathrm{Im}[S'(k)] \tag{2.17}$$

が成り立つ．このため，保持すべきスペクトル情報は，長さ $N/2+1$ の複素数系列である．音声信号の音素的な特徴は主として調音フィルタの振幅伝達特性に含まれる．したがって，音声信号の振幅スペクトル，あるいはその 2 乗値であるパワースペクトルが注目すべきスペクトル表現である．音声信号のパワースペクトル系列は，離散スペクトル系列から

$$|S'(k)|^2 = \mathrm{Re}S'(k)^2 + \mathrm{Im}S'(k)^2 \tag{2.18}$$

の手順により計算される．実際の音声をフレーム処理した後，パワースペクトル系列を計算する手順を図 2.2 に示す．

MFCC の計算では，周波数軸上に L 個の三角窓を配置したフィルタバンク分析により行う (図 2.3)．すなわち，窓の幅に対応する周波数帯域の信号のパワーを，単

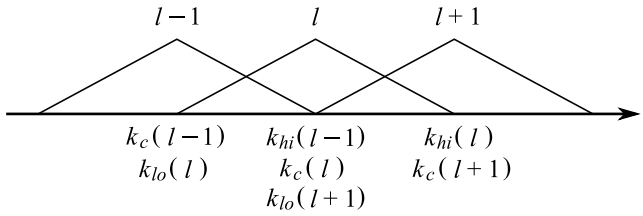

図 2.3: MFCC 分析用フィルタバンク

一スペクトルチャンネルの振幅スペクトル $|S'(k)|$ の重みづけの和で求める．

$$m(l) = \sum_{k=lo}^{hi} W(k;l)|S'(k)| \qquad (l = 1, \ldots, L) \tag{2.19}$$

$$W(k:l) = \begin{cases} \frac{k-k_{lo}(l)}{k_c(l)-k_{lo}(l)} & k_{lo} \le k \le k_c(l) \\ \frac{k_{hi}(l)-k}{k_{hi}(l)-k_c(l)} & k_c \le k \le k_{hi}(l) \end{cases} \tag{2.20}$$

ただし，$k_{lo}(l)$，$k_c(l)$，$k_{hi}(l)$ は，それぞれ l 番目のフィルタの下限，中心，上限のスペクトルチャンネルの番号であり，隣り合うフィルタ間で

$$k_c(l) = k_{hi}(l-1) = k_{lo}(l+1) \tag{2.21}$$

なる関係がある．さらに，$k_c(l)$ はメル周波数軸上で等間隔に配置される．メル周波数は

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2.22}$$

により計算される．ただし f の単位は [Hz] にとる．

最終的に，フィルタバンク分析により得られた L 個の帯域におけるパワーを離散コサイン変換することで，MFCC が求められる．

$$c_{mfcc}(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log m(l) \cos\left\{\left(l - \frac{1}{2}\right)\frac{i\pi}{L}\right\} \tag{2.23}$$

2.3 Latent Dirichlet Allocation

LDA (Latent Dirichlet Allocation) は，ひとつの文書が複数のトピックを含むことを表現できる確率的な文書モデルのひとつであり，図 2.4 のように表される．文

書 d のトピック (カテゴリ)z は，ディリクレ事前分布 α によって決まる，多項分布 θ によって決定され，文書内の単語 W は，同様にディリクレ事前分布 β によって決まる，多項分布 ϕ から発生するグラフィカルモデルとなる．

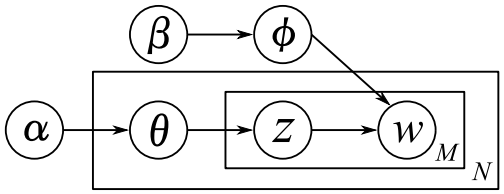

図 2.4: LDA のグラフィカルモデル

モデルにおいて，各文書は，その文書の長さだけの空欄で構成されていると考える．各文書を構成する空欄の各々について，まず，文書毎に異なる多項分布に従ってトピックを一つ選んで割り当て，次に，そのトピックに依存して決定される多項分布を用いて語彙を選択し，その空欄を満たすことで文書全体が構成される．

以降 LDA モデルを取り扱う際，語彙集合を $\mathcal{V} = \{v_1, \ldots, v_M\}$ とし，i 番目の文書の l 番目に v_j という語彙が表れていることを，$w_{il} = v_j$ という確率変数で表記する．また，i 番目の文書での語彙の現れ方を $\mathbf{w}_i$ という確率変数で表記し，これらを全文書についてまとめて $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_N\}$ という確率変数で表わすこととする．同様に，トピックの集合を $\mathcal{T} = \{t_1, \ldots, t_K\}$ とし，i 番目の文書の l 番目の空欄にトピック t_k が割り当てられていることを，$z_{il} = t_k$ という確率変数で表記する．i 番目の文書での各空欄へのトピックの割り当てを $\mathbf{z}_i$ という確率変数で表記し，これらを全文書についてまとめて $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ という確率変数で表わす．

つまり，i 番目の文書を構成する空欄にトピック t_k が割り当てられる確率を θ_{ik}，トピック t_k が割り当てられた空欄が語彙 w_j によって満たされる確率を ϕ_{kj} と書くことができる．これらの多項分布のパラメータ θ_{ik}，ϕ_{kj} には，それぞれ，全ての $i = 1, \ldots, N$ について $\sum_{k=1}^{K} \theta_{ik} = 1$ が，全ての $k = 1, \ldots, K$ について $\sum_{j=1}^{M} \phi_{kj} = 1$ が成り立つ．すなわち，LDA による学習とは，ディリクレ事前分布 (ハイパーパラメータ $\alpha = (\alpha_1, \ldots, \alpha_K)$) に従って決定される，文書におけるトピックの出現確率分布を表わす多項分布のパラメータ θ 及び，ディリクレ事前分布 (ハイパーパラメータ $\beta = (\beta_1, \ldots, \beta_M)$) に従って決定される，トピックが割り当てられた空欄に

おける語彙の出現確率分布を表わす多項分布のパラメータ ϕ を，観測可能な変数から推定することになる．

以上の変数を用いた LDA モデルでの学習については次節の Gibbs Sampling で詳しく説明する．

2.4 Gibbs Sampling

Gibbs Sampling とは，確率分布からサンプルを得る際に用いられるマルコフ連鎖モンテカルロ法の一種であり，サンプリングによって，目標の確率分布を近似的に求める手法である．

2.3 節で示した LDA モデルを Gibbs Sampling の原理をもとに式で表現すると，ベイズの定理より下記のようになる．

$$\begin{aligned} P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) &= P(\phi|\beta) \prod_{i=1}^{N} P(\theta_i|\alpha) P(\mathbf{z}_i|\theta_i) P(\boldsymbol{w}_i|\mathbf{z}_i, \phi) \\ &= \left\{ \prod_{k=1}^{K} \frac{\Gamma(W\beta_j)}{\prod_j \Gamma(\beta_j)} \prod_{j=1}^{M} \phi_{kj}^{\beta_j - 1} \right\} \\ &\quad \cdot \prod_{i=1}^{N} \left[\left\{ \frac{\Gamma(T\alpha_k)}{\prod_k \Gamma(\alpha_k)} \theta_{ik}^{\alpha_k - 1} \right\} \prod_{k=1}^{K} \prod_{j=1}^{M} (\theta_{jk}\phi_{kj})^{N_{ijk}} \right] \end{aligned} \tag{2.24}$$

N_{ijk} は，i 番目の文書において，語彙 v_j で満たされた空欄に，トピック t_k が割り当てられた回数，W は語彙の次元数，T はトピックの次元数と定義した．

ここで，N_{ik} を i 番目の文書中でトピック t_k が割り当てられた空欄の個数，N_{jk} を全文書集合中で，語彙 v_j で満たされた空欄のうちトピック t_k が割り当てられている空欄の個数として表記すると，式 (2.24) は

$$P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = \prod_{i=1}^{N} \frac{\Gamma(T\alpha_k)}{\prod_k \Gamma(\alpha_k)} \theta_{ik}^{N_{ik}+\alpha_k-1} \prod_{k=1}^{K} \frac{\Gamma(W\beta_j)}{\prod_j \Gamma(\beta_j)} \prod_{j=1}^{M} \phi_{kj}^{N_{jk}+\beta_j-1} \tag{2.25}$$

と書き直される．この時，式 (2.25) の右辺は二つの部分に分けることができ，

$$\begin{aligned} &P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta) = \\ &\qquad \left\{ \prod_{i=1}^{N} \frac{\Gamma(T\alpha_k)}{\prod_k \Gamma(\alpha_k)} \theta_{ik}^{N_{ik}+\alpha_k-1} \right\} \left\{ \prod_{k=1}^{K} \frac{\Gamma(W\beta_j)}{\prod_j \Gamma(\beta_j)} \prod_{j=1}^{M} \phi_{kj}^{N_{jk}+\beta_j-1} \right\} \end{aligned} \tag{2.26}$$

となる．

ベイズ推定を行うため，θ，ϕ で積分すると，

$$\begin{aligned} P(\boldsymbol{w}, \mathbf{z}|\alpha, \beta) &= \int P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta)\, d\theta d\phi \\ &= \int \prod_{i=1}^{N} \frac{\Gamma(T\alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{ik}^{N_{ik}+\alpha_k-1} d\theta \\ &\qquad \cdot \int \prod_{k=1}^{K} \frac{\Gamma(W\beta_j)}{\prod_j \Gamma(\beta_j)} \prod_{j=1}^{M} \phi_{kj}^{N_{jk}+\beta_j-1} d\phi \\ &= \prod_{i=1}^{N} \left\{ \frac{\Gamma(T\alpha_k)}{\prod_k \Gamma(\alpha_k)} \cdot \frac{\prod_k (\Gamma(N_{ik}) + \alpha_k)}{\Gamma(N_{i\cdot} + T\alpha_k)} \right\} \\ &\qquad \cdot \prod_{k=1}^{K} \left\{ \frac{\Gamma(W\beta_j)}{\prod_j \Gamma(\beta_j)} \cdot \frac{\prod_j (\Gamma(N_{jk}) + \beta_j)}{\Gamma(N_{\cdot k} + W\beta_j)} \right\} \end{aligned} \tag{2.27}$$

となる．

ここで，仮に i 番目の文書に，0 番目の空欄としてひとつ空欄を付け加えることを考える．この空欄にトピック t_k が割り当てられ，またそれが語彙 w_j によって満たされる確率 $P(z_{i0} = t_k, w_{i0} = v_j|\boldsymbol{w}, \mathbf{z}, \alpha, \beta)$ をを求めると，以下のように表わすことができる．

$$\begin{aligned} & P(z_{i0} = t_k, w_{i0} = v_j|\boldsymbol{w}, \mathbf{z}, \alpha, \beta) \\ = & \int P(z_{i0} = t_k, w_{i0} = v_j|\theta, \phi)\, P(\theta, \phi|\boldsymbol{w}, \mathbf{z}, \alpha, \beta)\, d\theta d\phi \\ = & \int P(z_{i0} = t_k|\theta)\, P(w_{i0} = v_j|\mathbf{z}_{i0} = t_k, \phi)\, P(\theta, \phi|\boldsymbol{w}, \mathbf{z}, \alpha, \beta)\, d\theta d\phi \\ = & \int P(z_{i0} = t_k|\theta)\, P(w_{i0} = v_j|\mathbf{z}_{i0} = t_k, \phi)\, \frac{P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta)}{P(\boldsymbol{w}, \mathbf{z}|\alpha, \beta)} d\theta d\phi \\ = & \frac{1}{P(\boldsymbol{w}, \mathbf{z}|\alpha, \beta)} \int P(z_{i0} = t_k|\theta)\, P(w_{i0} = v_j|z_{i0} = t_k, \phi)\, P(\boldsymbol{w}, \mathbf{z}, \theta, \phi|\alpha, \beta)\, d\theta d\phi \end{aligned} \tag{2.28}$$

式 (2.28) において，$P(z_{i0} = t_k|\theta)$ および $P(w_{i0} = v_j|z_{i0} = t_k, \phi)$ は，それぞれ定義

された多項分布のパラメータ θ_{ik}，ϕ_{kj} を意味している．従って式 (2.28) は

$$
\begin{aligned}
&P\left(z_{i0}=t_k, w_{i0}=v_j|\boldsymbol{w},\mathbf{z},\alpha,\beta\right)\\
&\quad=\frac{1}{P\left(\boldsymbol{w},\mathbf{z}|\alpha,\beta\right)}\int\theta_{ik}\phi_{kj}\left\{\prod_{i'=1}^{N}\frac{\Gamma\left(T\alpha_k\right)}{\prod_{k'}\Gamma\left(\alpha_{k'}\right)}\prod_{k'=1}^{K}\theta_{ik'}^{N_{i'k'}+\alpha_{k'}-1}\right\}\\
&\qquad\qquad\cdot\left\{\prod_{k'=1}^{K}\frac{\Gamma\left(W\beta_j\right)}{\prod_{j'}\Gamma\left(\beta_{j'}\right)}\prod_{j'=1}^{M}\phi_{kj'}^{N_{j'k'}+\beta_{j'}-1}\right\}d\theta d\phi\\
&\quad=\frac{1}{P\left(\boldsymbol{w},\mathbf{z}|\alpha,\beta\right)}\int\theta_{ik}\left\{\prod_{i'=1}^{N}\frac{\Gamma\left(T\alpha_k\right)}{\prod_{k'}\Gamma\left(\alpha_{k'}\right)}\prod_{k'=1}^{K}\theta_{ik'}^{N_{i'k'}+\alpha_{k'}-1}\right\}d\theta\\
&\qquad\qquad\cdot\int\phi_{kj}\left\{\prod_{k'=1}^{K}\frac{\Gamma\left(W\beta_j\right)}{\prod_{j'}\Gamma\left(\beta_{j'}\right)}\prod_{j'=1}^{M}\phi_{kj'}^{N_{j'k'}+\beta_{j'}-1}\right\}d\phi\\
&\quad=\frac{\prod_{i'}\left\{\frac{\Gamma(T\alpha_k)}{\prod_{k'}\Gamma(\alpha_{k'})}\cdot\frac{\prod_{k'}\Gamma(N_{i'k'}+\alpha_{k'}+\Delta(i'=i\wedge k'=k))}{\Gamma(N_{i'\cdot}+T\alpha_k+\Delta(i'=i))}\right\}}{\prod_{i'}\left\{\frac{\Gamma(T\alpha_k)}{\prod_{k'}\Gamma(\alpha_{k'})}\cdot\frac{\prod_{k'}\Gamma(N_{i'k'}+\alpha_{k'})}{\Gamma(N_{i'\cdot}+T\alpha_k)}\right\}}\\
&\qquad\qquad\cdot\frac{\prod_{k'}\left\{\frac{\Gamma(W\beta_j)}{\prod_{j'}\Gamma\left(\beta_{j'}\right)}\cdot\frac{\prod_{k'}\Gamma\left(N_{j'k'}+\beta_{j'}+\Delta(j'=j\wedge k'=k)\right)}{\Gamma(N_{\cdot k'}+W\beta_j+\Delta(k'=k))}\right\}}{\prod_{k'}\left\{\frac{\Gamma(W\beta_j)}{\prod_{j'}\Gamma\left(\beta_{j'}\right)}\cdot\frac{\prod_{k'}\Gamma\left(N_{j'k'}+\beta_{j'}\right)}{\Gamma(N_{\cdot k'}+W\beta_j)}\right\}}\\
&\quad=\frac{N_{ik}+\alpha_k}{N_{i\cdot}+T\alpha_k}\cdot\frac{N_{jk}+\beta_j}{N_{\cdot k}+W\beta_j}
\end{aligned}
\tag{2.29}
$$

と表すことができる．なお，式 (2.29) 中の $\Delta()$ は () 内の命題が成立する時のみ 1 となり，成立しない場合 0 となるものとする．

得られた結果は，二つの項の積として書かれている．前の項は，i 番目の文書に新たに付け加えられた空欄に，トピック t_k が割り当てられる確率であり，後の項は，空欄にトピック t_k が割り当てられてことに対応して，語彙 v_j が選択される確率である．つまり，前者は，ある文書を構成する空欄に割り当てられるトピックの予測分布

$$
P\left(z_{i0}=t_k|\boldsymbol{w},\mathbf{z},\alpha,\beta\right)=\frac{N_{ik}+\alpha_k}{N_{i\cdot}+T\alpha_k} \tag{2.30}
$$

を，後者は，トピックに割り当てられた空欄を満たす語彙の予測分布

$$
P\left(w_{i0}=v_j|z_{i0}=t_k,\boldsymbol{w},\mathbf{z},\alpha,\beta\right)=\frac{N_{jk}+\beta_j}{N_{\cdot k}+W\beta_j} \tag{2.31}
$$

を表わしており，それぞれ $P(z_{i0} = t_k|\boldsymbol{w}, \mathbf{z}, \alpha, \beta)$，$P(w_{i0} = v_j|z_{i0} = t_k, \boldsymbol{w}, \mathbf{z}, \alpha, \beta)$ が推定を行うべき多項分布のパラメータ θ，ϕ に他ならない．

これまでは，与えられた文書集合のうちある文書に単語が付け加わったと仮定したが，今度は，与えられた文書集合のうちのひとつの文書から，見えていた単語がひとつ見えなくなり空欄が生じたと仮定する．この空欄は，i 番目の文書の l 番目の位置で生じたとし，その空欄にトピック t_k が割り当てられ，その空欄が語彙 v_j で満たされる確率を求める．

計算方法は式 (2.29) と同様で，

$$\begin{aligned} &P(z_{il} = t_k, w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta) \\ &\quad = \frac{N_{ik} - 1 + \Delta(k \neq k') + \alpha_k}{N_{i\cdot} - 1 + T\alpha_k} \cdot \frac{N_{jk} - 1 + \Delta(k \neq k') + \beta_j}{N_{\cdot k} - 1 + \Delta(k \neq k') + W\beta_j} \end{aligned} \tag{2.32}$$

となる．なお $\boldsymbol{w}_{-il}$ は，$\boldsymbol{w}$ から w_{il} を取り除いた残りであり，$\mathbf{z}_{-il}$ は，$\mathbf{z}$ から z_{il} を取り除いた残りである．また取り除いた一語に割り当てられていたトピックが $t_{k'}$ だったとしている．

Gibbs Sampling を行う場合，完全条件付き確率 $P(z_{il} = t_k|\boldsymbol{w}, \mathbf{z}_{-il}, \alpha, \beta)$ が必要となる．これは

$$\begin{aligned} P(z_{il} = t_k|\boldsymbol{w}, \mathbf{z}_{-il}, \alpha, \beta) &= P(z_{il} = t_k|w_{il} = v_j, \boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta) \\ &= \frac{P(z_{il} = t_k, w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta)}{P(w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta)} \\ &= \frac{P(z_{il} = t_k, w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta)}{\sum_{k'=1}^{K} P(z_{il} = t_{k'}, w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta)} \end{aligned} \tag{2.33}$$

と計算することができる．ここで式 (2.33) 右辺の分母は k に依存しないので

$$P(z_{il} = t_k|\boldsymbol{w}, \mathbf{z}_{-il}, \alpha, \beta) \propto P(z_{il} = t_k, w_{il} = v_j|\boldsymbol{w}_{-il}, \mathbf{z}_{-il}, \alpha, \beta) \tag{2.34}$$

という単純な比例関係が成り立つ．よって

$$\begin{aligned} &P(z_{il} = t_k|\boldsymbol{w}, \mathbf{z}_{-il}, \alpha, \beta) \\ &\quad \propto \frac{N_{ik} - 1 + \Delta(k \neq k') + \alpha_k}{N_{i\cdot} - 1 + T\alpha_k} \cdot \frac{N_{jk} - 1 + \Delta(k \neq k') + \beta_j}{N_{\cdot k} - 1 + \Delta(k \neq k') + W\beta_j} \end{aligned} \tag{2.35}$$

を得ることができる．

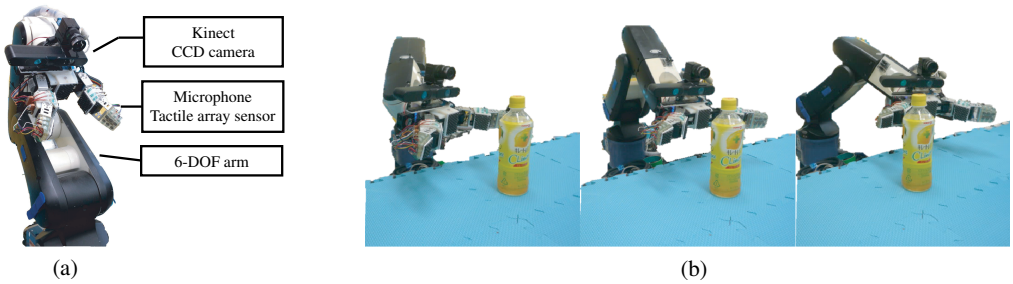

図 2.5: (a) 固定型単腕アームロボット　(b) 視覚情報の取得

式 (2.35) を用いて，各文書の各単語について，トピック割り当ての確率分布を求め，求められた分布を用いてトピックを一つ選択し，その単語へトピックを当て直す．この割り当てを反映するように N_{ik}，N_{jk}，$N_{\cdot k}$ を計算し直し，その結果を，さらに別の単語についてトピック割り当ての確率分布を得るために用いる．以上の過程を，結果が収束するまで繰り返し行うプロセスが Gibbs Sampling である．この収束結果から，改めて式 (2.30)，式 (2.31) を用いて計算を行うことで，推定すべき多項分布のパラメータ θ，ϕ を算出することが可能となる．

2.5　マルチモーダル情報の取得

本研究では，先述のマルチモーダル LDA で用いる情報の取得については先行研究の手法を用いた [7]．その具体的な手法について説明する．まず家庭用ロボットは，室内の机や棚を順に移動し，自律的に物体を検出する必要がある．但し，本研究ではマルチモーダル情報の取得に主眼を置いているため，ロボットの移動は行わないものとし，ロボットが学習するマルチモーダル情報として視覚情報，触覚情報，聴覚情報の 3 種類を想定する．視覚情報として，CCD(Charge Coupled Device) カメラ及び赤外線深度カメラによる複数視点からの画像 (色) 情報 [25]，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ー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．

ロボットが完全に自律的に未知物体のマルチモーダル情報を取得するには，(1)

(a)　　(b)

図 2.6: (a) 触覚情報の取得　(b) 聴覚情報の取得

未知物体の発見をどのように行うか，(2) 未知物体をどのように把持するか，(3) 未知物体のマルチモーダル情報をどのように観測するか，といった 3 つの問題が考えられる．一つ目の問題に関しては，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 [25] により解決する．また，この時点で物体認識を行い，検出した物体が既知ものであれば次の物体に対象を移す．一方物体が未知であれば，情報取得のフェイズに移行する．二つ目の問題には，視覚センサーによって得られる 3 次元情報をもとに把持位置や姿勢を決定することで対処する．三つ目の問題で特に重要なのは，視覚情報の取得方法である．これは，物体を複数方向から観測しておく必要があるためであり，そのためにはロボットが物体を把持して様々な方向から見る必要がある．しかし，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がある．そこで本論文では，一度検出した物体を把持し，アーム稼働領域の中心付近へ一度物体を移動させ，アームを物体に沿って周囲を移動させることで，複数視点からの観測を実現する．ロボットは，Affine Scale Invariant Feature Transform（ASIFT）を利用して物体のトラッキングを行いながら，物体の色・テクスチャ情報を取得する．実際にロボットが物体を複数の視点から視覚情報を取得している様子を図 2.5(b) に示す．

触覚情報は，前述した平面検出により未知物体を検出し，図 2.6(a) に示すような把持動作を 1 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ーの出力を取得する．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，物体を振る際に発生する音を取得する．ここで問題となるのは，腕を振ることによるモータ等のノイズの影響である．特に物体が音を鳴らすためには，かなりのスピードで振る必要があり，ノイズの影響は無視することができない．この問題を解決するために，ノイズ除去などの手法をとることも考えられる

が，ここでは何も持たずに腕を振った際の音を同様に取得しておき，特徴量のレベルでノイズを考慮することとする．図 2.6(b) にロボットが実際に音声取得を行っている様子を示す．

2.6 Bag of words モデル

自然言語処理では，LDA を用いて与えられた大量のテキストデータから意味のまとまりであるトピックを教師なしで見つけ出す手法が研究されている [10]．一つの重要な考え方は，トピックとは単語の出現頻度のパターンで定義されるというもので，Bag of words モデルと呼ばれる．従ってトピックは，単語の出現位置や順序に関係なく，その頻度を基にモデル化される．

LDA のモデル (図 2.4) の場合，文書中の i 番目の単語 w_i が局所的な特徴量に，トピック z がカテゴリに対応する．モデルの式は 2.3 節に記述したように

$$P(\mathbf{W}, \mathbf{Z}, \theta, \phi|\alpha, \beta) = P(\phi|\beta) \prod_{i=1}^{N} P(\theta_i|\alpha) P(\mathbf{z}_i|\theta_i) P(\mathbf{w}_i|\mathbf{z}_i, \phi) \quad (2.36)$$

で表わすことができる．本稿では，以降これらを視聴覚及び触覚を用いたマルチモーダルなカテゴリゼーションに拡張する．

2.7 マルチモーダルカテゴリゼーション

ロボットは物体を掴み，提案する情報取得システムの利用により，様々な角度からこれを観察することが可能である．ここでは，同一の物体を観測している間に得られる視覚情報，聴覚情報，触覚情報を位相情報を考慮することなく生起回数の情報として利用する．これは，各特徴量を「単語」と考えれば前述の Bag of words モデルであり，本論文ではこれをマルチモーダル情報に適用することで物体のカテゴリゼーションを行う．ここではこれを，マルチモーダル Bag of features モデルと呼ぶ．

図 2.7 にロボットによるマルチモーダルカテゴリゼーションシステムの概要を示す．ロボットはカメラ・マイク・アーム・ハンド・感圧センサを備えており，様々な物体を掴み，物体観測用テーブルを用いて観察する．その間に得られる，画像情

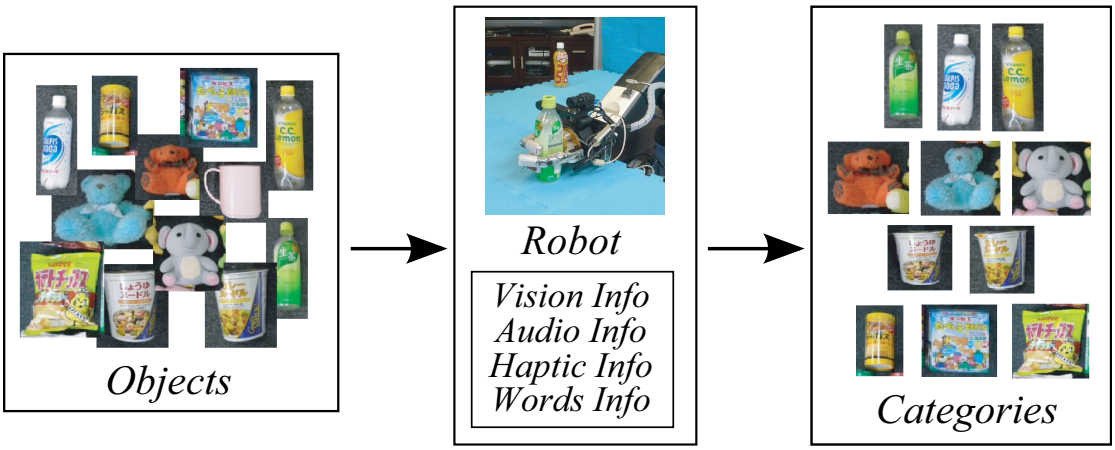

図 2.7: ロボットシステムの概要

報，音情報，触覚情報を用い，物体の性質の類似性から物体を分類する．この際物体の性質とは，物体の見た目や振った際の音，物体の硬さを意味している．

図 2.8 に提案するマルチモーダルカテゴリゼーション LDA におけるグラフィカルモデルを示す．w^v，w^a，w^h はそれぞれ視覚，聴覚，触角情報を示しており，それぞれハイパーパラメータ ϕ^v，ϕ^a，ϕ^h 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 の多項分布によって発生する．各情報については後述する．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．

図 2.8 から分かるように，これらのモデルでは各センサ情報は独立に出力される．つまりカテゴリ z が決まった場合，各センサ情報は他の情報とは無関係に決まることになる．しかし実際には，各センサ情報同士には何らかの関係性があると考えられ，例えばある視覚情報には，ある特定の音や硬さが関係している可能性がある．但しこれらの関係性を正しく捉えるためには，非常に精細なセンサ情報を得る必要があり，またグラフィカルモデルが複雑になることによる，学習や推論のための計算が複雑化が問題として挙げられる．そこで本論文では，図 2.8 のように各モーダル情報が独立なモデルを用いることとする．

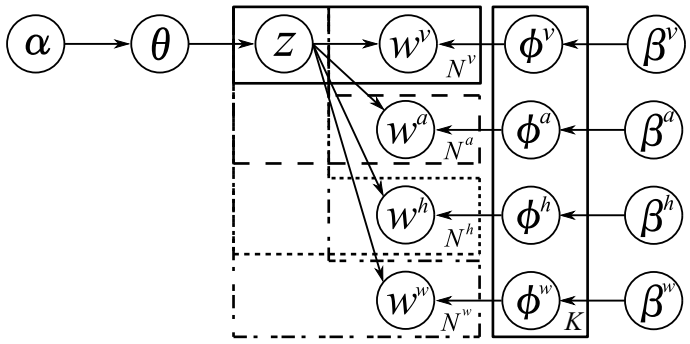

図 2.8: マルチモーダル LDA のグラフィカルモデル

2.8 マルチモーダル情報処理

この項ではロボットが取得した各知覚情報の処理について述べる．

2.8.1 視覚情報

ロボットは頭部にカメラを搭載しており，物体観測用テーブルを用いて様々な方向から物体を観察することで取得する画像を視覚情報として利用する．画像は，各物体毎に複数枚取得する (後に示す実験では，各物体に対して 7 枚の画像を用いた)．本稿では，各画像から特徴量として，色・テクスチャ情報，距離情報，反射強度情報を取得する．色情報としては照明変化の影響を受けにくい HSV 表色系の H（色相）と S（彩度）のヒストグラムを，テクスチャ情報として 128 次元の SIFT 記述子 [27] を用いる．これによって得られる特徴量は，回転や拡大縮小・照明の変化等に対する不変性を持ち，物体を様々な視点から観測する際の特徴量として優れている．また，位相情報を用いないため，オクルージョンの問題を回避することができる．

図 2.9 にロボットが同物体から取得した，CCD 画像情報，距離画像情報，反射強度情報，これらの情報から得られる CCD 画像を 3 次元情報にマッピングした画像情報を示す．

次に具体的な視覚情報の取得について述べる．まず前述した SIFT 記述子を用いた DSIFT により，1 枚の画像あたり約 300〜400 個程度の特徴ベクトルを得ること

(a) (b) (c) (d)

図 2.9: 取得された視覚情報の例 (左から CCD カメラの画像，距離画像，反射強度，CCD の画像を 3 次元情報にマッピングした画像)

ができる．すなわち 1 つの物体の 10 枚の画像から約 3000〜4000 個の特徴ベクトルを得る．これらの特徴ベクトルは，学習画像とは全く関係のない背景画像（室内シーンの画像 100 枚）から計算した 500 の代表ベクトル（コードブック）を用いてベクトル量子化する．従って，画像特徴量 w^v は実際にはコードブックのインデックスを表すことになる．この画像特徴量のインデックスの発生回数によりヒストグラムを作成する．従って最終的に，１つの物体につき 500 次元のヒストグラムが得られることになる．このヒストグラムのインデックスが Bag of words モデルの単語に相当し，ヒストグラムがその生起回数を表している．

2.8.2 聴覚情報

聴覚情報として，ロボットが実際に物体を掴み振ることで発生した音を指先に取り付けたマイクより取得する．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，フレーム毎の特徴量に変換する．なお後述する実験では，0.2[s] ごとのフレームで音声信号を分割し，FFT(Fast Fourier Transform) 及び，MFCC(Mel-Frequency Cepstrum Coefficient) 情報を取得した．実際にロボットが物体を振った際に取得した FFT，MFCC 情報を図 2.10 に示す．FFT 情報は横軸が周波数 [Hz]，縦軸が時間 [s] を，MFCC 情報は横軸がヒストグラムのインデックス，縦軸が時間 [s] を表わす．なお各情報はそれぞれ，物体を何も把持していない状態と，音が鳴る物体を把持した状態の 2 種類を示した．

特徴量としては，音声認識で最もよく使用されている MFCC を用いることとした．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，男女それぞれ 3 名の音声と 3 種類の雑音から計算した 50 の代表ベクトルを用

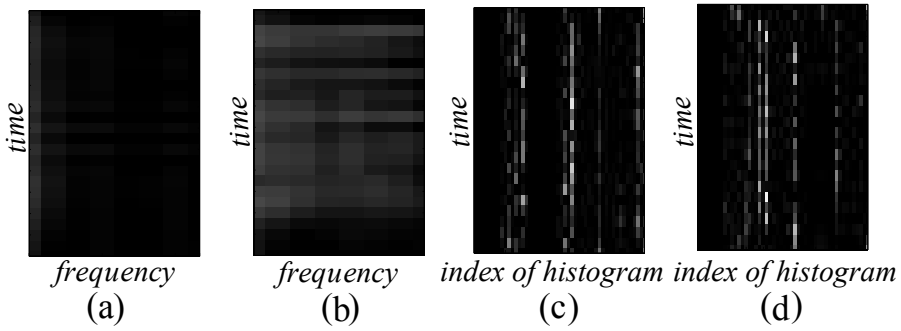

図 2.10: (a) 未把持状態での FFT 情報　(b) 物体把持状態での FFT 情報　(c) 未把持状態での MFCC 情報　(d) 物体把持状態での MFCC 情報

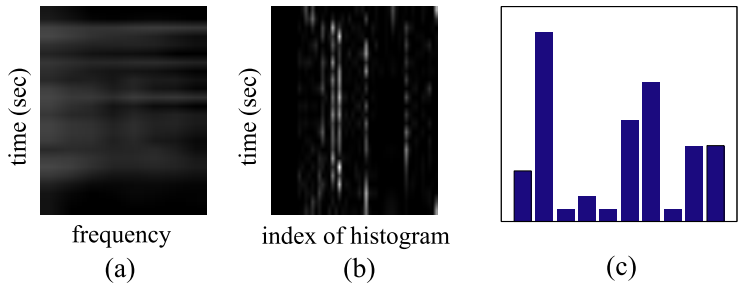

図 2.11: 聴覚情報における代表的な 6 次元ヒストグラム例

いてベクトル量子化する．従って，最終的に音声情報も，画像情報と同様に 50 次元のヒストグラムとなる．画像と同様，このヒストグラムのインデックスが Bag of words モデルの単語に相当し，ヒストグラムがその生起回数を表している．また，音声取得時の雑音を取り除くため，3.1 節で述べたように，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った．但し，ノイズ除去の際，MFCC 特徴量が引き算によって負になる場合は 0 とおくこととする．図 2.11 にヒストグラムの具体例を示す．但しこの図 2.11 は 50 次元のヒストグラムから，特徴的な 6 次元のみを表示したものである．

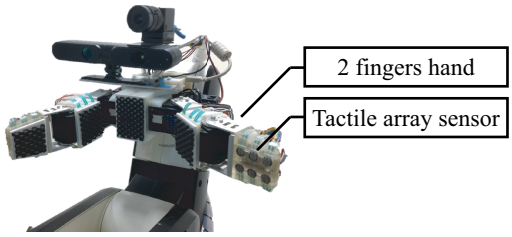

図 2.12: 実験に用いた触覚センサー付きのロボットハンド

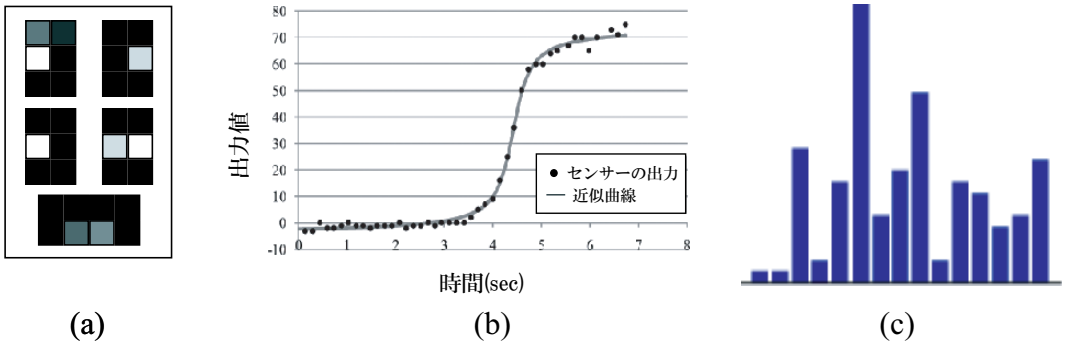

図 2.13: (a) 触角センサー出力　(b) センサー出力と近似曲線　(c) 触覚情報の 15 次元ヒストグラム

2.8.3　触覚情報

本稿の実験に際して触覚情報取得のために，図 2.12 に示すロボットハンドを使用した．このハンドは各指のトルク制御と角度のフィードバック及び，指先と根元，手の平部分に搭載された 32 個の感圧センサ情報を取得することが可能である．また物体の把持はトルク制御により，物体を一定速度で握り，負荷が一定値に達した時点でハンドを停止することとした．

触覚の特徴量としては，物体の把持を開始してからハンドが停止するまでの感圧センサーの出力値を用いることとした．実際に感圧センサーで圧力情報を取得した際の様子を図 2.13(a) に，把持開始から停止までの時系列における出力値データを図 2.13(b) に示す．なお図 2.13(b) において，横軸は時間 [sec]，縦軸は感圧センサーの出力値を表わす．この出力値は，ある決まった力で物体を掴んだ場合の物体からの応力，すなわち物体の硬さを表わしていると考えられる．また実際の把持作

業は，ロボットが自ら物体の位置を確認し，アームを制御し物体を掴むため，物体の把持位置や角度等にはばらつきが生じる．そこで学習に際には，各物体に対して 5 回分の把持を行うこととした．

しかし触覚アレイセンサーの情報はデータ数が多く，直接把持物体の特徴量を得ることが困難であるため，触覚情報として取り扱いやすくするために，各センサー出力値の時系列を以下の式で近似した．

$$p\left(t\right) = a\tan^{-1}\left(b\left(t+c\right)\right) + d \tag{2.37}$$

ただし，$p(t)$ は時刻 t でのセンサーの出力値である．センサーの出力値に対する近似曲線の例を図 2.13(b) に示す．この近似により，各素子から得られる時系列の変化を (a, b, c, d) という 4 つのパラメータで表現することができる．このうち，a は把持の際物体に掛かっている力に，b は物体に触れてから把持が停止するまでの時間に関係していると考えられる．また，c は把持する物体の大きさに，d は a に依存した値となるため，この 2 つの値を取り除き，(a, b) の 2 つのパラメータのみを，各センサーの情報として用いることとする．すなわち，1 回の把持で 32 組の (a, b) が得られることになる．さらにこの 2 次元の特徴ベクトルを予め計算した 15 の代表ベクトルを用いてベクトル量子化し，最終的に得られる 15 次元のヒストグラムを物体の触覚特徴量として扱う．触覚情報も同様に，ヒストグラムのインデックスが Bag of words モデルの単語に相当し，ヒストグラムがその生起回数を表しているため，多数・多次元の触覚情報から，把持位置や形状に依存しにくい特徴量として扱うことが可能である [28]．なお計算に用いる代表ベクトルは，多数の物体を把持することで得られる 2 次元の触覚特徴量を k 平均法により分類することで予め計算する．図 2.13(c) に触覚情報ヒストグラムの例を示す．

2.9　マルチモーダル LDA

既に述べたように，ここでのカテゴリゼーションの問題は，図 2.8 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．モデルのパラメータの学習は，与えられたデータに対して Gibbs sampling の原理を適応し，マルチモーダルに拡張することによって実現される．パ

ラメータの推定方法は，与えられたデータに対する目的関数を最大とする，EM アルゴリズムや変分ベイズ法等を利用することでも算出することができる．しかしこれらの手法の多くは，近似式の利用によりモデルの複雑化に対応できない，局所解に陥り易いという問題を孕んでおり，また計算の複雑化による処理時間の増加等の問題も挙げられる．そこで本研究では，近似式を用いず，計算過程も簡易である Gibbs Sampling の拡張によってパラメータの推定を行うこととする．マルチモーダル情報集合を $\mathbf{w^v}$, $\mathbf{w^a}$, $\mathbf{w^h}$ とし，モダリティのインデックスを m とすると，j 番目の物体の m 番目のモダリティ情報の i 番目に割り当てられるカテゴリ z_{mij} をサンプリングする式は以下のように書くことができる．

$$P\left(z_{mil}=k|\mathbf{W^m},\mathbf{Z^{-mil}},\alpha,\beta^m\right)\propto\left(n_{k,j}^{-mij}+\alpha\right)\cdot\frac{n_{m,w^m,k}^{-mij}+\beta^m}{n_{m,k}^{-mij}+W^m\beta^m} \tag{2.38}$$

ここで，W^m は m 番目のモーダル情報の次元数である．$n_{m,w^m,k,j}$ 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わしている．これは 2.4 節における式 (2.27) のマルチモーダルへの拡張であると言える．さらに，

$$n_{k,j} = \sum_{m,w^m} n_{m,w^m,k,j} \tag{2.39}$$

$$n_{mk} = \sum_{w^m,j} n_{m,w^m,k,j} \tag{2.40}$$

と置く．つまり，$n_{k,j}$ は j 番目の物体全ての情報に，カテゴリ k が割り当てられた回数を，$n_{m,k}$ は，全ての物体のモダリティ m の情報に，カテゴリ k が割り当てられた回数を表している．また，式 (2.38) 内の除算の添字は，その情報を除外していることを意味している．すなわち，$\mathbf{Z^{-mij}}$ は，j 番目の物体のモダリティ m の i 番目の情報へ割り当てられたカテゴリ z_{mij} を取り除いた残りである．マルチモーダルに拡張した Gibbs Sampling でも，基本的原理は 2.3 節及び 2.4 節に記述したものと同じである．実際のサンプリングは，式 (2.38) に従い，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行うことになる．これを繰り返すことで，n_* がある値へと収束する．最終的に，収束した結果から，パラメータの推定値 θ_{kj},

$\phi^m_{w^m k}$ は以下のようになる．

$$\theta_{kj} = \frac{n_{k,j}+\alpha}{n_j + K\alpha} \tag{2.41}$$

$$\phi^m_{w^m,k} = \frac{n_{m,w^m,k}+\beta^m}{n_{m,k}+W^m\beta^m} \tag{2.42}$$

ただし，K はカテゴリの総数である．式 (2.41)，式 (2.42) についても同様に，2.3 節における式 (2.30) 及び式 (2.31) のマルチモーダルへの拡張である．

2.10 カテゴリゼーションに基づく認識

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から，学習したパラメータを用いて，未知物体がそれぞれのカテゴリに属する確率を計算することになる．なお以下に述べる物体の認識において，視覚，聴覚，触覚の 3 種類の感覚情報による認識を行う．

未知物体のマルチモーダル情報 $\mathbf{w}^v_{obs}$,$\mathbf{w}^a_{obs}$,$\mathbf{w}^h_{obs}$ が与えられた場合，そのカテゴリは $P(z|\mathbf{w}^v_{obs},\mathbf{w}^a_{obs},\mathbf{w}^h_{obs})$ を最大とするカテゴリ z を選択すればよいこととなる．したがって，未知物体のカテゴリは，

$$\begin{aligned}\hat{z} &= \underset{z}{\operatorname{argmax}}\, P(z|\mathbf{w^v_{obs}},\mathbf{w^a_{obs}},\mathbf{w^h_{obs}}) \\ &= \underset{z}{\operatorname{argmax}} \int P(z|\theta)P(\theta|\mathbf{w^v_{obs}},\mathbf{w^a_{obs}},\mathbf{w^h_{obs}})\mathbf{d}\theta \end{aligned} \tag{2.43}$$

によって決めることができる．但し，$P(\theta|\mathbf{w^v_{obs}},\mathbf{w^a_{obs}},\mathbf{w^h_{obs}})$ は，学習時に推定した ϕ^v,ϕ^a,ϕ^h を固定し，前述の Gibbs Sampling を適用することで，α を再計算することにより求めることができる．

2.11 未観測情報の予測

これまで述べた提案手法により，ロボットはマルチモーダル情報を利用することで物体概念を構築し，その概念を通して未知物体に対するカテゴリの認識が可能となった．形成された概念による認識を行うことで，未知物体の未観測情報の予測を

行うことが可能となる．本研究では，ロボットが自律的に取得可能な視覚，聴覚，触覚情報を用いた単語情報の予測について考える．

図 2.8 のグラフィカルモデルにおいて，w^w は単語情報を表し，前述したように BoW モデルとして扱う．認識と同様に，未知物体のマルチモーダル情報 $\boldsymbol{w}^v_{obs}$，$\boldsymbol{w}^a_{obs}$，$\boldsymbol{w}^h_{obs}$ が与えられた場合，これらの情報について，$p(w^w|\boldsymbol{w}^{v,a,h}_{obs})$ を計算することで可能となる．すなわち

$$p(w^w|\boldsymbol{w}^{v,a,h}_{obs}) = \int \sum_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^{v,a,h}_{obs})d\theta \tag{2.44}$$

を計算することによって，未知物体の視覚，聴覚，触覚情報から単語情報を予測することが可能である．ただし，$p(z|\theta)$ は，前節同様にパラメータ推定により α を再計算することになり，その際パラメータ ϕ^v,ϕ^a,ϕ^h は学習時に推定した値を固定して用いる．

第 3 章

提案手法

本稿で提案する多様な概念を統合したグラフィカルモデルを図 3.1 に示す．このモデルは先述した確率的文書生成モデルである LDA をマルチモーダル化し、多層構造にしたものと考えることができる．

3.1 生成過程

まず多層マルチモーダル LDA の生成過程について説明する．このモデルは，上位カテゴリ z を生成する多項分布のパラメータ θ と，概念 C を生成する多項分布のパラメータ θ_z^C と，各モダリティ m の情報を生成する多項分布のパラメータ β_{zC}^m を，それぞれ α, α^C, ϕ^m をパラメータとするディリクレ事前分布から生成する．

$$\theta \sim \mathrm{Dir}(\alpha) \tag{3.1}$$

$$\theta_z^C \sim \mathrm{Dir}(\alpha^C) \tag{3.2}$$

$$\beta_{zC}^m \sim \mathrm{Dir}(\phi^m) \tag{3.3}$$

各概念の i 番目の情報 w_i^m を，以下の処理を繰り返すことで生成する
1. 上位カテゴリ z を θ をパラメータする多項分布にから生成する

$$z \sim \mathrm{Mult}(\theta) \tag{3.4}$$

2. 下位概念のカテゴリ z^C は θ_z^C をパラメータとする多項分布から生成する

$$z^C \sim \mathrm{Mult}(\theta_z^C) \tag{3.5}$$

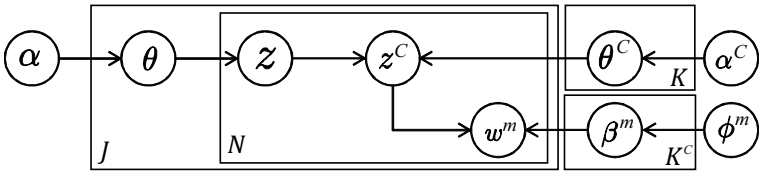

図 3.1: 多層 LDA

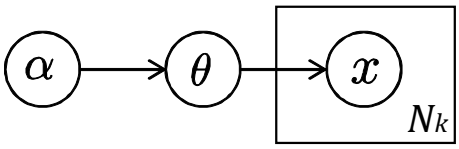

図 3.2: ディリクレ多項モデル

3. 概念 C の情報 w_i^m を $\beta_{z^C}^m$ をパラメータとする多項分布から生成する

$$w_i^m \sim \mathrm{Mult}(\beta_{z^C}^m) \tag{3.6}$$

3.2 事後分布の導出

次に学習を行うため，図 3.1 の事後分布について考える．まず，図 3.2 に示す一般的なディリクレ多項モデルの事後確率を考える．ディリクレ事前分布のパラメータ α，多項分布のパラメータ θ，観測されるデータ集合を $\mathbf{X}_k = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ とする．

ここで，データ集合 $\mathbf{X}_k = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ がクラス k に分類されたと仮定すると，

$$\begin{aligned} P(\mathbf{X_k}|\theta_k) &= Mult(\mathbf{X_k}|\theta_k) \\ &= \prod_n^{N_k} \prod_i^d \theta_{ki}^{x_{ni}} \\ &= \prod_i^d \theta_{ki}^{\sum_n^{N_k} x_{ni}} \end{aligned} \tag{3.7}$$

θ_k はディリクレ事前分布 α_k によって生成される．またインデックスを i，次元数

を d とする．よって，

$$\begin{aligned} P\left(\theta_k|\alpha_0\right) &= Dir(\theta_k|\alpha_0) \\ &= \frac{1}{Z(\alpha_0)}\prod_i^d \theta_{ki}^{\alpha_{0i}-1} \end{aligned} \tag{3.8}$$

ただし，ここで $Z(\alpha_0)$ は，

$$Z(\alpha_0) = \frac{\prod_i^d \Gamma\alpha_{\mathbf{0i}}}{\Gamma(\sum_i^d \alpha_{\mathbf{0i}})} \tag{3.9}$$

となる正規化項である．ここで，ベイズ推定のため多項分布のパラメータ θ を周辺化すると，

$$\begin{aligned} P\left(\mathbf{X}_k|\alpha_0\right) &= \int P\left(\mathbf{X}_k|\theta_k\right) P\left(\theta_k|\alpha_0\right) d\theta_k \\ &= \frac{1}{Z(\alpha_0)}\int \prod_i^d \theta_{ki}^{\alpha_{0i}+\sum_n^N x_{ni}-1} d\theta_k \\ &= \frac{Z(\alpha_0+\sum_n^N x_n)}{Z(\alpha_0)}\int \frac{1}{Z(\alpha_0+\sum_n^N x_n)}\prod_i^d \theta_{ki}^{\alpha_k-1} d\theta_k \end{aligned} \tag{3.10}$$

となる. また，ここで $\alpha_k = \alpha_0 + \sum_n^N x_n$ として前述の式 3.10 を考えると

$$P\left(\mathbf{X_k}|\alpha_k\right) = \frac{Z(\alpha_k)}{Z(\alpha_0)}\int \frac{1}{Z(\alpha_k)}\prod_i^d \theta_{ki}^{(\alpha_{0i}+\sum_n^N x_{ni})-1} d\theta_k \tag{3.11}$$

となる.

次に、クラス k に分類されたデータ集合 $\mathbf{X}_k = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ から更に新しいデータ $\mathbf{x}'$ が生成される確率を考えると，

$$\begin{aligned} P\left(\mathbf{x}'|\mathbf{X}_k, \alpha_0\right) &= \frac{P\left(\mathbf{x}', z\mathbf{X}_k\right)}{P\left(\mathbf{X}_k\right)} = \frac{Z\left(\alpha_{\mathbf{k}}+\mathbf{x}'\right)}{P\left(\alpha_{\mathbf{k}}\right)} \\ &= \frac{\prod_i^d \Gamma(\alpha_{ki}+x'_i)}{\Gamma\left(\sum_i^d (\alpha_{ki}+x'_i)\right)}\frac{\Gamma(\sum_i^d \alpha_{ki})}{\prod_i^d \Gamma(\alpha_{ki})} \end{aligned} \tag{3.12}$$

となる．ここで，$\mathbf{x}$ が特定の次元のみが 1 となる生起回数を表すデータであるとし，$x'_{i=j} = 1$, $x'_{i\neq j} = 0$ であった場合，上式 3.12 は，

$$P\left(\mathbf{x}'|\mathbf{X}_k, \alpha_0\right) = \frac{\Gamma(\sum_i \alpha_{ki})}{\Gamma\left(\sum_i \alpha_{ki}+1\right)}\frac{\Gamma(\alpha_{kj}+1)}{\Gamma(\alpha_{kj})} \tag{3.13}$$

と変形することができる．またガンマ関数の性質である $x\Gamma(x) = \Gamma(x+1)$ より，

$$P\left(\mathbf{x}'|\mathbf{X}_k, \alpha_0\right) = \frac{\alpha_{kj}}{\sum_i \alpha_{ki}} \tag{3.14}$$

となる．さらに，クラス k のデータ $\mathbf{X}$ の中で，i 次元目が 1 であったデータの個数を N_{ki}，さらに $N_k = \sum_i^d N_i$ とすると，$\alpha_{ki} = \alpha_{0i} + N_{ki}$ となるので，

$$P\left(\mathbf{x}'|\mathbf{X}_k, \alpha_0\right) = \frac{\alpha_{0j} + N_{kj}}{\sum_i \left(\alpha_{0i} + N_{ki}\right)} = \frac{\alpha_{0j} + N_{kj}}{\sum_i \alpha_{0i} + N_k} \tag{3.15}$$

と考えることができる．ここで，$\alpha_{00} = \alpha_{01} = \cdots == \alpha_{0d}$ とすると，

$$P\left(\mathbf{x}'|\mathbf{X}_k, \alpha_0\right) = \frac{\alpha_0 + N_{kj}}{d \cdot \alpha_{0i} + N_k} \tag{3.16}$$

となり，ディリクレ多項モデルの事後確率が導出できた．

つづいて，導出したディリクレ多項モデルの事後確率を図 3.1 の多層 LDA モデルに拡張する．このモデルはディリクレ事前分布パラメータ α によって多項分布のパラメータ θ が生成される．また多項分布のパラメータ β もディリクレ事前分布のパラメータ ϕ によって決定される．一般的にある事象が起こった条件下で別の事象が起こるとき，確率の乗法定理が成立するため，多層 LDA を応用すればよい．つまり，先ほど導出した式 (3.16) を用いれば,

$$P\left(z|\alpha, \boldsymbol{z}\right) = \frac{\alpha + N_{Dz}}{K\alpha + N_D} \tag{3.17}$$

となる．ここで文書 D にトピック z が割り当てられた数を N_{Dz} と表し, カテゴリ数を K, 文書数を N_D とする.

次に下位層の概念 z^C が発生する確率を求めると,

$$P\left(z^C|z, \alpha^C, \boldsymbol{z}\right) = \frac{\alpha^C + N_{zz^C}}{K^C\alpha^C + N_z} \tag{3.18}$$

となる. なお, ここで α^C は下位層のディリクレ事前分布パラメータを表す．また単語 w^m にカテゴリ z が割り当てられた回数を N_z，更に単語 w^m に上位カテゴリ z と下位カテゴリ z^C が割り当てられた回数を N_{zz^C} とした. 続いて, 下位層の概念 z^C から物体の単語情報 w^m が生成され,

$$P\left(w^m|z^C, \phi^m, \boldsymbol{z}, \boldsymbol{w}^m\right) = \frac{\phi^m + N_{z^C w^m}}{W^m\phi^m + N_z} \tag{3.19}$$

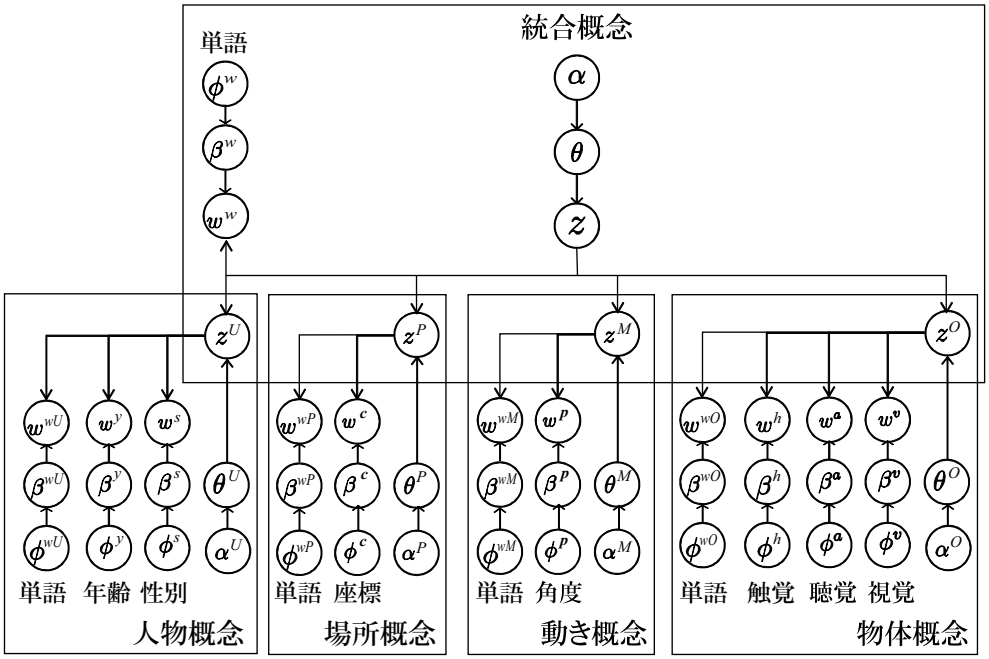

図 3.3: 統合概念形成 LDA のグラフィカルモデル

となる．ここで単語 w^m にカテゴリ z^C が割り当てられた回数を $N_{z^C w^m}$，単語数を W^m，トピック z^C が割り当てられた単語の総数を N_{z^C} とする．

式 (3.17)，(3.18), (3.19)3 つの式より，z,z^C,w^m の同時確率を求めると，

$$P\left(z, z^C, w^m | \alpha, \alpha^C, \phi^m, \boldsymbol{z}, \boldsymbol{z}^C, \boldsymbol{w}^m\right) = \frac{\alpha + N_{Dz}}{K\alpha + N_D} \frac{\alpha^C + N_{zz^C}}{K^C \alpha^C + N_z} \frac{\phi^m + N_{z^C w^m}}{W^m \phi^m + N_z^C} \quad (3.20)$$

となり，以上が多層 LDA の事後分布である．

3.3 多層マルチモーダル LDA

それでは今回提案する多層マルチモーダル LDA について詳しく説明する．図 3.3 より，このモデルは二層構造となっていることがわかる．上位層では z が物体，動き，人物と場所を統合した概念を表す上位カテゴリとなり，z^O, z^M, z^P, z^U がそれぞれ物体，動き，場所と人物概念のカテゴリであり，w^w は上位カテゴリ z から生成される単語情報である．また，下位層では物体概念 z^O から物体の特徴である視覚 w^v・聴覚 w^a・触覚 w^h・単語情報 $w^w O$ がそれぞれ生成され，動き概念 z^M から

動き w^p・単語情報 $w^w M$ が生成される．さらに，場所概念 z^P から場所の座標 z^c・単語情報 $w^w P$ が生成され，人物概念 z^U から性別 w^s・年齢 w^y・単語情報 $w^w U$ が生成される．

このモデルにおいて，統合概念のカテゴリ z と概念 $C \in \{O, M, P, U\}$ のカテゴリ z^C は直接観測できない隠れ変数であり，各モダリティ $m \in \{v, a, h, wO, p, wM, c, wP, s, y, wU, w\}$ の観測データ w^m 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ることで，各パラメータを推定する．w^c はハイパーパラメータ ϕ^c によって決まるディリクレ事前分布に従う β^c 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^C は，それぞれハイパーパラメータ α，α^C 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^C 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，導出した式 (3.20) をマルチモーダルに拡張し，Gibbs Sampling により推定する．

$$P(z_{jmi}, z^C_{jmi} | \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m) \propto$$
$$P(z_{jmi}|\boldsymbol{Z}_{-jmi}) P(z^C_{jmi}|z_{jmi}, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}) P(w^m_{ji}|z^C_{jmi}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m_{-ji}) \quad (3.21)$$

なお，右辺のそれぞれの確率分布は次のようなる．

$$P(z_{jmi} = k|\boldsymbol{Z}_{-jmi}) = \frac{\alpha + N^{-jmi}_{j,z=k}}{K\alpha + N_j{}^{-jmi}}, \quad (3.22)$$

$$P(z^C_{jmi} = l|z_{jmi} = k, \boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}) = \frac{\alpha^C + N^{-jmi}_{z=k,z^C=l}}{K^C\alpha^C + N^{-jmi}_{z=k}}, \quad (3.23)$$

$$P(w^m_{ji} = x|z^C_{jmi} = k, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{W}^m_{-ji}) = \frac{\phi^m + N^{-jmi}_{z^C=k,w^m=x,m}}{W^m\phi^m + N^{-jmi}_{z^C=k,m}}, \quad (3.24)$$

ただし，$\boldsymbol{Z}, \boldsymbol{Z}^C$ は，それぞれ全物体の全情報に割当たれた上位カテゴリと下位概念のカテゴリの集合を表し，$\boldsymbol{W}^m$ はモダリティ m の全物体の情報の集合である．N_{jz} は物体 j の全モダリティに上位カテゴリ z が割り当てられた回数であり，$N_{z^C w^m}$ はモダリティ m の特徴量 w^m に下位カテゴリ z^C が割り当てられた回数である．また，N_{z,z^C} は上位カテゴリ z と下位カテゴリ z^C の共起した回数を表しており，K, K^C, W^m はそれぞれ上位カテゴリのカテゴリ数，概念 C のカテゴリ数，モダリティ m の情報の次元数である．負の添字はその情報を除外することを表し，$-jmi$ は j 番目の物体のモダリティ m の i 番目の情報を除外することを表している．

モデルの学習は，隠れ変数である z, z^C を，収束するまで事後分布からサンプリングすることによって実現できる．しかし，隠れ変数が複数あり，複雑なモデルであるため，全てのパラメータを同時に求めると局所解に陥りやすいといった問題がある．そこで，図 3.3 の右側に示す下位カテゴリ z^C を個々の独立した MLDA として学習し，下位概念のパラメータ β^m を先に決定する．この時に，各カテゴリは $z^C \in \{z^O, z^M, z^P, z^U\}$ を次式を用いてサンプリングする．

$$\begin{aligned} z^C_{jmi} &\sim P(z^C_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ij}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_z P(z|\boldsymbol{Z}_{-jmi}) P(z^C_{jmi}|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z) \\ &\quad \times P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C_{jmi}). \end{aligned} \tag{3.25}$$

このサンプリングを収束するまで繰り返すことで，式 (3.24) を決定する．次に，式 (3.24) を固定し，上位カテゴリ z，下位カテゴリ z^C をサンプリングする．

$$\begin{aligned} z_{jmi} &\sim P(z_{jmi}|w^m_{ji}, \boldsymbol{W}^m_{-ij}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi}) \\ &\propto \sum_{z^C} P(z_{jmi}|\boldsymbol{Z}_{-jmi}) P(z^C|\boldsymbol{Z}_{-jmi}, \boldsymbol{Z}^C_{-jmi}, z_{jmi}) \\ &\quad P(w^m_{ji}|\boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, z^C). \end{aligned} \tag{3.26}$$

この時，下位カテゴリ z^C が上位概念の影響を受けて更新されることに注意が必要である．Algorithm1 と Algorithm2 がそれぞれ，下位概念のパラメータの決定と，モデル全体の学習アルゴリズムである．以上のような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 を上位カテゴリのカテゴリ数とする時，最終的なパラメータの推定値 $\hat{\beta}^m_{w^m z^C}$，$\hat{\theta^C}_{zz^C}$，$\hat{\theta}_{jz}$ は以下のようになる．

$$\hat{\beta}^m_{w^m z^C} = \frac{N_{z^C w^m m} + \phi^m}{N_{z^C m} + W^m \phi^m}, \hat{\theta^C}_{zz^C} = \frac{N_{zz^C m} + \alpha^C}{N_{zm} + K\alpha^C}, \hat{\theta}_{jz} = \frac{N_{jz} + \alpha}{N_j + K\alpha}, \tag{3.27}$$

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

Algorithm 1 Multi-layered MLDA (bottom-layer)

for all i, j, C, m **do**
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] +$
      $P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

Algorithm 2 Multi-layered MLDA (whole layer)

for all i, j, C, m **do**
  for $k \leftarrow 1$ to K **do**
    $P[k] \leftarrow P[k-1] +$
      $P(z_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K **do**
    if $u < P[k]/P[K]$ **then**
      $z_{jmi} = k$, break
    end if
  end for
  for $k \leftarrow 1$ to K^C **do**
    $P[k] \leftarrow P[k-1] +$
      $P(z^C_{jmi} = k | w^m_{ji}, \boldsymbol{W}^m_{-ji}, \boldsymbol{Z}^C_{-jmi}, \boldsymbol{Z}_{-jmi})$
  end for
  $u \leftarrow$ draw from Uniform [0,1]
  for $k \leftarrow 1$ to K^C **do**
    if $u < P[k]/P[K^C]$ **then**
      $z^C_{jmi} = k$, break
    end if
  end for
end for

3.4 未観測情報の予測

学習したモデルを用いることで，物体や動きの認識だけでなく，概念間の予測も可能となる．観測された情報 $\boldsymbol{w}^m$ から，以下の式を用いて，上位カテゴリ $\hat{z}$ と下位カテゴリ $\hat{z}^C \in \{\hat{z}^O, \hat{z}^M, \hat{z}^P, \hat{z}^U\}$ を予測することができる．

$$\hat{z}^C = \underset{z^C}{\operatorname{argmax}} \sum_{z} \sum_{z^{\bar{C}}} P(z) P(z^{\mathbf{C}}|z) P(z^C|\boldsymbol{w}^{m_C}) \tag{3.28}$$

$$\hat{z} = \underset{z}{\operatorname{argmax}} \sum_{z^{\mathbf{C}}} P(z) P(z^{\mathbf{C}}|z) P(z^C|\boldsymbol{w}^{m_C}) \tag{3.29}$$

但し，$z^{\bar{C}}$ は z^C 以外の下位カテゴリであり，$z^{\mathbf{C}}$ は全ての下位概念のカテゴリの集合を表す．

さらに，前章で述べた MLDA を用いた未観測情報の予測 2.11 と同様に，形成された概念による認識を行なうことで，未知物体の未観測情報の予測を行なうことが可能となる．例えば，単語情報の予測である．前述したように単語情報を BoW モデルとして扱う．そこで，mMLDA によって形成された概念を用いて，以下の式によって様々な概念における単語情報を予測することができる．

$$P(w^w|\boldsymbol{w}_{obs}^{v,a,h}) = \int \sum_{z} p(w^w|z) p(z|\theta) p(\theta|\boldsymbol{w}_{obs}^{v,a,h}) d\theta \tag{3.30}$$

3.5 近似多層マルチモーダル LDA

前述のように，全ての下位概念は，統合概念を無視すれば，独立した MLDA と等価なモデルと考える事ができる．さらに，単語情報 w^w 以外の w^m を無視することで，統合概念は z^O，z^M，z^P，z^U と w^w を生成する MLDA と等価なモデルと見なすことができる．すなわち各概念を独立した MLDA として学習し，フィードフォワード的に接続することで，簡易的に多様な概念を統合することができる．図 3.4 が mMLDA を分解し，独立した 5 つの MLDA として考えた場合のグラフィカルモデルである．このモデルでは，下位カテゴリ $z^C \in \{z^O, z^M, z^P, z^U\}$ を独立した MLDA で学習後，上位カテゴリ z をもう一つの独立な MLDA で学習すればよい．

近似モデルの学習ではまず，下位層を独立した MLDA として学習し後に，z^C を多項分布 $P\left(z^C \mid \boldsymbol{w}^{m_C}\right)$ からサンプリングする．近似モデルの上位層に相当する

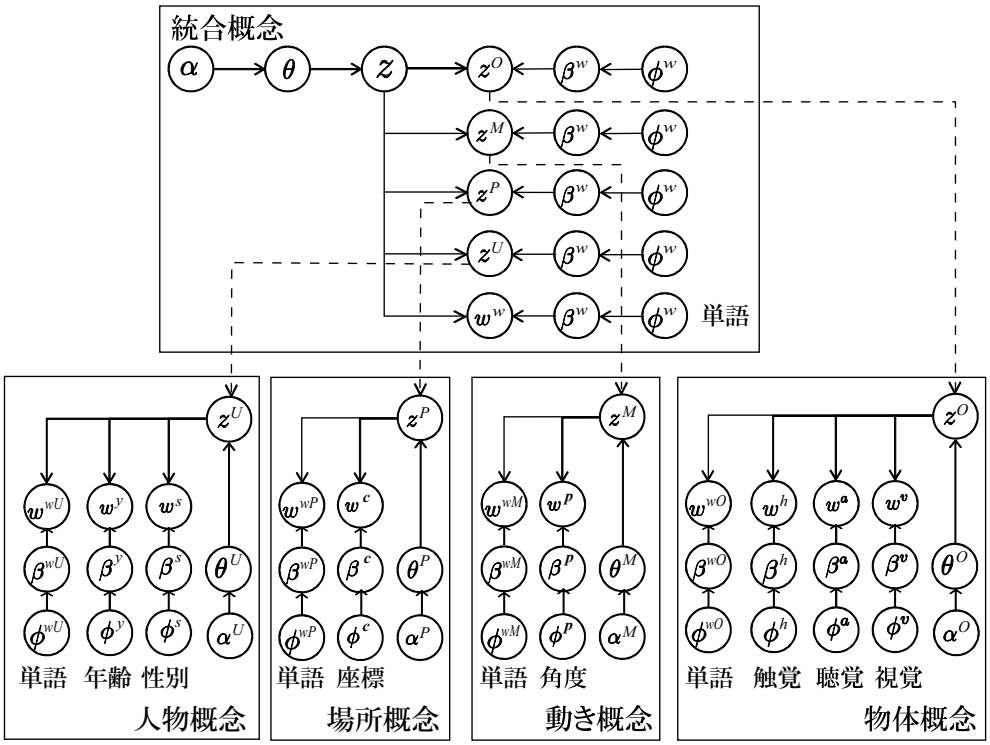

図 3.4: 近似多層マルチモーダル LDA

MLDA では，生成された z^C をそれぞれ図 2.4 に示した w として考えることで学習することができる．従って，下位概念の関係性は，モデルにおける隠れ変数 z によって学習され，この z が前下位概念の統合的な概念を表現するカテゴリとなる．ただし，z は固定された z^C の関係性を表現するだけであり，逆に z^C に影響を与えることはない．

近似モデルにおいても，学習したモデルを用いて未観測情報を予測することが可能である．例えば，物体情報 $\boldsymbol{w}^m$ から，確率の高い動きカテゴリを次のように予測することができる．

1. 物体情報から物体カテゴリをサンプリングする

$$\hat{z}^O \quad \sim \quad P\left(z^O \mid \boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h\right) \tag{3.31}$$

2. 次式によりサンプリングされた物体カテゴリ $\hat{\boldsymbol{z}}^O$ から，動作カテゴリが発生

する確率を計算する

$$P(\hat{z}^M|\hat{\boldsymbol{z}}^O) = \int \sum_z P(\hat{z}^M|z)P(z|\theta)P(\theta|\hat{\boldsymbol{z}}^O)d\theta \tag{3.32}$$

さらに，$\hat{z}^M$ に $\hat{z}^P$，$\hat{z}^U$ を代入することで，他概念のカテゴリも予測することができる．

後に示す実験の結果からもわかるように，mMLDA と近似モデルの定性的差異は明らかである．近似モデルは単純かつ容易に実装できるところがメリットではあるが，明らかな欠点をもつ．それは，上位概念が下位概念に一切影響を与えないことである．各概念を独立に学習することになるため，下位概念での分類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ることにつながる．一方，mMLDA は，各概念が同時に形成されるために，下位層での分類が相互に影響を及ぼし合い，モデル全体として分類や予測の精度を向上させることができると期待できる．本論文では，提案する mMLDA の有効性を評価するために，近似モデルとの比較を行う．

3.6　語意獲得のための単語選択

教師ありクラスタリングを行う際に，その精度を向上させるために，クラスタリングに重要な特徴量を選択する，特徴選択を行うことがある．ここでは，クラスタを前章の mMLDA によって教師なしで形成された各概念と考え，特徴量を単語と考えることで，単語の選択を行う．このようにして，単語選択を行うことによって，カテゴリ分類精度が向上するだけでなく，各カテゴリを表す単語が選択されるため，語意の獲得が可能となる．

3.6.1　相互情報量に基づく単語選択

特徴選択の代表的な方法の一つとして，相互情報量という尺度を用いる手法がある．本稿では，図 3.3 に示したように，物体，動き，統合概念に教示発話から得られる全ての単語情報を与えて学習を行う．各概念を表現する適切な単語が存在する

と考え，ここで，単語とカテゴリの結び付きの強さの尺度として，単語とカテゴリ間の相互情報量を用いる．単語 w^w と概念 $i \in$ { 物体概念，動き概念，場所概念，人物概念，統合概念 } のカテゴリ k との間の相互情報量は以下の式となる．

$$I(w^w, k|i) = \sum_{K \in (k, \bar{k})} \sum_{W \in (w^w, \bar{w^w})} P(W, K|i) \log \frac{P(W, K|i)}{P(W|i)P(K|i)} \tag{3.33}$$

但し，$\bar{k}$ は k 以外のカテゴリを表し，$\bar{w^w}$ は w^w 以外の単語を表している．相互情報量とは，二つの確率変数の共有する情報量であり，相互依存の尺度である．したがって単語とカテゴリ間の相互情報量が大きい場合，その単語はそのカテゴリを表現しているといえる．

本稿では，単語によって，複数の概念を表す可能性があると考え，式 3.33 を用いて求められた相互情報量を単語の各概念に対する重みとして考える．その重みを $weight(i, w^w)$ とし，次式で求められる．

$$weight(i, w^w) = \max_k I(w^w, k|i) \tag{3.34}$$

$$\hat{p}_i(w^w|\boldsymbol{w}^m_{obs}) = weight(i, w^w) p_i(w^w|\boldsymbol{w}^m_{obs}) \tag{3.35}$$

このように，単語の各概念に対する重みを求め，概念 i の単語予測 $p_i(w^w|\boldsymbol{w}^m_{obs})$ の際に重みをつけることで，各概念から生成される単語の予測精度を向上させることが可能である．

3.7 文章生成

前章で述べた各単語に対する概念選択を用いて，次に行うのは観測された情報を表現する文章の生成である．そのための基本的な考えとしては，文章を構築する単語は特定の概念と結びつき，文章を様々な概念を用いて構築されると考える．例えば，

教示文章 母はキッチンで野菜を切っています.

単語－概念 母－人物； キッチン－場所； 野菜－物体； 切る－動き；

概念の発火 （人物）（場所）（物体）（動き）.

Algorithm 3 Sentence Generation

$i \leftarrow 1, C_0 =$ "BOS''
for i **do**
  $C_i \leftarrow$ draw from $P(C_i|C_{i-1})$
  if $C_i =$ "EOS'' **then**
    break
  end if
  $w_i = \text{argmax}_{w^w} P(w^w|w^m, C_i)$
  $i \leftarrow i + 1$
end for

上記に示したように，教示文章を多様な概念が順番に発火すると考えることができる．よって，観測した情報を表現する文章の生成過程は次のように説明することができる．

1. 学習したモデルを用いて，観測情報に関する様々な概念におけるカテゴリ認識を行う．
2. 文章を生成する概念の発火順を生成する．
3. カテゴリ認識の結果を用いて，各概念に関係する単語の予測を行う．
4. 概念の発火順に予測した単語を並べる．

但し，"は" や "を" といった助詞や機能語は含まれない．

文章を生成する概念の発火順においては，学習データの教示発話から確率的に学習することができる．ここでは，教示文章中の概念間の遷移確率を学習データから求められる．

$$P(C_i|C_{i-1}) = \frac{N_{C_{i-1},C_i}}{N} \tag{3.36}$$

但し，C_i は文章中の i 番目の単語に該当する概念である．また，$N_{C_{i-1}C_i}$ と N はそれぞれ C_{i-1} から C_i に遷移する数と概念間遷移の総数である．最終的に文章を生成するためには，次の Algorithm3 に従って行うとする．但し，w_i は生成文章の i 番目の単語を表す．

第 4 章

実験

提案モデルの有効性を検証するために，大きく分けて 2 つの実験を行った．mMLDA を用いた概念形成及び統合に関する実験と，提案手法による語意の獲得及び文章生成に関する実験である．まず，実験にしようするデータに説明する．

4.1 情報取得

4.1.1 マルチモーダル情報

実験には，図 4.1 に示す実生活で使われている計 132 個の物体を使用した．これらの物体をしようする動作を行い，実験のためのデータを取得した．全てのデータの組み合わせを表 4.1 に示した．また，今回の実験においては，人物のデータとして，大人・子供の男女の画像をネットから取得し，OKAO Vision が提供した既存の画像センシング技術による年齢・性別推定を使用した．図 4.3 に使用した画像及びヒストグラム化した性別と年齢推定の結果を示す．さらに，場所の情報としては，図 4.4 に示したような家の間取りを仮定し，玄関，リビング，キッチン，ダイニング，浴室，庭の 6 つの場所を想定し，座標のデータを取得した．動き情報として，Kinect を用いてキャプチャーし，人体の各関節の角度を取得した．図 4.2 に取得した動き情報の一部と Kinect を用いて取得したボーン情報を示した．

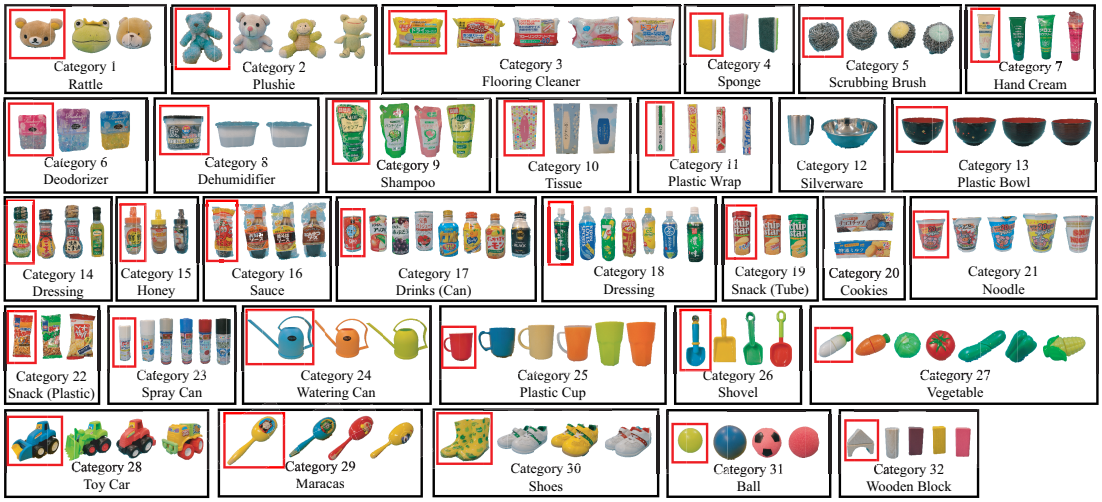

図 4.1: 実験で使用した物体

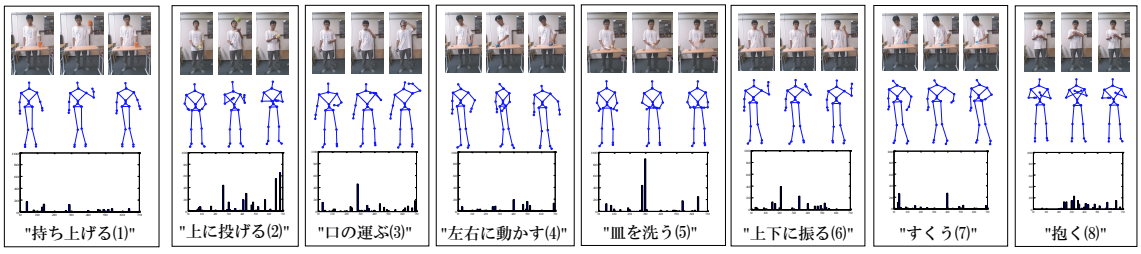

図 4.2: 物体に対して行った動きの例 (上) Kinect の画像 (中) 実際の動き（下）作成したヒストグラム

4.1.2 単語情報

本研究では人が表 4.1 に示した各データに対して，5 つの教示発話を与えることとする．取得した全ての教示発話は形態素解析器を用いて，単語分割を行い，他の知覚情報と同様に BoW モデルとしてヒストグラム化し生起回数の情報として取り扱う．表 4.2 に教示発話に一部を示す．

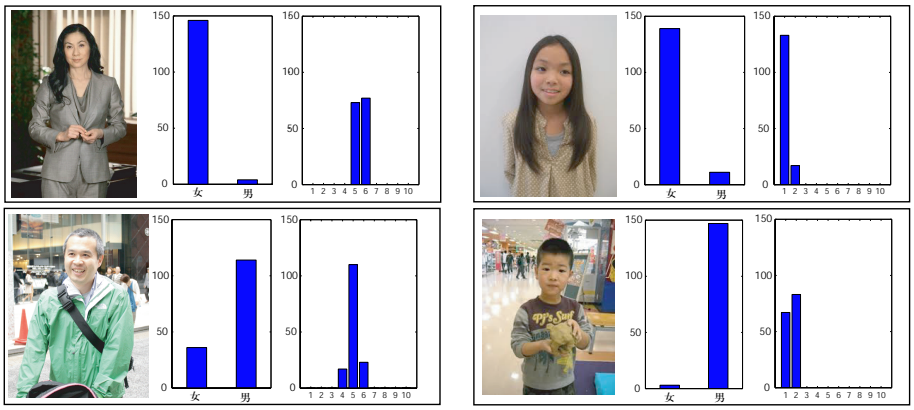

図 4.3: 人物の情報の例

図 4.4: 場所の情報の例

4.2 概念形成

4.2.1 カテゴリ数決定

LDA ではカテゴリ数を予め与えなければならず，このカテゴリ数の決定は LDA における重要な問題である．本論文で提案する mMLDA は LDA を拡張したモデルであるため，同様の問題が存在する．予備実験などを通して経験的に決定することも可能ではあるが，特に上位層の分類に対する正解を決めることは人手であっても困難であるため，ここでは自動的にカテゴリ数を決めることを考える．

下位層については，ノンパラメトリックベイズ手法であるマルチモーダル階層ディリクレ過程 (Multimodal Hierarchical Dirichlet Process: MHDP) [24] による

表 4.1: 動き，物体，場所，人物データの対応表（カッコ内の数字はカテゴリ ID）

動き	物体	場所	人物	動き	物体	場所	人物
持ち上げる (1)	茶碗 (13)	ダイニング (4)	全員 (1,2,3,4)	抱く (8)	ぬいぐるみ (2)	リビング (2)	子供の女 (3)
	飲み物（缶）(17)			積み重ねる (9)	積み木 (32)		子供 (3,4)
	カップヌードル (21)			置く (10)	消臭剤 (7)		大人の女 (1)
	プラスチックカップ (25)				除湿剤 (8)		
	スプレー缶 (23)	庭 (5)	男性 (2,4)		積み木 (32)		子供 (3,4)
上に投げる (2)	ぬいぐるみ (2)	リビング (2)	子供 (3,4)		プラスチックカップ (25)	ダイニング (4)	
	マラカス (29)			手に塗る (11)	ハンドクリーム (6)	リビング (2)	女性 (1,3)
	ボール (31)			取り出す (12)	ティッシュ箱 (10)		全員 (1,2,3,4)
口に運ぶ (3)	金属の食器 (12)	ダイニング (4)	全員 (1,2,3,4)		クッキー (20)		
	飲み物（缶）(17)				フローリングワイパー (3)	ダイニング (4)	大人の女 (1)
	ペットボトル (18)			ナイフで切る (13)	野菜（玩具）(27)	キッチン (3)	
	プラスチックカップ (25)			中身をかける (14)	ドレッシング (14)	ダイニング (4)	全員 (1,2,3,4)
	茶碗 (13)				蜂蜜 (15)		
	野菜（玩具）(27)				ソース (16)		
	カップヌードル (21)			中身を注ぐ (15)	シャンプー (9)	浴室 (6)	大人 (1,2)
	スナック (19)	リビング (2)			じょうろ (24)	庭 (5)	大人の男 (2)
左右に動かす (4)	車（玩具）(28)				飲み物（缶）(17)	ダイニング (4)	全員 (1,2,3,4)
	フリーリングワイパー (3)	ダイニング (4)	大人の女 (1)		ペットボトル (18)		
皿を洗う (5)	スポンジ (4)	キッチン (3)		包む (16)	ラップ (11)		大人女 (1)
	たわし (5)			塗る (17)	スプレー缶 (23)	庭 (5)	大人の男 (2)
上下に振る (6)	ガラガラ (1)	リビング (2)	子供 (3,4)	履く (18)	靴 (30)	玄関 (1)	全員 (1,2,3,4)
	マラカス (29)			袋を開ける (19)	スナック (19)	リビング (2)	
	ドレッシング (14)	ダイニング (4)	全員 (1,2,3,4)				
	ソース (16)						
	飲み物（缶）(17)						
	ペットボトル (18)						
	スプレー缶 (23)	庭 (5)	大人の男 (2)				
すくう (7)	ショベル (26)						

決定手法がそのまま利用できる．実際に MHDP によって下位層のカテゴリ数を推定したところ，図 4.5 に示したように物体，動き，場所，人物のカテゴリ数はそれぞれ 32，19，6，4 である．この結果を用いて以降の実験を行い，更に次の上位層のカテゴリ数を推定するために用いた．

上位層のカテゴリ数は，MHDP を直接適用して推定することができない．そこでここでは，近似モデルの上位 MLDA に MHDP を適用することでカテゴリ数を推定することとする．MHDP はサンプリングにより学習を行なっているため，初期値によって推定されるカテゴリ数が変わってしまう．そこで，MHDP を用いた分類を 100 回行い，100 個のモデルを学習した．図 4.5 が 100 個のモデルのカテゴ

表 4.2: 教示発話の例

教示発話
女の子はリビングで腕を上下に動かしてガラガラを振って音を聞く
女の子はリビングでガラガラを上下に動かして、振って音を聞く
父は浴室でシャンプーをもって、中身を注いで詰め替える
母はダイニングでピンク色のフローリングワイパーを開けて中身を取り出す
女の子はリビングでぬいぐるみを上に投げて遊んでいる

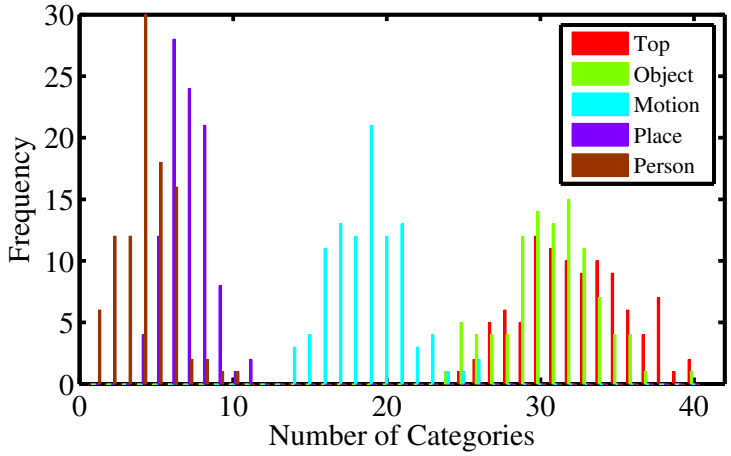

図 4.5: MHDP を用いた各概念のカテゴリ数の発生頻度

リ数のヒストグラムであり，横軸と縦軸はそれぞれ，推定した上位カテゴリ数とその頻度を示している．すなわち，このグラフはカテゴリ数の発生確率と考えることができ，カテゴリ数 30 が最も高い確率で発生していることが分かる．

以上の結果から，上位カテゴリ数を 30，物体，動き，場所，人物のカテゴリ数はそれぞれ 32，19，6，4 として，mMLDA と近似モデルによって概念形成を行い，各概念の評価を行った．

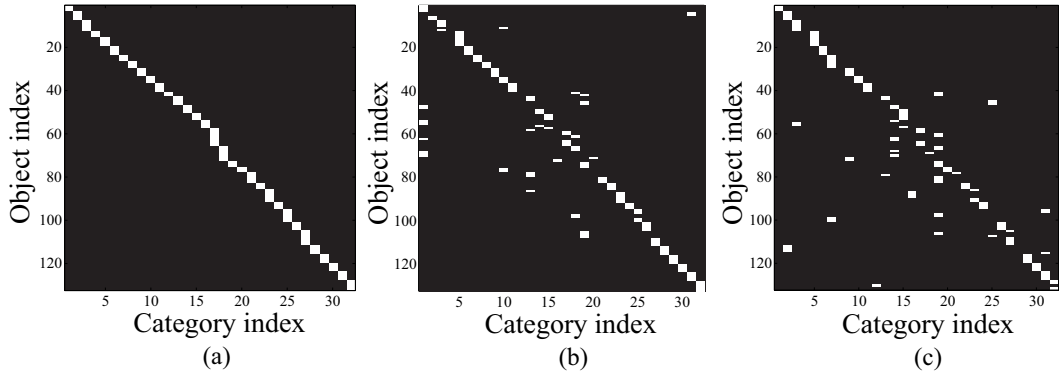

図 4.6: 物体の分類結果：(a) 正解，(b)mMLDA，(c) 近似モデル

4.2.2 物体概念

提案モデルと近似モデルによって形成された物体概念を評価した．物体概念の形成結果は図 4.6 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 4.6(a) が人手による分類であり，これを正解として各手法の分類結果を評価した．図 4.6(b) が提案手法 (mMLDA) による分類結果であり，図 4.6(c) が近似モデルによる分類結果である．これらの分類結果から，図 4.6(a) を正解として，次式により分類精度を計算した．

$$Acc = \frac{100}{J} \sum_{j}^{J} \delta(c_{\text{correct}}(j), c_{\text{result}}(j)) \tag{4.1}$$

ただし，J はデータ数，$c_{\text{correct}}(j)$，$c_{\text{result}}(j)$ はそれぞれ j 番目のデータの正解のカテゴリと，実際に分類されたカテゴリの ID である．$\delta(a, b)$ は，$a = b$ の時 1，さもなくば 0 となる関数である．分類精度を計算した結果，mMLDA では 74.24%，近似モデルでは 65.15%となり，提案モデルである mMLDA の方がより正解に近い分類ができている．

mMLDA の分類では，“飲み物 (缶)(17)” は一つのカテゴリに分類することができたのに対して，近似モデルでは，この物体を 3 つのカテゴリに分類してしまっている．同じ “飲み物 (缶)(17)” でも，異なる柄や形を持つため，近似モデルでは異なるカテゴリに分類されてしまったのに対して，mMLDA では “飲み物 (缶)(17)” と関係する動きも考慮して分類を行なうため，正しく一つのカテゴリに分類するこ

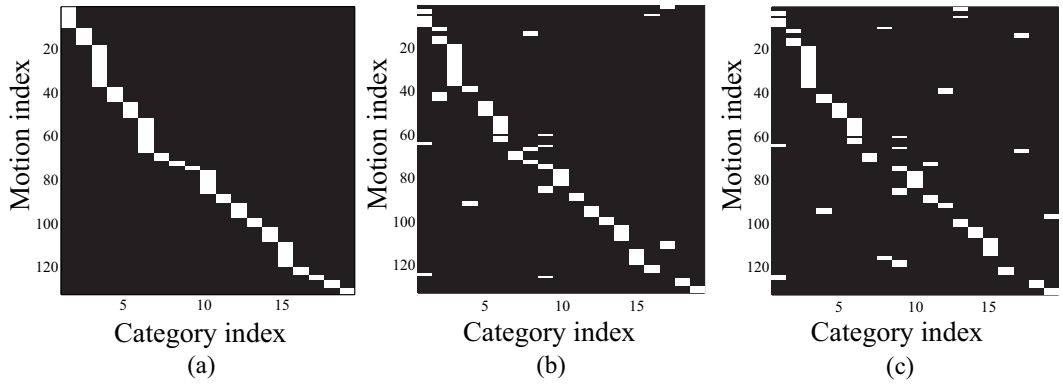

図 4.7: 動きの分類結果：(a) 正解，(b)mMLDA，(c) 近似モデル

とができたと考えられる．

4.2.3 動き概念

次に，提案モデルと近似モデルによって分類された動き概念を評価した．図 4.7 が分類結果であり，縦軸が実際の動きのカテゴリ番号，横軸が分類されたカテゴリ番号である．図 4.7(a) が人出による分類であり，物体と同様，この分類を正解として，各種法の分類を評価した．図 4.7(b) が mMLDA による分類結果，図 4.7(c) が近似モデルによる分類結果である．正解の分類 (図 4.7(a)) と比較すると，mMLDA(図 4.7(b)) の分類精度は 81.06%となり，近似モデル (図 4.7(c)) の分類精度は 75.00%となった．

mMLDA と近似モデルによる動き概念の形成結果の差異は，“中身をかける (14)” の分類結果で見ることができる．mMLDA の分類結果では，この動きを一つに分類することができた．一方，近似モデルではこの動きを 2 つのカテゴリに分類してしまい，“一部は中身を注ぐ (15)” と同一のカテゴリとなった．この違いは，2 つの動きは似通っている動きを持つが，扱う物体が異なるため，mMLDA では 2 つのカテゴリに分類することができたと考えられる．このように，mMLDA は近似モデルに比べて，物体と動きがそれぞれ影響しあうため，より正解に近い分類が可能となる．

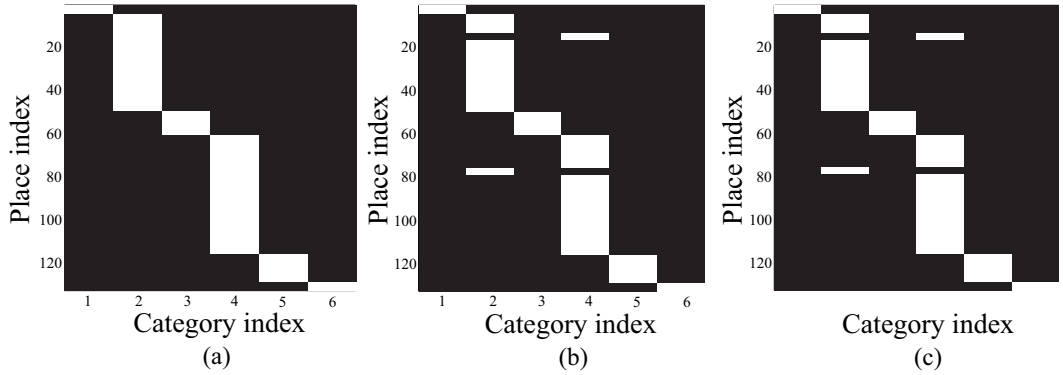

図 4.8: 場所の分類結果 (a) 正解となる分類 (b)mMLDA (c) 近似モデル

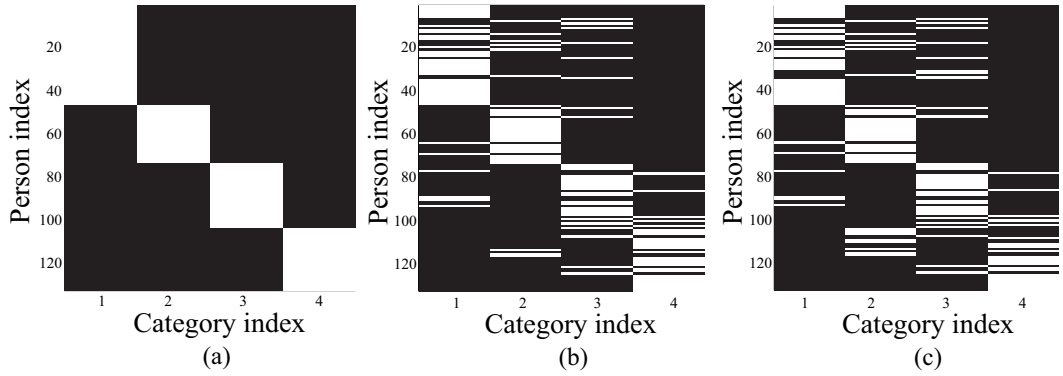

図 4.9: 人物の分類結果 (a) 正解となる分類 (b)mMLDA (c) 近似モデル

4.2.4 場所概念

物体と同様，図 4.8 が人手による場所概念の分類であり，この分類を正解として，提案モデルと近似モデルの分類を評価した．図 4.8 が mMLDA による分類であり，図 4.8 が近似モデルによる分類結果である．正解の分類 (図 4.7(a)) と比較すると，mMLDA(図 4.7(b)) と近似モデル (図 4.7(c)) の分類精度は両方 96.97%となった．場所のカテゴリにおいて，本実験で用いた学習データとして，かなり分かりやすいデータだったため，提案手法と近似モデルの分類結果に差が出なかったと考えられる．

4.2.5 人物概念

次に，mMLDA と近似モデルによって形成された人物概念を評価した．図 4.9(a)，図 4.9(b)，と図 4.9(c) がそれぞれ，人手による分類，mMLDA の分類結果と近似モデルの分類結果を示す．他の概念と同様，図 4.9 を正解の分類として，両方のモデルによる分類結果と比較すると，それぞれ 75.75%と 71.21%となった．人物のカテゴリ分類の結果では，mMLDA と近似モデルの差が見られないが，分類精度を比較した結果，mMLDA はより正解の分類が可能とする．

4.2.6 統合概念

4.2.6.1 形成された統合概念

mMLDA の上位層では物体，動き，場所，と人物の関係性を表すカテゴリが形成されており，その中には人にとって意味のあるカテゴリも形成されている．表 4.3 が実際に形成された物体・動き・場所・人物概念が組み合わさり形成された統合概念である．例えば，統合概念 21 では，動きの “口に運ぶ (3)” と物体の “飲み物 (缶)(17)” や，“ペットボトル (18)”，“プラスチックカップ (25)” 等が 1 つのカテゴリに分類された．これは，“何かを飲む” という概念が形成されたと考えられる．他にも，統合概念 25 では “口に運ぶ (3)” の動きと，“茶碗 (13)”，“カップヌードル (21)”，“野菜（玩具）(27)” 等が組み合わさった概念が形成され，これは “何かを食べる” という概念であるといえる．さらに，統合概念 9 と 22 では，“左右に動かす (4)” が，物体によって異なる上位カテゴリに分類された．1 つは “車 (玩具)(28)” と関係し，もう一つは “フローリングワイパー (3)” と関係する上位カテゴリであり，これらはそれぞれ “車の玩具を走らせる” という概念と，“フローリングワイパーで掃除をする” といった概念であるといえる．このように同じ動きに対しても，異なる物体によって，意味の異なる統合概念が形成できているといえる．

また，統合概念 3，8 と 19 では，“中身を注ぐ (16)” が物体または場所によって異なる上位カテゴリに分類された．統合概念 3 では，“庭 (5)” と “じょうろ (24)” と関係し，“水遣りをする” といった概念であると考えられる．統合概念 8 では，“浴室 (6)” と “シャンプー (9)” と関係し，統合概念 19 では，“ダイニング (4)”，“ペットボトル (18)” と “飲み物（缶）(17)” が組み合わさった概念であった．

表 4.3: mMLDA を用いた統合概念の形成結果

No	動き	物体	場所	人物	No	動き	物体	場所	人物
1	上下に振る	スプレー缶	庭	大人の男	18	中身をかける	ドレッシング	ダイニング	全員
	塗る						ソース		
2	上に投げる	ぬいぐるみ	リビング	子供			蜂蜜		
		ボール			19	中身を注ぐ	ペットボトル	ダイニング	全員
3	中身を注ぐ	じょうろ	庭	大人の男			飲み物（缶）		
4	上下に振る	ガラガラ	リビング	女子	20	口に運ぶ	金属の食器	ダイニング	全員
5	取り出す	ティッシュ箱	リビング	全員	21	口に運ぶ	ペットボトル	ダイニング	全員
		クッキー					飲み物（缶）		
6	手に塗る	ハンドクリーム	リビング	大人の女			プラスチック		
7	皿を洗う	スポンジ	キッチン	大人の女			カップ		
		たわし			22	左右に動かす	車（玩具）	リビング	男子
8	中身を注ぐ	シャンプー	浴室	大人	23	積み重ねる	積み木	リビング	子供
9	左右に動かす	フローリングワイパー	ダイニング	大人の女		置く			
10	取り出す	フローリングワイパー	ダイニング	大人の女	24	抱く	ぬいぐるみ	リビング	女子
11	上に投げる	マラカス	リビング	子供	25	口に運ぶ	カップヌードル	ダイニング	全員
	上下に振る						野菜（玩具）		
12	履く	靴	玄関	全員			茶碗		
13	開ける	スナック	リビング	全員					
14	包む	ラップ	ダイニング	大人の女	26	口に運ぶ	スナック	リビング	子供
15	持ち上げる	茶碗	ダイニング	全員	27	置く	消臭剤	リビング	大人の女
		カップヌードル					除湿剤		
		プラスチックカップ			28	上下に振る	ドレッシング	ダイニング	全員
		飲み物（缶）					ソース		
		スプレー缶	庭				ペットボトル		
16	置く	カップヌードル	ダイニング	大人	29	すくう	ショベル	庭	大人の男
17	手に塗る	ハンドクリーム	リビング	女の子	30	ナイフで切る	野菜（玩具）	キッチン	大人の女

一方，統合概念 23 では，異なる 2 つの動き，“積み重ねる (9)” と “置く (10)”，が同じ物体，場所，人物と組み合わさった概念が形成された．これは，“積み木で遊ぶ” といった統合概念であるといえる．このように，物体・動き・場所・人物概念の関係性を表す統合概念が形成されると考えられる．

4.2.6.2 概念間の関係性の学習

以上のように，意味のある統合概念が形成できたと言えるが，統合概念は正解を定義することが難しいため，定量的に mMLDA と近似モデルを比較することができない．そこで，ここでは物体，動き，場所と人物概念の関係性を正確に表現で

きているかどうかで評価する．物体カテゴリ z^O，動きカテゴリ z^M，場所カテゴリ z^P，と人物カテゴリ z^U の関係性は，その同時確率 $P(z^O, z^M, z^P, z^U)$ で表現することができる．正解となる同時確率 $\hat{P}(z^O, z^M, z^P, z^U)$ は，表 4.1 に示した各物体と動きの関係の学習サンプル数から，次式を用いて求めた．

$$\hat{P}(z^O, z^M, z^P, z^U) = \frac{N_{z^O, z^M, z^P, z^U}}{N} \tag{4.2}$$

ただし，N_{z^O, z^M, z^P, z^U} は，物体カテゴリ物体カテゴリ z^O，動きカテゴリ z^M，場所カテゴリ z^P，と人物カテゴリ z^U の共起したデータ数であり，表 4.1 から求めることができる．また，N はデータの総数である．また，mMLDA と近似モデルで学習された同時確率は $P(z^O, z^M, z^P, z^U)$ は，次のように計算可能である．

$$P(z^O, z^M, z^P, z^U) = \sum_z P(z^O|z)P(z^M|z)P(z^P|z)P(z^U|z)P(z|\alpha) \tag{4.3}$$

ここでは学習された同時確率 がどれだけ正解 $\hat{P}(z^O, z^M)$ に近いかで評価し，その評価基準として次式で定義される Kullback-Leibler(KL) 距離を用いた．

$$D_{KL}\left(P(z^O, z^M, z^P, z^U) \| \hat{P}(z^O, z^M, z^P, z^U)\right) = \tag{4.4}$$
$$\sum_{z^O}\sum_{z^M}\sum_{z^P}\sum_{z^U} P(z^O, z^M, z^P, z^U) \log \frac{P(z^O, z^M, z^P, z^U)}{\hat{P}(z^O, z^M, z^P, z^U)}$$

KL 距離は 2 つの確率分布に対して，それらの間の差異を測るものであり，各モデルと正解基準との違いを表している．近似モデルの結果と mMLDA の結果の正解との KL 距離を求めた結果，それぞれ 11.34 と 8.53 となり，mMLDA の学習結果が正確に近いという結果となった．すなわち，mMLDA の方が近似モデルに比べ，より正確に物体と動きの関係性を捉えられているといえる．

また実験では，上位カテゴリ数はノンパラメトリックな MHDP によって推定された 30 を用いた．しかし，この上位カテゴリ数によっても形成された上位カテゴリは変化してしまう．そこで，KL 距離を用い正解の同時確率と比較することで，上位カテゴリ数の妥当性について評価する．mMLDA により，上位カテゴリ数を変化させて概念形成を行い同時確率 $P(z^O, z^M, z^P, z^U)$ を計算し，正解となる同時確率 $\hat{P}(z^O, z^M, z^P, z^U)$ との KL 距離を計算した．その結果が図 4.10 であり，横軸がカテゴリ数，縦軸が正解との KL 距離である．カテゴリ数が少ない場合，少ない

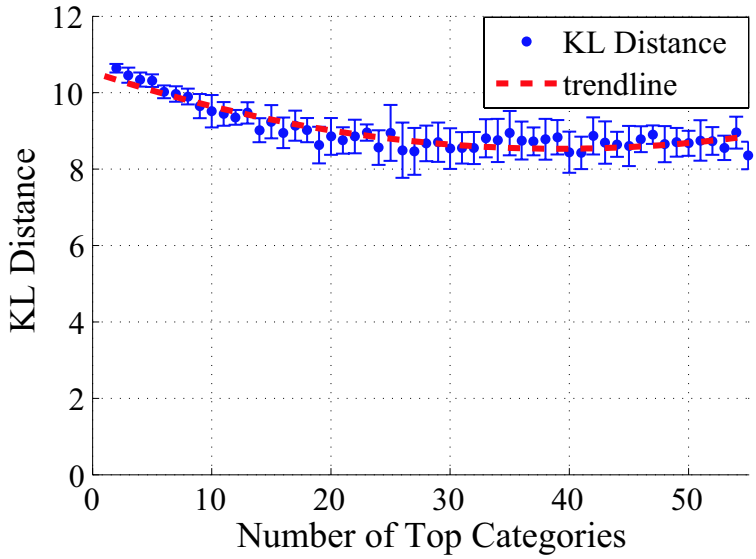

図 4.10: 上位カテゴリ数に対する同時確率分布の正解との KL 距離

パラメータで物体と動きの関係を表現するため，正しく学習できず，正解との KL 距離が大きくなっている．一方，カテゴリ数が多くなると，多くのパラメータで表現できるため，正しくその関係を捉えることができ，正解との KL 距離が小さくなる．さらに，上位カテゴリ数がある一定以上大きくなると，KL 距離は収束し変化しなくなるが，細かく分類しすぎてしまうために，正しい概念が形成できない恐れがある．そのため，図 4.10 より，上位カテゴリ数は 30〜40 の範囲が妥当であると考えられ，今回 MHDP で推定された上位カテゴリ数 30 は適当であるといえる．

4.3 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，可観測の情報から未観測の概念の予測を行った．実験は次の 4 つのパターンにおいて行った．

1. 物体の視・聴・触覚情報から，動き・場所・人物のカテゴリを予測
2. 動きの角度情報から，物体・場所・人物のカテゴリを予測
3. 場所の座標情報から，物体・動き・人物のカテゴリを予測

表 4.4: 未観測情報のデータ

No	動き	物体	場所	人物	No	動き	物体	場所	人物
1	上下に振る	ガラガラ	リビング	女の子	16	持ち上げる	飲み物（缶）	ダイニング	男の子
2	上に投げる	ぬいぐるみ	リビング	女の子	17	口に運ぶ	ペットボトル	ダイニング	大人の女
3	左右に動かす	フローリングワイパー	ダイニング	大人の女	18	口に運ぶ	スナック	リビング	男の子
4	皿を洗う	スポンジ	キッチン	大人の女	19	持ち上げる	カップヌードル	ダイニング	大人の男
5	皿を洗う	たわし	キッチン	大人の女	20	開ける	スナック	リビング	大人の男
6	手に塗る	ハンドクリーム	リビング	大人の女	21	持ち上げる	スプレー缶	リビング	女の子
7	テーブルに置く	消臭剤	リビング	大人の女	22	中身を注ぐ	じょうろ	庭	大人の男
8	テーブルに置く	除湿剤	リビング	大人の女	23	持ち上げる	プラスチックカップ	ダイニング	大人の女
9	中身を注ぐ	シャンプー	浴室	大人の男	24	すくう	ショベル	庭	大人の男
10	取り出す	ティッシュ箱	リビング	大人の男	25	口に運ぶ	野菜（玩具）	ダイニング	男の子
11	包む	ラップ	ダイニング	大人の女	26	左右に動かす	車（玩具）	リビング	男の子
12	持ち上げる	茶碗	ダイニング	大人の男	27	上に投げる	マラカス	リビング	男の子
13	上下に振る	ドレッシング	ダイニング	大人の男	28	履く	靴	玄関	大人の男
14	中身をかける	蜂蜜	ダイニング	男の子	29	上に投げる	ボール	リビング	男の子
15	上下に振る	ソース	ダイニング	男の子	30	積み重ねる	積み木	リビング	男の子

表 4.5: 未観測情報の予測精度

観測した情報	mMLDA				近似モデル			
	物体	動き	場所	人物	物体	動き	場所	人物
視・聴・触覚	-	76.67%	80.00%	73.33%	-	66.67%	70.00%	70.00%
角度	86.67%	-	80.00%	90.00%	76.67%	-	73.33%	80.00%
座標	76.67%	76.67%	-	100%	70.00%	76.67%	-	90.00%
性別・年齢	80.00%	83.33%	86.67%	-	76.67%	73.33%	80.00%	-

4. 人物の性別・年齢情報から，物体・動き・場所のカテゴリを予測

実験のデータは表 4.4 に示した．未観測情報の予測は提案モデルの mMLDA と近似モデルによって行い，両方の結果を比較した．予測結果の評価は，表 4.1 に基づいて，観測した情報に関係する全ての未観測概念のカテゴリを正解とする．例えば，観測した物体が‘飲み物 (缶)(17)” である時，表 4.6 に示したカテゴリが正解とした．表 4.5 に mMLDA と近似モデルを用いた未観測予測の全ての結果を示した．その結果，mMLDA では，近似モデルに比べ，下位概念の関係性を正しく捉えられているため高い精度となった．

表 4.6: 飲み物（缶）に関係する物体，場所，人物のカテゴリ（カッコ内の数字はカテゴリ番号）

動き	物体	場所	人物
持ち上げる (1)	飲み物（缶）(17)	ダイニング (4)	女の子 (3)
口に運ぶ (3)	飲み物（缶）(17)	ダイニング (4)	大人の男 (2)
口に運ぶ (3)	飲み物（缶）(17)	ダイニング (4)	女の子 (3)
上下に振る (6)	飲み物（缶）(17)	ダイニング (4)	大人の女 (1)
上下に振る (6)	飲み物（缶）(17)	ダイニング (4)	大人の男 (2)
中身を注ぐ (15)	飲み物（缶）(17)	ダイニング (4)	女の子 (3)
中身を注ぐ (15)	飲み物（缶）(17)	ダイニング (4)	男の子 (4)

物体の情報から未観測情報を予測する実験では，図 4.1 に示した赤い矩形で表示された物体を認識用データとして用い，残りの物体を学習用のデータとし，観測された物体のマルチモーダル情報 $(\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ から動きカテゴリ z^M，場所カテゴリ z^P と人物カテゴリ z^U の予測を行った．図 4.11 が，“飲み物 (缶)(17)” から予測された未観測である動きカテゴリ，場所カテゴリ，人物カテゴリが発生する確率 $P(z^M|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$，$P(z^P|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ と $P(z^U|\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h)$ である．

mMLDA を用いた動きカテゴリの予測結果 (図 4.11(a)) では，正しく “持ち上げる (1)” や “口に運ぶ (3)” といった動きが予測することができる．一方，近似モデルを用いた予測の結果 (図 4.11(d)) では，“中身をかける (14)” といった動きが高い確率で予測されている．これは，近似モデルの分類結果では，物体の “飲み物 (缶)(17)” と “ドレッシング (14)” が同じカテゴリに分類されてしまったため，“ドレッシング (14)” に関係する “中身をかける (14)” が予測された考えられる．このように，近似モデルでは，物体と動きが独立しており相互に影響を及ぼさないため，誤分類を修正することができず，予測精度が低下したといえる．

また，mMLDA を用いた場所カテゴリの予測結果 (図 4.11(b)) では，正しく “ダイニング (4)” を予測することができている．一方，近似モデルの結果 (図 4.11(e)) では，誤った “キッチン (3)” が最も高い確率で予測されている．しかし，人物カテゴリの予測結果では,，表 4.6 に示したに示した通り，全ての人物カテゴリに関係するため，mMLDA と近似モデルの予測結果 (図 4.11(c) と図 4.11(f)) が異なったと

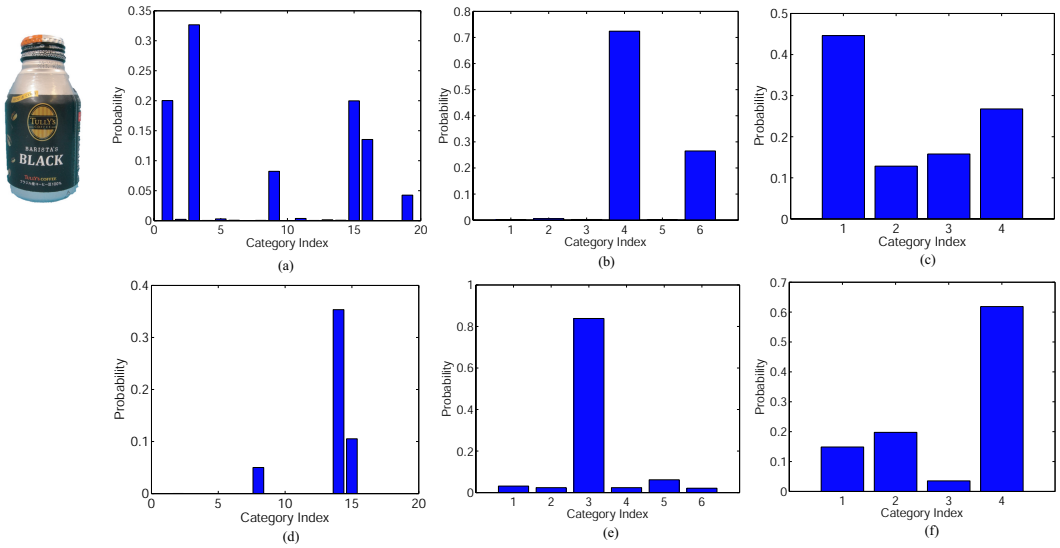

図 4.11: “飲み物（缶）(17)” から mMLDA と近似モデルを用いた各概念のカテゴリの発生確率：(a) mMLDA で動きカテゴリ, (b) mMLDA で場所カテゴリ, (c) mMLDA で人物カテゴリ, (d) 近似モデルで動きカテゴリ, (e) 近似モデルで場所カテゴリ, (f) 近似モデルで人物カテゴリ

しても，正しく予測できる．

以上のように，近似モデルに比べ mMLDA の予測性能が高いことが分かる．

4.4 単語情報に関する概念選択

相互情報量を用いて，学習データの単語情報に各概念に対する重みを求め，図 4.12 に示した．図 4.12 において，横軸が各概念を表し，縦軸が単語を表す．図 4.12(a) は事前に定義した各単語と概念の正解となる結び付けである．図 4.12(b) は mMLDA の学習結果から求めた各概念との相互情報量を表す．評価のために，単語選択の結果をあらかじめ作成した表 4.7 の正解の単語リストと比較した．但し，各単語について相互情報量が最も高い概念を，その単語の概念として選択した．その結果，正解率は 68.75%となった．また，各概念の正解率をより詳しく見ると表 4.8 に示した正解率となった．以上の結果から分かるように，場所と人物概念の正解率が一番高くなった．これは，表 4.8 に示したとおり，両概念に接地する単語の数が少なく，

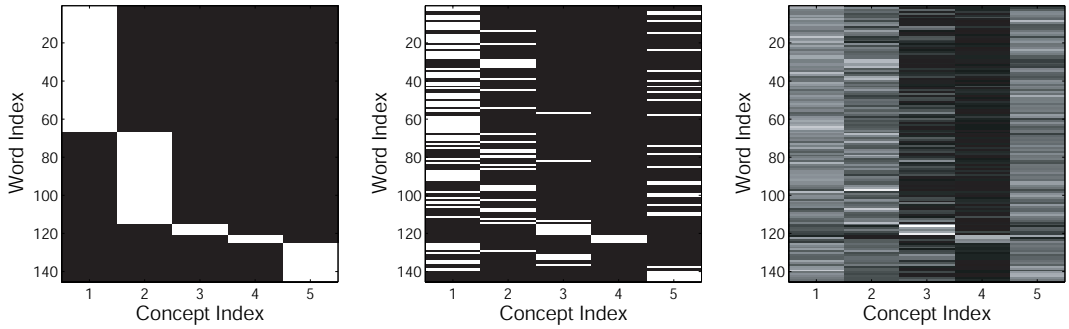

図 4.12: 概念選択の結果

表 4.7: 各概念を表現する単語の一部

物体	動き	場所	人物	統合
ガラガラ	かける	キッチン	女の子	塗料
スナック	運ぶ	ダイニング	男の子	飲む
飲み物	塗る	リビング	父	食べる
ペットボトル	動かす	玄関	母	拭く
ぬいぐるみ	投げる	庭		遊ぶ

形成された分類精度も高かったため，概念が正しく選択される．

4.5 未観測情報の単語予測実験

次に，単語予測実験について説明する．まずは，物体概念から単語の予測を行った．物体の “ぬいぐるみ” から予測された単語が図 4.13 である．図 4.13(a) は “ぬいぐるみ” の視・聴・触覚情報が観測された時の単語の発生確率を表し，統合概念を表す “リビング” という単語が一番高い確率を持つと予測された．一方，本稿で提案した相互情報量を各概念に対する重みづけとして求めた結果が図 4.13(b) である．重みづけの結果から，物体概念を表す “ぬいぐるみ” の単語が一番高い確率を持つことが分かる．他にも，“スプレー缶” から予測された単語の発生確率では，“庭” という単語が一番高くなった．この予測結果に単語の相互情報量による重み付けを

表 4.8: 各概念における概念選択の正解率

	物体	動き	場所	人物	統合	全概念
単語数	91	48	6	4	32	181
正解率	78.78%	53.33%	100%	100%	56.52%	68.75%

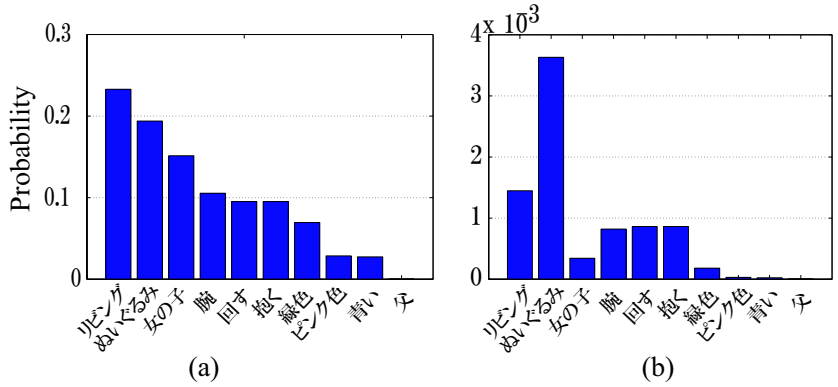

図 4.13: “ぬいぐるみ” からの単語予測：(a) 単語の発生確率，(b) 相互情報量による重み付けをした単語発生確率

することで，“スプレー” と “缶” という単語が正しく予測された．このように，相互情報量の重みづけによって，単語を正しく予測することが可能である．

同様に，動き情報のみが観測された時の単語予測において，“持ち上げる” の動き情報から単語の予測を行った結果を図 4.14 に示した．図 4.14(a) から “ダイニング” といった単語が高い確率で予測された．一方，図 4.14(b) の結果から，動き概念以外に関係する単語の確率は，相互情報量の重みづけによって低くなり，“持ち上げる” や “もつ” が高く予測された．しかし，今回の学習データにおいて，“持ち上げる” という単語は統合概念を表す単語と設定したのにも関わらず，相互情報量の重みづけにおいても，動き概念と統合概念との相互情報量の値がほぼ同じとなった．その為，“持ち上げる” という動き情報に対して，動き概念を表す “もつ” が 2 番目に高く予測される結果となった．他の例として，“口に運ぶ” という動きに対する単語の予測では，“ダイニング” という単語が一番高く発生すると予測された．これに，単語と各概念との相互情報量により重み付けすることで，“口” と “運ぶ”

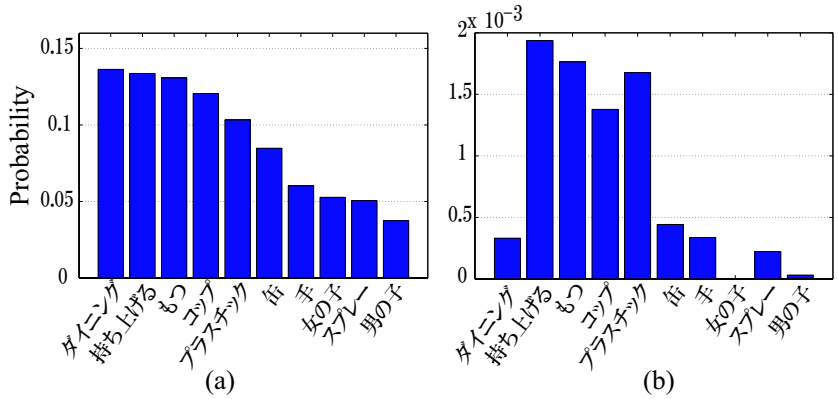

図 4.14: “持ち上げる” からの単語予測：(a) 単語の発生確率，(b) 相互情報量による重み付けをした単語発生確率

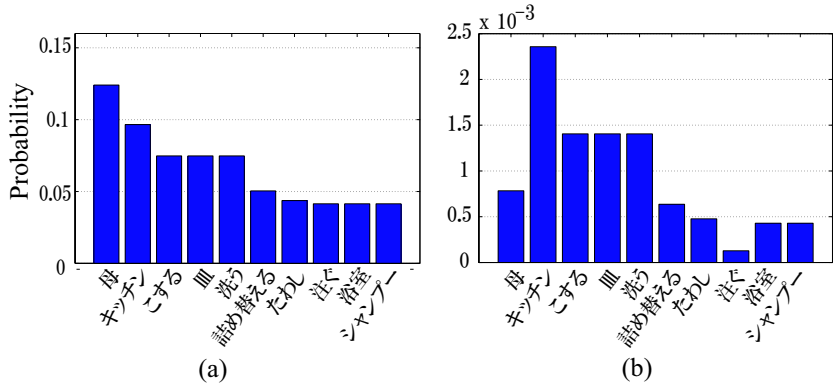

図 4.15: “キッチン” からの単語予測：(a) 単語の発生確率，(b) 相互情報量による重み付けをした単語発生確率

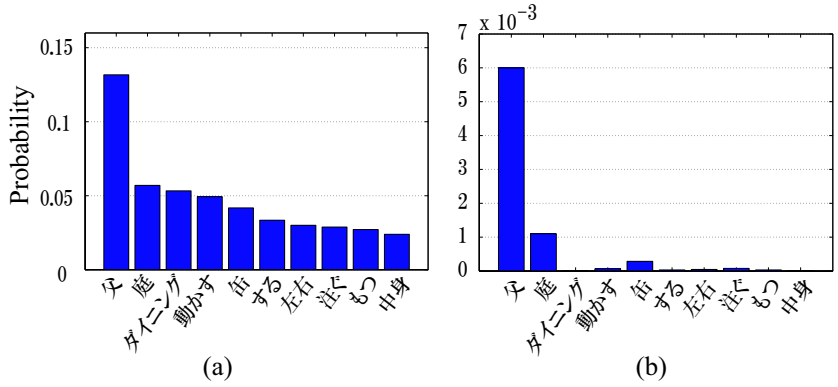

図 4.16: “大人の男” からの単語予測：(a) 単語の発生確率，(b) 相互情報量による重み付けをした単語発生確率

表 4.9: 文章生成用のデータ

No	動き	物体	場所	人物
1	上げる	スプレー缶	庭	父
2	上に投げる	ボール	リビング	男の子
3	片手で口に運ぶ	ペットボトル	ダイニング	母
4	左右に動かす	車（玩具）	リビング	男の子
5	皿を洗う	たわし	キッチン	母

といった単語が予測され，正しい予測であると言える．

図 4.15 に示した “キッチン” の場所情報から予測された単語の結果も提案手法の有効性を示した．単語発生確率 (図 4.15(a)) の結果では，“母” という単語が高く予測された．一方，提案手法を用いた結果 (図 4.15(b)) では，正しく “キッチン” といった場所概念に関係する単語が予測された．但し，人物の予測結果では，今回のデータにおいて，単語発生確率と提案手法は同様な結果を示した (図 4.16)．

表 4.10: 生成された文章の結果

No	生成された文章
1	{ 父，庭，スプレー，缶，ダイニング，持ち上げる }
2	{ 男の子，リビング，投げる }
3	{ 母，ダイニング，ペットボトル，ダイニング，運ぶ，}
4	{ 男の子，リビング，車，動かす，遊ぶ }
5	{ 母，キッチン，たわし，洗う }

4.6 未観測情報の文章生成

次に，文章生成の実験について説明する．言語モデルは 660 個の教示文章から学習した．実験は表 4.9 に示したデータの組み合わせを用いた．

文章生成の結果を表 4.10 に示した．ここで，様々構成をもつ文章が生成され，助詞などの機能語を除いて，正しく文章を生成することができた．但し，表 4.10 の 1 番に示した結果において，生成された単語 “ダイニング” は “スプレー” と “缶” に続いて物体概念に関係する単語ですが，間違った単語予測の結果です．これは，本論文で提案した手法を用いた各概念の単語予測においては，その概念に関係する単語を一番高い確率で予測できますが，2 番目や 3 番目には間違った単語予測の可能性があります．例えば，図 4.13(b) に示した物体概念の単語予測の結果では，“ぬいぐるみ” といった物体を表す単語が予測できますが，2 番目の予測結果は “リビング” であった．これと同様に，表 4.10 の 1 番の結果にも，間違った単語予測結果が文章の中に生成された．

第 5 章

結論

本論文では，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．今後の課題として，mMLDA を用いることでより多様な概念の統合を行い，ロボットによる複雑な概念形成を実現することが挙げられる．また，MHDP を階層化することでノンパラメトリックベイズ手法への拡張を行うことも今後の課題である．

更に，語意の獲得に関して，本論文で提案した相互情報量を用いて単語の各概念に対する重みを計算することで，mMLDA における単語の予測性能を向上させることが可能であることを実験を実験結果により示した．また，提案手法による生成された文章は定性的な評価により有効性を示したが，今後の課題として，定量的な評価も行う必要があると考えられる．

謝辞

本研究において，多大なる助言，ご指導を頂ました長井隆行准教授に心から感謝致します．研究内容やプログラミング，論文の執筆に至るまで，丁寧にご指導頂きました．また，ロボカップや国内外での学会発表など，数多くの活躍の場を与えて頂き，非常に貴重な経験をすることができました．この場を借りて，心より御礼申し上げます．ありがとうございました．

また，研究がお忙しい中，確率統計の基礎に始まり，プログラミングや具体的な実験内容など，幾度となくご助力を頂きました中村友昭先輩に心より感謝致します．研究に関して知識がなかった学部の頃から，修士卒業に至るまで，本当にお世話になりました．数多くの実績を上げることができたのは，長井先生，そして中村先輩のご支援があってのことだと思います．本当にありがとうございました．

また，研究室に在籍した 3 年間，共に研究に励み，何度も相談に乗って下さったムハンマドアッタミミ先輩，そして様々な場面で応援して頂いた長井研究室の皆様に心より感謝致します．

参考文献

[1] R. Fergus, P. Perona, and A. Zisserman, “Object Class Recognition by Unsupervised Scale-Invariant Learning”, in Proc. of CVPR 2003, vol.2, pp.264–271, 2003

[2] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T. Freeman, “Discovering Object Categories in Image Collections”, in Proc. of ICCV 2005, pp.370–377, 2005

[3] L. Fei-Fei, “A bayesian hierarchical model for learning natural scene categories,” IEEE Conference on Computer Vision and Pattern Recognition, pp.524–531, 2005.

[4] C. Wang, D. Blei, and L. Fei-Fei, “Simultaneous image classification and annotation,” IEEE Conference on Computer Vision and Pattern Recognition, vol.0, pp.1903–1910, 2009.

[5] E. Torres-Jara, L. Natale, and P. Fitzpatrick, “Tapping into Touch,” Lund University Cognitive Studies, pp.22–24, 2005

[6] J. Sinapov, and A. Stoytchev, “Object Category Recognition by a Humanoid Robot Using Behavior-grounded Relational Learning”, in Proc. of ICRA 2011, pp.184–190, 2011

[7] 中村, 長井, 岩橋, “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.J92-D, no.10, pp.2507–2518, Oct. 2008

[8] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, “Grounding of Word Meanings in LDA-Based Multimodal Concepts”, Advanced Robotics, 25, pp.2189-2206, 2011

[9] 長井, 中村, “マルチモーダルカテゴリゼーション: 経験を通して概念を形成し言葉の意味を理解するロボットの実現に向けて”, 人工知能学会誌, Vol.27, No.6, pp.555-562, Nov.2012

[10] Blei, D. M., Ng, A. Y. and Jordan, M. I., “Latent Dirichlet Allocation”, Journal of Machine Learning Research, pp. 993–1022, 2003.

[11] W. Takano, H. Imagawa, and Y. Nakamura, “Prediction of Human Behaviors in the Future through Symbolic Inference”, in Proc. of ICRA 2011, pp.1970–1975, 2011

[12] W. Takano, and Y. Nakamura, “Bigram-Based Natural Language Model and Statistical Motion Symbol Model for Scalable Language of Humanoid Robots”, in Proc. of ICRA 2012, pp.1232–1237, 2012

[13] T. Taniguchi, and S. Nagasaka, “Double Articulation Analyzer for Unsegmented Human Motion using Pitman-Yor Language Model and Infinite Hidden Markov Model”, in Proc. of SII 2011, pp.250–255, 2011

[14] T. Ogata, S. Nishide, H. Kozima, K. Komatani, and H. Okuno, “Intermodality Mapping in Robot with Recurrent Neural Network”, Pattern Recognition Letters, vol.31, no.12, pp.1560–1569, 2010

[15] L. Montesano. M. Lopes, A. Bernardino, and J. S.-Victor, “Learning Object Affordances: From Sensory-Motor Coordination to Imitation”, IEEE Trans. on Robotics, vol.24, no.1, Feb. 2008

[16] B. Moldovan, P.Moreno, M. Otterlo, J. S.-Victor, and L. D. Raedt, “Learning Relational Affordance Models for Robots in Multi-Object Manipulation Tasks”, in Proc. of ICRA 2012, pp.4373–4378, 2012

[17] A. Gupta, A. Kembhavi, and L. S. Davis, “Observing Human-Object Interactions: Using Spatial and Functional Compatibility for Recognition”, IEEE Trans. on PAMI, vol.31, no.10, pp.1775–1789, Oct. 2009

[18] B. Yao, and L. Fei-Fei, “Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses”, IEEE Trans. on PAMI, vol.34, pp.1691–1703, Sep. 2012

[19] T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, T. Toda, H. Okada, and T. Omori, “Learning Novel Objects for Extended Mobile Manipulation”, Journal of Intelligent and Robotic Systems, vol.30, pp.1–18, 2011

[20] A. Vedaldi, and B. Fulkerson, “Vlfeat: An open and portable library of computer vision algorithms,” ACM International Conference on Multimedia, pp.1469–1472, 2010

[21] 中村，西田，長井，“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

[22] O. Mangin, and P.-Y. Oudeyer, “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

[23] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M., “Hierarchical Dirichlet Processes”, Journal of the American Statistical Association, vol.101, 1566–1581, 2006

[24] 中村, 荒木, 長井, 岩橋, “階層ディリクレ過程に基づく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”, 計測自動制御学会論文集, Vol.49 No.4, Apr.2013

[25] M.Attamimi, A.Mizutani, T.Nakamura, T.Nagai, K.Funakoshi, M.Nakano, “Real-Time 3D Visual Sensor for Robust Object Recognition”, in Proc. of IROS, pp.4560–4565, 2010

[26] 伊東慶輔，中村友昭，長井隆行, “RRT を用いたコンフィギュレーション空間における双腕同時経路計画”, 日本ロボット学会学術講演会論文集，1M2-2，2010

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

[28] 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, 2010.03

発表実績

(1) M. Fadlil 他, “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”, 人工知能学会全国大会, 2G4-OS-19a-3, 2013

(2) M. Fadlil 他, “HDP-HMM を用いたロボットによる物理的知識の獲得”, 日本ロボット学会学術講演，会 3N1-5, Sep.2012

(3) M. Fadlil 他, “階層ディリクレ過程隠れマルコフモデルを用いたロボットによる物理的知識の獲得”, 創発システムシンポジウム ポスター発表, Sep.2012 (**優秀ポスター賞受賞**)

(4) M. Fadlil 他, “HDP-HMM を用いた物理的知識のモデル化”，情報処理学会第 74 回全国大会，6R-3，Mar.2012

(5) M. Fadlil, *et al.*, “Integrated Concept of Objects and Human Motions Based on Multi-layered Multimodal LDA”, in Proc. IROS2013, Nov. 2012.

(6) M. Fadlil 他, “多層マルチモーダル LDA と相互情報量による語意の獲得”, 日本ロボット学会学術講演会, Sep. 2013.