RF-002

グラフィカルモデルを用いた物体概念モデル

A Computational Model of Object Concept Using Graphical Model

中村友昭† 新地康人† 長井隆行†
Tomoaki Nakamura Yasuhito Shinchi Takayuki Nagai

1 まえがき

物体認識は，物体の視覚的特徴に基づくものが多く研究されている [1]–[3]．しかし，これは物体の理解という観点から見れば不十分であると考えられる．なぜなら，多くの物体には使用目的や用途が存在しており，これらの情報は視覚的特徴と同様に重要である [4]．そこで本稿では物体 (道具) の理解を，経験を通して習得した物体概念の適用による機能，使い方の予測であると定義する．物体を視覚特徴および機能，使い方の関係性によりモデル化することでシステムに学習させる．提案する物体概念のモデルは，グラフィカルモデルに基づいており，全体として物体の概念を表現している一方，機能の概念を表現する部分，使い方の概念を表現する部分，視覚的な特徴を表現する部分に分けることができる．機能概念は，物体が影響を及ぼす対象物の視覚的な変化を特徴ベクトルとしたガウス分布でモデル化する．使い方概念は，物体の把持位置や対象物との接触位置，使う際の手形状を特徴とした多項分布でモデル化される．これらの概念は，物体概念に先立って変分ベイズ法によって学習される．また，視覚特徴としては SIFT (Scale Invariant Feature Transform)[3] を用いる．これは，道具の視覚特徴として，局所的な特徴を取得することが必要となるためである．こうした基本的な概念を構築した後，これらの関係性を利用して物体概念全体を構築する．最終的には，学習したモデルを用いることで，例えば視覚特徴から経験的に道具の認識や機能，使い方の推定を行うことが可能となる．これは，既に述べた本稿での物体 (道具) の理解の定義につながる．

従来，物体を機能で認識する試みがなされている [5][6]．しかし，これらの研究では機能の辞書や単純な形状テンプレートと機能との関係性を表した辞書などを人手により与えており，学習の枠組みは取り入れてはいない．また，視覚的情報と人間の動作を利用した物体認識に関する研究も行われている [7]．この場合，視覚特徴と機能との関係性を学習することは行っていないため，物体の本質である機能と視覚特徴の関係性については考慮されていない．

2 物体概念モデル

本稿では道具の理解を，経験を通して習得した道具の概念の適用による機能や使い方の予測であると考える．道具の概念は，視覚情報及び，使い方と機能の関係性をモデル化することで構築する．

図 1(a) に，提案する物体概念のグラフィカルモデルを示す．図において，O, V, F, U, I, V_f はそれぞれ，物体のカテゴリ，視覚特徴，機能，使い方，物体の ID，対象物の変化に関する視覚情報 (これについては後述する)，を

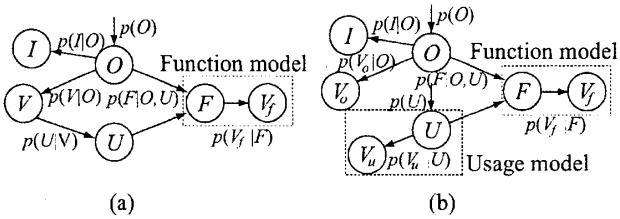


図 1 物体概念モデル (a) 視覚特徴，機能，使い方に基づくモデル (b) 使い方の視覚情報を分離したモデル

表している．このモデルは，使い方が道具の視覚的な特徴によって決まり，物体のカテゴリとその使い方が，発揮する機能に影響を及ぼすことを意味している．ここでこのモデルをより扱いやすくするために，視覚的特徴 V を，物体 (道具) 自体の視覚的特徴 V_o と，実際に使っているシーンを観測した際に得られる視覚的特徴 V_u に分割する．これによりグラフィカルモデルは，図 1(b) のようになる．本稿ではさらに，図の点線で囲まれた U と V_u からなる使い方概念モデルと，F と V_f からなる機能概念モデルを独立してモデル化 (学習) し，その後に全体の物体概念モデルとして組み合わせる．従って，モデル全体における同時確率は，

$$p(O,I,V_o,F,U,V_u,F_f) = p(O)p(I|O)p(V_o|O)p(F|O,U)p(U)p(V_u|U)p(V_f|F) \quad (1)$$

となるが，$p(V_u|U)$ 及び，$p(V_f|F)$ は全体とは独立に学習されるものであり，物体概念モデルのパラメータを学習する際には固定されることになる．学習と認識の詳細については，後で述べる．

3 視覚的特徴・使い方・機能

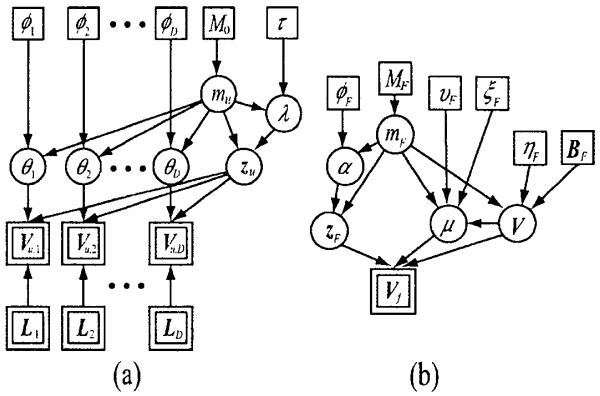
3.1 物体の視覚特徴 (V_o)

ここでは道具の視覚的特徴として，SIFT (Scale Invariant Feature Transform) [3] を用いる．SIFT は画像中の特徴点を抽出し，各特徴点とその周囲の点の方向と勾配を計算する．そして，それらのヒストグラムを特徴ベクトルとする手法である．従って，SIFT は物体の局所的な視覚特徴量を抽出できる．SIFT による特徴ベクトルは，シーン全体の明るさや回転の影響を受けにくい．しかし，画像によって抽出される特徴点の数が異なり，道具ごとの視覚特徴量としては扱いにくい．そのため本稿では，様々な道具の画像から SIFT の特徴ベクトルを抽出し，それらの特徴ベクトルを k 平均法でクラスタリングすることで代表ベクトルを求め，各画像における代表ベクトルのヒストグラムを視覚特徴量とする．

3.2 道具の使い方モデル (U)

道具の使い方は，次の 3 つの情報で定義できると考える．道具の把持位置，対象物に変化を与える道具の部分，

† 電気通信大学大学院電気通信学研究科電子工学専攻, UEC



(a) (b)

図 2 各概念の詳細なグラフィカルモデル (a) 使い方モデル (b) 機能モデル

及び道具を持つ際の手形状である．これらの情報は，実際に人が道具を使用しているシーンから観測する．具体的には，道具を使用する前に SIFT 特徴を抽出し，道具を使用中のシーンにおいてその特徴がどのような役割か (把持されているのか，対象物に接触しているのか，それ以外か) を観測することで把持位置，対象物に変化を与える部分の SIFT 特徴を求める．また，道具を持つ際の手形状は，あらかじめいくつかの手形状を様々な角度から撮影し，手形状モデルを作成しておく．そして，道具使用中の手形状とそのモデルを比較する．これにより，尤もらしい手形状モデルを求め，道具を持つ際の手形状の観測データとする．観測した道具の把持位置，対象物に変化を与える部分，および手形状を多項分布でモデル化し，使い方のモデルとする．道具の把持位置，対象物に変化を与える部分においては，SIFT の代表ベクトルごとに役割 (把持されているのか，対象物に接触しているのか，単に観測されただけ，観測されない) が観測されるので，4 つの事象が発生しうる多項分布が代表ベクトルの数の次元数だけ存在するモデルとなる．さらに手形状においては，手形状モデル数だけ事象が発生しうる多項分布が存在するモデルとなる．共役事前分布を考慮した使い方のグラフィカルモデルは，図 2(a) のようになる．図 2(a) において，$\boldsymbol{\theta}_d$, $\boldsymbol{z}_u$, m_u, $\boldsymbol{\lambda}$ はそれぞれ，多項分布のパラメータ，隠れ変数 (どのカテゴリに属するかという情報)，モデル数，混合比である．また，$\boldsymbol{\phi}$, M_0, $\boldsymbol{\tau}$ は，ハイパパラメータ (共役事前分布のパラメータ) である．さらに，$\boldsymbol{V}_{u,d}$ 及び $\boldsymbol{L}_d$ は可観測な情報であり，それぞれ，d 番目の SIFT に関する情報 (d 番目の SIFT の内，各事象が生起した回数)，d 番目の SIFT が画像フレーム内で合計いくつ観測されたかを示している．この使い方モデルの学習には，適切なモデル数の推定が可能である変分ベイズ法 [8] を用いる．

3.3 機能の概念モデル (F)

機能は，対象物の変化を観測し，統計的手法でクラスタリングすることでモデル化する．対象物に起こるどのような変化に着目するかは非常に重要であると考えられるが，ここでは一般的な道具を考慮し，対象物の変化を表すパラメータとして，対象物の個数変化，輪郭変化，色変化，重心位置変化の 4 つについて考える．そして，観測された対象物の変化を 4 次元のベクトルとして扱い，混合ガウス分布でモデル化する．モデル化には，変分ベイズ法を用いる．図 2(b) に機能モデルの詳細を示す．この図において，$\boldsymbol{V}_f$ は観測された対象物の特徴変化ベクトルであり，各機能の平均ベクトルを $\boldsymbol{\mu}$，各機能の分散を $\boldsymbol{V}$，観測データ内における各機能の混合比を α，分類される機能数を m_F，各データが分類される機能を示す潜在変数を Z_F とする混合ガウス分布としてモデル化されている．また，図の四角は仮定した共役事前分布のパラメータであり，それぞれ，ϕ_F はディリクレ分布，M_F は一様分布，ν_F, ξ_F はガウス分布，η_F, $\boldsymbol{B}_F$ はウィシャート分布のパラメータである．

4 モデルの学習と認識

4.1 学習

物体の学習とは，モデルのパラメータである条件付確率の推定を意味する．実際に観測される情報は，物体の視覚特徴 V_o，機能の視覚特徴 V_f，使い方の視覚特徴 V_u である．ただし，機能と使い方に関してはそれぞれの概念がすでに形成されているので，各々のモデルを適用した中で最尤の機能 F とその尤度 $p(F|V_f)$ 及び，使い方 U とその尤度 $p(U|V_u)$ を使用する．また物体のカテゴリ O は非観測である．これらの不完全データに対して EM アルゴリズムを適用することで，モデルの学習を行う．提案する物体概念モデルの尤度関数は次式となる．

$$L(D) = \log \sum_U \sum_F \sum_O p(I, V_o, F, U, O|\theta) p(F|V_f) p(U|V_u) \quad (2)$$

ここでは簡単のために，さらに次のように近似する．

$$L(D) = \log \sum_O p(I, V_o, F, U, O|\theta) p(F|V_f) p(U|V_u) \quad (3)$$

これは，I, V_o, F, U を観測データ D とし，O を隠れ変数 Z としている．つまり，U, F は本来，非観測であるが，機能と使い方のモデルを適用した際の最尤のものを F, U の観測結果とする．ここで，上式に Jensen の不等式を適用すると，

$$L(D) \geq F(q(O), \theta) = p(F|V_f) p(U|V_u) \times \sum_O q(O|I, V_o, F, U, \hat{\theta}) \log \frac{p(I, V_o, F, U|\theta)}{q(O|I, V_o, F, U, \hat{\theta})} \quad (4)$$

と書くことができる．よって尤度関数を最大化するために，下限である $F(q(O), \theta)$ を $q(O)$ と θ について交互に最大化する．ここで，θ をパラメータとし，$\hat{\theta}$ をパラメータの推定値とする．式 (4) の等号は次式のときに成立する．

$$q(O|I, V_o, F, U, \hat{\theta}) = p(O|I, V_o, F, U, \theta) \quad (5)$$

従って，$F(q(O), \theta)$ の $q(O)$ に関する最大化は，

$$p(O|I, V_o, F, U) = \frac{p(O)p(I|O)p(V_O|O)p(F|O, U)}{\sum_O p(O)p(I|O)p(V_O|O)p(F|O, U)} \quad (6)$$

となり，これが E-Step である．また，θ に関する最大化が M-Step であり，これらのステップを繰り返すことで最終的なパラメータを得る．

表 1　実験に用いた道具

物体カテゴリ	番号	A set [個]	B set [個]	合計
はさみ	T1	7	3	10
ペン	T2	8	3	11
ペンチ	T3	2	2	4
ピンセット	T4	3	2	5
カッター	T5	3	2	5
ホッチキス	T6	4	2	6
のり	T7	5	3	8
セロテープ	T8	4	3	7
ビニールテープ	T9	2	2	4

4.2　認識・予測

学習したモデルを用いることで，不完全な情報を周辺化し確率的に確からしい結果を導くこと（認識，予測）が可能となる．これは，経験を通して習得した道具の概念の適用による機能の予測であり，道具の理解につながる．

例えば，視覚的特徴 V_o，機能 F，使い方 U の全てを観測した上で，物体のカテゴリ認識を行うには，

$$\underset{O}{\operatorname{argmax}}\, P(O|\hat{I}, V_o, F, U) = \frac{p(O)p(\hat{I}|O)p(V_o|O)p(F|O,U)}{\sum_O p(O)p(\hat{I}|O)p(V_o|O)p(F|O,U)} \tag{7}$$

とすればよい．ただし $p(\hat{I}|O)$ は，新たな入力 $\hat{I}$ に適用するために EM アルゴリズムを用いて再計算する．また，視覚的特徴 V_o のみから機能 F を予測する場合には，

$$\underset{F}{\operatorname{argmax}}\, P(F|\hat{I}, V_o) = \frac{\sum_O \sum_U \left\{p(O)p(\hat{I}|O)p(V_o|O)p(F|O,U)\right.}{\sum_O \sum_U \sum_F \left\{p(O)p(\hat{I}|O)p(V_o|O)p(F|O,U)\right.} \times \frac{\left.p(U)p(F|V_f)p(U|V_u)\right\}}{\left.p(U)p(F|V_f)p(U|V_u)\right\}} \tag{8}$$

を計算することになる．その他の組み合わせも，同様に計算することができる．

5　実験

5.1　実験環境

実験に用いたのは，9 種類 60 個の道具である．これらの道具を学習用と認識用の 2 セットに分けた．道具の内訳は表 1 のようになる．実験では，1 つの道具につき 10 回観測データを取っており，計 600 個のデータで実験を行う．また便宜上，道具の種類ごとに番号をつけた．全ての情報は CCD カメラを用いて取得した．

5.2　機能概念のモデル化

A セットの道具を 10 回ずつ使用し，使用前後の対象物（紙）の変化を観測して，変分ベイズ法により機能の学習を行った．図 3(a) より，最適な機能のモデル数は 6 と推定されているのが分かる．また，観測データを機能で分類した結果は図 3(b) のようになる．これより，機能 1 は対象物を分裂させる機能，機能 2 は対象物の色を変える機能，機能 3 は対象物の形状を変化させる機能，機能 4 は対象物の位置の移動を行う機能，機能 5 は対象物を接着させる機能，機能 6 は対象物を接着させ色も付ける機能であることが分かる．

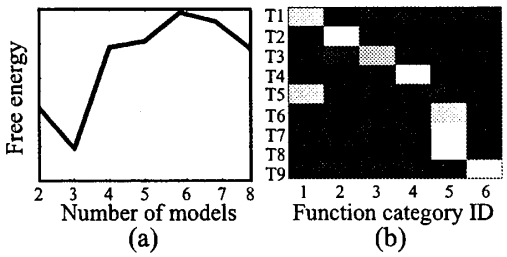


図 3　機能概念 (a) カテゴリ数と自由エネルギー (b) データの分類結果

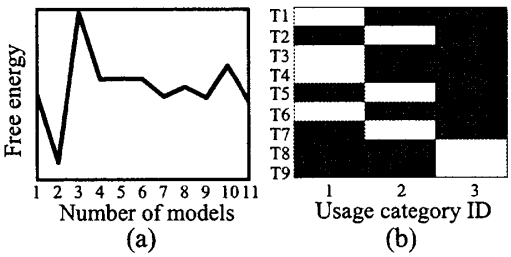


図 4　使い方概念 (a) カテゴリ数と自由エネルギー (b) データの分類結果

5.3　使い方モデルの構築

A セットの道具を使用し，道具の把持位置，対象物に接触する位置，手形状を観測した．そして，それらの情報から変分ベイズ法により使い方モデルの学習を行った．図 4(a) より，最適な使い方のモデル数は 3 と推定されていることが分かる．道具を使い方で分類した結果が図 4(b) である．ペン，カッター，のりは細長い形状をしており，把持する手形状も非常に似ているため同じ使い方に分類された．また，セロハンテープとビニールテープは道具の把持する位置の視覚特徴が類似していること，道具の本体が対象物に接触しないという共通点があることから同じ使い方に分類された．1 番の使い方に分類されたはさみ，ペンチ，ピンセット，ホッチキスは把持する際の手形状が似ていること，道具の把持位置，対象物接触位置が道具特有の形状ではあるが，観測された数が少なかったため同じ使い方に分類されたと考えられる．

5.4　物体概念のモデル化

上述の機能，使い方モデル，および視覚特徴を用いて，A セットの道具について物体概念モデルを学習した．道具の種類が 9 種類であるため，EM アルゴリズムでモデル数を 9 としてモデル化を行った．視覚特徴として，60 個の道具から観測した 500 次元の SIFT の代表ベクトルヒストグラムを使用した．物体学習における機能および使い方の有効性を示すため，視覚特徴のみでの分類，視覚特徴と機能による分類，さらに視覚特徴と機能，使い方による分類の計 3 パターンでモデル化し比較する．結果を図 5 に示す．図 5 は，道具の種類ごとにカテゴリ分類された割合を色の濃淡で表したものである．図 5 中の T1 ～T9 は，表 1 の道具番号に対応している．図 5(a)，(b) はそれぞれ視覚特徴のみでの分類，視覚特徴と機能での分類の結果である．そして，図 5(c) は視覚特徴と機能，

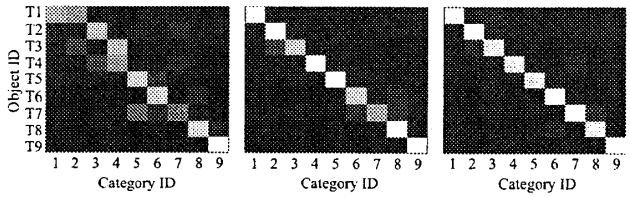


図 5　道具のカテゴリ分類結果 (a) 視覚的特徴のみによる分類結果 (b) 視覚的特徴と機能を用いた分類結果 (c) 視覚的特徴, 機能, 使い方を用いた分類結果

表 2　機能の予測結果

	分裂	色変化	変形	移動	接着	接着&色
T1	30	0	0	0	0	0
T2	0	28	2	0	0	0
T3	1	0	19	0	0	0
T4	0	0	0	20	0	0
T5	20	0	0	0	0	0
T6	0	0	0	0	20	0
T7	0	0	0	0	30	0
T8	0	0	0	0	30	0
T9	0	0	0	0	0	20

使い方での分類結果である．3 つの図を比較すると (a), (b), (c) の順に分類の精度がよくなっていることが分かる．(b) と (c) の分類の正解率を算出すると (b) は 89%に対し，(c) が 93%となり，より多角的な情報で道具を学習することで精度が向上すると言える．

5.5　視覚特徴による機能の予測

5.4 で構築した物体概念を適用し，視覚特徴から機能を予測する実験を行った．B セットの道具について，観測された視覚特徴から確率的に確からしい機能を推定した．各道具と予測された機能を，表 2 に示す．いくつかのペンとペンチで誤った機能が予測されたが，ほとんどの道具について正しい機能が予測された．正解率は 98 %であった．

5.6　視覚特徴による使い方の予測

B セットの道具の視覚特徴から把持位置，対象物に接触する位置を予測する実験を行った．結果の一部を図 6(a) に示す．図 6 は把持位置を灰色の円で，対象物に接触させる位置を白の四角で表示している．どの道具に対しても，概ね正しく把持位置や対象物に接触する位置が予測できていると言える．

5.7　機能的視覚特徴の抽出

物体の概念モデルを構築した後，各物体概念において共通して観測された視覚特徴を円で表示した結果が図 6(b) である．これらの視覚特徴は道具特有の視覚特徴であり，機能的視覚特徴であると言える．図 6(b) より提案モデルを適用することで道具特有の視覚特徴が抽出できていると言える．また，道具の変形や道具以外の物体が入ったシーンでも，道具特有の視覚特徴が抽出できており，様々なシーンにおいて，物体の認識や機能の予測ができる可能性を示している．

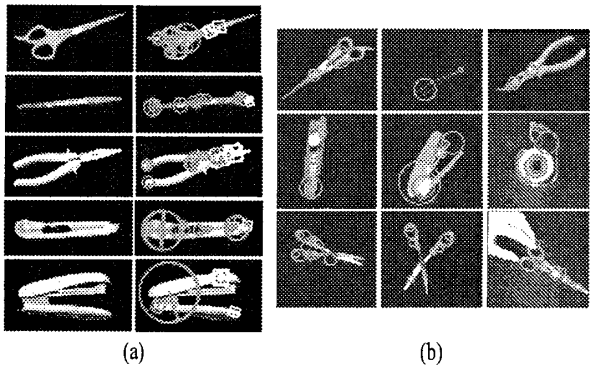


図 6　使い方の予測結果 (a) と機能的視覚特徴 (b)

6　むすび

本稿では，視覚特徴と機能，使い方に着目した物体の学習・認識の手法を提案した．機能を対象物の変化としてモデル化することにより，人手で機能を定義，拡張する必要がなく，汎用的なシステムが構築できたと言える．また，視覚的な情報による使い方のモデル化手法も提案した．最終的には，これらのモデルを組み合わせることにより，物体概念全体を構成した．物体概念のモデルを用いることで，視覚情報のみから，機能や使い方といった観測していない情報を推定可能であることを示した．今後の課題としては，更なるモデル化精度の向上，対象物のモデル化，物体概念数の自動決定などが挙げられる．また，ロボットや画像検索への応用も今後の課題である．

参考文献

[1] A.A.J.Sivic *et.al.*, " Discovering object categories in image collections, " AI Memo, 2005-005:1–12, Feb.2005.

[2] P.R.Fergus *et.al.*, " Object class recognition by unsupervised scale-invariant learning, " in Proc. of CVPR, Vol.2:264–271, Feb.2003.

[3] David G. Low, " Distinctive image features from scale-invariant keypoints, " International Journal of Computer Vision, Vol.60, No.2:91-110, Nov.2004.

[4] B. Landau *et.al.*, " Object Shape, Object Function, and Object Name", Journal of Memory and Language, " ML972533, **38**, pp.1–27,1998

[5] A.D.L.Stark *et.al.*, " Recognizing object functioin through reasoning about partial shape descriptions and dynamic physical properties, " Roceedings of The IEEE, 84(11):1640–1656, Nov.1996.

[6] K.Woods *et.al.*, "Learning membership functions in a function-based object recognition system, " Journal of Artificial Intelligence Research, 3:187-222, Oct.1995.

[7] A.Kojima *et.al.*, " Toward a Cooperative Recognition of Human Behaviors and Related Objects, " The Second International Workshop on Man-Machine Symbiotic Systems:195–206, Nov.2004.

[8] H. Attias, " Infering Parameters and Structure of Latent Variable Models by Variational Bayes, " in Proc. of Uncertainty in Artificial Intelligence, 1999