Deep Belief Nets を使った手の姿勢のマルチモーダル表現の獲得

Acquiring multimodal representation of hand grasping based on Deep Belief Nets

○　奈倉　敬典　(阪大)　　学　ジェイコブ　パイク　(阪大)
正　荻野　正樹　(JST ERATO)　　正　浅田　稔　(阪大，JST ERATO)

Takanori NAGURA, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Jacob PYKE, Osaka University, 2-1, Yamadaoka, Suita, Osaka
Masaki OGINO, JST ERATO Asada Project
Minoru ASADA, Osaka University, JST ERATO Asada Project

This paper proposes a hierarchical model to model the multi-modal information in grasping based on deep belief network. In the network, one modal information is self-organized to extract statistical information of given data, and different modal information are easily integrated in the hierarchical architecture. For the learning data, the joint angles, tactile sensors, images of hand and object to be grasped are given based on the dynamics simulator. After learning, the proposed network can recall the sensor information when grasping from the object image.

Key Words: Deep belief nets, multimodal representation, grasping

1　はじめに

近年，人間の手先のように多くのセンサーを配置したロボットハンドの開発が盛んである　[1]．ロボットがロボットハンドなどを用いて環境中の物体と関わる，例えば把持するには環境中に存在する様々な情報を適切に統合することが重要である．人間やサルは視覚などの単一の情報だけでなく．複数のモダリティから得られた情報を段階的に処理した後に統合し判断することで適切な行動を選択していると示唆されていることに着目し，本研究では，Deep Belief Nets　[2] (以下 DBN) というニューラルネットワークの一種を用いる．物体の把持におけるマルチモーダルの情報を段階的に学習させ，あるモダリティの情報に対して，それに対応した他のモダリティの適切な情報が取得できるモデルを提案する．

脳科学の知見では，サルが物体を把持しようとするとき，網膜に映った画像は大脳皮質下に送られ階層的に処理される．視覚情報は，「物体の情報」と自分の「手の情報」に分けられ，それぞれ階層的な処理の中で重要な特徴を抽出される．これらの情報はのちに AIP 野という脳部位において統合され，対応する運動コマンドが呼び出されていると考えられている [3]．

これらの知見をもとに物体の把持をモデル化した先行研究として，図 1 に示した Oztop et al. らの研究がある [4]．円柱を画像情報を入力層として用い，出力層を運動コマンドとするニューラルネットを使ったモデルを発表した．出力層である F5 野（運動と関連する脳部位）に力強い把持など，その性質で分けられた 16 のグループの運動コマンドを用意しておく．各物体を呈示したときに，バックプロパゲーションによって対応する運動コマンドを適切に結びつける学習をさせる．学習の結果，中間層は入力画像（円柱）の決まった直径と高さ，あるいはどちらかに反応する層が得られ，この中間層を AIP 野のモデルとして発表した．このモデルは物体の画像と運動コマンドの関係性を学習したものであるが，人間が把持を獲得する過程では触覚などによる体性感覚などの他の情報も関わっているとされる [5]．モダリティの異なる情報を結びつけての学習が比較的容易で，各層を特徴抽出器のようなものに自己組織化することにより階層的に構築できるニューラルネットワークの 1 種に Deep Belief Nets がある．本研究では，この Deep Belief Nets に着目し，手とオブジェクトの入力画像に対して適切な関節角度を想起できるようなマルチモーダルな情報処理システムを目指す．

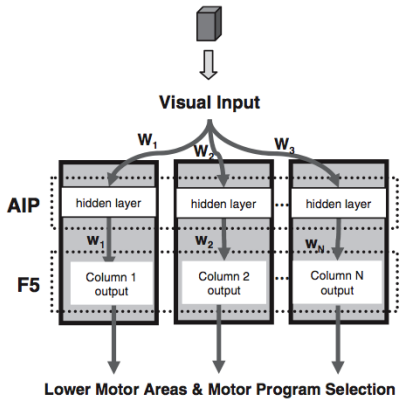


Fig.1 The model of AIP area by Oztop et al. [4]

2　Deep Belief Nets

Deep Belief Nets (以下 DBN) [2] は多数の層からなるニューラルネットワークであり，その層は Restricted Boltzmann Machine (以下 RBM) による学習を階層的に積み重ねていくことで構築される．

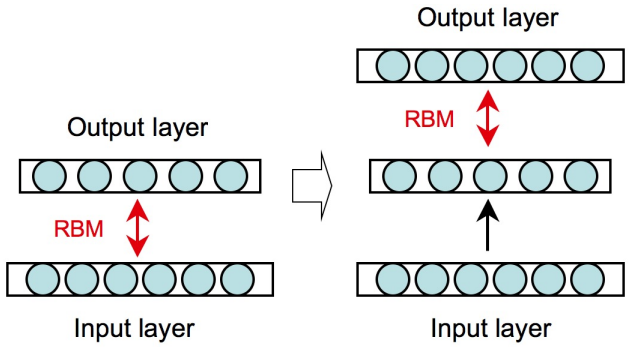
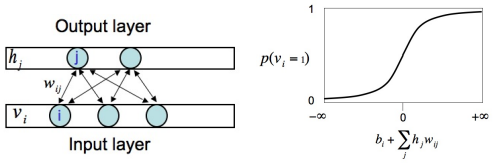


Fig.2 DBN consisting of RBNs

2.1 Restricted Boltzmann Machine

Restricted Boltzmann Machine（RBM）はユニット間に対称的な結合荷重を持つ相互結合型のニューラルネットワークである．入力層の各ユニット v_i は、出力層のユニット h_j，それぞれに対して対称的な結合荷重 w_{ij} を持つ．ただし，入力層のユニット同士，出力層ユニット同士は結合を持たない (図 3(a))．入力層のユニット，出力層のユニットはそれぞれバイアス b_j,c_i を持ち，それぞれ以下の式 (1)，式 (2) に従って確率的に出力が 1 となる (図 3(b))．



(a) Connection weights　(b) Activation Probability

Fig.3 Connection weights and activation probability in RBN

$$p(h_j = 1) = \frac{1}{1 + \exp(-b_j - \sum_i v_i w_{ij})} \quad (1)$$

$$p(v_i = 1) = \frac{1}{1 + \exp(-c_i - \sum_j h_j w_{ij})} \quad (2)$$

RBM の学習においては図 2.1 に示すような reconstruction と呼ばれる計算プロセスを使って進められる．RBM の入力層に入ってきたデータ v_i に従って出力層のユニット h_j の値が求められ，またそのときの値に従って結合荷重を使って入力層の値を再び求める．このデータを reconstruction data と以下では呼ぶ．

下記の更新式を用いて RBM の学習はなされる．

$$\Delta w_{ij} = \epsilon\big(v_i^0\, p(h_j^0 = 1) - p(v_i^1 = 1)\, p(h_j^1 = 1)\big) \quad (3)$$

$$\Delta b_j = \epsilon\big(p(h_j^0 = 1) - p(h_j^1 = 1)\big) \quad (4)$$

$$\Delta c_i = \epsilon\big(p(v_i^0 = 1) - p(v_i^1 = 1)\big) \quad (5)$$

RBM の学習が進むと入力データと reconstruction data は近いものとなっていく．学習の結果として，上の層のユニット数が下の層のユニット数より少ない場合，入力データにおける確率分布が学習され，下の層の特徴抽出を行うユニットが現れる．この特徴抽出器は，双方向性のある結合を利用して，上の層のユニットに例えば 1 を入力して下の層に出力されたものを確認することで，各層のユニットがどのような情報をコードしているかを確認することができる．

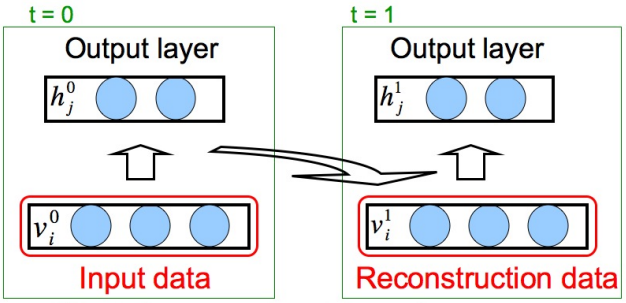


Fig.4 Calculation process during learning in RBN

2.2 Deep Belief Nets

Deep Belief Nets (DBN) は RBM による学習を積み重ねてできる多層のニューラルネットワークであり，異なるモーダル間のデータの対応関係を学習させたいときは，それぞれの RBM を組み合わせた RBM を構築することで二つのデータを統合することができる．RBM により，それぞれのデータが統計分布にしたがってあらかじめ自己組織化されているため，異なるデータ間の対応関係の学習がしやすくなっている．これはバックプロパゲーションを使って多層のネットワークの学習を行う際の前処理としても有効であり，多くのパラメータを学習させる時に局所解に陥ることを防ぐ効果がある．本研究では一部の層で，学習誤差を小さくするためにバックプロパゲーション法を用いている．この効果は人間の学習方法と似ていると Hinton らは主張している [2]．例えば，人間は多くの物体を見るが，実際にそれらの名前を教えてもらうことは少ししかないけれども，名前を教えてもらった物体と視覚情報の類似した物体にもその名前を適応することができる．これは，あらかじめ多くの物体を見るなかで物体の視覚情報が構造化されていくためである．

3 DBN を用いた pre-shaping model

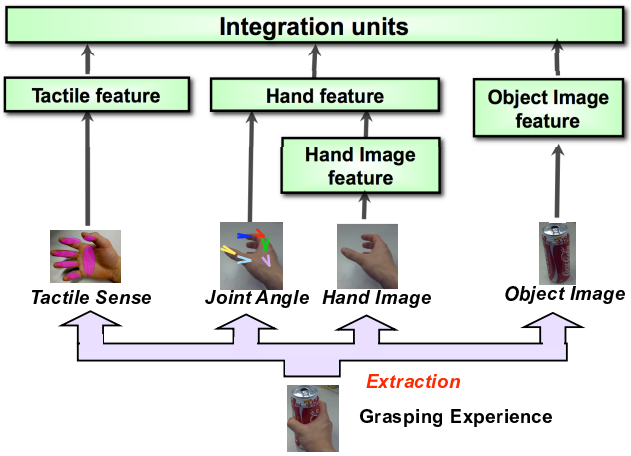


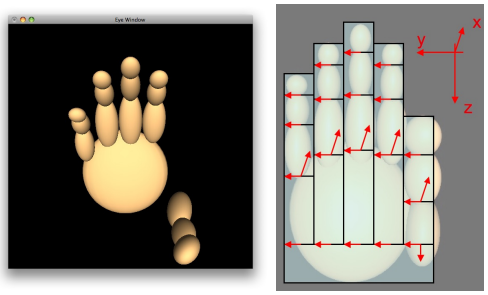
Fig.5 Overview of the proposed model

本研究では，物体把持時のマルチモーダルな情報，すなわち，「手の視覚情報」，「物体の視覚情報」，「手の関節角度情報」，「接

触情報」の 4 つのセンサー情報を統合するニューラルネットワークを提案する．図 5 に提案モデルの概要を示す．物体を把持しているときの，それぞれのモダリティにおける情報は，それぞれ RBM によって表象される．手の関節角度と画像情報に関しては，把持を行わないときの情報をもとに RBM によってあらかじめ Hand feature として統合される．そして把持時の情報をもとに，それぞれの隠れ層の出力が，上位の RBM によって統合される．学習後は，呈示された物体に対して正しい把持が想起される．通常，把持時には物体のみの情報，手のみの画像情報を得ることはできないが，本研究ではそれぞれ別の情報として入力している．つまり，物体の視覚情報はあらかじめ把持前に取得することができ，手の視覚情報は関節角度情報をもとに推定することができると仮定している．

3.1 シミュレーター

本研究では，動力学シミュレーションを使って把持のシミュレーションを行い，そのときのセンサー情報を学習に用いた．図 6 に構築した手のシミュレーションモデルを示す．手は 25 自由度からなり，1 つのリンクにつき 9 つの接触センサーを持つ．動力学計算，接触計算はオープンソースの動力学計算ライブラリである Open Dynamics Engine [6] を使用した．



(a) Input image　(b) Degree of Freedoms of simulated hand

Fig.6 Hand simulator

シミュレータではある目標の「手の関節角度」の目標値を与えると，その目標値に向かう過程での関節角度とそのときの「手の画像」のデータのセットを得ることができる．

またシミュレータのロボットは 19 の関節に図 7(b) のようにそれぞれ 9 個のタッチセンサーを持ち，ある一定以上の入力を境にして 0 もしくは 1 の離散値を「タッチセンサーから得られる入力」として獲得する．

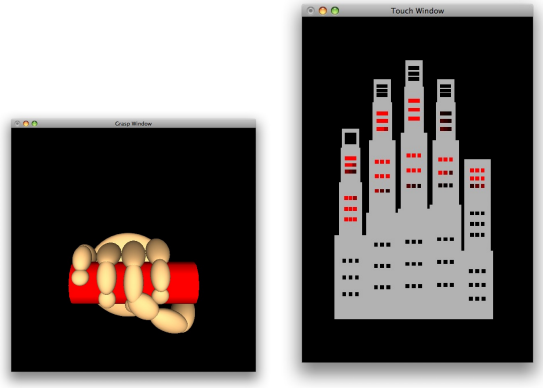
各関節角度の可動範囲を 10 等分し，関節角度の入力層として同数のノードを用意した．対応する関節角度に応じたノードを 1 とし，それ以外は 0 となるようにしている．

手の画像は入力画像にキャニーフィルターを用いてエッジ情報を抽出し，二値化を行って，画素と同数のノードを持つ入力層へ入力した．

ロボットが把持するオブジェクトは，図 8 に示すように，長さの可変な「直方体」，「円柱」，「球」を用意した．問題を簡略化するために把持において背景やロボット自身の手から分離して画像を獲得できるという仮定を置いている．

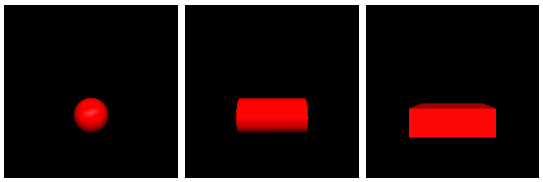
3.2 マルチモーダル情報の統合

モデル構築のための学習の手順として，まずロボットの「手の画像」と「手の関節角度」はロボットの「手の姿勢」を表現することになる．これらは 1 対 1 の関係となるため，「対象物体の画像」がこの 2 つの関係性の重みに影響を与えないように学習した後に固定する．「手の姿勢」と「対象物体の画像」の情報



(a) Grasping cylinder　(b) Tactile information

Fig.7 Grasping and Tactile Information



(a) Sphere　(b) Cylinder　(c) Box

Fig.8 Objects used in experiments

は別々の経路で処理された後に統合することになる．この関係の学習後に，「手の姿勢のモデル」と「対象物体の画像」と把持成功時の「タッチセンサーから得られる入力」との関係を学習させる．

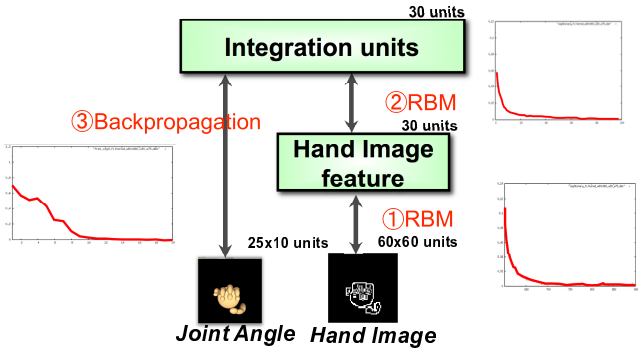


Fig.9 Integration of joint angles and image of hand

まずはロボットの「手の画像」に RBM による学習をすることによって，上位層が手の画像の特徴抽出器となる．その上に「手の姿勢」を表す層として上位に統合層を設ける．「手の画像」の情報のうち「関節角度」との関係に重要である特徴を確認するために中間層を設けた．次に「手の画像」と「手の関節角度」のデータセットを用いてバックプロパゲーション法によって，それらの対応関係を獲得する．図 9 に手の統合モデルの概要を示す．赤で示した数字は学習の順番であり，それぞれの学習段階で，適切に目標関数であるクロスエントロピー（RBN）と学習誤差（バックプロパゲーション）が小さくなっていることが

わかる．

「手の姿勢」と「対象物体の画像」，「タッチセンサーから得られる入力」との関係を構築するために．前述の「手の姿勢」の層の上にマルチモーダルの情報を統合する層となる Top layer を設ける．モダリティごとに中間層を置くことによって，各モダリティ内において情報を圧縮し，マルチモーダル情報の統合を学習が容易になる．また，中間層を置くことによって，把持に関する学習において各モダリティ毎にどのような情報が重要かを確認することができる．それぞれ RBM によりデータを圧縮した上で，「手の姿勢」「タッチセンサーから得られる入力」「対象物体の画像」の 3 つを 1 つの層のように扱い，データセットを 1 つの入力として RBM により学習することにする．以後この RBM を Supervised RBM をして表記する．この Supervised RBM の学習後，モデルの学習を終了する．

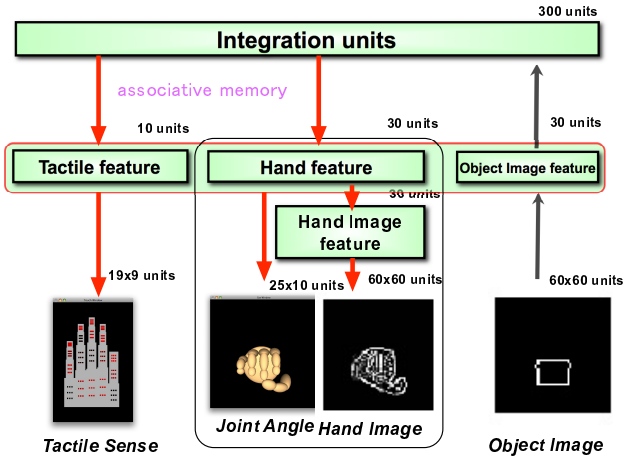


Fig.10 Reconstructing the tactile sensing and hand features from an object image

提案モデルは，連合型記憶モデルのネットワークになっているため，学習後に，「対象物体の画像」の入力により，それを把持しているときの他のモダリティの情報を想起することができる．図 10 に直方体の画像を入力したときの例を示すが，「対象物体の画像」に対応した「手の姿勢」や「タッチセンサーから得られる入力」の情報を獲得できていることがわかる．各モダリティの入力「タッチセンサーから得られる入力」「手の画像」「対象の物体」の 3 つの入力の中から 1 つ，もしくは 2 つを選んで入力をすると，入力していない層にも対応する情報が出力される．「手の画像」だけでなく「タッチセンサーから得られる入力」のデータも合わせて入力することにより，より適切な手の姿勢が想起される可能性が高くなるという結果も確認された．また確率的に出力をさせることでゆらぎを持った出力がなされ，ある入力に対して様々な出力を出すことができる．これは単なる 1 対 1 の対応関係の学習ではなく，1 つの入力に対して複数の例示を含む情報を獲得できると解釈できる．

4　考察

本研究では DBN のアルゴリズムを利用して，手の把持におけるマルチモーダル情報の統合モデルを提案した．このモデルにおいて連想記憶を利用して，他のモダリティの想起が適切に行えることを確認した．今回は把持に成功したときのデータセットを与えたが，実際は失敗したときのデータも学習に大変影響を及ぼしていると考えられる．今後は，「タッチセンサーから得られる入力」を用いて把持を評価することで学習すべきデータをネットワークに評価させるなど，与える情報を有効に使った実験をしていきたいと考える．また，本モデルでは教師あり RBM によってマルチモーダルの情報を統合したが，この結合の後に双方向のバックプロパゲーションによってモダリティ間の結合荷重を変化させることによって，中間層を入力における確率分布ではなく同時に学習させるモダリティの情報に応じて変化させていくことができる可能性がある．

参考文献

[1] Hiroshi Yokoi, Alejandro Hernandez Arieta, Ryu Katoh, Wenwei Yu, Ichiro Watanabe, and Masaharu Maruishi. Mutual adaptation in a prosthetics application. In *Embodied Artificial Intelligence*, pp. 146–159. Springer verlag, 2004.

[2] Geoffrey Hinton. To recognize shapes, first learn to generate images. In *Computational Neuroscience:Theoretical Insights into Brain Function.*, 2007.

[3] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos, and G. Rizzolatti. Object representation in the ventral premotor cortex (area f5) of the monkey. *The Journal of Neurophysiology*, Vol. 78, No. 4, pp. 2226–2230, 1997.

[4] Erhan Oztop, Michael A. Arbib, and Nina Bradley. The development of grasping and the mirror system. *Action to Language via the Mirror Neuron System*, pp. 397–423, 2006.

[5] Jane Case-Smith and Charlane Pehoski. *Development of Hand Skills in the Child.* 協同医書出版社, 1996.

[6] Russell Smith. Open dynamics engine. *http://www.ode.org/*.