3H4-OS-24b-3

Deep neural network による映像・音響・運動データの統合と共起

Intersensory synchrony modeling and memory retrieval using deep neural networks

野田 邦昭　有江 浩明　菅 佑樹　尾形 哲也
Kuniaki Noda　Hiroaki Arie　Yuki Suga　Tetsuya Ogata

早稲田大学 理工学術院 基幹理工学研究科
Graduate School of Fundamental Science and Engineering, Waseda University

Humans are known to succeed in enhancing perceptual precision and reducing ambiguity by combining inputs from multiple modalities, including vision, audition, and somatic sensation. This paper presents a novel computational framework for modeling intersensory synchrony of multimodal temporal sequences based on a deep learning framework. To evaluate our proposed model, we designed a bell ring task by a humanoid robot. Our experimental results demonstrate that the cross-modal memory retrieval function of the proposed method succeeds in generating visual sequence from the corresponding sound and bell ring motion.

1. はじめに

人は，視覚，聴覚，体性感覚など複数のモーダルから得られた情報を統合化して環境を理解することにより，知覚の曖昧さを低減し，堅牢な認識を実現している．しかし，これまでのロボティクスにおいては，画像認識，音声認識など，それぞれの目的に特化したセンサ情報を独立に用いることで情報処理を行うことが慣例となってきた．一方，機械学習の分野では，Deep learning が近年注目を集めており，特に音声認識への応用においては唇画像および音響情報を統合化して認識することで雑音に対する耐性の高い音声認識を実現できることが報告されている [Ngiam 2011]．本研究では，Deep learning の学習フレームワークをロボットの感覚運動統合学習に導入し，実環境下での行動経験によって得られた時系列情報から複数モーダル間の共起性を自己組織的に獲得することが可能な学習フレームワークを提案する．これにより，クロスモーダルな記憶連想を実現し，身体運動情報とそれに伴う音響情報から，モーダル間の共起性を正しく反映した画像情報の生成行う．また，獲得された内部表現の解析より，提案モデルは視聴覚運動情報の共起性を自己組織的に構造化し，記憶学習していることを示す．

2. マルチモーダル時系列記憶連想システム

2.1 システム全体の構成

本研究で提案するマルチモーダル時系列記憶連想システムの概要を図 1 に示す．本システムでは，音響，画像および関節角度から構成される時系列情報の次元圧縮と統合学習のために，多段階層型ニューラルネットワークを用いる．ニューラルネットワークの構造は，入出力次元に対し，中央の中間層のニューロン数が最も小さくなるような砂時計型を用いる．さらに，入出力が同一になるような制約の下で学習パラメータを更新することにより，恒等写像（オートエンコーダ）の学習を行う．ネットワークの学習には，Martens によって提案されている学習手法を用いる [Martens 2010]．これにより，最狭部の中間層を境として，もとの入力情報を次元圧縮し，特徴ベクトルを生成するための変換と，次元圧縮された特徴ベクトルから元の情報を復元するための変換を行う写像変換が，階層型ネットワークに同時に獲得される．

連絡先: 野田邦昭，早稲田大学理工学術院，東京都新宿区大久保 3-4-1，03-5286-2742，kuniaki.noda@akane.waseda.jp

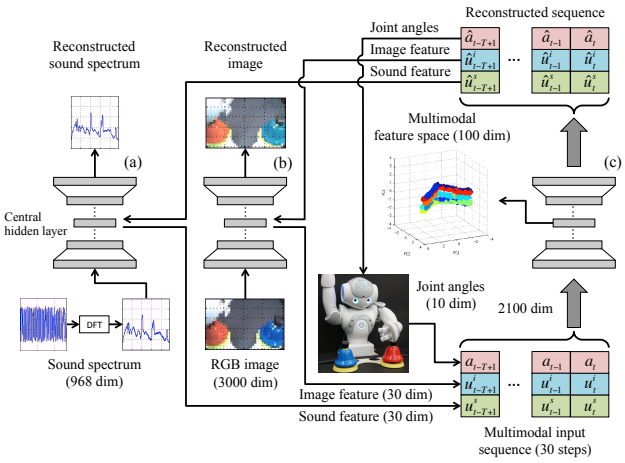

図 1: マルチモーダル時系列記憶連想システム

2.2 マルチモーダル時系列の学習と記憶連想

音響情報は，短時間フーリエ変換によって周波数スペクトル時系列に変換した後，オートエンコーダを用いて次元圧縮し，音響特徴ベクトルを生成する．画像情報は，直接オートエンコーダによって次元圧縮し，画像特徴ベクトルを生成する．ロボットから得られた関節角度ベクトル時系列と，次元圧縮の結果得られた音響特徴ベクトルおよび画像特徴ベクトル時系列を合わせたものを入力ベクトル時系列とし，恒等写像学習を行う．再帰結合を持たない階層型ニューラルネットワークで多次元時系列を学習するために，タイムディレイ型ニューラルネットワークの学習 [Lang 1990] を行う．これは，一定の時間幅を持ったウィンドウで多次元時系列をサンプルし，学習器への入力情報を生成する手法である．クロスモーダルな記憶連想は，音響特徴ベクトル時系列，画像特徴ベクトル時系列，および関節角度時系列のうち 2 つのモーダルを入力とし，残りのモーダルの時系列を出力層から取得することにより行う．なお，入力に外部からデータを与えないモーダルに関しては，ネットワークの出力として得られたベクトルを再帰的に入力にフィードバックすることにより，時系列を内部生成する．

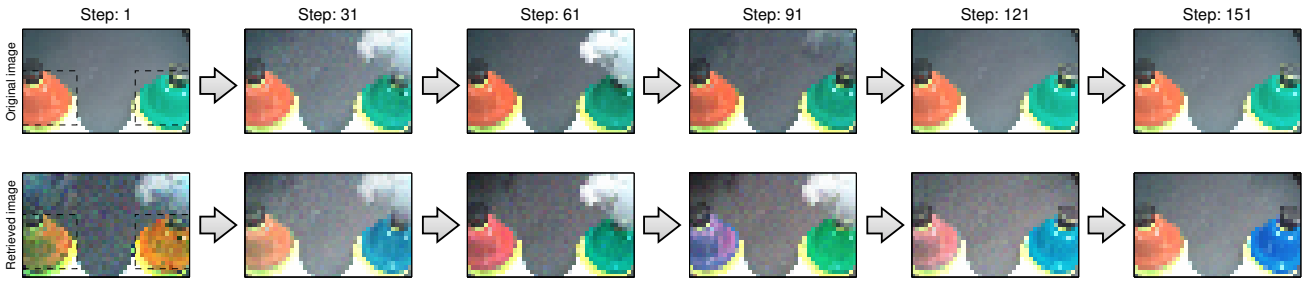

図 2: 音響情報および関節角度時系列から生成された画像時系列の例．上段：正解の画像時系列，下段：想起された画像時系列．画像中の黒い破線で囲まれた領域（13×13 ピクセル）は，画像想起の性能評価に用いた左右のベル画像領域を示す．

表 1: データサンプル数と実験パラメータ

	学習データ	入出力	エンコーダ次元
音響	5352	968	1000-500-250-150-80-30
画像	2688	3000	1000-500-250-150-80-30
時系列	8736	2100	1000-500-250-150-100

3. 評価実験

3.1 実験設定

提案する学習フレームワークを評価するため，Aldebaran Robotics 社の NAO を用いてベル叩き行動の感覚運動統合学習を行った．色および音程の対応で一意に特定できる 3 種類のベルを用い，ベル叩き行動を生成した際の音響，画像，運動に関する時系列データを用いた．2 つのベルを選択してロボットの正面の左右に置き，手先位置指令によって関節角度軌道を生成し，6 種類のベル配置に対してベル叩きを行った．学習には頭部正面のマイクロフォンで記録した音響情報（周波数スペクトル），頭部のカメラから取得した画像を切り出し，40×25 の解像度にリサンプルした RGB 画像，および左右の腕の関節角 10 自由度を用いた．各ネットワークの学習に用いた学習データサンプル数，各ネットワークの入出力次元およびネットワークの構造に関するパラメータを表 1 に示す．学習に用いたネットワークの構造は，音響特徴学習と画像特徴学習については共に 12 階層（結合重みの層数）とし，時系列学習については 10 階層とした．なお，これらの数値は経験的に設定した．

3.2 実験結果

音響特徴ベクトルと関節角度ベクトルからなる時系列を入力とし，画像時系列をクロスモーダルな記憶連想によって生成した結果の例を図 2 に示す．ベルが実際にたたかれる前は，ベルを特定するための音響情報が得られないため，実際の画像に対し，想起された画像のベルの色はランダムに初期化されている（図 2（b）1 ステップ）．その後，実際にベルが叩かれると，叩かれたベルに対する音響情報が入力されるため，関節角度時系列から得られた情報と対応し，実際に叩いた側のベルの色が正しく想起できている（図 2（b）61 ステップ）．なお，叩かれた側と反対のベルに関しては，ベルを特定できる音響情報が得られないため，残された 2 種類のどちらかの色からランダムに対応づけられることが，複数の実験結果から確認された．

3 つのモーダルの情報からなる時系列を統合学習した際にネットワークの中間層から得られた 100 次元の特徴ベクトルに対して主成分分析を行い，第 3 主成分までの情報を 2 つの主成分からなる平面に射影して視覚化した感覚運動特徴空間を図 3 に示す．結果，第 1 主成分と第 2 主成分からなる平面には，左右のベル叩きの運動と対応した特徴空間が自己組織化されていることが確認された（図 3（a））．一方，第 3 主成分はベルの配置と対応していることが確認された（図 3（b））．

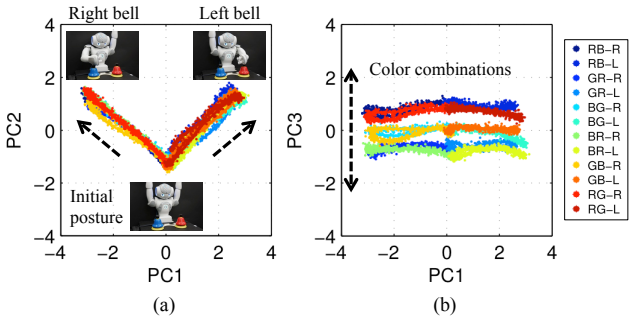

図 3: 自己組織化された感覚運動統合特徴空間

4. まとめ

本研究では，ロボットの感覚運動情報を統合化して学習することにより，モーダル間の共起性を自己組織的にモデル化し，獲得された共起性を反映した記憶連想を実現する学習フレームワークを提案した．評価実験として，ロボットを用いたベル叩きタスクを設定し，画像情報，音響情報，関節角度時系列の統合学習を行った．さらに，音響情報と関節角度情報から画像を想起する記憶連想実験を行い，獲得されたモデルを評価した．想起画像を評価した結果，身体運動と対応した側のベルの色が音響情報と対応したものに正しく変化することが確認され，提案モデルが複数のモーダル間の共起性を正しくモデル化し，獲得された共起性に基づいて記憶連想することが可能であることを示した．今後は画像の想起だけでなく，音響データの再現や運動生成など他のモーダルの組み合わせによる記憶連想実験を行う予定である．

謝辞　本研究は，さきがけ領域研究「情報環境と人」及び科研費新学術領域研究「構成論的発達科学」(24119003）の助成を受けた．

参考文献

[Ngiam 2011] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng,“ Multimodal deep learning, ”Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011.

[Martens 2010] J. Martens: Deep Learning via Hessian-free Optimization, Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010.

[Lang 1990] K. Lang, A. Waibel, and G. Hinton: A Time-Delay Neural Network Architecture for Isolated Word Recognition, Neural networks, vol. 3, pp. 23-43, 1990.