1R-2

多層神経回路モデルによる共感覚現象の学習と連想

山口 雄紀 † 野田 邦昭 ‡ 西出 俊 † 奥乃 博 † 尾形 哲也 ‡

† 京都大学 大学院情報学研究科 知能情報学専攻 ‡ 早稲田大学 基幹理工学部表現工学科

1. はじめに

人間社会で活躍されることが期待されるロボットは多様なモダリティ情報を統一的な枠組みで扱うことが必要である．服部らは共感覚現象の一例であるブーバ・キキ効果に注目し，再帰結合型神経回路モデルを用いた視聴覚間モダリティ変換を実現し，音から連想される視覚変化をロボットの動作生成に応用した [1]．この手法の問題点は，学習能力の制約による入力自由度やデータ数の制限とそれに付随する表現力不足であった．本研究では，多層神経回路モデルを導入することでモデルを改良し，学習対象として共感覚現象を再現することが異種モダリティ変換において重要であると考え，共感覚研究によって取得されたデータを利用して検証を行った．

2. 共感覚現象

人のモダリティ変換に関連する研究として共感覚研究がある．共感覚とは，文字に色が見えたり，音からイメージを連想するなど 1 つの感覚刺激から，複数の感覚が引き起こされる知覚現象である．共感覚現象は，一部の人にのみ起こる現象であるが，一般的に人がクロスモーダルな認知を行っていることは Ramachandran らによって示されたブーバ・キキ効果によって分かっている [2]．Ramachandran らは，尖った直線からなる図形と滑らかな曲線の図形を被験者に見せ，どちらが“ ブーバ”でどちらが“ キキ”であるかを問う実験を行った．そして，約 95 %の被験者が，前者が“ キキ”で後者が“ ブーバ”であると答えたという結果を得た．このことから，人の知覚において異種モダリティ間の変換が行われていることが分かる．

3. 視聴覚間モダリティ変換システム

提案するモダリティ変換システムの概要を図 1 に示す．本システムでは画像圧縮および時系列学習のために多段型階層ニューラルネットワークを用いる．圧縮された画像特徴量と音響特徴量からなる時系列を恒等写像変換となるように学習することにより，一方のモダリティ情報を入力すれば，もう一方のモダリティ情報が連想できるようになる．本システムの学習および連想は以下の 3 つの手順で行う．

1. 画像圧縮用ニューラルネットワークの学習
2. 時系列用ニューラルネットワークの学習
3. 音響情報時系列からの画像情報時系列の生成

3.1 画像圧縮用ニューラルネットワークの学習

本研究では各時刻における音響情報として 12 次元の Mel-frequency cepstrum coefficient (MFCC)，画像入力として RGB 画像を用いる．画像と音のモーダル間で次元の調整をするため，前処理として画像情報の次元圧縮を行．具体的には学習器としてフィードフォワード型の

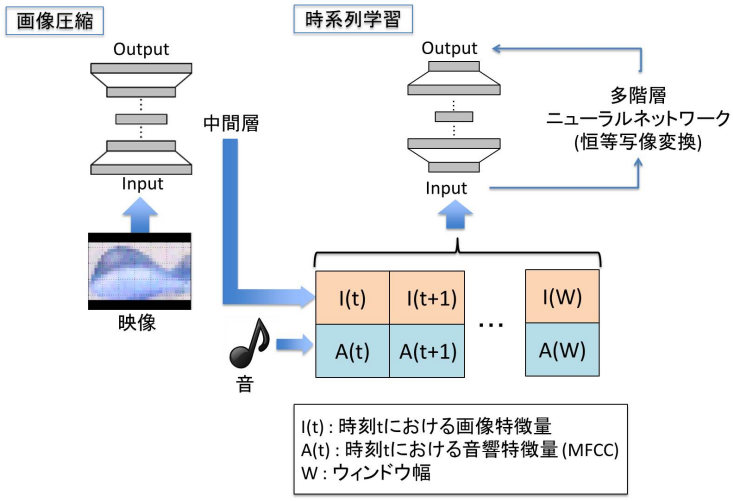


図 1: モダリティ変換システムの概要図

多層ニューラルネットワークを用い，学習手法として，Martens によって提案されている学習アルゴリズムを用いる [3]．入力データの次元圧縮と復元を実現するため，ニューラルネットワークの構造は，中央の中間層のニューロン数が最も小さくなる砂時計型の構造をとるものとし，入出力が同一になる恒等写像変換の学習を行った．これにより，中間層の値が各画像の特徴ベクトルとして取り出せる．なお，画像特徴ベクトルは元の情報を保持したまま圧縮されているので，元の画像を復元することが可能である．

3.2 時系列用ニューラルネットワークの学習

視聴覚間モダリティ変換を行うために音および画像特徴量からなるマルチモーダルな時系列の学習を行う．まず，時系列データを一定の時間幅を持ったウィンドウで 1 時刻ごとにシフトしながらサンプリングする．得られる各ウィンドウをそれぞれ 1 つのデータと見なして画像圧縮と同様の手法で砂時計型の多層ニューラルネットワークが恒等写像変換となるように学習を行う．

3.3 音声情報時系列からの画像情報時系列の生成

音声情報時系列から画像情報時系列を生成する際には，時系列用ニューラルネットワークに時系列の初期時刻からウィンドウ幅分の音声情報のみを入力し，恒等写像により初期ウィンドウの画像情報を生成する．以降の時刻については，ウィンドウ幅分の音声情報と 1 時刻前の画像出力を再帰的に入力に戻すことで，時系列全体を生成する．生成された画像情報を視覚的に確認するため，画像圧縮用ニューラルネットワークの最狭部に生成された画像情報を入力し，残りの階層でフィードフォワード計算をすることで，画像の復元を行う．

4. 評価実験

構築したモダリティ変換システムの評価のために，共感覚患者を対象に取得された映像データを用いた実験を

Learning and Association of Synethesia Phenomenon Using Multilayered Neural Netowork Model: *Yuki Yamaguchi (Kyoto Univ.), Kuniaki Noda (Waseda Univ.), Shun Nishide (Kyoto Univ.), Hiroshi G. Okuno (Kyoto Univ.), and Tetsuya Ogata (Waseda Univ.)*

行った．映像データを元にシステムを学習し，学習に使用した音および未学習の音に対して共感覚患者が連想した画像とシステムが生成する画像の比較を行った．

4.1 学習データ

本研究では，共感覚患者に数種類の楽器音を聞かせ，それぞれの音から連想されるイメージをアニメーションにした動画 (http://www.youtube.com/watch?v=O8Die3XX_NY) を学習データに用いた [4]．本実験ではこの動画の中から 7 種類の楽器音を選択し，そのうち 6 種類の楽器音と映像で学習を行い，残りの 1 つを未知の音として評価に用いた．これらの楽器音は高さや長さ，大きさなどが異なっている．学習データは 1 時刻につき，解像度が 32 × 24 の RGB 画像 2304 次元と MFCC のベクトル 12 次元からなる．各データは 40msec ごとにサンプリングされる．

4.2 ネットワーク構成

学習に使用した各ネットワークの構成を表 1 に示す．中間層のニューロン数は経験的に，1000，500，250，150，80，30/50，80，150，250，500，1000 に設定した．画像圧縮により，30 次元の画像特徴量を求め，MFCC のデータと合わせ，合計 42 次元ベクトルの時系列が得られた．時系列学習の際はウィンドウの時間幅を 10step としたので，入力次元は 420 次元となった．

表 1: 各ネットワークの構成

	階層数	入出力層	最狭中間層
画像圧縮	12	2304 dim	30 dim
時系列学習	12	420 dim	50 dim

4.3 実験結果

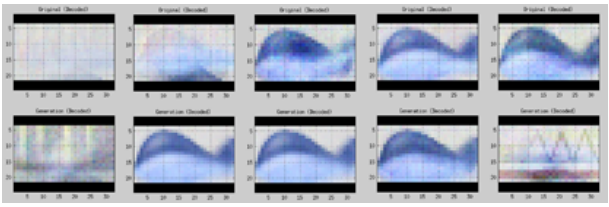
図 2 に既知の音から連想された映像の一部を示す．各図では上段に共感覚患者が楽器音から実際に連想した正解映像を，下段にモダリティ変換システムによって生成された映像を 200msec ごとに並べている．この図より，楽器音 1 の一部フレームで正解と異なる映像が連想されているが，それ以外のフレームではかなり正解に近い画像が生成されている．この結果から音のみからでも映像をほぼ再現できていることが分かる．

また，図 3 に未知の楽器音から連想された映像と共感覚患者が連想した正解映像を示す．連想された映像は，楽器音 1 と楽器音 3 の映像が順に現れている．これらの映像は学習した映像の中でも正解映像と類似する部分があるものである．楽器音 1 の映像には，正解映像に現れる特徴的な曲線形状と似た図形が現れている．一方，楽器音 3 の映像には，正解映像と同様の色が現れている．このことから，未知の音に対しても共感覚患者の知覚する映像に近い映像が連想できていると考えられ，モダリティを超えた感覚ダイナミクスの記述が可能であることが示唆される．

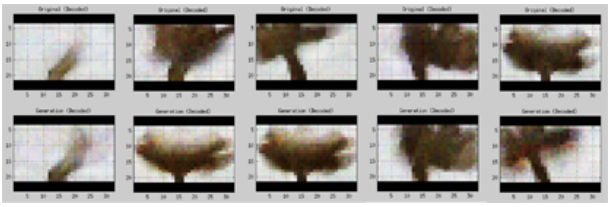
5. おわりに

本研究では，モダリティ変換を実現するために共感覚現象に注目した．共感覚患者が楽器音から連想した映像と音からなるデータを多層神経回路モデルを用いて学習することで，共感覚者と同様の知覚が再現できた．

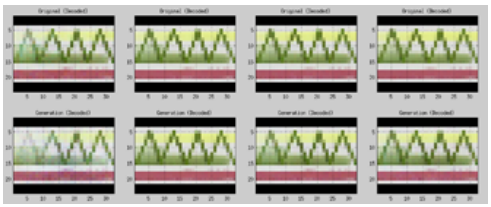
今後は本研究で実現したシステムをロボットのモダリティ変換に応用し，行動生成に結び付ける予定である．



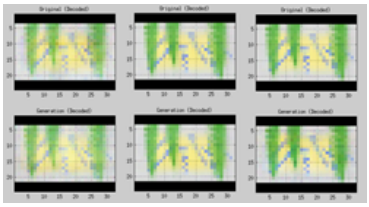
(a) 楽器音 1　左から 0, 200, 400, 600, 800msec



(b) 楽器音 2　左から 0, 200, 400, 600, 800msec



(c) 楽器音 3　左から 0, 200, 400, 600msec



(d) 楽器音 4　左から 0, 200, 400msec

図 2: 既知の楽器音から生成された映像 (下) と共感覚患者が連想した正解映像 (上)

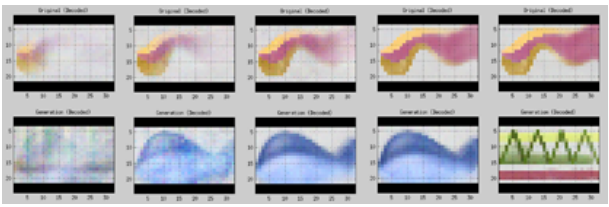


図 3: 未知の楽器音から生成された映像 (下) と正解映像 (上)　左から 0, 200, 400, 600, 800msec

謝辞　本研究は，さきがけ領域研究「情報環境と人」，科研費新学術領域研究「構成論的発達科学」(24119003)，科研費若手研究活動スタート支援 (23800033)，栢森情報科学振興財団の助成を受けた．

参考文献

[1] 服部 佑哉, 駒谷 和範, 尾形 哲也, 小嶋 秀樹, 奥乃 博: RNNPB を用いたモダリティ間マッピングによるロボットの動作生成, 情報処理学会第 68 回全国大会, 6L-8, 2006.
[2] Ramachandran, V. S., and Hubbard, E. M.: Synaesthesia: A window into perception, thought and language, Journal of Consciousness Studies, Vol. 8, No. 12, pp. 3-34, 2001.
[3] James Martens: Deep Learning via Hessian-free Optimization, Proceedings of the 27th International Conference on Machine Learning (ICML), 2010
[4] Jamie Ward: The Frog Who Croaked Blue: Synesthesia and the Mixing of the Senses, Taylor & Francis, 2008.