5R-08

深層モデルを用いた画像系列の特徴量自己組織化と映像生成

川野 真子 † 西出 俊 ‡ 康 鑫 ‡ 任 福継 ‡

† 徳島大学 理工学部 ‡ 徳島大学 大学院社会産業理工学研究部

1. はじめに

我々の日常生活には様々なモダリティが飛び交っており，それらを統一的に扱うマルチモーダルな情報を扱った研究が盛んに行われている．中でも，あるモダリティから別のモダリティに変換する研究が注目されており，例として唇画像系列から音声パラメータを推定する研究が挙げられる [1]．本研究でも同様にモダリティの変換に注目し，図 1 のように映像と音のマルチモーダルな情報変換を行うシステムの構築を目指している．目標システムでは映像と音からそれぞれの特徴量系列を抽出し，神経力学モデル Multiple Timescale Recurrent Neural Network (MTRNN) によって両モダリティを統合的に学習する．本稿ではこれまで構築した図 1 の緑枠の映像に関する特徴量系列の自己組織化と時系列学習と生成について報告する．

2. 提案モデル

本研究では図 1 の映像の特徴量系列の自己組織化モデルとして畳み込みオートエンコーダを用い，時系列の学習モデルとして MTRNN を用いる．本章ではそれぞれのモデルについて紹介する．

2.1 畳み込みオートエンコーダ

画像の特徴量を自己組織化するモデルとしてオートエンコーダ (図 2) が挙げられる [2]．オートエンコーダは，多層ニューラルネットワークを用いた恒等写像学習モデルであり，入力層から中間層までがエンコーダ，中間層から出力層までがデコーダの役割を果たす．これにより，入力データの情報量を小さくした特徴量を中間層に自己組織化することが可能となる．本研究では，モデル内部に畳み込み層を付加した畳み込みオートエンコーダを用いて，映像から切り出した画像系列の特徴量を自己組織化する．

2.2 神経力学モデル MTRNN

MTRNN は現時刻の状態 $IO(t)$ を入力することによって，次時刻の状態 $IO(t+1)$ を出力する予測器である．入力部と出力部は同様の構造になっており，入出力 (IO) 層と Fast Context(C_f) 層，Slow Context(C_s) 層の 3 つの層から

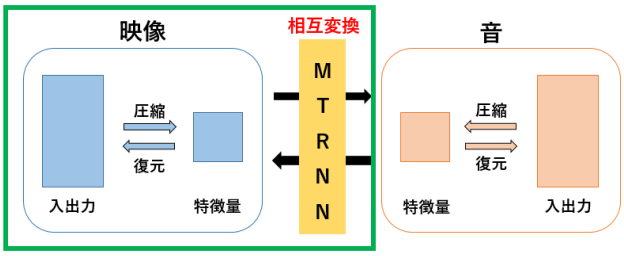

図 1: 目標システム

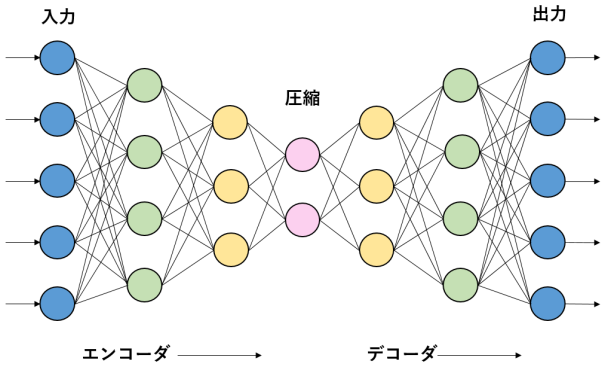

図 2: オートエンコーダ

ら構成される．各層には時定数と呼ばれるパラメータを設定し，時定数によって層内のニューロンの発火速度が制御される．この発火速度の違いによって，MTRNN は階層的に異なるレベルの情報を学習可能とする．MTRNN の構成を図 3 に示す．

MTRNN には学習・認識・生成の 3 つの基本的な機能が存在する．学習では IO 層に与えた時系列データに対して順計算を行うことで各時刻における出力値を計算する．出力値と教師信号の誤差をもとに Back Propagation Through Time (BPTT) を用いて，重みと C_s の初期値 $C_s(0)$ を更新する．認識では認識する時系列データを IO 層に入力し，BPTT を用いることによって $C_s(0)$ のみを更新する．これによって，時系列データを表現する $C_s(0)$ が求まる．生成では，任意の $C_s(0)$ を MTRNN に入力し，順計算することによって $C_s(0)$ が表現する時系列データを出力する．

本モデルを用いて $C_s(0)$ から映像を生成する際，MTRNN の生成機能によって各時刻における画像特徴量を計算し，その特徴量を畳み込みオートエンコーダに入力することで各時刻における画像を生成する．生成した画像を統合することで映像を生成する．映像生成の概略図を図 4 に示す．

3. 実験設定

本実験では，畳み込みオートエンコーダで画像系列データの特徴量を抽出し，MTRNN を用いて特徴量の学

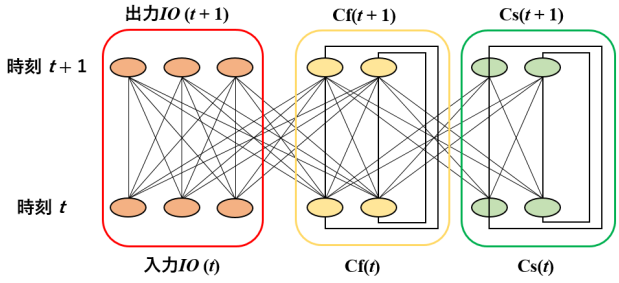

図 3: MTRNN の構成

Self-organization of Image Feature Sequences and Video Generation Using Deep Network *Mako Kawano (Tokushima Univ.), Shun Nishide (Tokushima Univ.), Xin Kang (Tokushima Univ.), and Fuji Ren (Tokushima Univ.)*

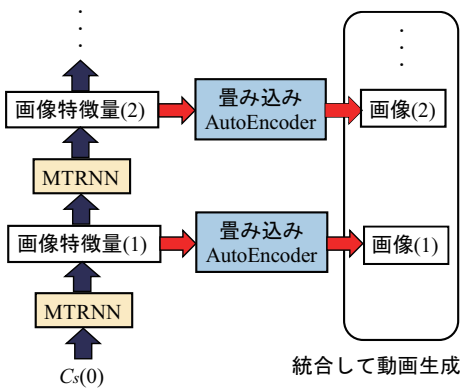

図 4: 映像生成の概略図

表 1: 実験で用いた映像の動作パターン

動作パターン	キノコを取ったとき フラワーを取ったとき 敵 (クリボー/カメ) を踏んだとき マリオが死んだとき

習・認識・生成を行った．評価として (1) 元画像，(2) 畳み込みオートエンコーダのみを用いた場合の復元画像，(3) 畳み込みオートエンコーダと MTRNN で生成された特徴量に基いた復元画像の 3 つを比較した．

本実験で用いる映像データはマリオのプレイ動画において映像と音に関連性のある部分を切り出したものとした．これは映像と音の間に関連性があるデータを用いる方がマルチモーダルな変換を行う際に評価がしやすいためである．本実験では表 1 に示す 4 種類の映像を切り出した．切り出した映像の長さはそれぞれ 2 ～ 4 秒であり，30[frames/sec] で画像系列 (128 × 128pixel，RGB の 3 チャンネル) に分割した．学習には 42 個の動画 (4212 枚の画像) を用いた．

畳み込みオートエンコーダは畳み込み層とプーリング層を対にし，エンコーダ部とデコーダ部それぞれに 5 層ずつ設定した．入力（出力）ニューロン数は 128 × 128 × 3 であり，特徴量を自己組織化する中間層のニューロン数を 128 とした．MTRNN のニューロン数は *IO* 層が 128，C_f 層を 100，C_s 層を 30 とした．*IO* 層は畳み込みオートエンコーダの中間層で自己組織化した特徴量を入力し，C_f 層と C_s 層のニューロン数は実験的に決定した．

4. 実験結果

本章では実験の結果について，主観的に評価した結果を報告する．

本実験では各動作パターンに対し，(1) 元映像，(2) 畳み込みオートエンコーダで圧縮・復元した画像列を組み合わせることで作成した映像，(3) 畳み込みオートエンコーダで圧縮した特徴量を MTRNN で認識・生成し，復元した映像をそれぞれ比較した．(1) と (2) を比較すると，42 個の動画ほぼすべてについて背景部分の復元には成功した．一方，マリオや敵などの詳細な部分の復元は困難であり，詳細部はぼやけた映像として生成された．(1) と (3) を比較すると，42 個中 17 個の動画は背景の復元に成功し，25 個の動画については背景の生成も困難であった．詳細部については (2) と同様ほとんどぼやける結果となった．(3) の動画について，復元に失敗した動画は動画の初期では復元に成功していたが，ステップ数は増えるにつれ，ノイズが乗るものが多かった．

5. 考察

本実験の結果より，1. 学習性能と映像復元と 2. 詳細部分への対応の 2 点について考察する．

5.1 学習性能と映像復元

本実験では 42 個の動画のうち 17 個の動画は大まかな背景の復元に成功し，25 個の動画は復元に失敗した．画像の復元に失敗した原因としては，MTRNN モデルの構成が大きく影響をしていると考えられる．MTRNN の生成機能においては前時刻の出力を逐次的に入力に入れるため，誤差が積算する．そのため，通常のニューラルネットワークの学習と比較し，十分に低い学習誤差を得る必要がある．モデルの構成や学習方法の改良などが課題として残されている．

5.2 詳細部分への対応

本実験の結果，背景部分よりもマリオや敵などの詳細部分の復元が困難であった．その原因として，畳み込みオートエンコーダに入力する際に画像を圧縮したことや特徴量に落とし込むことで情報量が欠損してしまうことが挙げられる．一方，MTRNN で学習する際はデータの次元数が大きいと学習時間が増大する上に学習性能が劣化するため，特徴量の次元数とモデル全体の性能はトレードオフの関係にある．詳細部に対応する方法として，詳細部のみを抽出して畳み込みオートエンコーダで学習する手法が挙げられる．また，本実験では MTRNN で復元した映像は元映像より変化が遅かった．これは MTRNN が急激な変化を学習することが困難であるという特性に起因すると考えられ，今後はこれらの問題点を解決することが必要である．

6. おわりに

本稿では映像と音のマルチモーダルな情報変換を行うモデルの基盤システムとして，畳み込みオートエンコーダと神経力学モデル MTRNN を用いた映像の特徴量自己組織化と生成について報告した．実験では畳み込みオートエンコーダによって画像系列の特徴量を自己組織化し，MTRNN で特徴量の学習・認識・生成を行った．実験結果より，本モデルによって大まかな情報の復元が可能であることを確認した．

今後は，本モデルの改良と共に，音特徴量の抽出や映像と音の統合的な学習に展開していきたいと考えている．

謝辞

本研究は科学研究費補助金，若手研究 (A)(課題番号 16H05877) の支援を受けた．

参考文献

[1] 伊藤大貴，滝口哲也，有木康雄，「LipNet 構造を用いた唇画像から音声への変換」 日本音響学会研究発表会講演論文集，ROMBUNNO.2-Q-30，2018.

[2] 野田 邦昭，有江 浩明，菅 佑樹，尾形 哲也，「Deep neural network による映像・音響・運動データの統合と共起」 人工知能学会全国大会，3H4-OS-24b-3，2014.

[3] Y. Yamashita and J. Tani, “Emergence of Functional Hierarchy in a Multiple Timescale Recurrent Neural Network Model: a Humanoid Robot Experiment,” PLoS Computational Biology, Vol.4, No.11, e1000220, 2008.