Recurrent Attention Module による動画ベース表情認識の内部構造分析

○三好遼†, 長田典子‡, 橋本学†
†: 中京大学工学研究科機械システム工学専攻
‡: 関西学院大学理工学研究科/感性価値創造研究センター
{miyoshi, mana}@isl.sist.chukyo-u.ac.jp

概要：本研究では，空間的な微細な情報と時間情報を考慮できる Recurrent Attention Module（RAM）および，それを Convolutional RNNs に適用した表情認識手法を提案する．また，Attention map からモデルがどのような特徴を捉えていたかを分析する．提案手法の有効性を CK+データセットを用いて検証した．実験の結果，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 86.85%，90.21%であるのに対して，RAM を加えることによってそれぞれ 6.12%，4.9%向上した．また，ConvLSTM+RAM から得られた Attention map は，発火がばらついているのに対して，Enhanced ConvLSTM+RAM では，表情が表現されている領域が集中して発火していることがわかった．このことから Enhanced ConvLSTM の方が有効な特徴が抽出できていると考えられる．

＜キーワード＞ Facial Expression Recognition, Attention Mechanism, 表情分析

1. はじめに

表情は，感情や意図を伝達するための重要な非言語的情報の 1 つであり，モビリティやエンタテインメントなど幅広い分野で活用されている．

表情には，オンセット，ピーク，オフセットの 3 つの状態がある．オンセットは表情の開始の瞬間，ピークは表情が最も強く表れている瞬間，オフセットは，表情が消える瞬間を表す．そして，表情はオンセットからオフセットまでの変化および変化に伴う顔の皺やテクスチャによって表現される．そのため，このような顔の時空間における動的情報は重要であり，それをどのようにしてモデリングするかが表情認識において重要である．

昨今，画像認識分野においてさまざまな Attention 機構が提案されており，それらは空間方向の Attention 機構[1]とチャンネル方向の Attention 機構[2]に大別される．空間方向における Attention 機構[1]は，画像中の識別に有効である領域をより強調させ，そうでない領域は抑えるような機構である．また，モデルが画像中のどの領域に着目していたかを可視化することができる．チャンネル方向における Attention 機構[2]は，特徴マップの各チャンネルを重み付けることによって注目すべきチャンネルを強調する機構である．

表情認識において，表情変化に伴う顔の皺やテクチャは有効な情報である．しかし，それらは微細な情報であるため，捉えることが難しい．ここで，空間方向における Attention 機構は，空間の微細な特徴を捉えるのに有効である．また，表情は時空間の変化によって表現されるため，時間情報も考慮する必要がある．そこで，本研究では，上記の特徴や時間情報を考慮できる Recurrent Attention Module（RAM）を提案する．また，Convolutional RNNs(Conv-RNNs)にその機構を導入することによって精度向上を図る．また，RAM から得られる Attention Map を分析することによってモデルがどのような特徴を捉えているのか分析する．

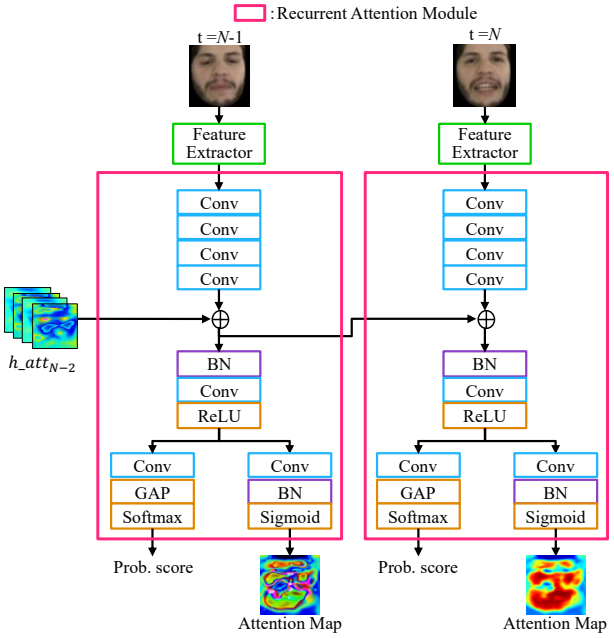


図1 Recurrent Attention Module の概要

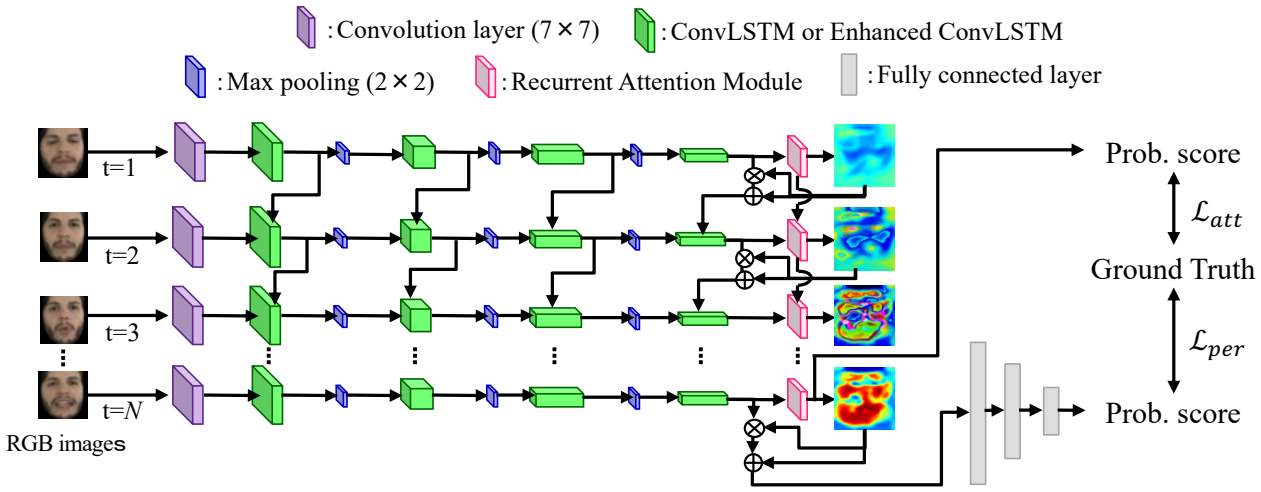


図 2 提案する表情認識手法の概要

2. Recurrent Attention Module (RAM)

図 1 に Recurrent Attention Module (RAM) の概要を示す．提案するモジュールは，Attention Branch Network (ABN) [1]をベースとし，時間情報を考慮できるように改良したモジュールである．

ABN は，Feature extractor，Attention branch，Perception branch の 3 つのモジュールから構成される．まず，画像を Feature extractor に入力し，特徴マップを得る．次に，その特徴マップを Attention branch に入力する．Attention branch では，識別に有効である領域を強調させた Attention map とクラス尤度を出力する．そして， Feature extractor から得られた特徴マップに残差機構によって Attention map を反映させることによって識別に有効な領域をより強調した特徴マップを得る．最後に，Attention を反映させた特徴マップを Perception branch に入力し，クラス尤度を算出する．ABN は，Attention branch，Perception branch のそれぞれから得られるクラス尤度と教師信号との学習誤差が最小になるように学習する．ABN は空間的な特徴を捉えるのに適しているが，feed-forward のモジュールであるため，動画からの表情認識において重要な時間的関係を学習することができない．

そこで，本研究では，Attention branch を再帰的にすることによって，時間的関係を学習可能にする Recurrent Attention Module (RAM) を提案する．RAM では，Attention Map の生成およびクラス尤度を算出する際に利用する中間層の情報を内部状態として保持する．そして，その情報を次のタイムステップの入力とすることによって時間情報を扱えるように改良する．RAM は，Attention branch と同様にクラス尤度と教師信号の学習誤差が最小になるように学習する．

3. 提案する表情認識手法

3.1. 提案手法する表情認識の概要

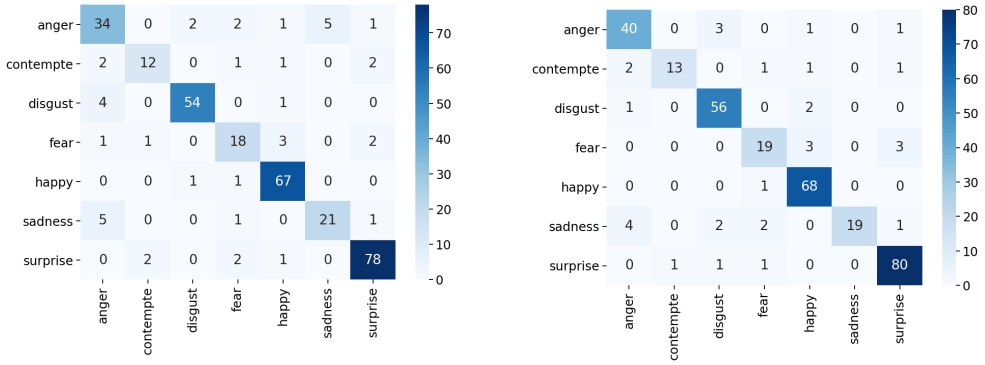
提案する表情認識手法の概要を図 2 に示す．提案手法は，1 層の畳み込み層，3 層の Max pooling 層，4 層の ConvLSTM もしくわ Enhanced ConvLSTM で特徴を抽出し，得られた特徴マップを RAM に入力することによって，クラス尤度と Attention map を得る．また，得られた Attention map を残差機構によって特徴マップに反映させ，全結合層に入力することによって表情を認識する．提案手法は，RAM と全結合層から得られるクラス尤度と教師信号との損失を学習誤差とする．

3.2. Convolutional LSTM (ConvLSTM)

ConvLSTM[3]は，時間的、空間的な関係を同時に表現した時空間的特徴を抽出するために，LSTM の各ゲートでおこなわれる演算を畳み込みに変更した手法である．各ゲートにおいて，タイムステップ t における入力 x_t，タイムステップ $t-1$ における隠れ層の状態 h_{t-1} に畳み込み処理を適用することにより時空間特徴を抽出する。以下に ConvLSTM の重要な式を記載する．

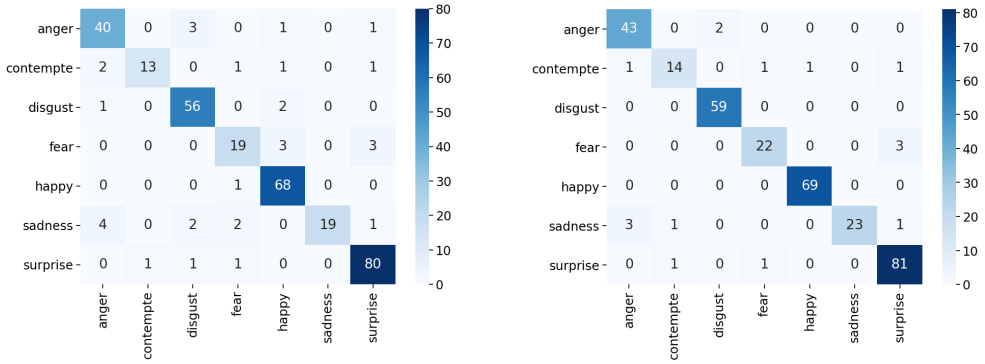
$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{xf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\,(W_{xc} * x_t + W_{xc} * h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{ci} \circ c_t + b_o) \\
h_t &= o_t \circ \tanh\,(c_t)
\end{aligned} \tag{1}
$$

ここで，σ はシグモイド関数，tanh はハイパボリックタンジェント関数，$*$ は畳み込み演算，$\circ$ はアダマール積を表す．また，x_t は入力，h_t は隠



(a) ConvLSTM

(b) Enhanced ConvLSTM



(c) ConvLSTM + RAM

(d) Enhanced ConvLSTM + RAM

図 3 各モデルの Confusion matrix

れ層の出力，c_t は ConvLSTM の内部状態，W は重み行列，b はバイアスである．

ConvLSTM は，Input gate i_t，Forget gate f_t を用いて，c_t を制御する．すなわち，i_t が 1 であるときゲートは開かれ入力が通され，0 であるときゲートが閉ざされ，入力が遮断される．f_t においても同様である．そして，それらの出力に基づいてc_t が更新される．さらに，更新された c_t は，Output gate o_t によって制御される．

3.3. Enhanced ConvLSTM

Enhanced ConvLSTM[4]は，表情の時空間変化と勾配消失を抑制するために ConvLSTM を改良した手法である．Enhanced ConvLSTM は，動画における表情変化が隣接フレーム間で乏しいことに着目し，タイムステップ $t-2$ の有効な情報を利用可能にするための Temporal gates，また，勾配消失を抑制するための Spatial skip connections，Temporal skip connections を ConvLSTM に加えたアルゴリズムである．以下に重要な式を記載する．

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{xf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\left(W_{xc} * x_t + W_{xc} * h_{t-1} + b_c\right) \\
\tilde{s}_t &= tanh(W_{x\tilde{s}} * x_t + W_{h\tilde{s}} * h_{t-2} + b_{\tilde{s}}) \\
s_t &= \sigma(W_{xs} * x_t + W_{hs} * h_{t-2} + b_s) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{ci} \circ c_t + b_o) \\
h_t &= g(o_t \circ \tanh(c_t) + s_t \circ \tilde{s}_t + W_{xr} * x_t) \qquad (2)
\end{aligned}
$$

ここで g は，Group Normalization [5] を表す．$\tilde{s}_t$ は，Temporal modulation gate，s_t は，Temporal skip gate と呼ぶ．このゲートによって，より古い有効な特徴が抽出される．そして，$s_t \circ \tilde{s}_t$ と $W_{xr} * x_t$ を最終出力 h_t に加えることによって，時間および空間方向に勾配が消失しにくいルートを生成する．

4. 実験

4.1. 実験条件

本実験では，CK+[6]を用いて実験をおこなった．CK+は，123 名の被験者から得られた 593 本の画像シーケンスを含んだデータセットである．また，それらの画像シーケンスのうち，118 名から得られた 327 本の画像シーケンスに anger，contempt，disgust，fear，happy，sadness，surprise の 7 種類のラベルのどれかが教師信号として付与されている．本実験では，提案手法および比較手法を 10-fold person-independence cross-validation によって評価した．本実験では，ConvLSTM[3]および Enhanced ConvLSTM[4]とそれらに RAM を加えたモデルを比較する．本実験では，1 つの動画を 8 フレームのビデオクリップに分割して学習をおこなった．ま

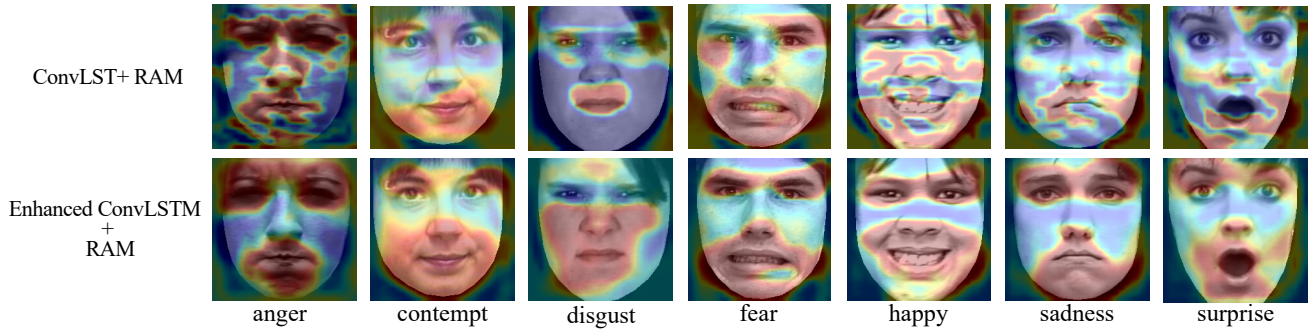


図 4 ConvLSTM+RAM および Enahced ConvLSTM+RAM から得られた Attention map

た，連続するビデオクリップは，前後 6 フレームが重なるようにして分割された．識別も学習時と同様に分割したビデオクリップを用いた．最終的な表情の認識は，分割されたビデオクリップの識別結果の多数決によって決められる．

4.2. 認識率の比較

表 1 に各モデルの認識率，図 3 に Confusion matrix を示す．表 1 より，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 86.85%，90.21%であるのに対して，RAM を加えることによってそれぞれ 6.12%，4.9%向上した．この結果から，RAM が認識精度の向上に寄与していることが確認された．

次に，RAM がどのような表情の認識の精度向上に効果があった分析する．図3の(a) ConvLSTM と(c) ConvLSTM+RAM および(b) Enhanced ConvLSTM と(d) Enhanced ConvLSTM+RAM を比較する．それぞれに RAM を加えることによってほとんどのクラスにおいて精度が向上していることがわかる．精度が大きくが向上しているクラス(anger，disgust，fear)の表情は眉間の皺や微細な表情変化によって表現される．RAM を加えることによってこれらの表情の微細は特徴を捉えることができたと考えられる．

4.3. Attention map の分析

RAM から得られた Attention map から ConvLSTM および Enhanced ConvLSTM がどのような領域を着目していたかを定性的に分析する．図 4 に各表情の動画中の最も表情が喚起されているフレームから得られた Attention map の例を示す．ConvLSTM+RAM から得られた Attention map は，表情認識に有効な領域が発火している場合もあるが，多くの場合で発火がばらついていることがわかる．それに対して，Enhanced ConvLSTM+RAM から得られた Attention map は表情を喚起している領域が集中して発火しており，それ以外の領域はあまり発火していない．表情変化において口元の変化より目元の変化の方が小さいが，Enhanced ConvLSTM+RAM は，目元の領域に着目できている．これらのことから，Enhanced ConvLSTM は ConvLSTM より細かな変化を捉えられており，表情認識に有効な特徴を抽出できていると考えられる．Enhanced ConvLSTM は，表情の時空間変化を捉えるというモチベーションで設計されているが，今回の実験から正しく改良できていると考えられる．

表 1　提案手法と従来手法の認識精度

Method	Accuracy
ConvLSTM[3]	86.85%
ConvLSTM + RAM	92.97%
Enhanced ConvLSTM[4]	90.21%
Enhanced ConvLSTM + RAM	95.11%

5. まとめ

本研究では，空間的な微細な情報と時間情報を考慮できる Recurrent Attention Module (RAM) および，それを Convolutional RNNs に適用した表情認識手法を提案した．また，RAM から得られる Attention map からモデルがどのような特徴を捉えていたかを分析した．提案手法の有効性を CK+データセットを用いて検証した．実験の結果，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 86.85%，90.21%であるのに対して，RAM を加えることによってそれぞれ 6.12%, 4.9%向上した．また，ConvLSTM+RAM から得られた Attention map は，発火がばらついているのに対して，Enhanced ConvLSTM+RAM では，表情を喚起している領域が集中して発火しており，それ以外の領域はあまり発火していなかった．また，表情変化において口元の変化より目元の変化の方が小さいが，Enhanced ConvLSTM+RAM は，目元の領域に着目できていた．これらのことから，Enhanced ConvLSTM は ConvLSTM より細かな変化を捉えられており，表情認識に有効な特徴を抽出できていると考えられる．

謝辞　JST 研究成果展開事業 COI プログラム「感性とデジタル製造を直結し，生活者の創造性を拡張するファブ地球社会創造拠点」の支援によっておこなわれた．

参考文献

[1] Fukui et al., “Attention Branch Network: Learning of Attention Mechanism for Visual Explanation”, CVPR, 2019.
[2] J. Hu et al., “Squeeze-and-Excitation Networks”, CVPR, 2018
[3] X. Shi et al. “Convolutinal LSTM Network: A Machine Learning Approach for Prediction Nowcasting”, NIPS, 2015.
[4] 三好ら，” 動画からの表情認識のための表情の時空間的変化に着目した Enhanced Convolutional LSTM の提案”, MIRU, 2020.
[5] Y. Wu et al., “Group normalization”, ECCV, pp.3-19, 2018
[6] Kanade et al., “Comprehensive database for facial expression analysis”, FG, 2000.

三好遼: 2019 年 4 月中京大学大学院工学研究科機械システム工学専攻に入学．機械学習，ヒューマンセンシングに興味を持つ．

長田典子：1983 年京都大学理学部数学系卒業．同年三菱電機(株)入社．1996 年大阪大学大学院基礎工学研究科博士後期課程修了．2003 年より関西学院大学理工学部情報科 学科助教授、2007 年教授．2009 年米国パデュー大学客員研 究員．2013 年感性価値創造研究センター長．2015 年革新的 イノベーション創出プログラム「感性とデジタル製造を直結し，生活者の創造性を拡張するファブ地球社会創造拠点」サテライトリーダー．専門は感性工学、メディア工学等．

橋本学：1987 年大阪大学大学院修了．同年三菱電機（株）入社．生産技術研究所，先端技術総合研究所に勤務．2008 年より中京大学教授．2017 年より工学部長．3 次元物体認識，ロボットビジョン，ヒューマンセンシングの研究などに従事．2012/2017 年度画像センシングシンポジウム優秀学術賞，2015 年精密工学会小田原賞など受賞．