論文

微小な顔表情変化を捉えるための Recurrent Attention Module の提案 *

三好 遼** 長田 典子 †† 橋本 学 †

Proposal of Recurrent Attention Module for Capturing Subtle Facial Expression Changes

Ryo MIYOSHI, Noriko NAGATA and Manabu HASHIMOTO

Facial expressions are expressed by subtle changes in the shape of the eyes and mouth, wrinkles between the eyebrows, etc. However, it is difficult to capture these changes. In this study, we propose a Recurrent Attention Module (RAM) and a facial expression recognition method using RAM to capture subtle changes in facial expressions. In our experiments, we used CK+ and eNTEFRACE05 databases. In CK+, the recognition accuracies of ConvLSTM and Enhanced ConvLSTM are 93.0% and 95.7%, respectively, while the addition of RAM improves the accuracies by 2.7% and 1.2%, respectively. Furthermore, the recognition accuracies of ConvLSTM and Enhanced ConvLSTM in the eNTERFACE05 database were 39.8% and 49.3%, respectively, while adding RAM improved the accuracies by 4.6% and 0.6%, respectively. In comparison with the conventional method, the proposed method could not outperform the conventional method in CK+. On the other hand, the proposed method improves the accuracy of the eNTERFACE05 database by 0.6% over the conventional method.

Key words: facial expression recognition, deep learning, recurrent neural networks, attention module, video processing

1. 序 論

顔表情は，感情や意図を伝達するための重要な非言語的情報の 1 つである．Ekman らによって定義された 6 つの基本顔表情（anger, disgust, fear, happiness, sadness, surprise）[1] は，個人間で普遍的であり，ヒューマンコンピュータインタラクション（HCI）[2] や医療 [3] といった幅広い分野で活用されている．

表情は，**図 1** 上段に示すように，オンセット，ピーク，オフセットの 3 つのフェーズに分けることができる．オンセットは表情の開始を表し，ピークは表情が最も強く表れている瞬間を表す．また，オフセットは表情が消える瞬間を表す．このように，表情はオンセットからオフセットまでの一連の変化として表現される．心理学分野では，人は単一の静止画像より動画からの方が正確に表情を認識できることが示されており，顔の動的情報を利用することによって表情を認識していることが示唆されている [4]．近年では，上記のような顔の動的変化を捉えることが有効であると考えられ，動画ベースの表情認識が活発に取り組まれるようになっており，大仰な表情の認識について良い成果を示している [5]．しかし，実環境下における人の精神状態は，大仰な表情ではなく微細な表情により表現される [6]．そのため，表情認識によって精神状態や感情を読み取るためには，微細な表情を認識することが重要である．そこで，そのような表情を捉えるための手法が提案されている [7–11]．しかし，微細な表情変化は図 **1** 下段の赤枠で囲われた目元や口元の形状変化，眉間の皺やテクスチャによって表現され，この変化は小

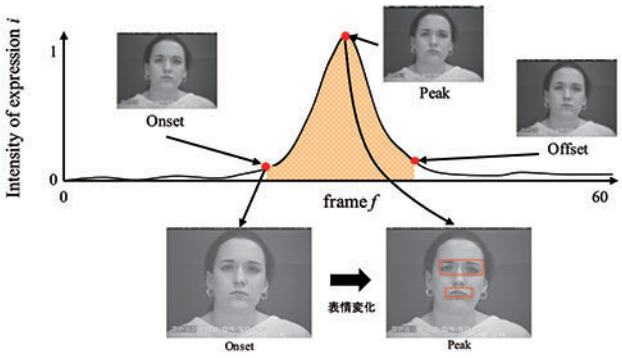

Fig. 1 Features of facial expressions

領域の時空間的変化であることが多いため，画像処理で捉えることが難しい．そこで，本研究では，微細な表情変化を捉えることによる表情認識の精度向上を目的とする．

画像認識において，識別に有効な空間領域により注目することによって，精度向上を示した Attention Branch Networks（ABN）[12] が提案されている．ABN は，Feature extractor，Attention branch，Perception branch の 3 つのモジュールから構成される．Feature extractor では，画像から特徴マップを抽出する．Attention branch では，その特徴マップを用いて Attention map とクラス尤度を算出する．そして，Attention map を特徴マップに反映させることによって，識別に有効な領域を強調させた特徴マップを得る．最後にその特徴マップを用いてクラス識別をおこなう．ABN は，識別に有効な空間領域を強調するため，空間的に微細な特徴を捉えるのに適している．

ABN のような Attention map を活用し，空間的特徴を強調および抑制するアプローチは，空間の微細な特徴を捉えるのに有効であると考えられる．そこで，本研究では，表情認識におけ

* 原稿受付 令和 3 年 5 月 28 日
掲載決定 令和 3 年 10 月 18 日
** 中京大学大学院（愛知県名古屋市昭和区八事本町 101-2）
† 正 会 員 中京大学大学院
†† 関西学院大学（兵庫県三田市学園 2 丁目 1）

る課題である表情変化に伴う顔の皺やテクチャといった空間的に微細な情報を捉えるために ABN のような Attention map を活用するアプローチをとる．ここで，表情は図 **1** 上段に示すようにオンセットからオフセットまでの一連の変化によって表現される．そのため，動画からの表情認識では，空間だけでなく時間情報を考慮する必要がある．しかし，ABN は表情変化を捉えるたに有効な時空間的な特徴を抽出することができない．そこで，本研究では時間情報を考慮し，視覚的説明モデルから得られる Attention map を活用することによって，微細な表情変化を捉える Recurrent Attention Module（RAM）を提案する．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．

2. 提案する Recurrent Attention Module（RAM）および RAM を用いた表情認識手法

2.1 Recurrent Attention Module（RAM）

表情はオンセットからオフセットまでの変化および目元や口元の形状変化，眉間の皺やテクスチャによって表現されるが，この変化は微細であることが多いため，画像処理で捉えることが難しい．このような問題に対して，微細な変化を示す部分（目元や口元，眉間の皺）により注目する処理を加えることによって変化を捉えることができると考えられる．ここで，画像認識において，識別に有効な空間的な領域により注目する手法として Attention Branch Network（ABN）[12] が提案されている．ABN では，Attention branch から得られる Attention map を特徴マップに反映させることによって識別に有効な空間領域に注目している．そこで，表情変化を示す領域に注目するためには，ABN のようなアプローチが有効であると考えられる．しかし，ABN では動画からの表情認識において重要な時間情報を扱うことができない．本研究では，表情変化を示す領域に注目するために，時空間情報を扱うことのできる Recurrent Attention Module を提案する．RAM は，ABN の Attention branch において時間情報を扱えるように改良したモジュールである．

RAM の概要を図 **2** に示す．まず，各タイムステップにおいて得られる特徴マップに対して，畳み込み（Conv）によって特徴を抽出する．つぎに，その特徴マップと一つ前のタイムステップの特徴マップを要素和によって統合する．そして，統合した特徴マップに対して正規化（BN），畳み込み，非線形変換 (ReLU) を適用する．最後に，非線形変換した特徴を用いて，クラス尤度と Attention map を算出する．クラス尤度は，畳み込み，Global Average Pooling，Softmax を適用することによって算出される．Attention map は，畳み込み，正規化，Sigmoid 関数を適用することによって算出される．

RAM は，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．そうすることによって，表情を捉えるために有効な Attention map を得ることができる．また，一つ前のタイムステップにおける，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から得られる特徴マップを内部状態として保持し，現在のタイムステップの特徴マップに反映させることによって時間情報を捉える．

ここで，RAM は畳み込みを伴う時系列ネットワークである畳み込み RNN（ConvLSTM[13]，Enhanced ConvLSTM[14]）のみしか適用できない．畳み込みを伴い，時間情報を扱える手法として，3D CNN 系のモデルが挙げられるが，それらは各タイムステップを逐次的に扱わないため，RAM を適用することはできない．

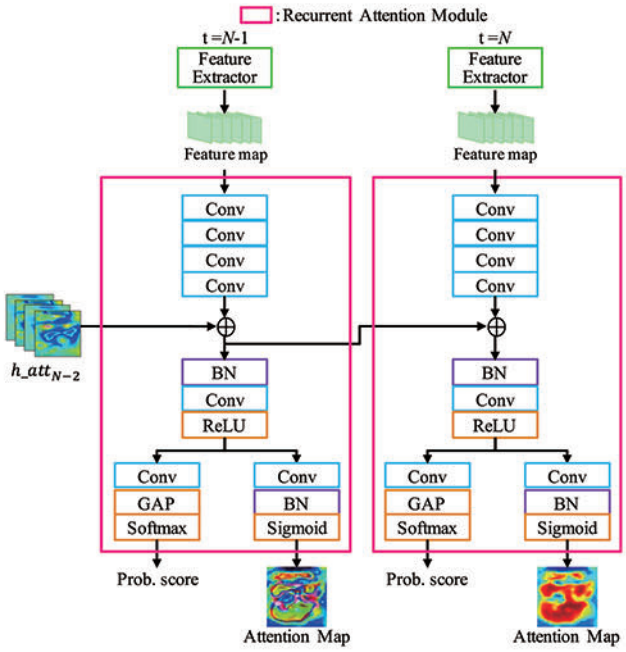

Fig. 2 Overview of recurrent attention module

2.2 RAM を用いた表情認識手法の概要

RAM を用いた表情認識手法の概要を図 **3** に示す．この手法は，Feature extractor，Recurrent attention branch，Perception branch から構成される．Feature extractor は，1 層の畳み込み層，3 層の Max pooling 層，4 層の Convolutional RNNs（Conv-RNN）によって構成される．Recurrent attention branch は，RAM と RAM から得られた特徴マップを反映させる機構で構成される．Perception branch は，3 層の全結合層によって構成される．Feature extractor には，RGB，Optical flow 画像シーケンスが入力される．

提案手法の流れを以下に示す．

1. 各タイムステップごとに画像を Feature extractor に入力することによって，特徴マップを取得する．
2. 得られた特徴マップを Recurrent attention branch に入力し，クラス尤度および Attention map を算出する．
3. 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る．
4. Attention map が反映された特徴マップを Perception branch に入力することによって，クラス尤度を算出する．

この手法は，Recurrent attention branch から得られるクラス尤度と教師信号との誤差 $\mathcal{L}_{att}$，Perception branch から得られるクラス尤度と教師信号との誤差 $\mathcal{L}_{per}$ を基に学習される．以下に，提案手法の損失関数を記載する．

$$\mathcal{L} = \mathcal{L}_{att} + \mathcal{L}_{per} \tag{1}$$

また，識別の際は，Perception branch のクラス尤度のみで識別をおこなう．

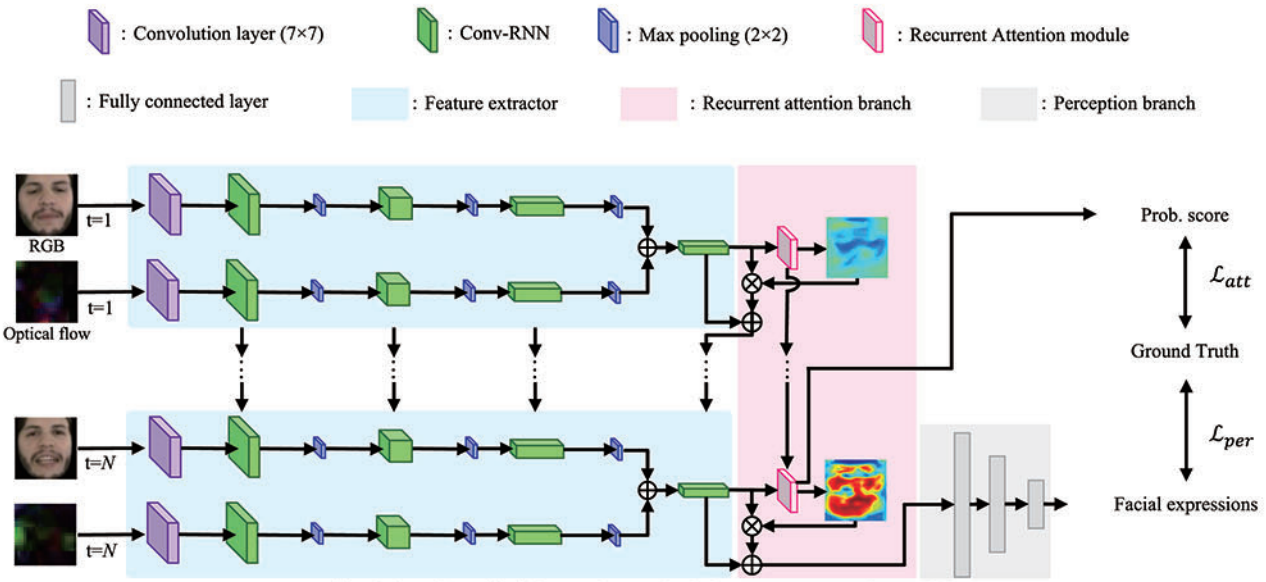

Fig. 3 Overview of facial expression method using recurrent attention module

2.3 本手法で用いる Conv-RNN

2.3.1 Convolutional LSTM（ConvLSTM）

ConvLSTM[13] は，LSTM[15] の内部の行列演算を畳み込み演算に変更することによって，時空間を同時に表現した特徴を抽出できるようにした手法である．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式（2）に ConvLSTM の重要な式を記載する．

$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
\tilde{c}_t &= \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o) \\
h_t &= o_t \circ \tanh(c_t),
\end{aligned}
\tag{2}
$$

ここで，σ と tanh はそれぞれ sigmoid 関数とハイパボリックタンジェント関数，$*$ と $\circ$ はそれぞれ畳み込み演算とアダマール積を表す．さらに，t は現在のタイムステップ，$t-1$ は現在より一つ前のタイムステップ，x_t はタイムステップ t における入力データ，h_t および h_{t-1} は各タイムステップにおける隠れ状態，c_{t-1} はタイムステップ $t-1$ における memory cell を表す．また，W は weight matrix，b は offset matrix を示す．

2.3.2 Enhanced ConvLSTM

Enhanced ConvLSTM[14] 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．式（3）に Enhanced ConvLSTM において追加された新たなゲートや変更された式について記載する．ConvLSTM と Enhanced ConvLSTM の間で，i_t，f_t，o_t，$\tilde{c}_t$，c_t は同一である．

$$
\begin{aligned}
\tilde{s}_t &= \tanh(W_{xp} * x_t + W_{hp} * h_{t-2} + b_p) \\
s_t &= \sigma(W_{xs} * x_t + W_{hs} * h_{t-2} + b_s) \\
h_t &= g(o_t \circ \tanh(c_t) + s_t \circ \tilde{s}_t + W_{xr} * x_t),
\end{aligned}
\tag{3}
$$

ここで，g は Group Normalization[16] を表す．また，$t-2$ は現在より二つ前のタイムステップ，h_{t-2} は 2 つ前のタイムステップにおける隠れ状態を表す．

3. 表情認識実験と考察

3.1 実験条件

本実験では，CK+[17] および eNTEFRACE05 database[18] という 2 つの公開データセットを用いた．CK+ は実験室環境下における表情喚起を収めたデータセットであり，一般的な表情認識のベンチマークテストに用いられているため，本実験においても使用する．また，提案手法の最終適用先は店舗における顧客と店員のインタラクションの際の基本 6 感情の認識を想定している．eNTERFACE05 database は，発話を含んだマルチモーダルな感情喚起を収めたデータセットであり，本研究の適用先の想定シーンに近いデータセットであるため，本手法の評価に用いる．ここで，微細な表情認識を評価するデータセットとして，LSEMSW[19] や CASME[20] といったデータセットがある．LSEMSW は subtle facial expression に焦点を当てており，本研究と目的は近いが，静止画像であるため提案手法の評価に用いることができない．また，CASME は micro-expression に焦点を当てており，本研究の目的とは異なるため，提案手法の評価に用いない．

CK+[17] は，123 名の被験者から 593 本の画像シーケンスを取得したデータセットである．さらに，それらの画像シーケンスの内，118 名から得られた 327 本の画像シーケンスにおいて，7 つの表情のクラス（anger，disgust，fear，happiness，sadness，surprise，contempt）のいずれかが教師号として付与されている．図 **4**-(a) にデータベース内の画像の例を示す．各動画の長さは，1-2 秒程度である．各動画は，RGB 画像もしくは，グレスケール画像から構成され，解像度は，縦 490pix，横 640pix で

(a) CK+

(b) eNTRERFACE05 database

Fig. 4 Examples of face images obtained from each dataset

Table 1 Comparison of accuracy with and without RAM on CK+

Method	Accuracy
ConvLSTM	93.0%
Enhanced ConvLSTM[14]	95.7%
ConvLSTM + RAM	95.7%
Enhanced ConvLSTM + RAM	**96.9%**

Table 2 Comparison of accuracy with and without RAM on eNTERFACE05

Method	Accuracy
ConvLSTM	39.8%
Enhanced ConvLSTM[14]	49.3%
ConvLSTM + RAM	44.4%
Enhanced ConvLSTM + RAM	**49.9%**

ある．このデータベースは，表情分析に焦点が当てられて作成されている．各被験者に対して，指定された表情を表現するようにしてデータセットは作成された．各画像シーケンスは，無表情から始まり，最終フレームに最も表情が喚起されたシーンが含まれている．本研究では，RGB 画像，グレイスケール画像どちらともを利用した．また，グレイスケール画像を RGB 画像と同じく 3 チャンネル画像として扱った．具体的には，グレイスケール画像のチャンネル情報をコピーしたチャンネルを 2 つ追加することによって 3 チャンネル画像として扱った．

eNTERFACE05 database[18] は，43 名の被験者から得られた 1290 本の動画が格納されたデータベースである．このデータベースは，各動画に基本 6 表情（anger, disgust, fear, joy, sadness, surprise）のうちのどれかが付与されている．データベース内の画像の例を図 **4**-(b) に示す．各動画の長さは，1-5 秒程度である．各動画は，RGB の 3 チャンネルから構成され，解像度は，縦 570pix，横 720pix である．このデータベースは，被験者にある文章を指定された感情で表現するようにして作成されている．そのため，表情のみではなく，マルチモーダルな情報によって感情が喚起されている．

各データセットにおいて手法を評価する方法として，leave-one-subject-group-out（LOSGO）を採用した．この方法は，被験者を k 個のグループに分割して，交差検証する評価方法である．そのため，学習用のデータセットに含まれている被験者のデータは，検証用のデータセットに含まれない．各データセットにおける評価の際の k は，従来研究に従って，CK+ は $k = 10$，eNTERFACE05 は $k = 5$ として評価実験を実施した．

Table 3 Comparison with state-of-the-art on CK+

Method	Accuracy
ST network[23]	98.5%
DTAGN[24]	97.3%
CNN+Island loss[25]	94.4%
LOMo[26]	92.0%
FAN[27]	**99.7%**
Enhanced ConvLSTM[14]	95.7%
Enhanced ConvLSTM + RAM	96.9%

Table 4 Comparison with state-of-the-art on eNTERFACE05

Method	Accuracy
Mansoorizadeh et al.[28]	38.0%
Fejani et al.[29]	39.28%
Zhalahpour et al.[30]	42.2%
Pan et al.[31]	43.0%
Meng et al.[27]	48.0%
Miyoshi et al.[14]	49.3%
Enhanced ConvLSTM + RAM	**49.9%**

ネットワークに入力する顔画像は，OpenFace[21] によってアライメントされた顔画像を用いた．また，Optical flow は，Dense な Optical flow を抽出できる手法 [22] を用いた．

3.2 実験結果

3.2.1 RAM の有無による認識率の比較

Recurrent Attention Module の有無による認識率を比較する．**表 1** に CK+，表 **2** に eNTERFACE05 database における各手法の認識率を示す．表 **1** より CK+ において，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 93.0%，95.7% であるのに対して，RAM を加えることによってそれぞれ 2.7%，1.2% 向上した．また，表 **2** より eNTERFACE05 database において，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 39.8%，49.3% であるのに対して，RAM を加えることによってそれぞれ 4.6%，0.6% 向上した．

この結果から，ConvLSTM，Enhanced ConvLSTM の両 Conv-RNN および表情認識のベンチマークテストに一般的に用いられている CK+，マルチモーダルな感情喚起のシーンを収めたデータセットである eNTEFRACE05 database の両データセットにおいて，RAM が認識精度の向上に寄与していることが確認された．また，ConvLSTM では eNTERFACE05 database，Enhanced ConvLSTM では CK+ の精度がより改善された．ConvLSTM の eNTERFACE05 database における精度がより改善された理由として，より表情変化を捉えられたためだと考えられる．eNTERFACE05 database は感情が 1 秒〜5 秒程度で表現され，特定にキーフレームというよりは，時間区間によって感情が表現される．一方，CK+ は，最終フレームに最も表情が強く喚起されており，そのキーフレームのみでも表情が識別できる．そのため，より変化を捉える必要のある eNTERFACE05 database における精度が改善されたと考えられる．一方，Enhanced ConvLSTM が CK+ における認識率がより改善している理由として，空間的な変化を捉えられたためだと考えられる．Enhanced ConvLSTM は，時間的な変化をより捉えるため

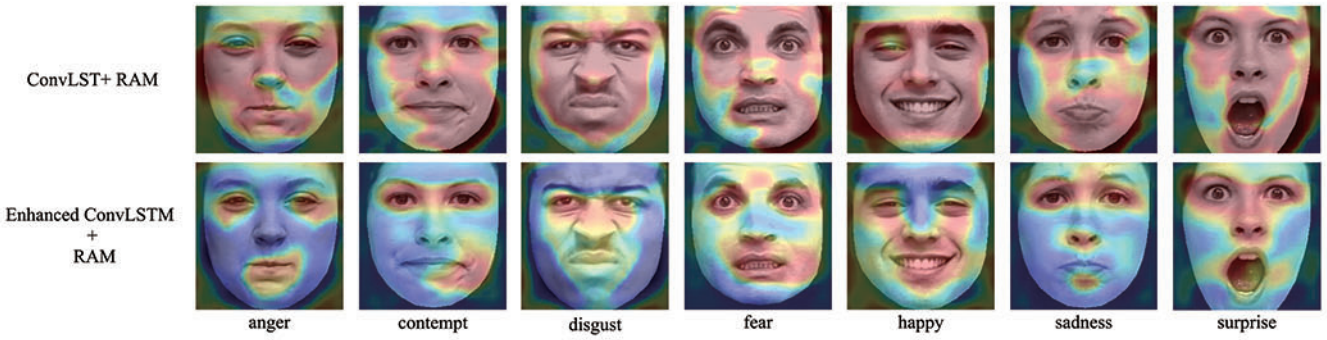

Fig. 5 Example of an attention map obtained from each method

に ConvLSTM を改良した手法である．ここで，RAM を加えることによって，空間的な特徴も捉えられるようになり，空間的な情報が支配的である CK+ における精度がより改善されたと考えられる．

次に各手法の RAM から得られた Attention map を可視化することによって，ネットワークが表情を認識するために顔のどのような領域を注視していたかを定性的に分析する．本実験では解釈性の観点から，表情のみによって感情が喚起されている CK+ における実験から得られる Attention map を分析する．図 **5** に ConvLSTM+RAM，Enhanced ConvLSTM+RAM において，最も表情が喚起されたフレームから得られた Attention map の例を示す．得られた Attention map から，ConvLSTM は顔全体を注視して表情を認識していることがわかった．一方，Enhanced ConvLSTM は目元や口元，眉間といった表情が喚起されている領域を注視していることがわかった．これらの注視領域の違いから，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．

3.2.2 従来手法との認識率の比較

表 **3** に CK+ における state-of-the-art と提案手法との比較を示す．提案手法の認識率は，96.9% であり，従来手法より高い精度を示すことができなかった．ここで，ST network[23]，DTAGN[24] は，静止顔画像と顔キーポイント情報，CNN+Island loss[25]，FAN[27] は静止顔画像，LOMo[26] は画像シーケンスを入力とする手法である．ここで，静止画像を入力とする手法である FAN の認識率が最も高い．CK+ は表情のみで感情が喚起され，最も強く表情が喚起されたフレームが最終フレームにくる．そのため，最終フレームのみでも表情認識が可能であり，時間的情報より空間的情報の方が支配的である．そのため，時間的な変化を捉える提案手法より静止顔画像を入力とする手法の認識率が高かったと考えられる．表 **4** に，eNTERFACE05 database における state-of-the-art と提案手法との比較を示す．実験の結果，提案手法の認識率は，49.9% であり，従来の最も精度が高い手法より 0.6% 高いことを確認した．ここで，eNTERFACE05 database は，発話を伴うマルチモーダルな感情喚起のデータセットであり，2 秒〜5 秒で感情が表現される．一方，CK+ は顔表情のみによって感情が喚起され，喚起される時間は 1 秒〜2 秒で表現される．CK+ より eNTERFACE05 database の方が，長期的に感情が喚起されているため，一定の時間区間を見ることが重要である．ここで，CK+ において最も精度が高い手法である FAN は静止顔画像を入力としており，時間的な変化を利用していない．一方，提案手法は時間的変化を利用しているため，時間区間によって感情が喚起されているデータセットにおいて高精度であったと考えられる．

4. 結　論

本研究では，微細な表情変化を捉える Recurrent Attention Module（RAM）および RAM を用いた表情認識手法を提案した．実験では，表情認識のベンチマークテストに一般的に用いられている CK+ および，マルチモーダルな感情喚起のシーンを収めた eNTEFRACE05 database という 2 つの公開データセットを用いた．RAM の有無による認識率の比較では，CK+ において，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 93.0%，95.7% であるのに対して，RAM を加えることによってそれぞれ 2.7%，1.2% 向上した．さらに eNTERFACE05 database において，ConvLSTM および Enahnced ConvLSTM の認識精度はそれぞれ 39.8%，49.3% であるのに対して，RAM を加えることによってそれぞれ 4.6%，0.6% 向上した．従来手法との比較では，CK+ においては，従来手法を上回ることができなかったが，eNTERFACE05 database においては，従来手法より 0.6% 精度向上したことを確認した．

謝　辞

本研究は，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) P. Ekman and W. Friesen: Constants across cultures inthe face and emotion, Journal of personality and social psychology, **17**, 2 (1971) 124.
2) M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan: Real time face detection and facial expression recognition: Development and applications to human computer interaction, In 2003 Conference on computer vision and pattern recognition work-shop, **5**, (2003) 53.
3) P. Ekman and W. Friesen: A new pan-cultural facial expression of emotion, Motivation and emotion, **10**, 2 (1986) 159.
4) Z. Ambadar, J .Schooler and J. Cohn: Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions, Psychological science, **16**, 5 (2005) 403.
5) S. Li and W. Deng: Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing, (2020) 1.
6) G. Warren, E. Schertler and P. Bull: Detecting deception from emotional and unemotional cues, Journal of Nonverbal Behavior, **33**, 1 (2009) 59.
7) G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales and N. Robertson: Deep multi-task learning to recognise subtle facial expressions of mental states, In Proceedings of the European Conference on Computer Vision (ECCV), (2018) 103.
8) Y. Zhao and J. Xu: A convolutional neural network for compound micro-expression recognition, Sensors, **19**, 24 (2019) 5553.

9) D. Choi and B. Song: Facial micro-expression recognition using two-dimensional landmark feature maps, IEEE Access, **8**, (2020) 121549.
10) H. Shahar and H. Hel-Or: Micro expression classification using facial color and deep learning methods, In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, (2019).
11) B. Allaert, I. Bilasco and C. Djeraba: Micro and macro facial expression recognition using advanced local motion patterns, IEEE Transactions on Affective Computing, (2019).
12) H. Fukui, T. Hirakawa, T. Yamashita and H. Fujiyoshi: Attention branch network: Learning of attention mechanism for visual explanation, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2019) 10705.
13) X. SHI, Z. Chen, H. Wang, D. Yeung, W. Wong and W. WOO: Convolutional lstm network: A machine learning approach for precipitation nowcasting, Advances in Neural Information Processing Systems 28, (2015) 802.
14) R. Miyoshi,N. Nagata and M. Hashimoto: Enhancedconvolutional lstm with spatial and temporal skip connections and temporal gates for facial expression recognition from video, Neural Computing and Applications, (2021) 1.
15) F. Gers and E. Schmidhuber: LSTM recurrent networks learn simple context-free and context-sensitive languages, IEEE Transactions on Neural Networks, **12**, 6 (2001) 1333.
16) Y. Wu and K. He: Group normalization, In Proceedings of the European Conference on Computer Vision (ECCV), (2018) 3.
17) P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, In 2010 ieee computer society conference on computer vision and pattern recognition-workshops, (2010) 94.
18) O. Martin, I. Kotsia, B. Macq and I. Pitas: The enterface' 05 audio-visual emotion database, In 22nd International Conference on Data Engineering Workshops (ICDEW' 06), (2006) 8.
19) G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson and Y. Yang: Deep multi-task learning to recognise subtle facial expressions of mental states, In Proceedings of the European Conference on Computer Vision (ECCV), (2018).
20) W. Yan, Q. Wu, Y. Liu, S. Wang and X. Fu: Casme database: A dataset of spontaneous micro-expressions collected from neutralized faces, In 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), (2013) 1.
21) T. Baltrusaitis, A. Zadeh, Y. Lim and L. Morency: Openface 2.0: Facial behavior analysis toolkit, In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), (2018) 59.
22) G. Farnebäck: Two-frame motion estimation based on polynomial expansion, In Scandinavian conference on Image analysis, (2003) 363.
23) K. Zhang, Y. Huang, Y. Du and L. Wang: Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Transactions on Image Processing, **26**, 9 (2017) 4193.
24) H. Jung, S. Lee, J. Yim, S. Park and J. Kim: Joint fine-tuning in deep neural networks for facial expression recognition, In Proceedings of the IEEE international conference on computer vision, (2015) 2983.
25) J. Cai, Z. Meng, A. Khan, Z. Li, J. Reilly and Y. Tong: Island loss for learning discriminative features in facial expression recognition, In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), (2018) 302.
26) K. Sikka, G. Sharma and M. Bartlett: Lomo: Latent ordinal model for facial analysis in videos, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2016) 5580.
27) D. Meng, X. Peng, K. Wang and Y. Qiao: frame attention networks for facial expression recognition in videos, In 2019 IEEE International Conference on Image Processing (ICIP), (2019) 3866.
28) M. Mansoorizadeh and N. Charkari: Multimodal information fusion application to human emotion recognition from face and speech, Multimedia Tools and Applications, **49**, 2 (2010) 277.
29) M. Bejani, D. Gharavian and N. Charkari: Audiovisual emotion recognition using anova feature selection method and multi-classifier neural networks, Neural Computing and Applications, **24**, 2 (2014) 399.
30) S. Zhalehpour, O. Onder, Z. Akhtar and C. Erdem: Baum-1: A spontaneous audio-visual face database of affective and mental states, IEEE Transactions on Affective Computing, **8**, 3 (2017) 300.
31) X. Pan, G. Ying, G. Chen, H. Li and W. Li: A deep spatial and temporal aggregation framework for video-based facial expression recognition, IEEE Access, **7**, (2019) 48807.