Enhanced ConvLSTM を用いた動画からの表情認識

○三好 遼†, 長田 典子‡, 橋本 学†
†: 中京大学大学院工学研究科機械システム工学専攻
‡: 関西学院大学大学院理工学研究科/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，従来の Convolutinal LSTM（ConvLSTM）の時空間方向それぞれに skip connection を導入した Enhanced Convolutional LSTM（Enhanced ConvLSTM），およびこれを用いた動画からの表情認識手法を提案する．提案手法は，2 つの Enhanced ConvLSTM ストリームと 2 つの ResNet ストリームから構成される．実験では，skip connection の有無による認識率の比較，および提案手法と従来手法の認識率を比較した．ConvLSTM に skip connection を導入することにより，認識率が 4.44%向上した．また，提案手法の認識率は 45.29%であり，従来手法より 2.31%認識率が高いことが示された．
＜キーワード＞ 表情認識，Convolutional LSTM，Skip Connection

1. はじめに

表情は，コミュニケーションにおいて感情や意図を伝達するための重要な非言語的行動の 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．また，表情認識のためのデータベース[4-6]が公開されるなど，動画からの表情認識のニーズはより一層高まっている．

表情認識において，時間情報は重要な情報である．表情は，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．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によって表現される．心理学分野において，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唆されており，人は顔の動的情報を利用し表情を認識していることが明らかになっている[7]．これらのことから，時間的情報は，表情認識するうえで重要な情報である．そのため，どのようにして表情認識に有効な時空間的特徴を抽出するかが重要となる．

これまで表情認識に関する研究において，時間情報を用いた手法と時間情報を用いない手法が提案されてきた[8-19]．時間情報を用いない手法では，Hand-craft 特徴を用いた手法[8-11]や，Convolutional Neural Network (CNNs)を用いた手法[12-14]などが提案されてきた．しかし，これらの手法は静止画像から表情を認識しており，表情間の遷移における時間的関係を考慮していない．また，動画から表情認識する際，動画内には表情とは関係のないフレームが多く含まれており，静止画像からの表情認識手法ではこれらの影響を強く受けてしまう．そのため，これらの手法をそのまま動画からの表情認識に利用することは困難である．時間情報を用いた手法では，工学的な特徴を用いた手法[17-19]や，Deep Neural Networks (DNNs)を用いた手法[15,16]などが提案されてきた．Pan ら[16]は，CNNs を用いて静止画像から空間的特徴を抽出し，得られた特徴の時間的特徴を Long Short-Term Memory (LSTM) によって学習することにより，表情を認識している．この手法では，空間的特徴と時間的特徴が別のモジュールによって抽出されている．そのため，時空間を考慮した特徴を得ることがで

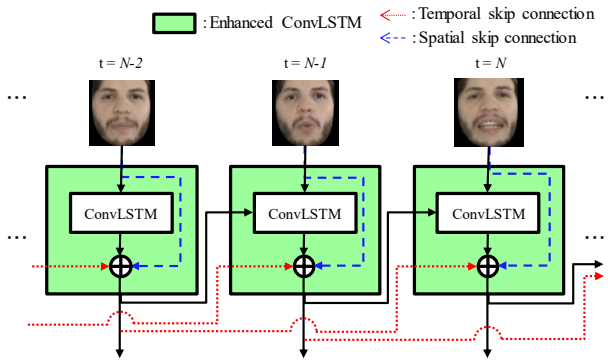


図 1 Enhanced ConvLSTM の概要

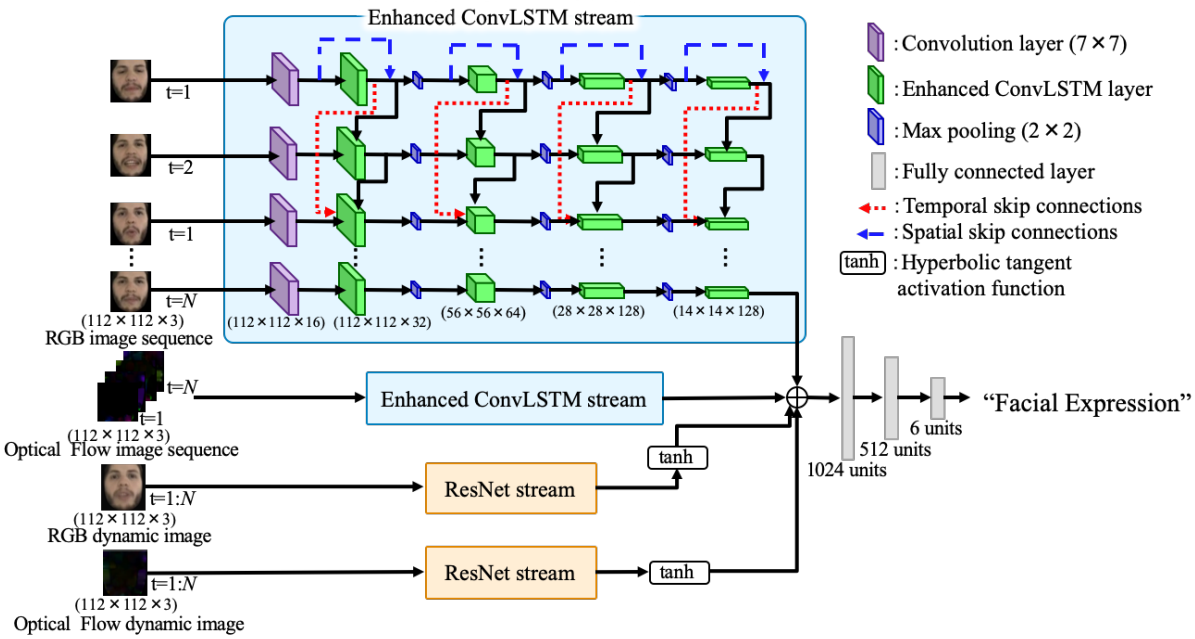


図 2 提案する表情認識手法の概要

きない．また，LSTM は，時間方向にネットワークが展開されるため，層が時間方向に深くなる．そのため，勾配消失が起きやすい．

時間情報を考慮した他の手法として，行動認識で用いられている 3D CNNs を用いた手法[20-22]が考えられる．しかし，表情認識のためのデータベースは，行動認識のためのデータベース[23-25]に比べ，規模が著しく小さい．そのため，表情認識のためのデータベースでは，行動認識で有効性が示されている 3D CNNs 手法のパラメータを学習することが困難であり，それらの手法を適用することはできない．

本研究では，Enhanced convolutional LSTM (Enhanced ConvLSTM) を用いた 2D ベースの表情認識手法を提案する．提案手法は，2 つの Enhanced ConvLSTM ストリームと 2 つの ResNet ストリームの計 4 つのストリームから構成される．Enhanced ConvLSTM ストリームでは，多層の Enhanced ConvLSTM layer によって時空間的特徴を抽出する．Enhanced ConvLSTM は，勾配消失の抑制および，より古い情報を利用可能にするために，従来の Convolutional LSTM（ConvLSTM）の時空間方向それぞれに skip connection を導入した．実験では，Enhanced ConvLSTM の有効性を調査するために skip connection の有無による認識率の比較をおこなった．また提案する表情認識手法と従来手法との認識率の比較をおこなった．実験の結果，ConvLSTM に skip connection を導入することによって，認識率が 4.44%向上した．また，提案手法の認識率は 45.29%であり，従来手法より 2.31%認識率が高いことを確認した．

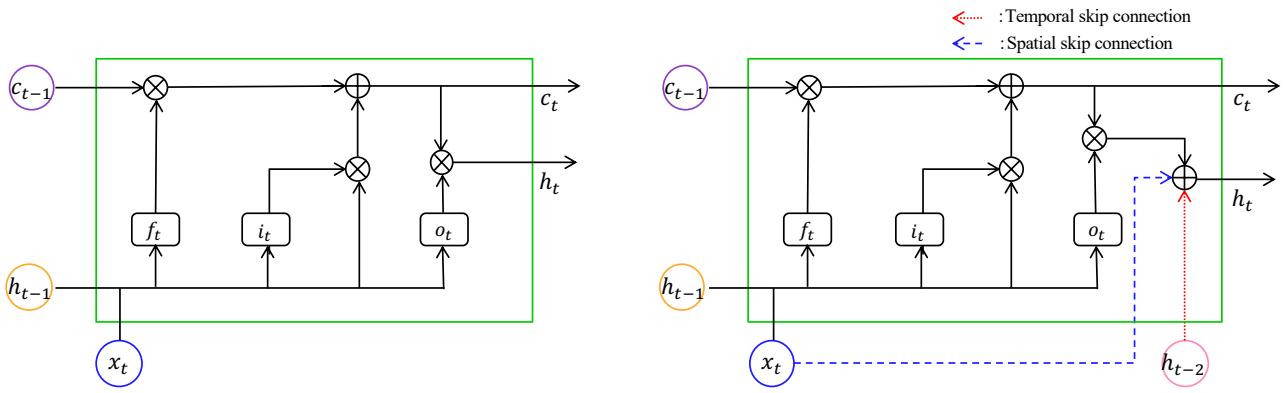
2. 時間情報を用いた表情認識手法

2.1. Hand-craft 特徴を用いた手法

Hand-craft 特徴量を用いた手法として[17-19]がある．S. Zhalehpour ら[19] は，動画の中から最も表情が表出されたフレームを選択し，選択された顔画像を用いて表情認識をおこなう．また，Local Phase Quantization (LPQ) 特徴量を用いて顔のテクスチャを特徴量化している．この手法は，eNTERFACE05 database において Hand-craft 特徴量を用いた手法の中で最も高い性能を示した．しかし，提案されている特徴量は，DNNs を用いた手法に比べ，low-level な特徴量である．

2.2. DNNs を用いた手法

DNNs を用いた手法として[15,16]がある．Pan ら[16]は，VGG-19[26]と LSTM[27]を組み合わせた手法を提案した．VGG-19 は，画像認識に用いられる CNNs の有効な手法の 1 つである．LSTM は，長期的な時間的相関を学習することができる Recurrent Neural Networks (RNNs) の 1 つの手法である．Pan らの手法では，空間的な特徴を抽出するために VGG-19, 時間的特徴を抽出するために LSTM を用いている．そのため，空間的特徴と時間的特徴が別のモジュールによって抽出されており，時空間的特徴を抽出することができない．



(a) ConvLSTM (Conventional)　　(b) Enhanced ConvLSTM (Proposed)

図 3 従来の ConvLSTM と Enhanced ConvLSTM の概要

3. 提案手法

本研究では，従来の ConvLSTM を勾配消失の抑制および，より古い情報を利用できるように改良する．そして，Enhanced ConvLSTM を用いた動画からの表情認識手法を提案する．

提案する表情認識手法の概要を図 2 に示す．提案手法は，2 つの Enhanced ConvLSTM ストリームと 2 つの ResNet ストリームの計 4 つのストリームと 3 層の全結合層から構成される．Enhanced ConvLSTM ストリームには，RGB，optical flow の画像シーケンスをそれぞれ入力し，ResNet ストリームでは，RGB，optical flow の dynamic image をそれぞれ入力する．そして，4 つのストリームから得られた特徴マップを足し合わせ，全結合層に入力することによって表情を認識する．各ストリームに入力する顔画像は，OpenFace[28]によって取得する．また，optical flow は，Gunnar ら [29]の方法によって算出される．Enhanced ConvLSTM ストリームによって，細かな動きの特徴を抽出し，ResNet ストリームでは，大きな動きの特徴を抽出する．次に，Enhanced ConvLSTM ストリーム，ResNet ストリームのそれぞれの詳細について説明する．

3.1. Enhanced ConvLSTM ストリーム

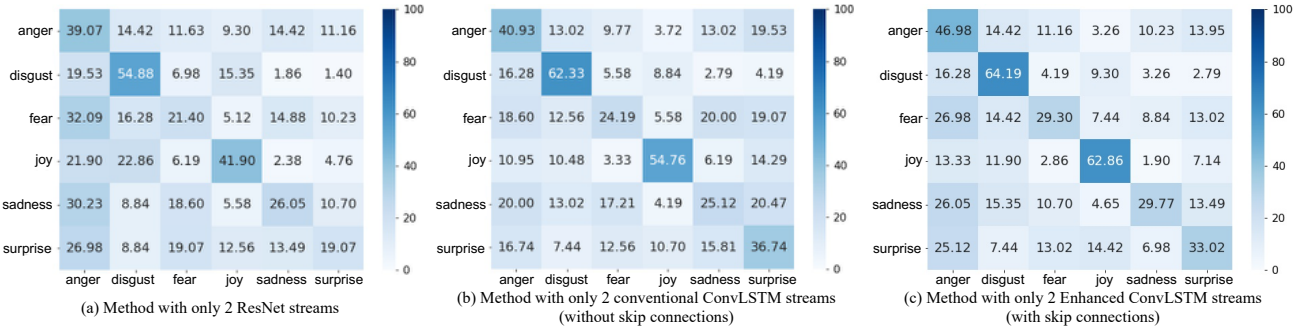
本章では，Enhanced ConvLSTM ストリームおよび Enhanced ConvLSTM について説明する．CNNs と LSTM を組み合わせた手法では，CNNs による空間的特徴の抽出と LSTM による時間的特徴の抽出が異なるモジュールであるため，時空間的特徴を抽出することができない．そこで，Enhanced ConvLSTM ストリームでは，Enhanced ConvLSTM layer を積み重ねることによって，時空間的特徴を抽出する．多層の Enhanced ConvLSTM layer によって高次の特徴を抽出するために，多くの CNNs と同様に Enhanced ConvLSTM layer の後に，max pooling を適用する．また，各 Enhanced ConvLSTM layer において用いたカーネルのサイズは，1，2 層目：5×5，3，4 層目：3×3 とした．浅い層と深い層でカーネルサイズを変えた理由は，浅い層ではより空間的に大域的な特徴，深い層ではより細かい特徴を抽出するためである．

ConvLSTM[30]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。以下に従来の 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



(a) Method with only 2 ResNet streams

(b) Method with only 2 conventional ConvLSTM streams (without skip connections)

(c) Method with only 2 Enhanced ConvLSTM streams (with skip connections)

(d) Method with 2 ResNet streams and 2 conventional ConvLSTM streams ((a) and (b))

(e) Proposed method with 2 ResNet streams and 2 Enhanced ConvLSTM streams ((a) and (c))

図 4 提案する表情認識手法の構成要素の各組合せの認識率

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．ここで ConvLSTM は，タイムステップt と $t-1$ に基づいて制御される．

Y. Wang ら[31]の研究でも述べられているが，従来の LSTM は過去の情報を保持することはできるが，長期的な過去の情報を保持することができない．その理由は，タイムステップ $t-1$ に強く反応するためである．そのため，系列が長くなるほどより古い情報は失われ，過去の情報を参照することができない．また，LSTM は，時間方向に展開されるため，層が時間方向に深くなる．そのため，勾配消失が起きやすい．また，LSTM では，活性化関数にシグモイド関数やハイパボリックタンジェント関数が用いられているため，多層の LSTM では，空間方向での勾配消失が発生する．そこで本研究では，タイムステップ $t-2$ の情報をダイレクトに使用することのできるパスと ResNet[33]のような空間方向にスキップするパスを追加することにより，従来の ConvLSTM より勾配消失の抑制および，より古い情報を利用することができるように改良する．具体的には，ConvLSTM の時空間方向に skip connection を導入する．以下に提案する ConvLSTM の重要な式を記載する．

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * x_t + W_{xf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_{xc} * x_t + W_{xc} * h_{t-1} + b_c)$$
$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{ci} \circ c_t + b_o)$$
$$h_t = g(o_t \circ \tanh(c_t) + h_{t-2} + W_{xs} * x_t) \quad (2)$$

ここで，g は，Group Normalization[32]を示す．図 3 に従来の ConvLSTM と Enhanced ConvLSTM の概要を示す．図 3，式 1，2 からわかるように，従来の ConvLSTM と Enhanced ConvLSTM の違いは，隠れ層の出力のみである．$W_{xs} * x_t$ によって空間的 skip connection，h_{t-2} によって時間的 skip connection が実現される．また，Enhanced ConvLSTM では，隠れ層の出力 h_t に対して Group Normalization を適用することによって，出力値を正規化し，過学習および勾配の爆発を抑制する．

3.2. ResNet ストリーム

ResNet ストリームでは，dynamic image を入力とし，Enhanced ConvLSTM ストリームに比べ，大きな動きの特徴を抽出する．ResNet は，画像認識において高い性能を示している CNNs の手法である．ResNet は，空間方向に skip connection を導入したことにより，勾配消失を抑制し，より大規模なモデルを学習できるようにした手法である．dynamic image は，画像シーケンスを時間方向で足し合わせた画像である．この画像は，時間的前後関係は失われているが，時間的情報が含まれた画像である．

ResNet ストリームでは，dynamic image を ResNet に入力することによって，Enhanced ConvLSTM よ

表 1 ResNet-50 の畳み込みカーネル

Layer name	Convolution kernel
Conv1	3×3，16
Conv2_x	$\begin{bmatrix}1\times1,\ 16\\3\times3,\ 16\\1\times1,\ 32\end{bmatrix}\times3$
Conv3_x	$\begin{bmatrix}1\times1,\ 32\\3\times3,\ 32\\1\times1,\ 64\end{bmatrix}\times4$
Conv4_x	$\begin{bmatrix}1\times1,\ 64\\3\times3,\ 32\\1\times1,\ 128\end{bmatrix}\times6$
Conv5_x	$\begin{bmatrix}1\times1,\ 128\\3\times3,\ 64\\1\times1,\ 128\end{bmatrix}\times3$

りマクロな特徴を抽出する．そして，得られた特徴量を tanh によって活性させる．その理由は，Enhanced ConvLSTM ストリームから得られた特徴量と同じスケールにするためである．また，本手法では, 層の総数が 50 である ResNet-50 を用いる．表 1 に ResNet の各ブロックのパラメータを記載する．

4. 実験

本研究では，eNTERFACE05 database[5]を用いて 1 つの実験をおこなった．1 つ目は，skip connection の有無による認識率の比較実験，2 つ目は，提案手法の構成要素の各組合せの性能比較実験，3 つ目は，従来の表情認識手法との比較実験である．

eNTERFACE05 database は，43 名の被験者から 1290 本の動画を取得したデータベースであり，anger, disgust, fear, joy, sadness, surprise の 6 種類の表情が教師信号として与えられている．各動画の長さは，1 秒から 4 秒程度である．動画の各フレームは，横幅 720 pixel，縦幅 570 pixel の 3 チャンネル (RGB) である．図 5 に eNTERFAFE05 database から得られる顔画像の例を示す．

本実験では，1 つの動画を 16 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．また，連続するビデオクリップは，前後 8 フレームが重なるようにして分割された．提案手法の最適化手法には，Momentum SGD を用いた．実験では，被験者を 5 グループに分割し，Leave-One-Subject-Group-Out (LOSGO)によって評価した．5 回の認識率の平均値によって性能を比較した．

図 5 eNTERFACE database から得られる顔画像

表 2　skip connection の有無による認識率

Method	Accuracy
(a) Conventional ConvLSTM	39.84%
(b) ConvLSTM with only temporal skip connection added	42.80%
(c) ConvLSTM with only spatial skip connection added	43.42%
(d) ConvLSTM with both temporal skip connection and spatial skip connection added (Enhanced ConvLSTM)	44.28%

4.1. skip connectionの有無による性能比較実験

skip connection の有無による実験結果を表 1 に示す．（a）は，2 つの ConvLSTM ストリームから構成され, それぞれ顔画像シーケンス, optical flow 画像シーケンスを入力として受け取る．(a) をベースとし，(b) は Temporal skip connection のみを加え．（c）は Spatial skip connection のみ，（d）には Temporal skip connection，Spatial skip connection の両方を加えた手法である．実験結果を表 2 に示す．実験の結果，Temporal skip connection のみを加えた手法（b），Spatial skip connection のみを加えた手法（c）ともに従来の ConvLSTM を用いた手法（a）より高い認識率を示した．また，（b）より（c）の方が高い認識率を示した．そして，Temporal skip connection，Spatial skip connection 両方加えた手法（d）が最も高い認識率を示した．(b) より（c）の方が高い認識率を示した要因として，LSTM は，c_t を一次線形和で保持することによって時間方向における勾配消失を抑制しているが，活性化関数にシグモイド関数やハイパボリックタンジェント関数を用いている．そのため，時間方向より空間方向における勾配消失の方が大きかったため，このような結果になったと考えられる．

4.2. 提案手法の各構成要素の性能比較実験

表 3 に提案手法の各構成要素の認識率を示す．

まず，Enhanced ConvLSTM ストリームのみの手法 (c) と従来の ConvLSTM ストリームのみの手法 (b) の認識率を比較する．改良型 ConvLSTM ストリームのみ（b）の認識率は，44.28%であった．そ

表 3 提案手法の各構成要素の認識率

Method	Accuracy
(a) 2 ResNet streams	33.70%
(b) 2 ConvLSTM streams (without skip connection)	39.84%
(c) 2 ConvLSTM streams with skip connection	44.28%
(d) 2 ResNet streams and 2 ConvLSTM streams ((a) and (b))	41.48%
(e) 2 ResNet streams and 2 ConvLSTM streams with skip connection ((a) and (b))	45.29%

れに対して，従来の ConvLSTM ストリームを使用した場合（c）の認識率は，39.84%であった．この結果から，Enhanced ConvLSTM は，従来の ConvLSTM に比べ，認識率が 4.44%向上したことが確認された．各クラスにおける認識率の向上を確認するために，図 4 の（b）と（c）を比較する．"surprise"では，低下しているが，その他のクラス（"anger", "disgust", "fear", "joy", "sadness"）においては，認識率が向上していることがわかる．"joy"クラスの精度が最も向上しており，約 8%向上している．

次に，提案手法(e)と Enhanced ConvLSTM ストリームのみを用いた場合(c)を比較する．提案手法(e)の認識率は，45.29%であった．それに対して，Enhanced ConvLSTM ストリームのみ(c)の場合の認識率は，44.28%であった．提案手法は，Enhanced ConvLSTM ストリームのみの場合に比べ，認識率が 1.01%向上したことを確認した．各クラスにおける認識率の向上を確認するために，図 5 の(e)と(c)を比較する．"disgust", "fear", "surprise"クラスにおいて若干の認識率の低下がみられるが，"anger"，"joy", "sadness"クラスにおいては，認識率が向上している．また，"anger"クラスにおいて最も精度が向上しており，約 7%の向上が確認できた．

最後に，ResNet ストリームと Enhanced ConvLSTM のどちらが認識率の向上に寄与しているかを確認する．従来の ConvLSTM ストリーム(b)に ResNet ストリーム(a)を追加した(d)の認識率は，41.48%であった．それに対して，Enhanced ConvLSTM(c)の認識率は，44.28%であった．この結果から ResNet ストリームより，Enhanced ConvLSTM の方が認識率の向上に寄与していることがわかった．

4.3. 提案手法と従来手法の性能比較実験

提案手法を従来手法と比較した結果を表 4 に示す．手法[15-17]は，Hand-craft 特徴量を用いた手法であり，手法[17]は，Hand-craft 特徴量を用いた手法の中で最も高精度な手法である．手法[14]は，

表 4 従来手法と提案手法の認識率

Method	Accuracy
Mansoorizadeh et al. [15]	38.00%
Fejani et al. [16]	39.28%
Zhalahpour et al. [17]	42.16%
Pan et al. [14]	42.98%
Ours	45.29%

VGG-19 と LSTM を組み合わせた End-to-End の DNNs 手法である．実験の結果，提案手法の認識率は 45.29%であり，従来手法と比べ，3.82%認識率が向上したことを確認した．

5. まとめ

本研究では，ConvLSTM の改良および，それを用いた動画からの表情認識手法を提案した．提案手法は，2 つの Enhanced ConvLSTM ストリームと 2 つの ResNet ストリームから構成される．Enhanced ConvLSTM ストリームでは，多層の Enhanced ConvLSTM layer によって時間的，空間的な関係を同時に表現した時空間的特徴を抽出した．Enhanced ConvLSTM は，従来の ConvLSTM の時間方向，空間方向それぞれに skip connection を追加することによって，時間，空間方向における勾配消失の抑制とより過去の情報を利用できるように改良された．

実験では，eNTERFACE05 database を用いて，skip connection の有無による性能比較，提案手法の構成要素の各組合せの性能比較および提案手法と従来手法の性能比較をおこなった．skip connection の有無による性能比較では，従来の ConvLSTM（skip connection なし）に比べ，Enhanced ConvLSTM（skip connection あり）の方が高い認識率を示した．提案手法の構成要素の各組合せの性能比較実験において，Enhanced ConvLSTM ストリームのみの場合の認識率は，44.28%であり，従来の ConvLSTM ストリームのみの場合に比べ，3.82%向上した．さらに，Enhanced ConvLSTM ストリームに ResNet ストリームを加えることによって認識率は 45.29%となり，Enhanced ConvLSTM ストリームのみの場合と比べ，1.01%認識率が向上した．従来手法との比較実験では，従来手法に比べ，2.31%認識率が向上したことを確認した．

今後の方針として，Enhanced ConvLSTM の最適化に最適かに取り組む．具体的には，どの程度の過去の情報を考慮すべきかを実験的に調査することによって，最適化をおこなう予定である．

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] P. Ekman et al., : “Constants across cultures in the face and emotion”, Journal of personality and social psychology, Vol.17, No.2 p.124, 1971

[2] M. S. Bartlett et al., : “Real time face detection and facial expression recognition:Development and applications to human computer interaction”, IEEE Conference on computer vision and pattern recognition workshop, Vol. 5, pp. 53-53, 2003

[3] P. Ekman et al., : “A new pan-cultural facial expression of emotion”, Motivation and emotion, Vol. 10, No. 2, pp. 159-168, 1986

[4] P. Lucey et al., : “The extended cohn-kanade dataset (ck+): A completedataset for action unit and emotion-specified expression”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94-101, 2010

[5] O. Martin et al., : “The enterface’05 audio-visual emotion database”, ICDEW’06, pp. 8-8, 2006

[6] M. Pantic et al., :“Web-based database for facial expression analysis”, IEEE International Conference on Multimedia and Expo, pp. 5, 2005

[7] Ambadar et al., "Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions.", Psychological science Vol.16, No.5, pp.403-410, 2005

[8] 佐々木康輔ら :“顔キーポイントの移動方向コードに基づく個人差の影響を受けにくい表情認識”, 電気学会論文誌 C (電子・情報・システム部門誌), Vol.138, No.5, pp.661-618, 2018

[9] W. Chao et al., : “Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection” Signal Processing, Vol. 117, pp.1-10, 2015

[10] P. Liu et al., : “Facial expression recognition via a boosted deep belief network” CVPR, pp. 1805‑1812, 2014

[11] F. Frade et al., : “Intraface” FG, 2015.

[12] A. Mollahosseini et al, : “Going deeper in facial expression recognition using deep neural networks” WACV, pp.1-10, 2016

[13] A. Lopes et al., : “Facial expression recognition with convolutionalneural networks: coping with few data and the training sample order”, Pattern Recognition, Vol.61, pp.610-628, 2017

[14] H. Ding et al., : “Facenet2expnet:Regularizing a deep face recognition net for expression recognition”, FG, pp. 118-126, 2017

[15] P. Khorrami et al., : “How deep neural networks can improve emotion recognition on video data ”, ICIP, pp. 619-623, 2016

[16] X. Pan et al., : “A deep spatial and temporal aggregation framework for video-based facial expression recognition”, IEEE Access, Vol. 7, pp.48807-48815, 2019

[17] M. Mansoorizadeh et al., : “Multimodal information fusion application to human emotion recognition from face and speech” Multimedia Tools and Applications, Vol. 49, No.2, pp.277-297, 2010

[18] M. Bejani et al., : “Audiovisual emotion recognition using anova feature selection method and multi-classifier neural networks”, Neural Computing and Applications,Vol.24, No.2, pp.399–412, 2014

[19] S. Zhalehpour al., : “Baum-1: A spontaneous audio-visual face database of affective and mental states”, IEEE Transactions on Affective Computing, Vol.8, No.3, pp.300–313, 2017

[20] K. Hara et al., : “Learning spatiotemporal features with 3d residual networks for action recognition”, CVPR, pp. 3154-3160, 2017

[21] D. Tran et al., : “Convnet architecture search for spatiotemporal feature learning”, arXiv, preprint arXiv:1708.05038, 2017

[22] K. Hara et al., : “Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?”, CVPR, pp. 6546–6555, 2018

[23] F. Heilbron et al., : “Activitynet: A large-scale video benchmark for human activity understanding”, CVPR, pp.961–970, 2015

[24] K. Soomro et al., : “Ucf101: A dataset of 101 human actions classes from videos in the wild”, arXiv, preprint arXiv:1212.0402, 2012

[25] W. Kay et al., : “The kinetics human action video dataset”, arXiv, preprint arXiv:1705.06950, 2017

[26] K. Simonyan et al., : “ery deep convolutional networks for large-scale image recognition”, arXiv preprint arXiv:1409.1556, 2014

[27] F. Gers et al., : “Lstm recurrent networks learn simple context-free and context-sensitive languages”, IEEE Transactions on Neural Networks, Vol.12, No 6, pp.1333-1340, 2001

[28] T. Baltrusaitis et al., : “Openface 2.0: Facial behavior analysis toolkit”, FG, pp.59-66, 2018

[29] G. Farneb¨ack et al., : “Two-frame motion estimation based on polynomial expansion”, In Scandinavian conference on Image analysis, pp.363–370, 2003

[30] X. Shi et al., : “Convolutional lstm network: A machine learning approach for precipitation nowcasting”, NIPS, pp.802-810, 2015

[31] Y. Wang et al., : “Eidetic 3d lstm: A model for video prediction and beyond.”, ICLR , 2019

[32] Y. Wu et al., “Group normalization”, ECCV, pp.3-19, 2018

[33] K. He et al., “Deep residual learning for image recognition”, CVPR, pp.770-778, 2016

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.