一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
MVE2019-12(2019-08)

改良型ConvLSTMを用いた動画からの表情認識手法の提案

三好　　遼† 　長田　典子†† 　橋本　　学†

† 中京大学工学研究科　〒466–8666 愛知県名古屋市昭和区八事本町 101–2
†† 関西学院大学理工学研究科/感性価値創造研究センター　〒669–1337 兵庫県三田市学園 2–1
E-mail: †{miyoshi,mana}@isl.sist.chukyo-u.ac.jp, ††nagata@kwansei.ac.jp

あらまし　本研究では，勾配消失の抑制および，より古い情報を利用するために，従来の ConvLSTM の時空間方向それぞれに skip connection を導入した改良型 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．Enhanced ConvLSTM ストリームでは，細かい動きの特徴の抽出，ResNet ストリームでは，大きな動きの特徴を抽出する．実験では，skip connection の有無による認識率の比較，および提案する表情認識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，提案する表情認識手法の認識率は 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

A Proposal of Facial Expression Recognition Method from Video Using Enhanced ConvLSTM

Ryo MIYOSHI†, Noriko NAGATA††, and Manabu HASHIMOTO†

† Graduate School of Engineering, Chukyo University　101–2 Yagoto-honmachi, Showa-ku, Nagoya-shi, Aichi, 466–8666 Japan
†† Graduate School of Science and Technology/Research Center for Kansei Value Creation, Kwansei Gakuin University　2–1 Gakuen, Shanda-shi, Hyogo, 669–1337 Japan
E-mail: †{miyoshi,mana}@isl.sist.chukyo-u.ac.jp, ††nagata@kwansei.ac.jp

Abstract　We propose an enhanced convolutional long short-term memory (ConvLSTM) algorithm, i.e., Enhanced ConvLSTM, by adding skip connections in the spatiotemporal directions to conventional ConvLSTM to suppress gradient vanishing and use older information and We propose a method that uses this algorithm to automatically recognize facial expressions from videos. The proposed facial expression recognition method consists of two Enhanced ConvLSTM streams and two ResNet streams. The Enhanced ConvLSTM streams extract features for fine movements, and the ResNet streams extract features for rough movements. We conducted experiments to compare a method using ConvLSTM with skip connections and a method without them. A method using Enhanced CovnLSTM had a 4.44% higher accuracy than the a method using conventional ConvLSTM. Also the proposed facial expression recognition method achieved 45.29% accuracy, which is 2.31% higher than that of the conventional facial expression recognition method.

Key words　facial expression recognition, convolutional LSTM, skip connection

1. まえがき

顔表情は，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チャンネルの 1 つである．Ekman らは，anger, disgust, fear, happiness, sadness, surprise という個人間で普遍な 6 つの基本顔表情を定義した [1]．これらの顔表情は，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への注目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．表情は，オンセットからオフセットまでの顔のアピアランスの変化によって表現される．そのため，顔のアピアランス（空間的特徴）と，それの変化（時間的特徴）が，表情認識のために重要な情報である．よって，それらをどのように特徴量化するかが重要である．

これまで表情認識において，2 タイプの研究がおこなわれてきた [7]～[17]（タイプ A：時間情報を利用しない表情認識手法，タイプ B：時間情報を利用した表情認識手法）．

タイプ A では，hand-craft 特徴を用いた手法 [7]～[9] や，Convolutional Neural Networks (CNNs) を用いた手法 [10]～[12] が提案されてきた．これらの手法は静止画像から表情を認識しており，表情認識に有用な時間的情報を考慮していない．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，これらの手法を動画からの表情認識に適用することは難しい．

タイプ B では，工学的な特徴を用いた手法 [13]～[15] や，Deep Neural Networks (DNNs) を用いた手法 [16], [17] などが提案されてきた．Pan ら [17] は，CNNs を用いて静止画像から顔の空間的特徴を抽出し，long short-term memory (LSTM) によって時間的特徴を抽出している．表情は，顔のアピアランスとそれの時間的変化によって表現され，それらは同時に変化する．そのため，時空間的特徴を捉えることが重要である．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．よって，時空間的特徴を得ることができない．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 [18]～[20] が考えられる．これらの手法は，3 次元（空間方向 2 次元，時間方向 1 次元）の畳み込みカーネルを用いた大規模なネットワークである．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 [21]～[23] に比べ，規模が著しく小さいため，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，これらの手法を表情認識に適用することはできない．

本研究では，改良型 convolutional LSTM (ConvLSTM) を用いた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，改良型 ConvLSTM を積み重ねることによって時空間的特徴を抽出する．改良型 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 ConvLSTM の時空間方向それぞれに skip connection を導入することによって改良された．実験では，skip connection の有無による認識率の比較および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，提案する表情認識手法の認識率は 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．

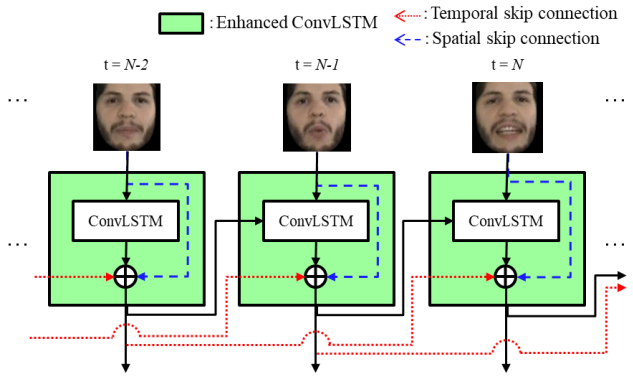


図 1　改良型 ConvLSTM の概要
Fig. 1　Outline of Enhanced ConvLSTM

2.　タイプ B の従来研究

2. 1　Hand-craft 特徴量を用いた手法

Hand-craft 特徴量を用いた手法として [13]～[15] がある．S. Zhalehpour ら [15] 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 である．

2. 2　DNNs を用いた手法

DNNs を用いた手法として [16], [17] がある．Pan ら [17] は，VGG-19 [24] と LSTM [25] 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．

3.　提案する表情認識手法

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，改良型 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

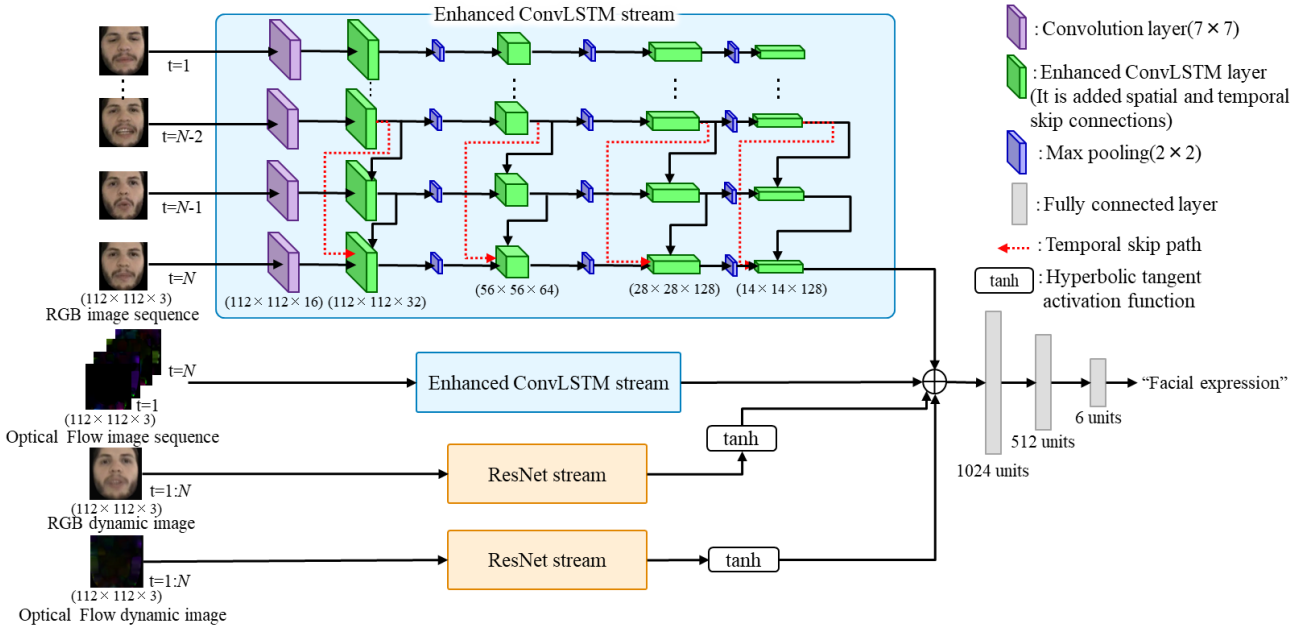


図 2　提案する表情認識手法の概要

Fig. 2　Outline of proposed facial-expression recognition method

ムに入力する顔画像は，OpenFace [26] によって取得される．また，optical flow は，Gunnar ら [27] 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 ストリームによって，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 ストリームおよび改良型 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，図 1 に示す改良型 ConvLSTM によって，時空間的特徴を抽出する．改良型 ConvLSTM のカーネルサイズは，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 [28] は，時空間的な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．以下に従来の ConvLSTM の重要な式を記載する．

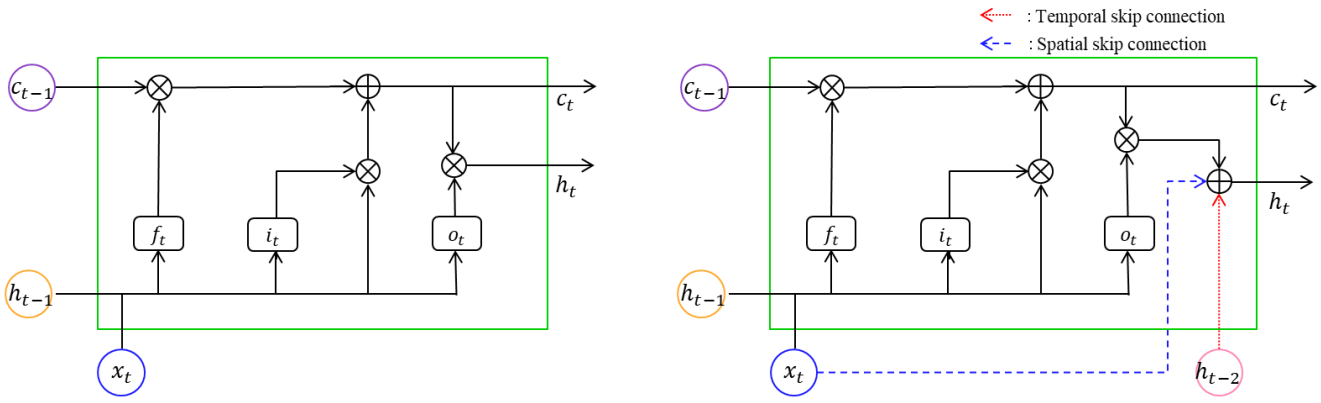
$$
\begin{aligned}
i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\
c_t &= f_t \circ c_{t-1} + i_t \circ tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1}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

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，$*$ 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き入力が通され，0 であるときゲートが閉じ，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．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 [29] 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，c_t がタイムステップ $t-1$ 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．

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．

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



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

図 3　従来の ConvLSTM と改良型 ConvLSTM の概要

Fig. 3　Outline of conventional ConvLSTM and Enhanced ConvLSTM

ここで，g は，Group Normalization [30] を示す．図 3 に従来の ConvLSTM と改良型 ConvLSTM の概要を示す．図 3，式 1，2 からわかるように，従来の ConvLSTM と改良型 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．勾配消失が起きる原因は，ネットワークが深くなることによって，層から層に伝播されえる誤差が小さくなるためである．skip connection を追加することにより，勾配がそのまま前の層に伝達されるパスができることによって勾配消失が抑制される．これは ResNet と同様のアイデアである．また，改良型 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 [31] 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，より大規模なネットワーク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，大きな動きを捉えるための特徴を抽出する．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 2 つの実験をおこなった．1 つ目は，提案手法の構成要素の各組合せの性能比較実験，2 つ目は，従来の表情認識手法との比較実験である．

表 1　ResNet-50 で使用したの畳み込みカーネル

Table 1　Convolution kernels used in ResNet-50

Layer name	convolution kernel
conv1	3 × 3, 16
conv2_x	[1×1,16; 3×3,16; 1×1,32] ×3
conv3_x	[1×1,32; 3×3,32; 1×1,64] ×4
conv4_x	[1×1,64; 3×3,32; 1×1,128] ×6
conv5_x	[1×1,128; 3×3,64; 1×1,128] ×3

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，RGB の 3 チャンネル 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 [32] 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．

4.1　提案手法の構成要素の各組合せの性能比較実験

まず，改良型 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

図 4　eNTERFACE05 database から得られる顔画像の例
Fig. 4　Example face images obtained from eNTERFACE05 database

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，改良型 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．

次に，提案手法（e）と改良型 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，改良型 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，改良型 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．

最後に，ResNet ストリームと改良型 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，改良型 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，改良型 ConvLSTM の方が認識率の向上に寄与していることがわかった．

4.2　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 3 に示す．手法 [13]〜[15] は，人手によって設計された特徴量を用いた手法であり，手法 [17] は，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．また，改良型 ConvLSTM のみを用いた場合の認識率は，44.28% であり，VGG-19 と LSTM を組み合わせた手法より，1.3% 認識率が高いことを確認した．この結果より，多層の改良型 ConvLSTM によって抽出された時空間的特徴が有用であることが確認された．

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．提案する表情認識手法は，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．ConvLSTM ストリームでは，細かな動きを捉えるための特徴，ResNet ストリームでは，大きな動きを捉えるための特徴を抽出する．改良型 ConvLSTM は，従来の ConvLSTM の時空間方向に skip connection を追加することによって，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．

表 2　提案手法の構成要素の各組合せの認識率
Table 2　Recognition rate of each combination of components of the proposed method

Method	Accuracy
(a) Method with only 2 ResNet streams	33.70%
(b) Method with only 2 conventional ConvLSTM streams (without skip connections)	39.84%
(c) Method with only 2 Enhanced ConvLSTM streams (with skip connections)	44.28%
(d) Method with 2 ResNet streams and 2 conventional ConvLSTM streams ((a) and (b))	41.48%
(e) Proposed method with 2 ResNet streams and 2 Enhanced ConvLSTM streams ((a) and (c))	45.29%

表 3　提案手法と従来手法の認識率
Table 3　Results of comparing proposed method with conventional methods

Method	Accuracy
Mansoorizadeh et al. [13]	38.00%
Fejani et al. [14]	39.28%
Zhalahpour et al. [15]	42.16%
Pan et al. [17]	42.98%
Method with only 2 Enhanced ConvLSTM streams	44.28%
Propsed	45.29%

実験では，eNTERFACE05 database を用いて，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．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，改良型 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，改良型 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，改良型 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．また，改良型 ConvLSTM のみを用いた場合の認識率は，44.28% であり，VGG-19 と LSTM を組み合わせた手法より，1.3% 認識率が高いことを確認した．この結果より，多層の改良型 ConvLSTM によって抽出された時空間的特徴が有用であることが確認された．

今後の方針として，改良型 ConvLSTM の効果が大きいことから，改良型 ConvLSTM の最適化に取り組む．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．

文　　献

[1] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and*

social psychology, Vol. 17, No. 2, p. 124, 1971.

[2] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R 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 workshop*, Vol. 5, pp. 53–53. IEEE, 2003.

[3] Paul Ekman and Wallace V Friesen. A new pan-cultural facial expression of emotion. *Motivation and emotion*, Vol. 10, No. 2, pp. 159–168, 1986.

[4] P. Lucey, J. F. 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*, pp. 94–101, June 2010.

[5] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp. 8–8. IEEE, 2006.

[6] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, pp. 5 pp.–, July 2005.

[7] Wei-Lun Chao, Jian-Jiun Ding, and Jun-Zuo Liu. Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. *Signal Processing*, Vol. 117, pp. 1–10, 2015.

[8] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1805–1812, June 2014.

[9] Fernando De la Torre Frade, Wen-Sheng Chu, Xuehan Xiong, Francisco Vicente Carrasco, Xiaoyu Ding, and Jeffrey Cohn. Intraface. In *Automatic Face and Gesture Recognition*, May 2015.

[10] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pp. 1–10. IEEE, 2016.

[11] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition*, Vol. 61, pp. 610–628, 2017.

[12] Hui Ding, Shaohua Kevin Zhou, and Rama Chellappa. Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 118–126. IEEE, 2017.

[13] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, Vol. 49, No. 2, pp. 277–297, 2010.

[14] Mahdi Bejani, Davood Gharavian, and Nasrollah Moghaddam Charkari. 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.

[15] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem. 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, July 2017.

[16] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang. How deep neural networks can improve emotion recognition on video data. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 619–623, Sep. 2016.

[17] 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*, Vol. 7, pp. 48807–48815, 2019.

[18] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3154–3160, 2017.

[19] Du Tran, Jamie Ray, Zheng Shou, Shih-Fu Chang, and Manohar Paluri. Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*, 2017.

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] Felix A Gers and E Schmidhuber. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, Vol. 12, No. 6, pp. 1333–1340, 2001.

[26] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

[27] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pp. 363–370. Springer, 2003.

[28] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 802–810. Curran Associates, Inc., 2015.

[29] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. 2018.

[30] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[32] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147, 2013.