Neural Computing and Applications (2021) 33:7381–7392
https://doi.org/10.1007/s00521-020-05557-4

S.I. : DICTA 2019

Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video

Ryo Miyoshi[1] · Noriko Nagata[2] · Manabu Hashimoto[1]

Received: 20 March 2020 / Accepted: 19 November 2020 / Published online: 2 January 2021

Abstract
We propose an algorithm that enhances convolutional long short-term memory (ConvLSTM), i.e., Enhanced ConvLSTM, by adding skip connections to spatial and temporal directions and temporal gates to conventional ConvLSTM to suppress gradient vanishing and use information that is older than the previous frame. We also 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. We conducted two experiments using eNTERFACE05 database and CK+. First, we conducted an ablation study to investigate the effectiveness of adding spatial and temporal skip connections and temporal gates to ConvLSTM. Ablation studies have shown that adding skip connections to spatial and temporal and temporal gates to conventional ConvLSTM provides the greatest performance gains. Second, we compared the accuracies of the proposed method and state-of-the-art methods. In an experiment comparing the proposed method and state-of-the-art methods, the accuracy of the proposed method was 49.26% on eNTERFACE05 database and 95.72% on CK+. Our proposed method shows superior performance compared to the state-of-the-art methods on eNTERFACE05.

Keywords Facial expression recognition · Deep learning · Recurrent neural networks · Long short-term memory

1 Introduction

Facial expression is an essential, nonverbal means of conveying one's emotions and intentions. Ekman et al. [1] defined six basic facial expressions—anger, disgust, fear, happiness, sadness, and surprise—which are universal among people. Automatic facial expression recognition is used in various fields such as human computer interaction (HCI) [2] and medical care [3].

A facial expression is produced over the course of three phases: onset, peak, and offset. Onset is the moment the expression begins, peak is the moment the expression appears most intense, and offset is the moment the expression disappears. An expression is represented by a spatiotemporal change from the onset to the offset. As dynamic facial information has been proven to be effective in the field of psychology for recognizing subtle facial expression [4], facial expression recognition systems need to be able to capture the dynamic changes mentioned above. However, it is difficult to capture changes in facial expression between adjacent frames because there is little to no change.

Studies on automatic facial expression recognition can be divided into two types. The first type uses a single image. The methods developed in these studies use handcrafted features [5–7] and convolutional neural networks (CNNs) [8–10]. However, they do not take into account dynamic facial information which is effective for facial expression recognition. The second type uses sequential images. The methods developed in these studies use handcrafted features and deep neural networks (DNNs). The handcrafted features used for the method [11–13] are

✉ Ryo Miyoshi
miyoshi@isl.sist.chukyo-u.ac.jp

Manabu Hashimoto
mana@isl.sist.chukyo-u.ac.jp

[1] Graduate School of Engineering, Chukyo University, Nagoya, Japan

[2] School of Science and Technology, Kwansei Gakuin University, Sanda, Japan

low level compared with the DNNs features. Two types of DNNs take an image sequence as input. The first type extracts spatial and temporal features with multi-modules [14, 15]. This type uses CNNs to extract spatial features and recurrent neural networks (RNNs) to learn the temporal information of the obtained spatial features. Since the spatial and temporal features are extracted by different modules, it is not possible to extract spatiotemporal features in which spatial and temporal relationships are simultaneously expressed. The second type of DNNs extracts spatial and temporal features with a single module. In this method, 3D CNNs have been shown to be effective for action recognition [16–18], and convolutional long short-term memory (ConvLSTM) has been shown to be effective for video prediction [19]. While 3D CNNs is an effective method, it has numerous parameters and requires large datasets, making it difficult to train. The database for facial expression recognition [20–22] is significantly smaller than those for action recognition [23–25]. It is difficult to apply 3D CNNs to facial expression recognition. In contrast, ConvLSTM can extract spatiotemporal features and has fewer parameters than 3D CNNs.

In this study, we propose a method based on ConvLSTM. However, there are two disadvantages to using ConvLSTM for facial expression recognition. The first is that older information cannot be referenced. Since the facial expression of the video does not change much between adjacent frames, it may be effective to use the information that older than the previous frame. As described in a previous study [26], LSTM cannot retain long-past information because it is affected by the information in the previous frame. Therefore, as the sequence becomes longer, older information is lost and cannot be referenced. The second is that the gradient tends to vanish. ConvLSTM is expanded in the temporal direction, so the layer becomes deeper in this direction, and gradient vanishing is more likely to occur. In addition, it uses sigmoid and tanh activation functions. As a result, the multilayer ConvLSTM may cause gradient vanishing between layers.

To use older information and suppress gradient vanishing, we added skip connections in the spatial and temporal directions and temporal gates to conventional ConvLSTM. The algorithm is named enhanced convolutional LSTM (Enhanced ConvLSTM). The addition of a temporal gates makes it possible to use older information, and temporal skip connections suppress gradient vanishing in the temporal direction. Furthermore, spatial skip connections suppress gradient vanishing between layers. We also propose a facial expression recognition method using Enhanced ConvLSTM. The proposed method consists of two Enhanced ConvLSTM streams in which spatiotemporal features are extracted by stacking the Enhanced ConvLSTM, as shown in Fig. 1.

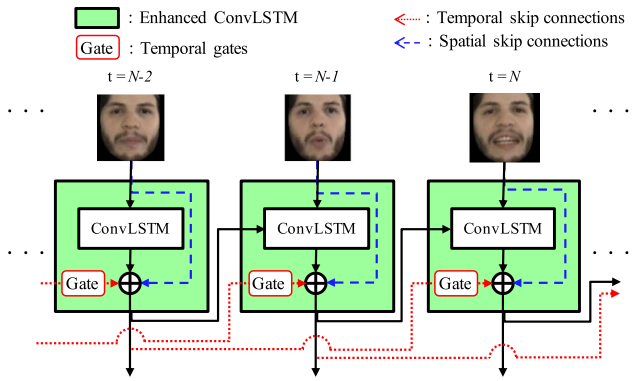


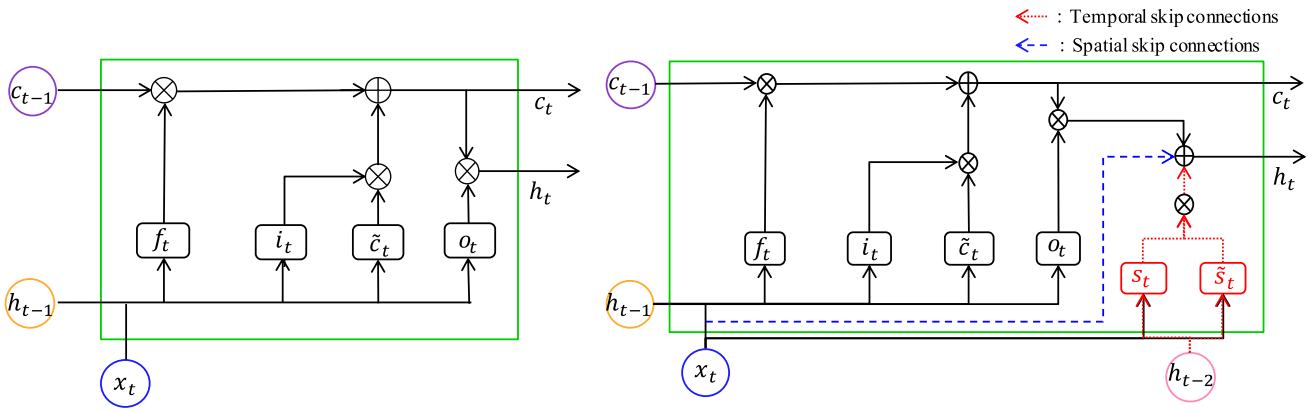
Fig. 1 Enhanced ConvLSTM. Two types of paths shown with red dotted and blue dashed lines were added to conventional ConvLSTM. In addition, temporal gates were added to capture older information more effectively. The red and blue paths make it possible to suppress gradient vanishing. Furthermore, the red paths and the gates enable the use of older information which was not previously possible with conventional ConvLSTM

We conducted two experiments using the eNTERFACE database and CK+. First, we investigated the effectiveness of adding skip connections and gates to ConvLSTM by conducting an ablation study. Second, we compared the accuracies of the proposed method and state-of-the-art methods. Ablation studies have shown that adding skip connections to spatial and temporal and temporal gates to conventional ConvLSTM provides the greatest performance gains. In an experiment comparing the proposed method and state-of-the-art methods, the accuracy of the proposed method was 49.26% on eNTERFACE05 database and 95.72% on CK+. Our proposed method shows superior performance compared to the state-of-the-art methods on eNTERFACE05.

2 Related work on facial expression recognition from image sequence

2.1 Handcrafted feature-based methods

Previously proposed methods [11–13] use handcrafted features for automatic facial expression recognition in videos. One such method [13] selects the video frame in which the facial expression is the most intense. It also uses local phase quantization features to characterize the texture of the face. Although this method was the highest performing among the three methods using handcrafted features in the eNTERFACE05 database, it is less accurate than DNN-based methods.



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

Fig. 2 Outline of conventional ConvLSTM and Enhanced ConvLSTM. Red dotted lines represent temporal skip connections and blue dashed lines represent spatial skip connections. Adding x_t s_t, and $\tilde{s}_t$ to conventional ConvLSTM enables spatial and temporal skip connections. s_t represents a temporal skip gate and $\tilde{s}_t$ represents a temporal skip modulation gate. These gates receive information from two steps before and the current information and learn features that effectively capture changes in facial expression

2.2 DNNs-based methods

Previous methods using DNNs (a combination of CNNs and RNNs) [14, 15] have been proposed for automatic facial expression recognition in videos. CNNs were used to extract spatial features and RNNs were used to extract temporal features. Pan et al. [15] proposed an effective method that combined the VGG-19 [27] with the LSTM [28] method. VGG-19 [27] is an image recognition technique using CNNs. LSTM [28] is a type of RNN capable of learning temporal dependencies. In this method, spatial and temporal features are extracted independently using the VGG-19 and LSTM, respectively. Because different modules are used for extraction, the method [14, 15] cannot extract features that simultaneously represent spatial and temporal relationships. In addition, LSTM is expanded in the temporal direction, causing the layer to be deeper in that direction. Therefore, gradient vanishing is likely to occur. In addition, LSTM is affected by the information in the previous frame, so it is difficult to retain a large amount of past information. Since the change in facial expression between adjacent frames is minimal, information from older frames is more effective for capturing changes in facial expression.

3 Enhanced convolutional LSTM (Enhanced ConvLSTM)

This section describes the proposed Enhanced ConvLSTM in detail.

Findings from psychological studies indicate that focusing on changes in facial expressions is effective for recognizing the expressions. Thus, it is important to pay attention to facial expression changes in automatic facial expression recognition. However, it is difficult to capture the changes on the video because the changes are minimal between adjacent frames. In addition, the use of ConvLSTM in video-based facial expression recognition is effective for capturing spatiotemporal changes in the face, but gradient vanishing may occur. Therefore, in this study, we propose an algorithm that suppresses gradient vanishing in ConvLSTM and focuses on facial expression changes.

Conventional ConvLSTM [19] is an algorithm that changes operations at each LSTM gate into a convolution operator to extract spatiotemporal features. The essential equations of conventional ConvLSTM are given below.

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

where σ is the sigmoid activation function, $*$ and $\circ$ denote the convolution operator and Hadamard product, respectively, x_t represents input data at time t, h_t represents the state of the hidden layer of the current frame, h_{t-1} represents the state of the hidden layer of the previous frame, i_t represents the input gates at time t, $\tilde{c}_t$ represents input modulation gates at time t, c_t represents the memory cell at t, f_t represents the forget gates at t, o_t represents output gates at t, W represents the weight matrix, and b represents the offset matrix.

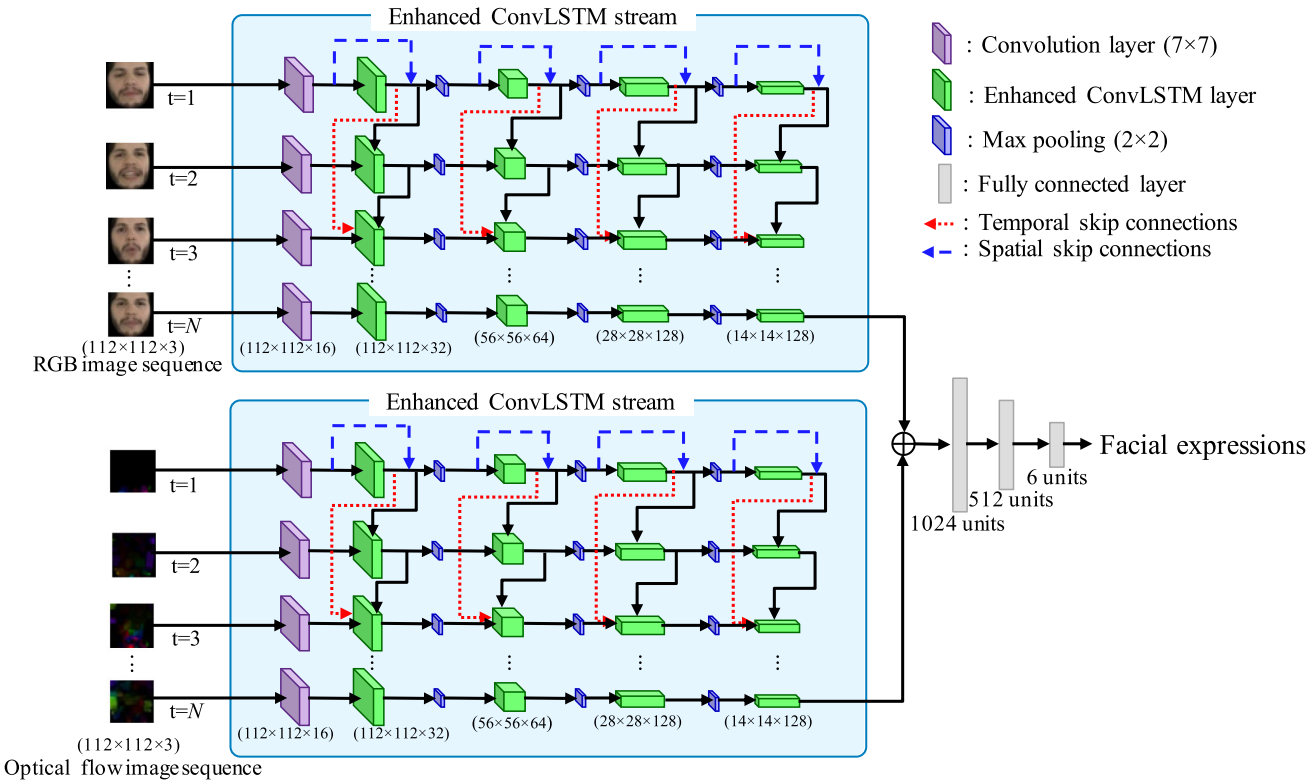


Fig. 3 Outline of proposed method consisting of two Enhanced ConvLSTM streams. Each stream receives RGB and optical flow image sequences. The facial expressions are recognized by fusing the output feature maps from each stream and inputting them to the fully connected layer. Feature maps are fused by calculating the element-wise sum of all feature maps

Conventional ConvLSTM controls c_t using i_t and f_t. When i_t is 1, the input gate is open, and when it is 0, the gate is closed and the input is blocked. The same is true of f_t. Then, c_t is updated on the basis of the outputs. c_t is controlled by the results of the output gates o_t on the basis of time steps t and $t-1$. As described in a previous study [26], LSTM cannot retain long-past information because it is affected by the information in the previous frame. Therefore, as the sequence becomes longer, older information is lost and cannot be referenced. In addition, LSTM is expanded in the temporal direction, so the layer becomes deeper in this direction, and gradient vanishing is more likely to occur. ConvLSTM uses sigmoid and tanh activation functions. As a result, the multilayer ConvLSTM may cause gradient vanishing between layers.

To suppress gradient vanishing and enable the use of older information, Enhanced ConvLSTM contains skip connections in the spatial and temporal directions and temporal gates in the temporal direction of conventional ConvLSTM. The essential equations of Enhanced ConvLSTM are given below.

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

where g represents group normalization [29]. Figure 1 shows an outline of Enhanced ConvLSTM, and Fig. 2 shows the details of the conventional and Enhanced ConvLSTM. As seen in Fig. 2 and Eq. 2, the difference between conventional ConvLSTM and Enhanced ConvLSTM is s_t, $\tilde{s_t}$ and h_t. s_t is referred to as the temporal skip gate and $\tilde{s_t}$ is the temporal skip modulation gate. These gates receive information from two steps before and the current information to learn effective features that capture changes in facial expression. For h_t in Enhanced ConvLSTM, add $W_{xr} * x_t$ and $s_t \circ \tilde{s_t}$. $W_{xs} * x_t$ does not pass through the gates in LSTM. Adding $W_{xr} * x_t$ to h_t creates a

route that propagates the gradient in the spatial direction. $s_t \circ \tilde{s}_t$ takes the information from two steps before as input. Thus, a route that propagates the gradient in the temporal direction is created by adding $s_t \circ \tilde{s}_t$ to h_t.

4 Proposed facial expression method

We developed our Enhanced ConvLSTM algorithm by adding temporal gates and skip connections to the spatial and temporal directions in conventional ConvLSTM. We also developed a facial expression recognition method that uses Enhanced ConvLSTM.

The outline of our proposed method is shown in Fig. 3. The method consists of two Enhanced ConvLSTM streams and three fully connected layers. RGB and optical flow image sequences are inputted to each Enhanced ConvLSTM stream. Then, the feature maps obtained from each stream are summed element-wise, and the sum is inputted to the fully connected layers to recognize the facial expression. A method that combines CNNs and LSTM such as the one developed by Pan et al. [15] cannot extract spatiotemporal features which simultaneously represent the relationship between spatial and temporal because spatial feature extraction using CNNs and temporal feature extraction using LSTM are different modules. Therefore, in the Enhanced ConvLSTM streams of the proposed method, spatiotemporal features are extracted by stacking Enhanced ConvLSTMs which can extract features that simultaneously represent the spatial and temporal relationship. As with many CNNs-based methods, max pooling is applied after the Enhanced ConvLSTM layer to extract high-level features by stacking the Enhanced ConvLSTM. The kernel size of the convolution filter in each Enhanced ConvLSTM is 5×5 in the first and second layers and 3×3 in the third and fourth layers.

The facial images in each stream were obtained using the OpenFace application module [30]. Optical flow was calculated using a common method [31].

5 Experiments

We conducted two experiments. First, we investigated the accuracy of each component of the proposed facial expression recognition method by conducting an ablation study. Second, we compared the accuracies of the proposed method and conventional methods. This section describes the experimental conditions and results.

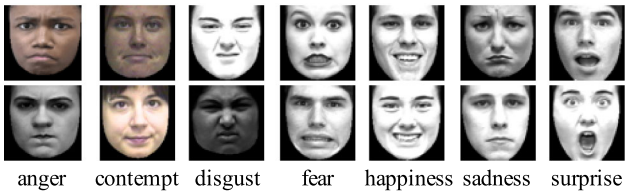
5.1 Experimental conditions

We investigated the effectiveness of the proposed method using the eNTERFACE05 database [21] and CK+ [20].

eNTERFACE05 [21] database contains 1,290 videos of 43 subjects. Six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) are given as supervised labels. The length of each video is about 1–4 s. Each video frame consists of three channels (RGB) and measures 570 pixels in height and 720 pixels in width. In the experiment with this dataset, we divided the video into 16-frame clips for network training and testing. The clip was split so that the previous 8-frame overlapped with the next 8-frame. The supervised label given to the clip was that given to the original video. Each clip was classified, and the class with the most clips was the final classification result. The proposed method is evaluated by leave-one-subject-group-out (LOSGO) which subjects from the eNTERFACE05 database were divided into 5 groups and cross-validation. This is same evaluation method used several previous works [11–13, 15].

CK+ [20] is a dataset containing 593 image sequences obtained from 123 subjects. In addition, among those image sequences, 327 image sequences obtained from 118 subjects were labeled with seven basic expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). This dataset is a database focused on the analysis of facial expressions. In the experiment with this dataset, we split the video into 8-frame clips and set the overlap to 6 frames for network training and testing. The classification method is the same as the experiment using the eNTERFACE05 database. The proposed method is evaluated by tenfold person-independence cross-validation experiments. This is

(a) eNTRERFACE05 database



(b) CK+

Fig. 4 Example face images obtained from eNTERFACE database and CK+

same evaluation method used several previous works [32, 33].

Figure 4 shows example facial images obtained from the eNTERFACE database and CK+. The images in each stream were obtained using the OpenFace module [30].

We utilized four Tesla V100 with 16 GB memory. For ConvLSTM and the proposed method, the mini-batch size was 6, and the learning rate was 0.000125. We also utilized Momentum SGD as the optimizer. The same hyperparameters and optimizer were used in all experiments.

5.2 Ablation study

In the ablation study, we investigated the effectiveness of spatial skip connections, temporal skip connections, and temporal gates.

5.2.1 Evaluation on eNTERFACE05 database

Table 1 shows the accuracy of the ablation study, and Fig. 5 shows the confusion matrix of accuracy.

Table 1 Results from ablation study using eNTERFACE05 database

Method	Accuracy (%)
(a) ConvLSTM (Conventional)	39.84
(b) ConvLSTM with spatial skip connections	45.68
(c) ConvLSTM with temporal skip connections and temporal gates	45.84
(d) Enhanced ConvLSTM (proposed)	49.26

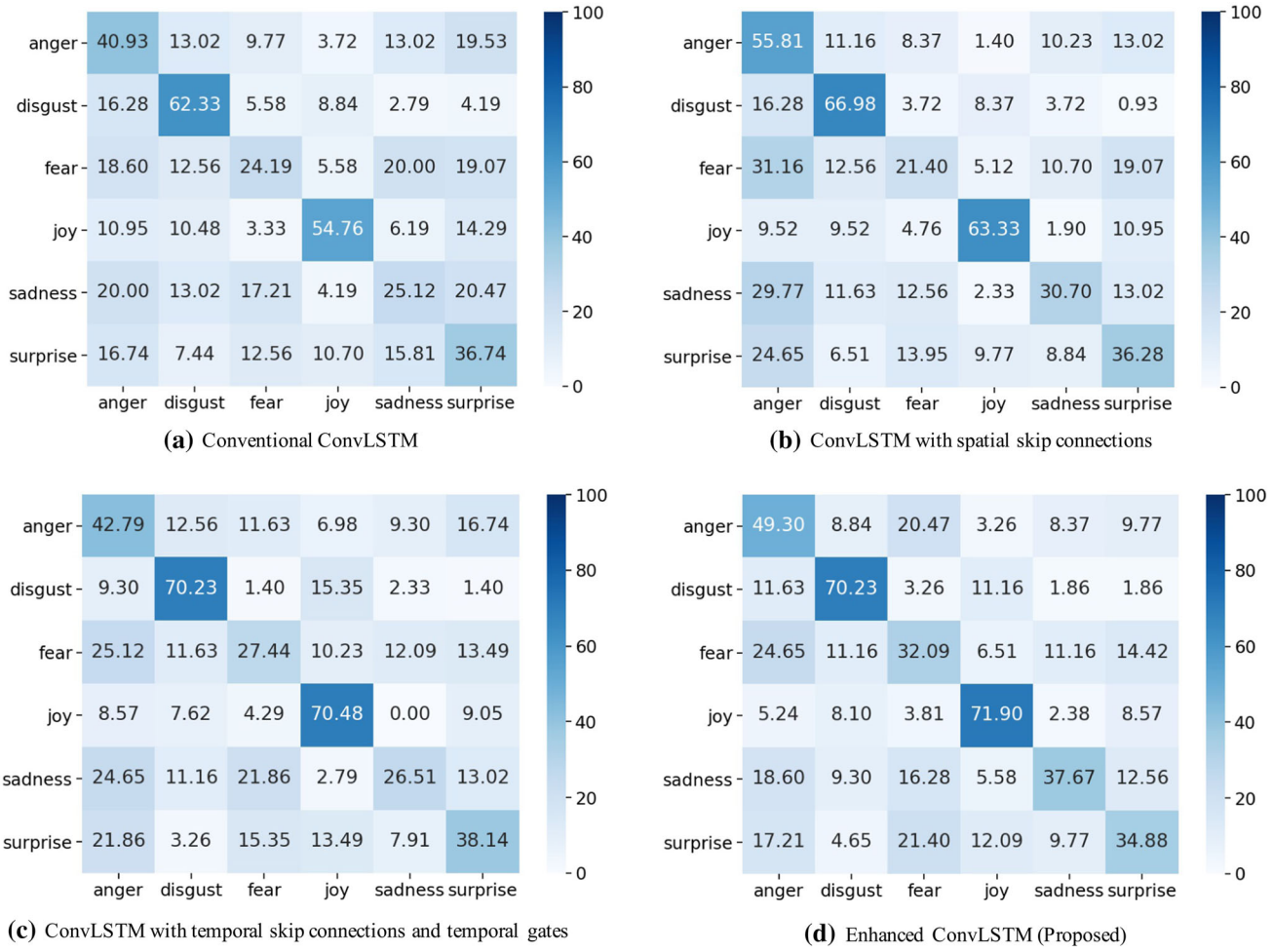


Fig. 5 Confusion matrix obtained from ablation study using eNTEFFACE05 database. A comparison between **a** ConvLSTM and **d** Enhanced ConvLSTM shows that the accuracy improved in most classes in **d**

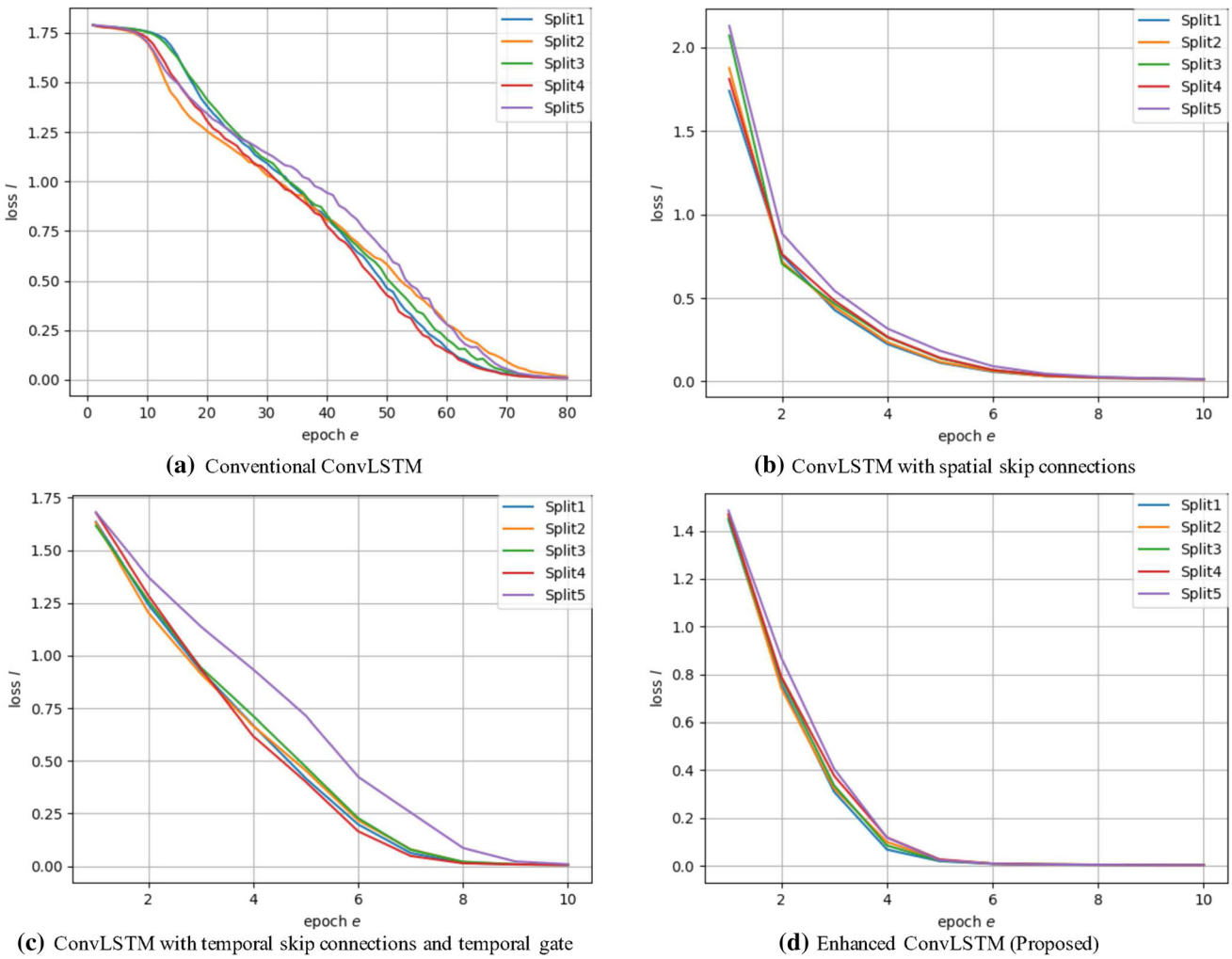


Fig. 6 Changes in loss during training. The proposed method is evaluated by leave-one-subject-group-out (LOSGO) which subjects from the eNTERFACE05 database were divided into 5 groups and cross-validation. Split1 represents the change in loss when trained on all groups except Group 1. Likewise, Split2 is the change in loss when trained on all groups except Group 2, Split3 excludes Group 3, Split4 excludes Group 4, and Split5 excludes Group 5

First, we compared the accuracies of conventional (a) ConvLSTM and (b) ConvLSTM with spatial skip connections. The accuracy of (a) was 39.84% and the accuracy of (b) was 45.68%, an improvement of 5.84%. In addition, the comparison between (a) and (b) in Fig. 5 showed improved accuracy in most classes. The accuracy of "fear" and "surprise" decreased slightly while the accuracy of the other classes ("anger," "disgust," "joy," "sadness") improved. The class "anger" improved the most, with an increase of 14.88%.

We then compared the accuracy of (a) conventional ConvLSTM with that of (c) ConvLSTM with temporal skip connections and temporal gates. The accuracy of (a) was 39.84% and the accuracy of (c) was 45.84%, an improvement of 6%. In addition, the comparison of (a) and (c) in Fig. 5 showed improved accuracy in all classes. The accuracy of the class "joy" improved the most, with an increase of 15.72%.

The above two results revealed that each skip connection and gate improved the accuracy of different classes. In other words, the skip connections and gates each captured different features that were effective for facial expressions recognition.

Finally, we compared the accuracy of (b) and (d) Enhanced ConvLSTM and compared the accuracy of (c) and (d). As shown in Table 1, the accuracy of Enhanced ConvLSTM was 49.26%, 3.58% higher than (b) and 3.42% higher than (c). Furthermore, comparing (b), (c), and (d) in Fig. 5, (d) appears to be the results of (b) and (c) complementing each other. For example, in the classes "fear" and "sadness," (b) was 21.40% and 30.70% and (c) was 27.44% and 26.51%, respectively, whereas (d) was 32.09%

Table 2 Results from ablation study using CK+

Method	Accuracy (%)
(a) ConvLSTM (conventional)	92.97
(b) ConvLSTM with spatial skip connections	94.50
(c) ConvLSTM with temporal skip connections and temporal gates	88.38
(d) Enhanced ConvLSTM (proposed)	95.72

and 37.67%, showing improvements in accuracy. This is likely because more precise features can be obtained by adding both spatial skip connections and temporal skip connections and temporal gates.

The spatial skip connections, temporal skip connections, and temporal gates were added to prevent gradient vanishing. Figure 6 shows the changes in the loss during training. As shown in Fig. 6, adding spatial skip connections (Fig. 6b) or temporal skip connections and temporal gates (Fig. 6c) to conventional ConvLSTM enables quick loss convergence. In addition, the loss of Enhanced ConvLSTM (Fig. 6d) converged more quickly than that of the other methods. We determined that these results were caused by better propagation of the gradient. Thus, spatial skip connections, temporal skip connections, and temporal gates can be considered effective for gradient vanishing. We also found that the use of both spatial and temporal skip connections and temporal gates converged the loss

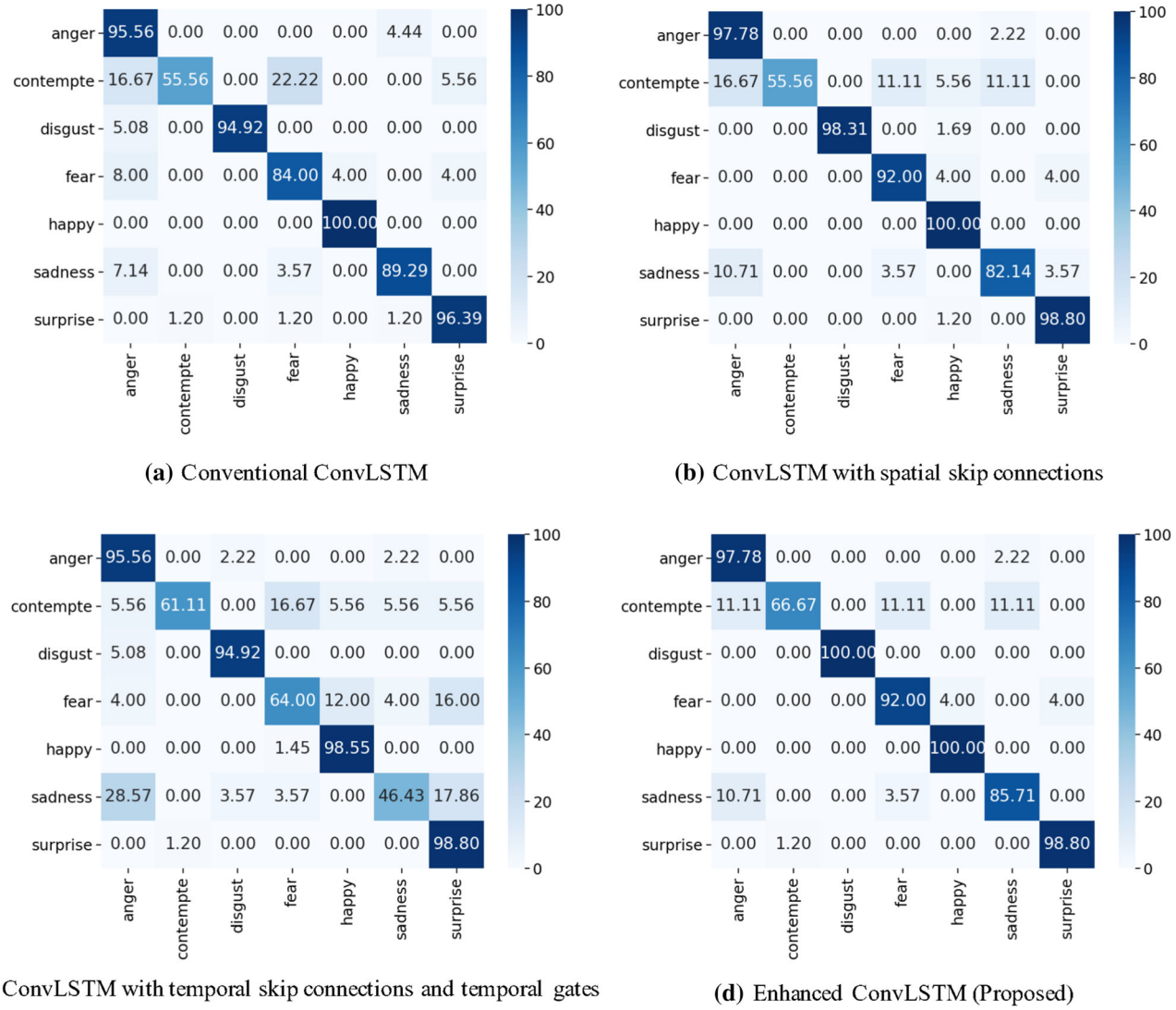
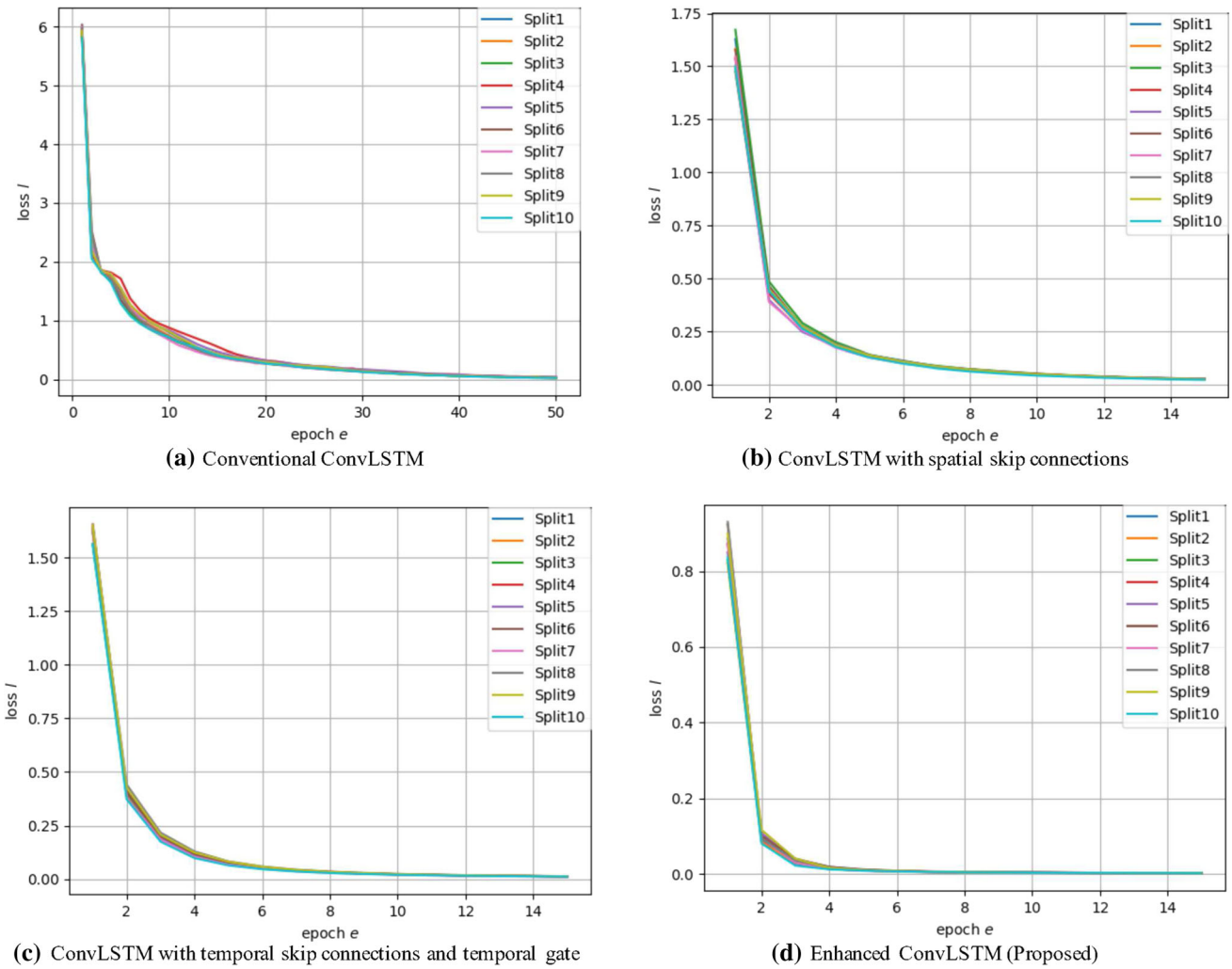


Fig. 7 Confusion matrix obtained from ablation study using CK+. A comparison between **a** ConvLSTM and **d** Enhanced ConvLSTM shows that the accuracy improved in most classes in **d**



(a) Conventional ConvLSTM

(b) ConvLSTM with spatial skip connections

(c) ConvLSTM with temporal skip connections and temporal gate

(d) Enhanced ConvLSTM (Proposed)

Fig. 8 Changes in loss during training. The proposed method is evaluated by tenfold person-independence cross-validation experiments. Split1 represents the change in loss when trained on all groups except Group 1. Likewise, Split2 is the change in loss when trained on all groups except Group 2, Split3 excludes Group 3, Split4 excludes Group 4, Split5 excludes Group 5, Split6 excludes Group 6, Split7 excludes Group 7, Split8 excludes Group 8, Split9 excludes Group 9, and Split10 excludes Group 10

most quickly. Therefore, the use of both spatial and temporal skip connections and temporal gates was most effective for preventing gradient vanishing.

5.2.2 Evaluation on CK+

Table 2 shows the accuracy of the ablation study, and Fig. 7 shows the confusion matrix of accuracy.

First, we compared the accuracies of conventional (a) ConvLSTM and (b) ConvLSTM with spatial skip connections. The accuracy of (a) was 92.97% and the accuracy of (b) was 94.50%, an improvement of 1.53%. In addition, the comparison between (a) and (b) in Fig. 7 showed improved accuracy in most classes.

We then compared the accuracy of (a) conventional ConvLSTM with that of (c) ConvLSTM with temporal skip connections and temporal gates. The accuracy of (a) was 92.97% and the accuracy of (c) was 88.38%, an decreased of 4.59%. However, the comparison of (a) and (c) in Fig. 5 showed some classes have improved accuracy.

The above two results, (b) contributes to the improvement of accuracy, but (C) may have an adverse affect recognition.

Finally, we compared the accuracy of (b) and (d) Enhanced ConvLSTM and compared the accuracy of (c) and (d). As shown in Table 1, the accuracy of Enhanced ConvLSTM was 95.72%, 1.22% higher than (b) and 7.34% higher than (c). Furthermore, comparing (b), (c), and (d) in Fig. 5, (d) appears to be the results of (b) and

Table 3 Comparison between proposed method and previous methods

Method	Accuracy (%)
Mansoorizadeh et al. [11]	38.00
Fejani et al. [12]	39.28
Zhalahpour et al. [13]	42.16
Pan et al. [15]	42.98
Meng et al. [34]	48.02
ConvLSTM (conventional)	39.84
ConvLSTM with spatial skip connections	45.68
ConvLSTM with temporal skip connections and temporal gates	45.84
Enhanced ConvLSTM (proposed)	49.26

(c) complementing each other also improved accuracy in most class. It was found that the recognition accuracy of (c) alone is lower than that of (a), but better features can be obtained by combining with (b).

Figure 8 shows the changes in the loss during training. As shown in Fig. 8, adding spatial skip connections (Fig. 8b) or temporal skip connections and temporal gates (Fig. 8c) to conventional ConvLSTM enables quick loss convergence. In addition, the loss of Enhanced ConvLSTM (Fig. 8d) converged more quickly than that of the other methods. Similar to Sect. 5.2.1, the proposed method is thought to converge learning faster and propagate the gradient better.

5.3 Comparison with state-of-the-art methods

We compare the state-of-the-art methods and the proposed method for the two databases.

5.3.1 Evaluation on eNTERFACE05 database

Table 3 compares the proposed method with conventional facial expression recognition methods, some of which use handcrafted features [11–13], combine CNNs and LSTM [15] and the CNN-based method [34]. The accuracy of the proposed method was 49.26%, which is 1.24% higher than that of the conventional methods. In addition, the accuracy of ConvLSTM was lower than that of the conventional method, but we verified that the accuracy becomes higher than that of the conventional method after adding skip connections in the spatial direction or skip connections and new gates in the temporal direction. The results indicate that feature extraction by stacking Enhanced ConvLSTM can extract effective features in comparison with CNN-based method and methods combining CNNs and LSTM.

5.3.2 Evaluation on CK+

Table 4 compares the proposed method with state-of-the-art methods for CK+. The accuracy of the proposed method was 95.72%. The proposed method showed higher accuracy than some conventional methods, but could not show the highest accuracy. However, the proposed method shows higher accuracy in the eNTEFFACE05 database than in FAN, which is the best in CK+. The eNTEFFACE05 database is a dataset that focuses on natural emotional expression. Therefore, the proposed method is superior in recognizing more natural emotional expressions.

6 Conclusion

We proposed Enhanced ConvLSTM, an effective algorithm for facial expression recognition. The algorithm suppressed gradient vanishing and enabled the use of older information with the addition of spatial and temporal skip connections and temporal gates. The spatial and temporal skip connections created a route in which the gradient did not vanish during back-propagation. The added temporal gates receive information from two steps before and the

Table 4 Comparison between proposed method and previous methods

Method	Accuracy (%)
ST network [35]	98.50
DTAGN [36]	97.25
CNN+Island loss [37]	94.35
LOMo [38]	92.00
FAN [34]	99.69
Enhanced ConvLSTM (proposed)	95.72

current information to learn effective features that capture changes in facial expression.

The proposed facial expression recognition method consists of two Enhanced ConvLSTM streams. We conducted two experiments using the eNTERFACE database. First, we investigated the effectiveness of adding skip connections and gates to ConvLSTM by conducting an ablation study. Second, we compared the accuracies of the proposed method and conventional facial expression recognition methods. Ablation studies have shown that adding skip connections to spatial and temporal and temporal gates to traditional ConvLSTM provides the greatest performance gains. In addition, it has been shown that learning converges fastest and is effective in preventing gradient vanishing when both spatial and temporal skip connections and temporal gates are used. In an experiment comparing the proposed method and state-of-the-art methods, the accuracy of the proposed method was 49.26% for the eNTERFACE05 database and 95.72% for the CK+. Our proposed method shows superior performance compared to the state-of-the-art methods on eNTERFACE05.

Acknowledgements This research was partially supported by the Center of Innovation Program (Grant No. JPMJCE1314) from the Japan Science and Technology Agency (JST).

References

1. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Personal Soc Psychol 17(2):124
2. Bartlett MS, Littlewort G, Fasel I, Movellan JR (2003) 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. IEEE, pp 53–53
3. Ekman P, Friesen WV (1986) A new pan-cultural facial expression of emotion. Motiv Emot 10(2):159–168
4. Ambadar Z, Schooler JW, Cohn JF (2005) Deciphering the enigmatic face: the importance of facial dynamics in interpreting subtle facial expressions. Psychol Sci 16(5):403–410
5. Chao W-L, Ding J-J, Liu J-Z (2015) Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection. Signal Process 117:1–10
6. Liu P, Han S, Meng Z, Tong Y (2014) Facial expression recognition via a boosted deep belief network. In: 2014 IEEE conference on computer vision and pattern recognition, pp 1805–1812
7. De la Torre Frade F, Chu W-S, Xiong X, Carrasco F V, Ding X, Cohn J (2015) Intraface. In: Automatic face and gesture recognition
8. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–10
9. Lopes AT, de Aguiar E, De Souza AF, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recognit 61:610–628
10. Ding H, Zhou SK, Chellappa R (2017) Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017). IEEE, pp 118–126
11. Mansoorizadeh M, Charkari NM (2010) Multimodal information fusion application to human emotion recognition from face and speech. Multimed Tools Appl 49(2):277–297
12. Bejani M, Gharavian D, Charkari NM (2014) Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks. Neural Comput Appl 24(2):399–412
13. Zhalehpour S, Onder O, Akhtar Z, Erdem CE (2017) Baum-1: a spontaneous audio-visual face database of affective and mental states. IEEE Trans Affect Comput 8(3):300–313
14. Khorrami P, Le Paine T, Brady K, Dagli C, Huang TS (2016) How deep neural networks can improve emotion recognition on video data. In: 2016 IEEE international conference on image processing (ICIP), pp 619–623
15. Pan X, Ying G, Chen G, Li H, Li W (2019) A deep spatial and temporal aggregation framework for video-based facial expression recognition. IEEE Access 7:48807–48815
16. Hara K, Kataoka H, Satoh Y (2017) Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE international conference on computer vision, pp 3154–3160
17. Tran D, Ray J, Shou Z, Chang SF, Paluri M (2017) Convnet architecture search for spatiotemporal feature learning. arXiv:1708.05038
18. Hara K, Kataoka H, Satoh Y (2018) 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
19. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W, Woo WC (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Inc, pp 802–810
20. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) 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
21. Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface'05 audio-visual emotion database. In: 22nd international conference on data engineering workshops (ICDEW'06). IEEE, pp 8–8
22. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and expo, p 5
23. Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) 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
24. Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402
25. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al (2017) The kinetics human action video dataset. arXiv:1705.06950
26. Wang Y, Jiang L, Yang MH, Li LJ, Long M, Fei-Fei L (2019) Eidetic 3d lstm: a model for video prediction and beyond. In: ICLR
27. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
28. Gers FA, Schmidhuber E (2001) LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Trans Neural Netw 12(6):1333–1340
29. Wu Y, He K (2018) Group normalization. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19

30. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). IEEE, pp 59–66
31. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. Springer, pp 363–370
32. Liu M, Shan S, Wang R, Chen X (2014) Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1749–1756
33. Kuo CM, Lai SH, Sarkis M (2018) A compact deep learning model for robust facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2121–2129
34. Meng D, Peng X, Wang K, Qiao Y (2019) Frame attention networks for facial expression recognition in videos. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 3866–3870
35. Zhang K, Huang Y, Yong D, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203
36. Jung H, Lee S, Yim J, Park S, Kim J (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE international conference on computer vision, pp 2983–2991
37. Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Island loss for learning discriminative features in facial expression recognition. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). IEEE, pp 302–309
38. Sikka K, Sharma G, Bartlett M (2016) Lomo: latent ordinal model for facial analysis in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5580–5589