Facial-Expression Recognition from Video Using Enhanced Convolutional LSTM

1st Ryo Miyoshi
Chukyo University
Graduate School of Engineering
Nagoya-shi, Japan
miyoshi@isl.sist.chukyo-u.ac.jp

2nd Noriko Nagata
Kwansei Gakuin University
Graduate School of Science and Technology
Sanda-shi, Japan
nagata@kwansei.ac.jp

3rd Manabu Hashimoto
Chukyo University
Graduate School of Engineering
Nagoya-shi, Japan
mana@isl.sist.chukyo-u.ac.jp

***Abstract*—We propose an enhanced convolutional long short-term memory (ConvLSTM) algorithm, i.e., Enhanced ConvLSTM, by adding skip connections in the spatial and temporal directions to conventional ConvLSTM to suppress gradient vanishing and use older information. 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 and two ResNet streams. The Enhanced ConvLSTM streams extract features for fine movements, and the ResNet streams extract features for rough movements. In the Enhanced ConvLSTM streams, spatio-temporal features are extracted by stacking the Enhanced ConvLSTM. We conducted experiments to compare a method using ConvLSTM with skip connections (proposed Enhanced ConvLSTM) and a method without them (conventional ConvLSTM). 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 method.**

***Index Terms*—Facial-expression recognition, ConvLSTM, skip connection**

I. Introduction

Facial expressions are the most important nonverbal way to communicate emotions and intentions. Ekman et al. defined six basic facial expressions: anger, disgust, fear, happiness, sadness, and surprise [1]. These facial expressions are universal among people and are used in fields such as human computer interaction (HCI) [2] and medical care [3]. The need for automatic facial-expression recognition from videos is increasing, and databases such as [4]–[6] have been released.

Temporal information is important in facial-expression recognition. A facial expression is divided into three phases: onset (the moment the expression begins), peak (the moment the expression appears most strongly), and offset (the moment the expression disappears). An expression is represented by the change from onset to offset, and that change needs to be captured in facial-expression recognition. Therefore, a method is needed for extracting spatio-temporal features that are effective for facial-expression recognition. Many facial-expression recognition methods have been proposed [7]–[14], which can be divided into two types (Types A and B)

Center of Innovation Program from the Japan Science and Technology Agency, JST

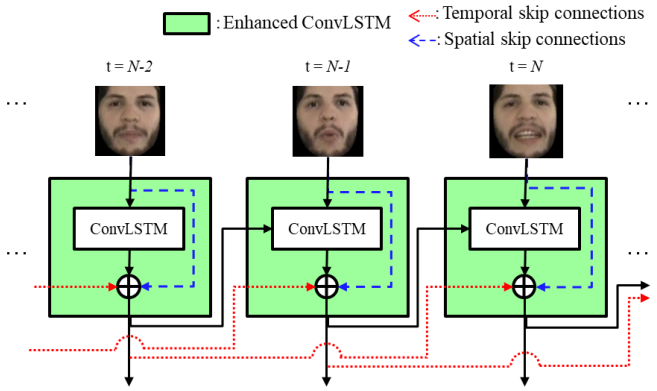


Fig. 1. Enhanced ConvLSTM. Two types of paths shown with red dotted and blue dashed lines were added to conventional ConvLSTM. Red and blue paths make it possible to suppress gradient vanishing, and red paths contribute to using older information than with conventional ConvLSTM.

Type-A are methods of facial-expression recognition from still images. They are the methods that using hand-crafted features [7]–[9] and using convolutional neural networks (CNNs) [10]–[12]. However, these methods do not take into account the temporal relationship in the transition between facial expressions. Videos contain many frames unrelated to facial expressions, and these frames negatively affect the accuracy of such methods to recognize facial expressions from still images. Therefore, these methods are difficult to use with videos.

Type-B methods are of facial-expression recognition from videos that take into account temporal information. They are the methods using hand-crafted features [15]–[17] and using deep neural networks (DNNs) [13], [14]. Pan et al. [14] used CNNs to extract features from still images and LSTM to recognize facial expressions by learning the temporal information of the obtained features. With this method, spatial and temporal features are extracted by different modules.

978-1-7281-3861-2/19/$31.00 ©2019 IEEE

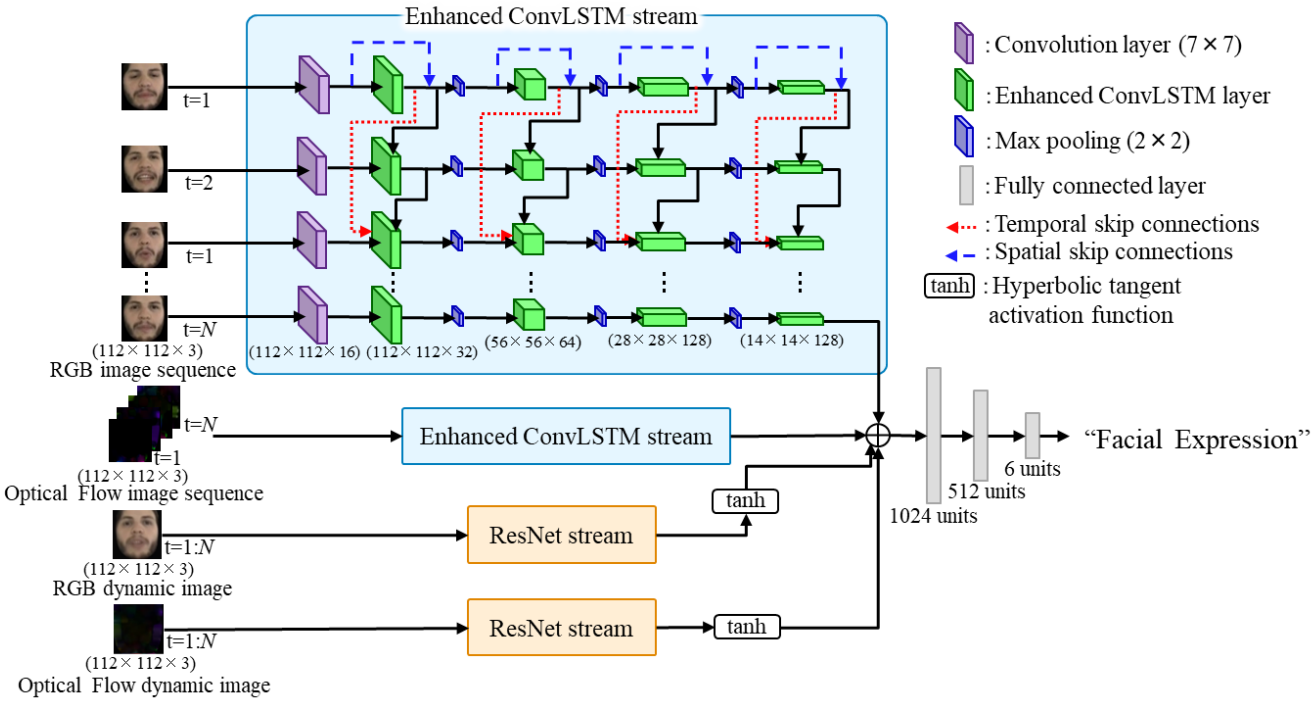


Fig. 2. Outline of proposed method. It consists of four streams, two Enhanced ConvLSTM streams and two ResNet streams. Enhanced ConvLSTM streams extract features for fine movements, and ResNet streams extract features for rough movements.

Therefore, it is impossible to extract spatio-temporal features. In addition, LSTM is expanded in the temporal direction, so the layer will be deeper in the temporal direction. Therefore, gradient vanishing is likely to occur. LSTM responds strongly to information one frame before; therefore, it is difficult to keep a large amount of past information. The change in facial expression between adjacent frames is not very large; therefore, older information is considered more effective for capturing changes in facial expression than information one frame before.

On the contrary, there are 3D-CNN-based methods that use 3D data (considering the video as three dimensions in the spatial and temporal directions) for human-action recognition [18]–[20]. We considered applying these methods to facial-expression recognition from videos. However, the database for facial-expression recognition is significantly smaller than those for action recognition [21]–[23]. With the facial-expression recognition database, it is difficult to learn the parameters of 3D CNN-based methods, which are effective in action recognition; thus, these methods cannot be used.

We propose an effective facial-expression recognition method and Enhanced ConvLSTM that has skip connections in the spatial and temporal directions to suppress grandient vanishing and uses older information. The proposed facial-expression recognition method consists of four streams: two Enhanced ConvLSTM streams and two ResNet streams. In the Enhanced ConvLSTM streams, spatio-temporal features are extracted by stacking the Enhanced ConvLSTM, as shown in Figure 1. We conducted experiments to compare a method using Enhanced ConvLSTM (with skip connections) and a method with conventional ConvLSTM (without the skip connections). Experimental results indicate that the a method using Enhanced ConvLSTM achieved a 4.44% higher accuracy than the a method using conventional ConvLSTM. The proposed facial-expression recognition method achieved 45.29% accuracy, which is 2.31% higher than that of the conventional method.

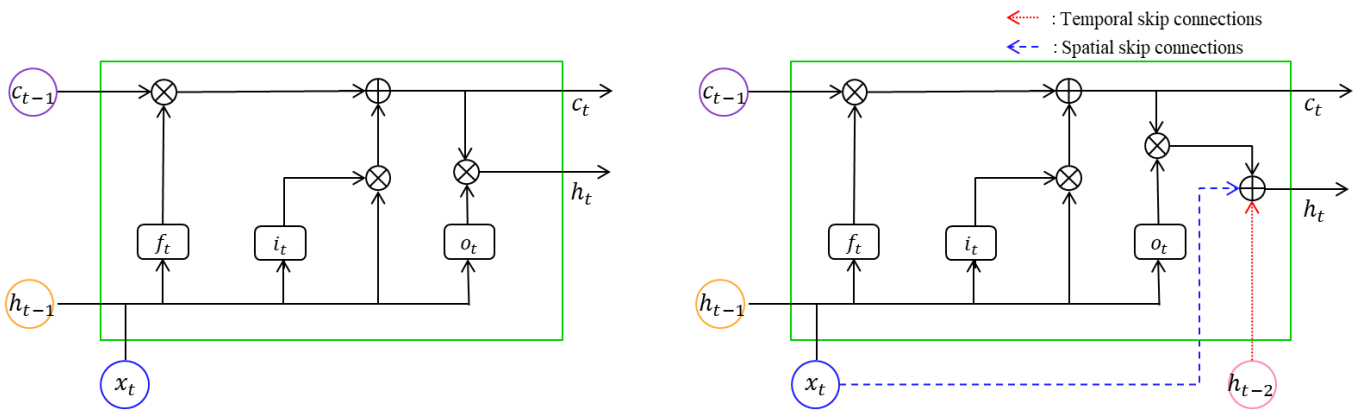
II. Related work regarding type-B methods

A. Hand-crafted feature-based methods

There are three major type-B methods [15]–[17] that automatic facial-expression recognition from videos using hand-crafted features. One such method [17] selects the frame in which the facial expression is the most expressed from the video. It also uses the local phase quantization features to characterize the texture of the face. This method achieved the highest performance among the three methods using hand-crafted features in the eNTERFACE05 database. However, it does not perform as well as DNN-based methods.

B. DNN-based method

Other type-B methods [13], [14] use DNNs. Pan et al. [14] proposed an effective method combining the VGG-19 [24] with LSTM [25] method. VGG-19 [24] is one of effective method for image recognition using CNNs. LSTM [25] is are a special kind of RNN, capable of learning long-term



(a) ConvLSTM (Conventional)

(b) Enhanced ConvLSTM (Proposed)

Fig. 3. Outline of conventional ConvLSTM and Enhanced ConvLSTM. Red dotted lines represent temporal skip connections, and blue dashed lines represent spatial skip connections. Spatial and temporal skip connections are enabaled by adding x_t and h_{t-2} to conventional ConvLSTM h_t calculation.

dependencies. With this method, a spatial feature is extracted using the VGG-19 and a temporal feature is extracted using LSTM independently. Therefore, no integrated spatio-temporal features can be extracted with this method.

III. Proposed Method

We enhanced conventional ConvLSTM by adding skip connections to both spatial and temporal directions. We call this algorithm "Enhanced ConvLSTM". We also developed a facial-expression recognition method that uses Enhanced ConvLSTM.

The outline of the proposed facial-expression method is shown in Figure 2. It consists of four streams, i.e., two Enhanced ConvLSTM streams and two ResNet streams, and three fully connected layers. RGB and optical flow image sequence are inputted to the Enhanced ConvLSTM streams, and the dynamic images of RGB and optical flow are transferred to the ResNet streams. Then, the feature maps obtained from the four streams are totaled and inputted to the fully connected layers to recognize facial expressions. The face images of each stream are easily obtained using the OpenFace application module [26]. Optical flow is also calculated using a common method [27]. The Enhanced ConvLSTM streams extract features for fine movements, and the ResNet extracts features for rough movements.

A. Enhanced ConvLSTM streams

This subsection describes the Enhanced ConvLSTM streams and the proposed Enhanced ConvLSTM in detail.

With a method that combines CNNs and LSTM such as Pan et al. [14] , spatio-temporal features cannot be extracted 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, spatio-temporal features are extracted by stacking Enhanced ConvLSTMs. The same as with many CNN-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 1st and 2nd layers and 3×3 in the 3rd and 4th layers. The reason for using large filters in shallow layers is to extract spatially global features. In deep layers, however, small filters are used to extract local features.

Conventional ConvLSTM [28] is an algorithm that changes operations at each gate of LSTM into the convolution operator to extract spatio-temporal features. The important equations of conventional ConvLSTM are given below.

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

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 results of input gates at time t, c_t represents the memory cell at t, f_t represents the results of input gates at t, o_t represents the results of input gates at t, W represents the weight matrix, and b represents the offset matrix.

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 based on those outputs. Furthermore, c_t is controlled by the results of the output gates o_t. It is controlled based on time steps t and $t-1$. As described in a previous study [29], LSTM cannot hold long-past information. The reason is that it responds strongly to the information one

frame before. 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 will be deeper in this direction. Therefore, gradient vanishing is likely to occur.

To suppress gradient vanishing and use older information, Enhanced ConvLSTM has skip connections in the spatial and temporal directions of conventional ConvLSTM. The important equations of Enhanced ConvLSTM are given below.

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

where g represents group normalization [30].

Figure 1 shows an outline of Enhanced ConvLSTM, and Figure 3 shows details of conventional and Enhanced ConvLSTMs. As you can see from Figure 3 and Equation 2, the difference between conventional ConvLSTM and Enhanced ConvLSTM is h_t. The spatial and temporal skip connections create a route where the gradient does not vanish during back-propagation. Also, older information can be used by the temporal skip connections. Enhanced ConvLSTM can be easily implemented by slightly changing conventional ConvLSTM.

B. ResNet streams

This subsection describes the ResNet streams in detail.

The ResNet streams extract features of rough movements. ResNet [31] is a method of suppressing gradient vanishing by adding skip connections and performs well in image recognition. A dynamic images have shown to be compact and effective representations of images in action recognition. A dynamic image is an image in which image sequences are totaled in the temporal direction. It contains temporal information and represents rough movements. Therefore, a dynamic image is inputted to ResNet, and we extract features for rough movements. Then, the obtained features are activated by $tanh$ that is hyperbolic tangent activation function. The reason for activation with $tanh$ is to make them the same scale as the features obtained from the Enhanced ConvLSTM streams.

IV. Experiments and results

We investigated the effectiveness of the proposed method and Enhanced ConvLSTM using the eNTERFACE05 database [5]. The eNTERFACE05 database contains 1,290 videos from 43 subjects, and 6 types of expressions (“anger”, “disgust”, “fear”, “happiness”, “sadness”, and “surprise”) are given as teacher signals. The length of each video is about 1 to 4 seconds. Each frame of the video is three channels (RGB) with 570 pixels in height and 720 pixels in width. Figure 5 shows an example face images obtained from the eNTERFACE database.

We conducted two experiments by using eNTERFACE database. First, we investigated accuracy of each component the proposed facial-expression recognition method by an ablation study. Second compared the accuracies of the proposed method and conventional facial-expression recognition methods. The subjects of eNTERFACE database were divided into five groups and evaluated the method using leave-one-subject-group-out. The accuracies of the proposed method and other facial-expression recognition methods were determined from the average value of five times.

A. Ablation study

Table I shows the accuracy of each component of the proposed method and the accuracy when the combination of components is changed, and Figure 4 shows the confusion matrix of accuracy.

TABLE I
Results from ablation study

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%

First, we compared the accuracy of a method using only conventional ConvLSTM streams (b) and a method using only Enhanced ConvLSTM streams (c). The accuracy of (b) was 39.84% and that of (c) was 44.28%, an improvement of 4.44%. We also compared (b) and (c) in Figure 4 to confirm the improvement in accuracy for each class. Although accuracy decreased for “surprise”, it improved for other classes (“anger”, “disgust”, “fear”, “joy”, “sadness”). The accuracy for “joy” improved the most; by about 8%.

We then compared the method using only the Enhanced ConvLSTM streams (c) with the proposed method (e). The accuracy of (c) was 44.28% and that of (e) was 45.29%, an improvement of 1.01%. We also compare (c) and (e) in Figure 4 to confirm the improvement in accuracy for each class. Although there was a slight decrease in accuracy for “disgust”, “fear”, and “surprise”, it improved for “anger”, “joy”, and “sadness”. The accuracy was the highest for “anger” class, an improvement of about 7%

Finally, we investigated which streams, ResNet or Enhanced ConvLSTM, contributes to improving accuracy. The accuracy of (d) was achieved by adding the ResNet streams (a) to the conventional ConvLSTM streams (b) 41.48%. The accuracy of the method using only Enhanced ConvLSTM streams (c) was 44.28%. Therefore, Enhanced ConvLSTM streams contributes more to accuracy than the ResNet streams.

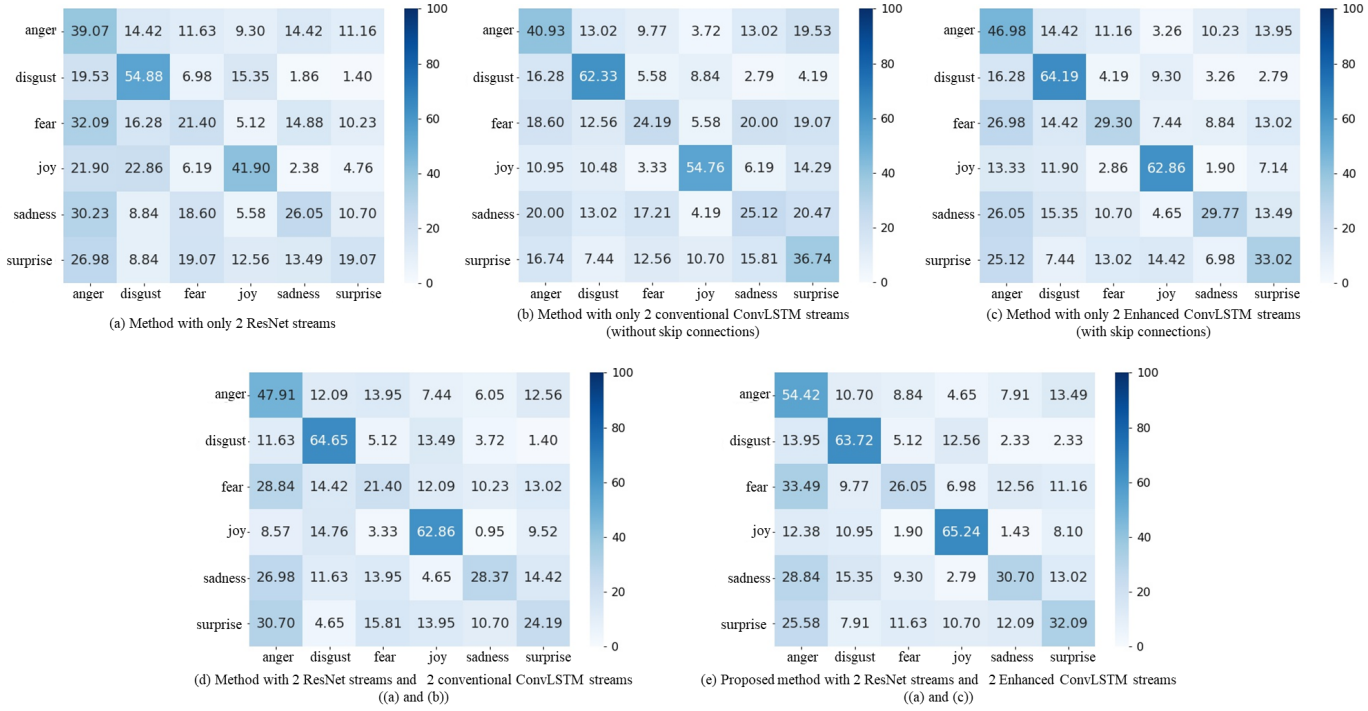


Fig. 4. Average value of the results of 5 experiments from ablation study

Fig. 5. Example face images obtained from eNTERFACE database

TABLE II
RESULTS OF COMPARING PROPOSED METHOD WITH PREVIOUS METHODS

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%
Method with only 2 Enhanced ConvLSTM streams	44.28%
Proposed	45.29%

B. Comparison with state-of-the-art methods

Table II compares the proposed method and conventional facial-expression recognition methods. Some use hand-crafted features [15]–[17] and combining CNNs and LSTM [14]. The accuracy of the proposed method was 45.29%, which is 3.82% better than the conventional methods. The accuracy improved 1.3% compared to that a method combining CNNs and LSTM when only Enhanced ConvLSTM streams were used. This indicates that feature extraction by stacking Enhanced ConvLSTMs can extract effective features compared to the feature extraction combining CNNs and LSTM.

V. CONCLUSION

We proposed an effective facial-expression recognition method and Enhanced ConvLSTM to suppress gradient vanishing and use older information by adding skip connections in the spatial and temporal directions. The proposed 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 two experiments by using eNTERFACE database. First, we investigated accuracy of each component the proposed facial-expression recognition method by an ablation study. Second compared the accuracies of the proposed method and conventional facial-expression recognition methods.

In the ablation study, the accuracy in the method using only the Enhanced ConvLSTM streams was 44.28%, which is 3.82% better than the conventional ConvLSTM streams. In the comparison experiments the proposed and conventinal facial-exression recognition methods, the accuracy of the proposed method was 45.29%, an improvement of 2.31% compared to the coventional methods. The accuracy of a method using only Enhanced ConvLSTM streams also improved 1.3% compared to a method combining CNNs and LSTM [14].

For future work, we will work on the optimization of

Enhanced. Specifically, we plan to experimentally investigate how much past information should be used.

VI. Acknowledgments

This research was partially supported by the Center of Innovation Program from the Japan Science and Technology Agency (JST).

References

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[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.