Facial Expression Recognition Based on Deep Spatio-Temporal Attention Network

Shuqin Li[1,2], Xiangwei Zheng[1,2(✉)], Xia Zhang[3], Xuanchi Chen[1,2], and Wei Li[4(✉)]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China
xwzhengcn@163.com
[2] Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan, China
[3] Internet Diagnosis and Treatment Center, Taian City Central Hospital, Taian, China
[4] Shandong Normal University Library, Shandong Normal University, Jinan, China
157953429@qq.com

Abstract. Facial expression recognition is extremely critical in the process of human-computer interaction. Existing facial expression recognition tends to focus on a single feature of the face and does not take full advantage of the integrated spatio-temporal features of facial expression images. Therefore, this paper proposes a facial expression recognition based on a deep spatio-temporal attention network (STANER) to capture the spatio-temporal features of facial expressions when they change subtly. A facial expression recognition with an attention module based on spatial global features (SGAER) is created firstly, where the addition of the attention module is able to quantify the importance of each part of the expression feature map and thus extract the spatial global appearance features at the time of subtle expression changes from a single frame expression image. Then, facial expression recognition with C-LSTM based on temporal local features (TLER) is built to process image sequences of facial regions linked to expression creation and extract dynamic local temporal information about expressions. Experiments are carried out on CK+ and Oulu-CASIA datasets. The results showed that STANER can achieve better performance with the accuracy rates of 98.23% and 89.52% on the two mainstream datasets, respectively.

Keywords: Facial expression recognition · Spatio-temporal features · Deep attention network

1 Introduction

With the rapid development and application of human-computer interaction [3] in various fields (e.g., healthcare [11], smart home [14]), facial expression recognition (FER) has gained more and more attention due to its important role

H. Gao et al. (Eds.): CollaborateCom 2022, LNICST 461, pp. 516–532, 2022.
https://doi.org/10.1007/978-3-031-24386-8_28

in human-computer interaction. Ekman *et al.* [5] in the 1970 s defined human facial expressions into six categories: anger, contempt, sadness, fear, happy and surprise. With the study of facial expressions, contempt was added as a basic expression by Matsumoto [25], so that emotions are now commonly classified into seven basic categories.

In some circumstances, changes in facial expressions appear to be very minor in the facial region and a fundamental difficulty in FER is capturing the features of minor changes in expression images. Although the process of facial expression production is very complex, it was found that it is mainly assessed by the main regions of the face (e.g., eyes, nose, mouth, etc.) after early studies [1,7,24]. Therefore, it is more helpful for FER to emphasize the detection of dynamic temporal change features of expressions from consecutive frames of important regions of the face.

According to the feature representation, FER is now classified into two categories: static image-based methods and dynamic sequence-based methods [18]. Static image-based methods [21,28,32] extract spatial appearance features well, but they ignore the dynamic temporal information generated during facial expression changes. On the contrary, dynamic sequence-based methods [13,39,40] extract dynamic temporal information well, but they overlook the change in the spatial appearance of the image. Hence, it is a difficult task to extract and apply the spatio-temporal features of facial expression images in FER activities.

In this paper, a facial expression recognition based on a deep spatio-temporal attention network (STANER) is developed to learn both subtle expression change features and dynamic change features of key facial regions. The first branch is FER with attention module based on spatial global features (SGAER), which is intended to leverage those subtle expression change features that are normally missed because certain tiny changes in expressions are difficult to capture. The second branch is FER with C-LSTM based on temporal local features (TLER), which is proposed to acquire dynamic local temporal features in consecutive expression frames. Based on a sequence of key facial regions that generate facial expressions, i.e., eyes, nose and mouth. These local consecutive frames are fed into the C-LSTM block to obtain high-level temporal features after extracting the shallow features, capturing the dynamic information of the local facial regions.

The contributions of this work are summarized as follows:

(1) A STANER is proposed to capture the spatio-temporal features of facial expression images to improve the robustness of facial expression recognition. The superiority of STANER has been demonstrated by extensive experiments on the mainstream datasets CK+ [23] and Oulu-CASIA [38].

(2) A SGAER branch is designed to solve the problem that when the magnitude of facial muscle changes is small and subtle facial expression change features are difficult to capture. The module optimizes the utilization of spatial features by using the attention created to track the subtle features of expression changes.

(3) A TLER branch is established to learn dynamic fine-grained temporal features of key local regions of the face. TLER detects spatio-temporal informa-

tion of local sequences using C-LSTM blocks constructed by convolutional neural networks (CNN) [17] and long short-term memory neural networks (LSTM) [10].

The rest of this paper is structured as follows. The survey of FER is discussed in Sect. 2. In Sect. 3, STANER method is described. In Sect. 4, the experimental analysis and their results are discussed to prove the efficiency of proposed method. Finally, the conclusion of the paper is presented in Sect. 5.

2 Related Works

2.1 Facial Expression Recognition

FER research has been conducted by traditional methods [21,39] and deep learning methods [13,28]. Traditional methods, including Local Binary Patterns (LBP) [32], Scale-invariant Feature Transform (SIFT) [22], Histogram of Oriented Gradients (HOG) [8], are adopted to extract facial expression image features, which are then fed into a classifier (e.g., Support Vector Machine SVM [2]) for the classification task. Pan *et al.* [29] proposed to bridge the gap between visual features and emotions by using both the use of CNNs and HOG to obtain more comprehensive VFER features and SVM for expression recognition. This method had good performance and outperformed the current level of conventional techniques. However, when the information becomes more complex, the representational capacity of hand-crafted features diminishes and classic approach models are unable to adequately fit large-scale complicated data. It was not until the introduction of deep learning [9], its subsequent rapid development in various fields and the great success of deep neural networks in many pattern recognition tasks based on large data and complex scenes that more and more researchers started to conduct experiments with deep learning-based facial expression recognition [18].

In recent years, CNNs have been increasingly popular in FER tasks. Kim *et al.* [15] combined multiple deep CNNs for training and won the FER international competition EmotoW2015. Liu *et al.* [20] used several CNNs with different structures for FER with good results. In addition, recurrent neural networks (RNNs) have also been employed in FER because they are better at predicting dynamic temporal aspects of arbitrary length sequences [4,6]. Researchers have discovered that LSTMs [10] are better at solving gradient disappearance and gradient explosion during training and they are commonly used to learn temporal features in FER. Zhang *et al.* [36] proposed a PHRNN-MSCNN, which consists of a partially hierarchy-based bidirectional RNN and CNN. In order to extract spatial appearance aspects and temporal order features of facial expression images, Liang *et al.* [19] suggested a network framework combining CNN and BiLSTM. Along with the development of FER, attention was introduced to FER because of its ability to better capture local regions. In FER, the attention focuses more on regional details of facial expression changes and filters out redundant information irrelevant to expression generation. Zhang *et al.* [37] proposed ECA-Resent for FER, using effective channel attention (ECA) to amplify the

weight of effective information and suppress the weight of invalid information. Sun *et al.* [33] combined feature attention weights of graphs and CNN to improve the accuracy of FER. Minaee *et al.* [26] used a deep learning method based on an attentional convolutional network that can focus on the regions most relevant to expressions, resulting in improved FER. Pei *et al.* [31] proposed an end-to-end spatially indexed attention model (SIAM) to extract valid potential appearance representations from CNN feature maps, which were then fed into the temporal attention layer constructed by LSTM to model temporal dynamics. Finally, the output feature vectors were weighted and averaged to improve the efficiency.

2.2 Extraction of Spatio-Temporal Features

FER methods can be classified into static image-based methods and dynamic sequence-based methods according to the feature representation [18]. In the static image-based methods, Zhao *et al.* [40] created a peak-piloted deep network (PPDN) to learn the association between peak expression images and non-peak expression images and capture their spatial appearance features. Yang *et al.* [34] designed a DeRL that can create neutral face images by training a face born on any input and learning the deposits (or residues) that remain in the middle of the generative model for FER. In the dynamic sequence-based methods, Liu *et al.* [21] proposed a facial expression recognition framework 3DCNN-DAP, which combined 3DCNN with deformable convolution to localize facial change regions and used a part-based representation for FER. Jung *et al.* [13] proposed DTAGN to capture temporal information of facial expression and automatically extract useful features from the raw data.

Both static image-based methods and dynamic sequence-based methods consider only one-sided facial expression features. To extract more efficient expression features and apply them to FER, more and more studies turn to the spatio-temporal properties of images for FER. Yu *et al.* [35] learned the spatio-temporal feature representation of facial expressions simultaneously by a DCPN network. Liang *et al.* [19] proposed to use CNN combined with BiLSTM to extract spatial features of each frame and dynamic temporal features of consecutive frames. Ryo Miyoshi *et al.* [27] proposed an enhanced convolutional long short-term memory (ConvLSTM) algorithm that could automatically recognize facial expressions in videos, which mainly used jump connections in spatio-temporal orientation and time gates to suppress gradient disappearance. Pan *et al.* [30] proposed a mainstream framework to fuse both spatial and temporal information to be utilized. The framework mainly consisted of CNN and LSTM. Jeong *et al.* [12] proposed a deep joint spatio-temporal feature recognition method for facial expressions. Firstly, the spatio-temporal features of facial expression images were extracted by 3DCNN. Then the whole facial signs are analyzed by using geometric network, and finally 23 facial sign points are selected to represent the dynamic muscle movements of the whole face. Zhu *et al.* [41] proposed a cascaded attentional facial expression recognition network with a pyramid structure and considering local spatial features, multi-scale-stereoscopic spatial context feature and temporal features to locate the changing features on dynamically changing regions (e.g., eyes, nose and mouth) as accurately as possible.

3 The Proposed Method

3.1 Overview

STANER is shown in Fig. 1. SGAER and TLER are the two key components of the approach. To begin, the facial key regions are cut from the input facial image sequence and input to TLER to extract high level temporal information. Then, the peak frame is chosen and fed into SGAER to learn spatial appearance features with an emphasis on subtle expression change features. At the end of the model, the recognition results from the two branches of the parallel structure are fused using decision-level fusion techniques, enabling the model to synthetically capture the spatio-temporal characteristics of subtle expression changes as they occur. The key branches of the proposed method are detailed in the following.

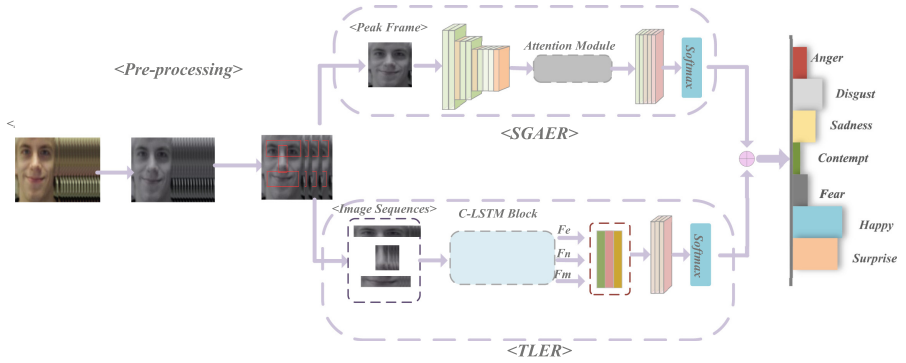


Fig. 1. Overview of the proposed method.

3.2 FER with Attention Module Based on Spatial Global Features

Subtle changes in distinct face regions that are linked together to make an expression are frequently used to create expressions. Capturing the intricacies of expressions is critical at FER. Therefore, SGAER is constructed to learn the nuances of expressions, in addition to quantifying the correlation of each position in the expression feature map and understanding the nuances of facial regions caused by expressions. The structure of the method is shown in Fig. 2.

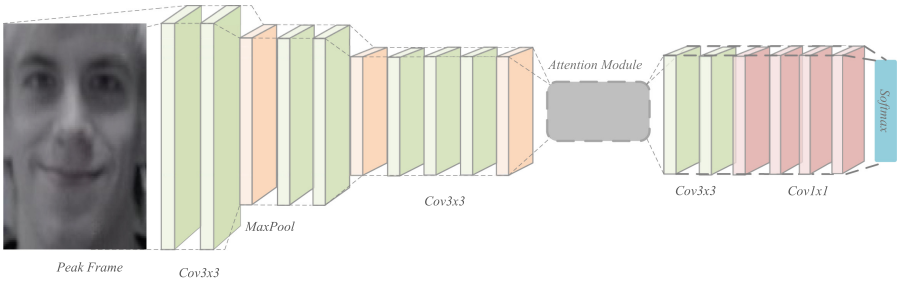


Fig. 2. Network structure of SGAER.

SGAER is composed of the front-end convolutional, the attention module and the end convolutional layer. The front-end convolution is made up of the first 10 layers of VGG-16. The structure of the constructed attention module is shown in Fig. 3. Shallow spatial appearance features are extracted from selected single-frame images through the front-end convolution, which is passed through the attention module to produce feature maps with attention. The processed feature maps are then transferred to the end convolution layer to obtain fine-grained appearance features. The attention is weighted in the range [0,1] using the *Sigmoid* function in the last layer of the attention module.

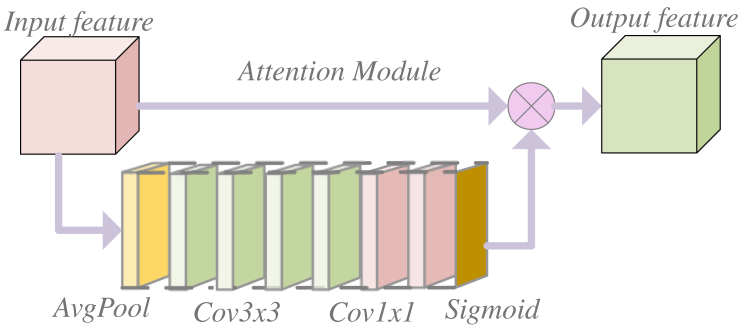


Fig. 3. Attention Module.The 3×3 convolutional kernel increases the received domain of feature information and acquires more spatial contextual information. The 1×1 convolutional kernel integrates multi-channel information into a single channel.

SGAER is designed in the following way. Firstly, the selected peak frame Ip is fed into the front-end convolution of SGAER to pick up the shallow spatial global features U of the expression image. Then, the spatial features at different locations after transmitting U to the attention module are given different weights and the main regional features generated by facial expressions are enhanced and

the expression features with subtle variations are amplified. Thus, the feature map M containing attention is formed.

$$M = \sigma\left\{Conv[w; AvgPool(U)]\right\} \tag{1}$$

where σ is the $Sigmoid$ activation function; $Conv$ represents the convolution operation; And w denotes the weight matrix; $AvgPool$ represents the global average pooling in the attention module.

Then, M is computed together with the originally obtained shallow spatial global features U to obtain the final feature map F.

$$F = (1 + M) \otimes U \tag{2}$$

where $\otimes$ denotes multiplication between elements.

F is sent to the end convolutional layer to extract the spatial appearance features F_G. And finally, F_G is input to the $SoftMax$ layer for final expression classification to generate expression classification result $P_G(C)$, where C is the number of facial expressions classification categories; x denotes the input vector of the $SoftMax$ function and $P_G(C)$ is defined as:

$$P_G(C) = S(x)_C = \frac{e^{x_C}}{\sum_i^c e^{x_i}} \tag{3}$$

where x_i denotes the computed output value of the ith category in the output vector, x_c denotes the current category output value to be computed, and the final loss function can be defined as:

$$Loss_S = -\sum_{i=1}^{C} y_i \ln\left(P_G(C)\right) \tag{4}$$

where y_i is the true value of the current facial expression.

3.3 FER with C-LSTM Based on Temporal Local Features

Since the generation of facial expressions is highly correlated with changes in only a few key regions of the face, TLER is created to enable the model to concentrate on learning the dynamic temporal changes in the facial regions associated with expression generation in consecutive frames. Figure 4 depicts the structure of TLER.

The TLER consists of a C-LSTM block and an end convolution. The shallow spatial features of the local sequences are extracted by the first few convolutional layers of the C-LSTM block, they are reconstructed into vectors and then passed through the LSTM to obtain the dynamic change information of the local sequences. Finally, the high-level semantic information of the expression features is learned by the end convolution. LSTM is proved to recover the temporal features of the expression sequence, and it is composed of an input gate, an output gate and a forget gate. Among them, the input gate determines which values the

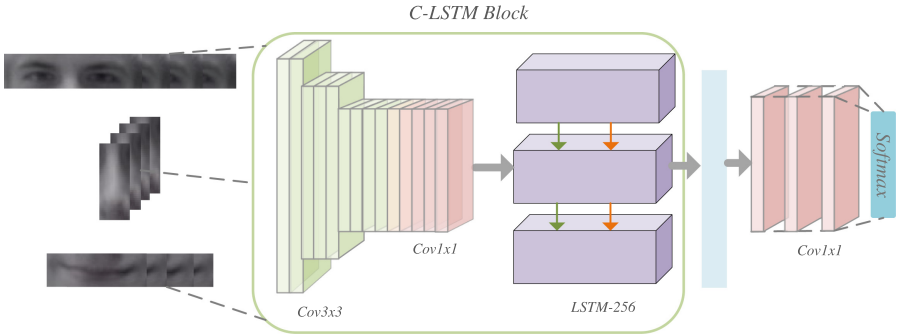


Fig. 4. Network structure of TLER.

unit will update; the output gate determines which values will be finally output and the forget gate defines which information the unit will discard. Where forget gate is defined as f_t:

$$f_t = \sigma\left(W_f x_t + W_f h_{t-1} + b_f\right) \tag{5}$$

input gate is defined as i_t:

$$i_t = \sigma\left(W_i x_t + W_i h_{t-1} + b_i\right) \tag{6}$$

$$g_t = \tanh\left(W_g x_t + W_g h_{t-1} + b_g\right) \tag{7}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{8}$$

output gate is defined as o_t:

$$o_t = \sigma\left(W_o x_t + W_o h_{t-1} + b_o\right) \tag{9}$$

$$h_t = o_t \otimes \tanh\left(c_t\right) \tag{10}$$

where h_{t-1} is the LSTM hidden layer output at moment $t-1$; x_t is the input vector. And c_t is the current update cell. g_t is the current alternative update cell, which contains all the update information of the current time node. W and b denote the weight matrix and bias value, respectively. σ is *Sigmoid* activation function and *tanh* is the hyperbolic tangent activation function; $\otimes$ denotes multiplication between elements.

The input to TLER is the extracted consecutive frames I_e,I_n, I_m of the three key facial regions of eyes, nose and mouth associated with the generated facial expressions. The processed local region image sequences are fed into the constructed C-LSTM block. Region-based low-level features are first generated by the CNN of the C-LSTM block and then the fine-grained dynamic temporal

features F_e, F_n, F_m of these local sequences are captured using the LSTM. At the end of the C-LSTM, F_e, F_n, F_m are connected to represent the global temporal features F':

$$F' = Concatenation\,(F_e : F_n : F_m) \tag{11}$$

F' is input to the end convolution layer to form the high-level local feature F_L. And finally, F_L is fed to the $Softmax$ layer for the final expression classification.

The result is obtained as $P_L(C)$, where C is the number of facial expression classification categories; x denotes the input vector of the $SoftMax$ function and $P_L(C)$ is defined as:

$$P_L(C) = S(x)_C = \frac{e^{x_C}}{\sum_i^c e^{x_i}} \tag{12}$$

where x_i denotes the computed output value of the ith category in the output vector; x_c denotes the current category output value to be computed.

The final loss function can be defined as:

$$Loss_T = -\sum_{i=1}^{C} y_i \ln\left(P_L(C)\right) \tag{13}$$

where y_i is the true value of the current facial expression.

Researchers have found that combining multiple networks for FER can yield more diverse information to ensure complementarity of features, often resulting in better recognition than individual networks. In the process of expression change, the overall spatial appearance features and local dynamic change features are equally important. In order to make the constructed model take into account the local dynamic change features while utilizing the spatial appearance change of expressions and without emphasizing or ignoring one feature, a simple average decision fusion method is used in the paper to fuse the expression classification results of SGAER and TLER, which is calculated as follows.

$$O(C) = argmax\left(\alpha P_G(C) + (1-\alpha)P_L(C)\right) \tag{14}$$

where α is 0.5. O represents the final output category of facial expression.

The loss function $Loss$ of the whole network is:

$$Loss = Loss_G + Loss_T \tag{15}$$

in which, $Loss_G$ and $Loss_T$ represent the final loss functions of SGAER and TLER, separately.

3.4 Algorithmic Description

The algorithm of STANER is described as follows:

Algorithm 1: STANER for FER

Require: Dataset D;
Output : Facial expression category: O

for *images in dataset* D **do**
    face cropping, grayscale processing and data augmentation obtain dataset D^*
end
for *images in dataset* D^* **do**
    Selecting peak frame: I_p and partial Sequences: I_e, I_n, I_m
end
for $I_p \in D^*$ **do**
    $P_G(C) \leftarrow$ SGAER (I_p);
    $Calculate \rightarrow Loss_G = -\sum_{i=1}^{C} y_i \ln (P_G(C))$
end
for I_e, I_n, $I_m \in D^*$ **do**
    $P_L\ (C) \leftarrow$ TLER (I_e, I_n, I_m);
    $Calculate \rightarrow Loss_L = -\sum_{i=1}^{C} y_i \ln (P_L(C))$
end
for $P_G(C)$,$P_L(C)$ **do**
    $Calculate$: $O(C)$=argmax($\alpha P_G(C)$+(1-α) $P_L(C)$)
end
$Calculate$: $Loss$=$Loss_T$+$Loss_G$
Output:O

4 Experiments

This section first describes the datasets and data preprocessing, followed by the implementation details of the proposed method and finally the experimental results and analysis.

4.1 Datasets

Two facial expression datasets named CK+ [23] and Oulu-CASIA [38] are used to evaluate this model.

The extended Cohn-Kanada database (CK+) [23] is made up of 593 sequences from 123 distinct people. Anger (An), disgust (Di), fear (Fe), happy (Ha), sadness (Sa) and surprise (Su) are the six basic emotion categories in this dataset. In addition, there is a unique term known as "contempt." Only 327 sequences out of 118 patients were tagged with seven expressions and they all started with a neutral expression and finished with a peak expression. The training and test sets are built using 10-fold cross-validation.

Oulu-CASIA [38] is more complex than CK+ and contains 480 image sequences recorded by 80 individuals under normal lighting conditions, with six basic emotion categories. Each of the six expressions has a sequence, each starting with a neutral expression and ending with a peak expression. The dataset is processed using the same 10-fold cross-validation.

Due to the fact that the experimental dataset is acquired from a real environment, all images in the CK+ and Oulu-CASIA datasets are first cropped and grayscale processed in order to prevent the effect of other expression-independent noise such as jewelry, lighting and color on the FER [18]. Then, the processed dataset is horizontally flipped to enhance the data to avoid the overfitting problem caused by the limited number of training samples. Finally, the peak frame of the facial expression image sequence is selected as the input of SGAER in the new dataset that had been processed. The 68 facial marker points of the face are identified using the Dlib [16], the eyes, nose and mouth regions are extracted. The consecutive frames of the local region are used as the input of TLER.

4.2 Evaluation Metrics and Implementation Details

The recognition accuracy *Acc* of expression classification is used to evaluate the model's overall classification capabilities. which is defined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

where FP represents false positive, FN represents false negative, TP represents true positive, TN represents true negative.

The parameter selection for the experiment is as follows:

The number of training networks is set to 150 epochs, and the training batch size is set to 64. The initial learning rate is 10^{-4}. The first 100 epochs remain the same and the learning rate is set to 10^{-5} after 100 epochs.

4.3 Ablation Experiments

Table 1 illustrates the *Acc* of ablation experiments on CK+ and Oulu-CASIA datasets. The result of C1 is to extract low-level features of the face from a single frame image using the first ten layers of VGG-16 and then input these features into the end convolutional layer to learn advanced feature information for FER. C2 is a FER that extracts spatio-temporal features of local facial sequences using only TLER. C3 is a FER that uses only SGAER. C4 is the fusion model of C1 and C2. STANER is the method proposed in this work, which utilizes both static-based and dynamic-based methods to capture the spatio-temporal features of the face for FER.

The experimental results of C1, C2 and C4 indicate that using only static image-based methods and dynamic sequence-based methods does not lead to good FER performance because they tend to consider only one-sided expression features. The fusion models of static image-based methods and dynamic

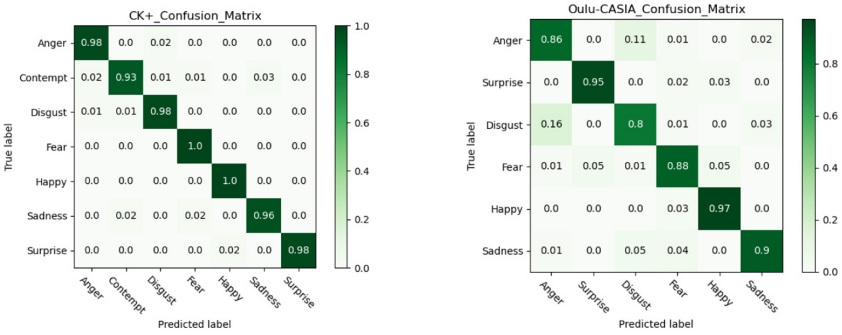
Table 1. *ACC* of ablation experiments on CK+ and Oulu-CASIA

Methods	CK+	Oulu-CASIA
C1	91.64%	79.40%
C2	92.70%	80.23%
C3	94.52%	82.66%
C4	96.19%	86.50%
STANER(Our method)	98.23%	89.52%

sequence-based methods can obtain complete cues of face expression changes and capture spatio-temporal features. The application of attention increases the performance of FER in both the static image-based methods and the fusion model, as seen by the comparison of C1, C3, C4 and STANER recognition results. This demonstrates that FER benefits from focusing on subtle expression change features, and that using attention not only allows the model to better explore the spatial appearance features of the original image, but also allows the model to capture more fine-grained high-level semantic features, greatly improving the model's proactivity.

4.4 Confusion Matrix

To measure the specific manifestation of STANER on each expression category, the confusion matrices of STANER on CK+ and Oulu-CASIA datasets are shown in Fig. 5(a) and (b). Figure 5(a) demonstrates that STANER is more accurate in recognizing disgust, fear, happy on CK+ and does well with contempt, which is not easily distinguished. Figure 5(b) illustrates that STANER not only has the highest recognition rate for surprise and happy on Oulu-CASIA, but also discriminates fear and anger, two frequently confused expression categories.



(a) (b)

Fig. 5. The confusion matrix. (a) The confusion matrix on CK+. (b) The confusion matrix on Oulu-CASIA.

4.5 Comparisons with Existing Methods

Tables 2 and 3 show the performance comparison of STANER with existing mainstream methods on CK+ and Oulu-CASIA, respectively. It can be seen that STANER exhibits better properties than most methods.

Table 2. Comparison of FER *Acc* on CK+ Dataset

Methods	Experiment Setting	*ACC*(%)
3DCNN-DAP [21]	10 folds	92.40%
ECA-ResNet [37]	10 folds	94.58%
DTAGN [13]	10 folds	97.25%
PPDN [40]	10 folds	99.30%
DeRL [34]	10 folds	97.30%
PHRNN-MSCNN [36]	10 folds	98.50%
Our method	10 folds	98.23%

Table 3. Comparison of FER *Acc* on Oulu-CASIA Dataset

Methods	Experiment setting	*ACC*(%)
ResNeXt-50 + pyramid + cascaded attention block + GRU [41]	10 folds	89.29%
DTAGN [13]	10 folds	81.46%
PPDN [40]	10 folds	84.59%
DeRL [34]	10 folds	88.00%
DCPN [35]	10 folds	86.23%
PHRNN-MSCNN [36]	10 folds	86.25%
Our method	10 folds	89.52%

Firstly, the comparison between STANER and the state-of-the-art methods on the CK+ dataset is shown in Table 2. The average accuracy of STANER is 98.23%. STANER does not perform as well as the static-based method PPDN [40], which only considers six expression categories in CK+ dataset and does not consider the expression contempt while STANER considers all expression categories in CK+. STANER performs slightly lower than PHRNN-MSCNN [36] on the CK+ dataset, where STANER only considers the appearance change feature of facial expressions, while both geometric and appearance information of expressions are utilized in PHRNN-MSCNN. Compared to the dynamic sequence-based method 3DCNN-DAP [13], STANER improved the accuracy by 4.03% in CK+, which led to the demonstration of the effectiveness of using spatial-temporal features in the FER task. STANER also improved in CK+ compared to the static image-based method using attention, ECA-Resnet [37], because it did not focus

only on single-sided facial expression features, but also considered information about the temporal variation of expressions.

Then, the comparison of STANER with other methods on the Oulu-CASIA dataset is shown in Table 3. The average accuracy of STANER is 89.52%. The accuracy of STANER is improved by 4.93% compared to that of PPDN [40] using a static image-based method. STANER is compared with the approach that exploits the spatio-temporal information of facial expression images. Compared with the DTAGN [13], STANER's recognition accuracy on Oulu-CASIA dataset is improved by 8.06%. STANER improves the recognition performance by 3.29% over the DCPN designed by Yu *et al.* [35] on Oulu-CASIA. It improved 3.27% over the PHRNN-MSCNN model proposed by Zhang *et al.* [36].STANER outperforms the Oulu-CASIA dataset compared to a cascaded attentional facial expression recognition network with a pyramidal structure that simultaneously considers local spatial features,multi-scale-stereoscopic spatial context feature and temporal features, as presented by Zhu *et al.* [41].

5 Conclusion and Future Work

In this study, STANER is designed to improve the performance of facial expression recognition when facial expressions change subtly. STANER can not only make full use of the spatio-temporal change information of incoming expressions, but also localize key parts of the face by attention to better utilize the features generated when facial expressions change subtly.

Specifically, SGAER is firstly constructed in order to learn the spatial appearance information of facial expressions when they change, and the attention module in SGAER can more precisely localize specific regions with significant dynamic changes. Secondly, TLER is constructed to extract temporal features from key local facial parts. For the input expression image sequences, the main parts related to facial expression changes (i.e., image sequences of eyes, nose, and mouth) are cropped, and temporal features are extracted using C-LSTM. Finally, SGAR and TLER are further fused using an average decision level strategy to obtain different expressions for recognition. Through extensive experiments, the excellent recognition properties of STANER on CK+ and Oulu-CASIA are demonstrated.

To apply STANER to more complex, realistic and natural environments, future work will include: (1) training and validating the model in complex natural environment datasets to further improve the accuracy and robustness of the model; (2) using a more accurate attention mechanism to capture changes in the appearance of key parts of the face during the spatial feature extraction stage; (3) using the model to design a real-time facial expression recognition system to improve the human-computer interaction capability of the model; and (4) improving the model network structure to reduce computation and save resource consumption.

Acknowledgements. This work is supported by Shandong Provincial Project of Graduate Education Quality Improvement (No. SDYJG21104, No. SDYJG19171, No. SDYY18058), the OMO Course Group "Advanced Computer Networks" of Shandong Normal University, the Teaching Team Project of Shandong Normal University, Teaching Research Project of Shandong Normal University (2018Z29), Provincial Research Project of Education and Teaching (No.2020JXY012), the Natural Science Foundation of Shandong Province (No. ZR2020LZH008, ZR2021MF118, ZR2019MF071).

References

1. Chen, L., Zhou, M., Su, W., Wu, M., She, J., Hirota, K.: Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Inf. Sci. **428**, 49–61 (2018)
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
3. Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., Yang, G.: cGAN based facial expression recognition for human-robot interaction. IEEE Access **7**, 9848–9859 (2019)
4. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2625–2634 (2015)
5. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124-129 (1971)
6. Graves, A., Mohamed, A.R., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645–6649. IEEE (2013)
7. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. **6**(1), 1–12 (2014)
8. Happy, S., Routray, A.: Robust facial expression classification using shape and appearance features. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), pp. 1–5. IEEE (2015)
9. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Ilyas, C.M.A., Haque, M.A., Rehm, M., Nasrollahi, K., Moeslund, T.B.: Facial expression recognition for traumatic brain injured patients. In: International Conference on Computer Vision Theory and Applications, pp. 522–530. SCITEPRESS Digital Library (2018)
12. Jeong, D., Kim, B.G., Dong, S.Y.: Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition. Sensors **20**(7), 1936 (2020)
13. Jung, H., Lee, S., Yim, J., Park, S., Kim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2983–2991 (2015)
14. Khowaja, S.A., Dahri, K., Kumbhar, M.A., Soomro, A.M.: Facial expression recognition using two-tier classification and its application to smart home automation system. In: 2015 International Conference on Emerging Technologies (ICET), pp. 1–6. IEEE (2015)

15. Kim, B.K., Lee, H., Roh, J., Lee, S.Y.: Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 427–434 (2015)
16. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
18. Li, S., Deng, W.: Deep facial expression recognition: a survey. In: IEEE Transactions on Affective Computing (2020)
19. Liang, D., Liang, H., Yu, Z., Zhang, Y.: Deep convolutional BiLSTM fusion network for facial expression recognition. Vis. Comput. **36**(3), 499–508 (2020)
20. Liu, K., Zhang, M., Pan, Z.: Facial expression recognition with CNN ensemble. In: 2016 International Conference on Cyberworlds (CW), pp. 163–166. IEEE (2016)
21. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
22. Liu, Y., Wang, J., Li, P.: A feature point tracking method based on the combination of SIFT algorithm and KLT matching algorithm. J. Astronautics **32**(7), 1618–1625 (2011)
23. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: 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. IEEE (2010)
24. Majumder, A., Behera, L., Subramanian, V.K.: Automatic facial expression recognition system using deep network-based data fusion. IEEE Trans. Cybern. **48**(1), 103–114 (2016)
25. Matsumoto, D.: More evidence for the universality of a contempt expression. Motiv. Emot. **16**(4), 363–368 (1992)
26. Minaee, S., Minaei, M., Abdolrashidi, A.: Deep-emotion: facial expression recognition using attentional convolutional network. Sensors **21**(9), 3046 (2021)
27. Miyoshi, R., Nagata, N., Hashimoto, M.: Enhanced convolutional LSTM with spatial and temporal skip connections and temporal gates for facial expression recognition from video. Neural Comput. Appl. **33**(13), 7381–7392 (2021)
28. Mollahosseini, A., Chan, D., Mahoor, M.H.: 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)
29. Pan, X.: Fusing hog and convolutional neural network spatial-temporal features for video-based facial expression recognition. IET Image Proc. **14**(1), 176–182 (2020)
30. Pan, X., Ying, G., Chen, G., Li, H., Li, W.: A deep spatial and temporal aggregation framework for video-based facial expression recognition. IEEE Access **7**, 48807–48815 (2019)
31. Pei, W., Dibeklioğlu, H., Baltrušaitis, T., Tax, D.M.: Attended end-to-end architecture for age estimation from facial expression videos. IEEE Trans. Image Process. **29**, 1972–1984 (2019)
32. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
33. Sun, W., Zhao, H., Jin, Z.: A visual attention based ROI detection method for facial expression recognition. Neurocomputing **296**, 12–22 (2018)

34. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2168–2177 (2018)
35. Yu, Z., Liu, Q., Liu, G.: Deeper cascaded peak-piloted network for weak expression recognition. Vis. Comput. **34**(12), 1691–1699 (2018)
36. Zhang, K., Huang, Y., Du, Y., Wang, L.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
37. Zhang, P., Liu, Y., Hao, Y., Liu, J.: Deep facial expression recognition algorithm combining channel attention. In: 2021 4th International Conference on Artificial Intelligence and Pattern Recognition, pp. 260–265 (2021)
38. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
39. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
40. Zhao, X., et al.: Peak-piloted deep network for facial expression recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 425–442. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_27
41. Zhu, X., He, Z., Zhao, L., Dai, Z., Yang, Q.: A cascade attention based facial expression recognition network by fusing multi-scale spatio-temporal features. Sensors **22**(4), 1350 (2022)