Deep Spatio-Temporal Decision Fusion Network for Facial Expression Recognition

Xuanchi Chen[1,2], Heng Yang[1], Xia Zhang[3], Xiangwei Zheng[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
liww72@sdnu.edu.cn

Abstract. Facial expression recognition (FER) becomes research focus in affective computing, as it already plays an important role in public security application scenarios such as urban safety management and safety driving assistance systems. Modeling the spatiotemporal information of facial expression sequences in a targeted manner, integrating and utilizing them appropriately is challenging. In this paper, a facial expression recognition method based on spatial-temporal decision fusion network (STDFN) is proposed. Firstly, the facial expression sequences are divided into four sub-sequences according to face regions, and BiLSTM are used for each of sub-sequences to extract local temporal features. The local morphological features of facial expressions can be captured in more detail to maximize the utilization of the temporal features of dynamic facial expressions. Then, VGG19 is utilized to extract the shallow spatial features of peak expression frame, and the channel weights of spatial features is assigned by squeeze-and-excitation module to attain the weighted spatial features. This allows valid spatial features to be purposefully retained to avoid overfitting. Finally, temporal features and spatial features are used separately calculating expression classification results. And a decision-level fusion module is designed to fuse the two results to obtain the final FER result. Extensive experimental results demonstrate that on three FER datasets CK+, Oulu-CASIA and MMI, achieves 98.83%, 89.31% and 82.86% accuracy, which proved that STDFN effectively improved the recognition accuracy of FER.

Keywords: FER · BiLSTM · SENet · Decision fusion

1 Introduction

Facial expressions, as the most commonly used fashion in human affective interaction, are the key factor for machines to perceive human emotions. With the exponential increase of computer computing power, facial expression recognition (FER) has gradually become a highlight in human-computer interaction and is widely used in areas such

Y. Xu et al. (Eds.): ML4CS 2022, LNCS 13657, pp. 106–120, 2023.
https://doi.org/10.1007/978-3-031-20102-8_9

as urban safety management, criminal investigation assistance and safety driving assistance [1, 2]. Due to previous ground-breaking research, FER can be classified into two types: static image-based approach that focuses more on spatial features and dynamic sequence-based approach that focuses more on temporal features [3]. The static image-based FER methods can effectively extract spatial information, but cannot model the dynamic information of facial expressions. Dynamic sequence-based methods extract temporal features by capturing the evolution of facial expressions and usually achieve good performance. However, FER is a rather complex facial analysis task, and even humans are unable to recognize the emotions of others by focusing only on a single facial feature information. It is difficult to improve the performance of FER based on spatial or temporal features only. Hence, a key challenge for FER is to extract the spatiotemporal features in a targeted manner and integrate the information of both complementarily.

To this end, a FER method based on spatial-temporal decision fusion network (STDFN) is proposed. In order to get the utmost out of the temporal and physical features of dynamic facial expression, we designed temporal feature extraction module to capture the expression evolution information between frames. The bi-directional long short-term memory (BiLSTM) [19] is used to learn the mode between frames in two directions, pay more attention to the context relationship between frames to obtain more accurate temporal information. In addition, the occurrence of facial expressions is often coupled with dynamic changes in the combination of facial parts [4]. For capture the morphological features of facial expressions in a more detailed way, our model first divides each frame of image sequence into four parts and produces four groups of sub-sequences. Then, using sub-sequence training four groups of BiLSTM to obtain local temporal features, and spliced to globe temporal features. At the same time, we designed spatial feature extraction module based on VGG [18] and SENet [5] (VGG-S). The shallow spatial feature maps of peak expression images are extracted by VGG19, channel weights are assigned to the shallow spatial feature maps using SENet, the obtained weighted feature maps are used to calculate the classification results of face expressions. Finally, an adaptive decision fusion module is designed to integrate the expression classification results of the two modules to obtain the final face expression classification results.

The main works of this paper are as follows:

(1) We peoposed FER method based on spatial-temporal decision fusion network, temporal and spatial feature extraction module were designed separately to capture spatiotemporal information, and an adaptive decision fusion module to integrate spatiotemporal information.
(2) We designed temporal feature extraction module based on BiLSTM, with divided facial landmarks into different areas to learn the details of facial expression evolution, which effectively utilizes the information before and after expression changes in various facial regions;
(3) We designed spatial feature extraction module based on VGG19-S, adopting VGG19 extract shallow features maps and SENet assign weight to shallow feature channels, which is helpful for accurate collection of useful spatial information.
(4) Numerous experiments have shown that STDFN achieved 98.83%, 89.31% and 82.86% accuracy on three FER datasets include CK+, Oulu-CASIA and MMI, demonstrating the effectiveness of the proposed method.

The rest of this paper is organized as follows. Section 2 presents the research progress of FER methods. Section 3 describes STDFN in detail, and Sect. 4 presents our experimental results. Finally, in Sect. 5 we summarize the conclusions and point out the direction of our future work.

2 Related Works

2.1 Facial Expression Recognition

As an essential task in computer vision, early FER methods extracted features by handcrafted and designed classifiers to classify the features to obtain expression recognition results. Klaser et al. [6] extended the traditional hand-designed feature approach by designing a local descriptor to model the temporal information contained in facial expression sequences, and effectively improves FER performance. Liu et al. [7] proposed STM-ExpLet for streaming modeling of videos, which models each video clip as a spatiotemporal streaming module (STM) to improve feature discrimination. However, most handcrafted features are susceptible to the external environment and has poor characterization capability. With the rapid development of deep learning algorithms, deep neural networks can draw on large amounts of data to learn the required features autonomously, effectively bypassing complex manual feature extraction.

2.2 Deep FER Based on Spatio-Temporal Features

The occurrence of facial expressions as a continuous human action possesses an innate temporal correlation, so researchers have proposed deep spatiotemporal networks based on expression sequences to model the more capable spatiotemporal features. Jung et al. [8] proposed a FER approach containing two sub-networks, which were used to extract the appearance features and temporal geometric features, and features were integrated for the expression recognition task, which effectively improved the expression recognition accuracy. Zhang et al. [9] proposed a temporal-spatial recurrent neural network to effectively improve the accuracy of FER tasks by building multiple RNNs to capture spatiotemporal information with high discriminative power by scanning from different angles. Zhao et al. [10] proposed a peak-piloted deep network (PPDN) for learning the evolutionary information between peak expressions and non-peak expressions, which improves models' generalization ability when facial differences between individuals are larger. However, existing methods do not capture the dynamic information of critical areas of the face in a targeted manner, and most of them are commonly based on the spatiotemporal information of facial expressions at the image level, which results in increased computational complexity of models and vulnerability to noise during image transmission.

3 The Proposed Method

Figure 1 shows the framework of our proposed approach. Our model starts from the input of the original image sequence and preprocesses the data first. To minimize the

complexity of the calculation, 16 key frames are selected from original sequences to represent the dynamic evolution of the whole sequence and these 16 images contain the first and last frames. For the sequences with insufficient frames, the last frame is used to fill backward until they are sufficient. Then, facial clipping is performed for the key frames, and the gray processing is to avoid the influence of light and facial color on the classification results. To avoid overfitting due to the small amount of data, rotation and flipping are used to expanding the dataset. The sequence data is divided into two modules, one is directly used to extract temporal information and the other selects a peak expression frame (the last frame) from the sequence data to extract spatial information. The two modules are used to predict facial expression classification respectively. In the end, classification by integrating the predicted values from two modules. In the following, we will introduce the details of the two modules and the method of decision fusion.

3.1 Temporal Feature Extraction Based on BiLSTM

The occurrence of facial expressions is coupled with the dynamic evolutionary process of key facial parts. For instance, happiness can be expressed as the corner of the mouth up and eyebrows open; sadness is the corner of the mouth down, eyes smaller; surprise is characterized that the eyes become larger and mouth opens. In fact, facial landmarks can well reflect these dynamic evolution processes. Therefore, the facial landmarks are divided into four parts for model training to attain four local features, which is employed to focus more on a particular part of the face over time. Local features of the four parts are fused to ensure the integrity of the extracted facial expression information. Then, BiLSTM is adopted to learn temporal feature, which can not only effectively avoid the gradient disappearance in long-term learning, but effectively capture the context information between image frames.

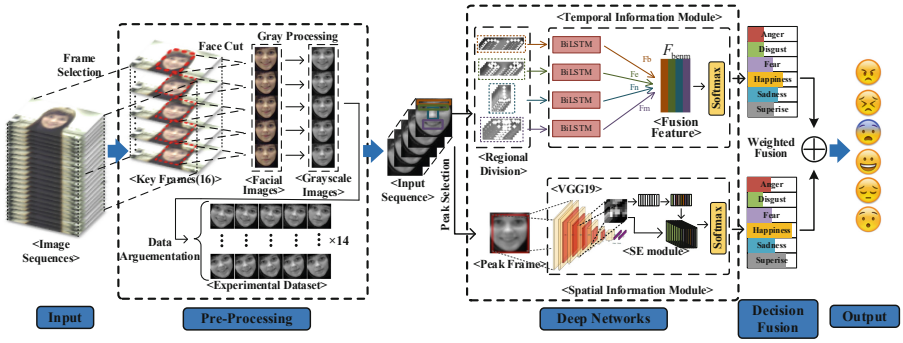


Fig. 1. Framework of spatial-temporal decision fusion network.

As shown in Fig. 2, the evolution process between image frames is abstracted as the coordinate change process of facial landmarks. First of all, the input layer of the network maps the facial landmark points of each frame into a one-dimensional vector, so that each facial region is represented as a matrix, which is respectively used as the input of the corresponding BiLSTM. Four partial local features are extracted by the

corresponding BiLSTM and the local features are stitched and fused to obtain the global features. Finally, Softmax is used as the classification layer.

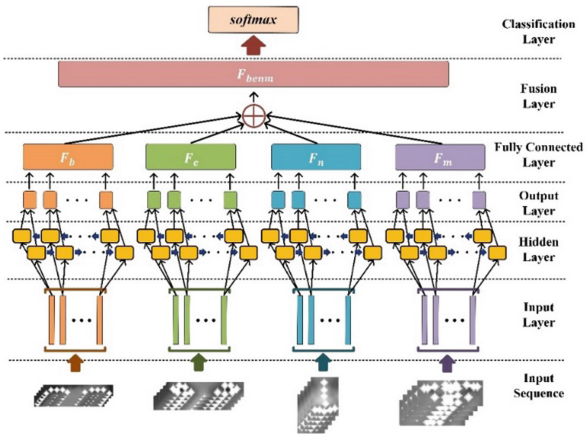


Fig. 2. BiLSTM structure for extracting temporal information.

BiLSTM extends the single-way LSTM by introducing a second layer, in which the connections between the hidden layer flow in reverse chronological order, exploiting the temporal information of "past" and "future". In our model, the original images are divided into four parts, including eyebrows, eyes, nose, and mouth, and they are represented as one-dimensional vectors as the input of BiLSTM. Partial based local features F_b, F_e, F_n, F_m are obtained from the output layer. Then the local features are combined by Formula (1), which obtain the global features F_{benm}:

$$F_{benm} = [F_b \oplus F_e \oplus F_n \oplus F_m] \tag{1}$$

where $F_b, F_e, F_n, and \ F_m$ are the features of eyebrows, eyes, nose and mouth, respectively, and $\oplus$ denotes the *concat* operation

Take F_b as an example to explain the hidden layer process of BiLSTM, the formula of forward propagation cell structure in BiLSTM is as Formula (2–7):

$$f_{bt} = \sigma\left[w_{bf}\left(h_{bt-1}, x_{bt}\right) + b_{bf}\right] \tag{2}$$

$$i_{bt} = \sigma\left[w_{bi}(h_{bt-1}, x_{bt}) + b_{bi}\right] \tag{3}$$

$$\hat{c}_{bt} = tanh(w_{b\hat{c}}(h_{bt-1}, x_{bt}) + b_{b\hat{c}}) \tag{4}$$

$$c_{bt} = \hat{c}_{bt} * i_{bt} + f_{bt} * c_{bt-1} \tag{5}$$

$$o_{bt} = \sigma\left[w_{bo}(h_{bt-1}, x_{bt}) + b_{bo}\right] \tag{6}$$

$$\overset{\frown}{h_{bt}} = o_{bt} * tanh(c_{bt}) \tag{7}$$

where, t denotes the number of frames, w denotes the weight matrix, x_{bt} is the input vector, f_{bt} is the oblivion gate, which decides to discard the information in the previous state. i_{bt} is the input gate, which determines the information currently to be retained. σ denotes the *sigmoid* activation function, *tanh* denotes the hyperbolic tangent activation function. $\hat{c}_{bt}$ expresses the alternative update units, and c_{bt} is the updated cell state, which multiplied by the old cell state and forgetting gate and added to the new candidate values. c_{bt} uses an *sigmoid* layer to determine the output of updated cell state. The cell states are processed by *tanh* and the multiplied with the output of the sigmoid layer to obtain $\overrightarrow{h_{bt}}$, where $\overrightarrow{h_{bt}}$ denotes the embedding representation of the forward-LSTM. The backward calculation process is similar to the forward process, and $\overleftarrow{h_{bt}}$ denotes the embedding representation of the backward-LSTM.

Then, the forward and backward outputs of the BiLSTM are combined by Formula (8–10):

$$\overrightarrow{h_{bt}} = \overrightarrow{LSTM}\left(h_{bt-1}, w_{bt}, c_{bt-1}\right) \tag{8}$$

$$\overleftarrow{h_{bt}} = \overleftarrow{LSTM}\left(h_{bt+1}, w_{bt}, c_{bt+1}\right) \tag{9}$$

$$F_b = h_{bt} = \overrightarrow{h_{bt}} \oplus \overleftarrow{h_{bt}} \tag{10}$$

where $\overrightarrow{LSTM}$ represents the forward LSTM and $\overleftarrow{LSTM}$ represents the backward LSTM. h_{bt} denotes the BiLSTM hidden layer output and is taken as F_b.

Finally, global feature F_{benm} is fed to *Softmax* for estimate the facial expression. The calculation process of *Softmax* is shown in Formula (11).

$$y_i = S(Z)_k = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \tag{11}$$

Among them, y_i represents the calculated softmax value and $k \in [1, K]$, K denotes the number of expression categories, Z represents the output of the previous layer.

The loss function uses cross entropy calculated by Formula (12), where $T_k \in [0, 1]$.

$$Loss = -\sum_{k=1}^{K} T_k \ \ln \ y_k \tag{12}$$

3.2 Spatial Feature Extraction Based on VGG19-S

To extract the spatial information of facial expressions, the peak frame from the sequence is selected as the input of VGG19-S. The shallow convolutional layer of pre-trained VGG19 is used to extract shallow features, and then SENet is employed to learn the channel weights of the shallow features, assign the weights to the shallow features to obtain a weighted feature map, and use Softmax for classification.

As shown in Fig. 3, the structure of the spatial feature extraction model VGG19-S is illustrated. The input dimension is 64 × 64 × 1 gray image, and the shallow features are obtained after convolution operations with two 3 × 3 × 64 layers, two 3 × 3 × 128 layers, four 3 × 3 × 256 layers and one 3 × 3 × 512 layer. Next, SENet is used to explicitly model the interdependence among feature channels, and the significance degree of each one is automatically obtained through learning. Then, according to this importance degree, channels with higher scores are promoted and those channels with lower scores are suppressed. With squeeze operation, feature compression is performed and each 2D feature channel is turned into a 1D constant, which has a global receptive domain. This is achieved by using the global average pool to generate channel statistics. For a shallow feature U, each channel has a spatial dimension $m \times n$, and a statistic V is generated for each channel, such that the c th element of V is calculated with Formula (13).

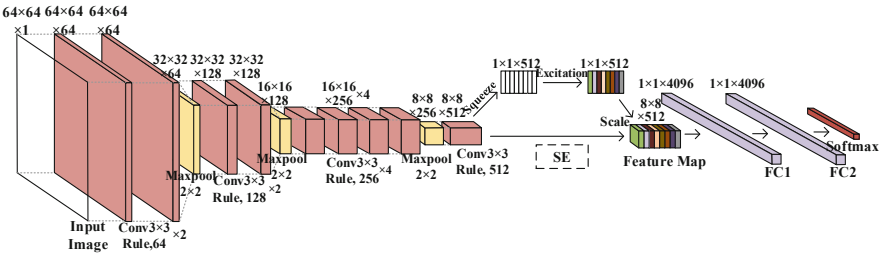


Fig. 3. VGG19-S structure for extracting spatial information.

$$V_C = Squeeze(U) = \frac{1}{m \times n} \sum_{a=1}^{m} \sum_{b=1}^{n} u_c(a, b) \tag{13}$$

Next, the *Excitation* is required to fully capture the channel dependencies. Weights are assigned to each feature channel by w to simulate channel correlation information. The calculation process as Formula (14).

$$S = Excitation(V, w) = \sigma(w_2 \delta(w_1 V)) \tag{14}$$

where δ denotes the ReLU function, and $w_1 \in R^{\frac{C}{r} \times C}$, $w_2 \in R^{C \times \frac{C}{r}}$, r is a hyperparameter, usually $r = 16$.

Finally, S is treated as the significance of each feature channel. The weighted result f_c of the feature channel c is calculated as Formula (15).

$$f_c = Scale(U_c, S_c) = S_c \cdot U_c \tag{15}$$

$$F_S = \{f_1, f_2, \ldots, f_c\} \tag{16}$$

where $Scale(U_c, S_c)$ denotes the channel-wise multiplication between S_c and U_c, and F_S denotes the spatial features of the peak frame.

3.3 Weighted Decision Fusion

For sequence images of facial expressions, there are two kinds of information: temporal domain and spatial domain. We use the two kinds of information to analyze facial expressions respectively. However, better results are often obtained by integrating the two kinds of information. A weighted fusion method of decision levels is used to integrate the two dimensions. In order to facilitate fusion, *Softmax* is used for classification of the two networks, $P_T(0 \leq P_T(k) \leq 1) = [P_T(1), P_T(2), \ldots, P_T(K)]$ represents the classification of the temporal network and $P_S(0 \leq P_S(k) \leq 1) = [P_S(1), P_S(2), \ldots, P_S(K)]$ represents the classification of the spatial network. The prediction results are calculated by Formula (17). Where α is the parameter.

$$Prediction = argmax(\alpha P_T(k) + (1 - \alpha) P_S(k)) \tag{17}$$

3.4 Algorithmic Description

The description of STDFN is shown in Algorithm 1.

Algorithm 1. STDFN(*Lb, Lb, Ln, Lm, Peak, label, λ, α, num_cell, r*)

Input : Landmark points set $\{Lb|0 \le Lb \le BS\}$, $\{Le|0 \le Le \le ES\}$, $\{Ln|0 \le Ln \le NS\}$, $\{Lm|0 \le Lm \le MS\}$; Peak frame image set $\{Peak|0 \le Peak \le S\}$; Lable set $\{lable|0 \le lable \le S\}$; Larning rate λ; Fusion weight α; Number of LSTM neurons num_cell; Hyperparameter r.

Output: Prediction result set $Prediction\{\}$.

Initialization: Initialize the parameters to be learned in the STDFN model.

Process:

For i =1 to I

  Temporal features are extracted by BiLSTM:

  $Lb^{i*} = \left(Lb_1^{i*}, Lb_2^{i*}, \dots, Lb_B^{i*}\right)$

  $Le^{i*} = \left(Le_1^{i*}, Le_2^{i*}, \dots, Le_E^{i*}\right)$

  $Ln^{i*} = \left(Ln_1^{i*}, Ln_2^{i*}, \dots, Ln_N^{i*}\right)$

  $Lm^{i*} = \left(Lm_1^{i*}, Lm_2^{i*}, \dots, Lm_M^{i*}\right)$

  $\left\{F_b^i, F_e^i, F_n^i, F_m^i\right\} = \text{BiLSTM}(Lb^{i*}, Le^{i*}, Ln^{i*}, Lm^{i*})$

  $F_{benm}^i = [F_b^i \oplus F_e^i \oplus F_n^i \oplus F_m^i]$

  Spatial features are extracted by VGG-S:

  $\mathrm{U}^{m\times n\times c} = VGG19(Peak(i))$

  $V_C = Squeeze(U) = \frac{1}{m\times n}\sum_{a=1}^{m}\sum_{b=1}^{n} u_c(a,b)$

  $S = Excitation(V, w) = \sigma(w_2\delta(w_1 V))$;

  $f_c = Scale(U_c, S_c) = S_c U_c$ // SENet to assign channel weights to shallow features.

  $F_S^i = [f_1, f_2, \dots, f_c]$ // get the spatial features

  Calculate classification results:

  $P_T^i = \left[P_T^i(1), P_T^i(2), \dots, P_T^i(K)\right] = softmax(F_{benm}^i)$;

  $P_S^i = \{P_S^i(1), P_S^i(2), \dots, P_S^i(K)\} softmax(F_S^i)$

  $Prediction(i) = argmax(\alpha P_T(k) + (1-\alpha)P_S(k))$

End For

$Prediction = \{Prediction(i)\}_{i=1}^{I}$

return $Prediction$

4 Experiments and Discussion

4.1 Datasets

TO extensively and objectively evaluate the performance of STDFN, we experimented on three widely used sequences collected under laboratory control facial expression datasets: the CK+ [12], the Oulu-CASIA [13], and the MMI [17]. The expression sequences are the evolution of neutral face frames to peak expression frames, where 6-class basic expression recognition tasks of anger, surprise, disgust, fear, happiness, and sadness are performed on both Oulu-CASIA and MMI datasets. In the CK+ dataset, contempt was added to perform the 7-class classification task. To ensure fairness, random 5-fold cross-validation was used on datasets to obtain the evaluation performance of the models.

4.2 Data Preprocessing

Face Clipping and Gray Processing. FER database is uncropped, and removing background can eliminate the influence of the surrounding environment in the image on the accuracy of FER. The dlib tool is used to crop the facial image out to 64 × 64 pixels. Then, to prevent the effects of factors such as illumination intensity on the prediction results, we convert the facial images to gray images.

Data Augmentation. FER based on deep learning algorithm needs enough effective training data, which can prevent overfitting in deep neural network training and ensure the generalization performance of FER. However, the current open FER database lacks sufficient training data. Therefore, data augmentation becomes an indispensable part of deep learning algorithm-based FER. as shown in Fig. 4, We use an off-line data augmentation method, rotated each pre-processed training image at angle of {−15°, −10°, −5°, 0°,15°, 10°, 15°}, then the rotated image is flipped on the X-axis. in the end, a single original image will produce a 14-fold image sample.

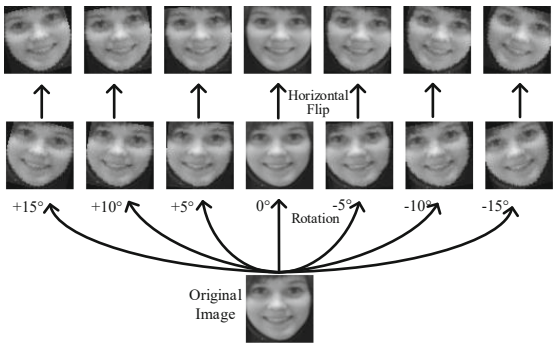


Fig. 4. Data augmentation in CK+. The first batch of images were obtained by rotation of 0°, 5°, 10°, 15°, −5°, −10°, −15°, then the second batch of images were obtained by inversion of the first batch.

4.3 Results and Analysis

Implementation Details and Parameters. We marked the whole face with 68 coordinates points, which were divided into 4 facial regions: eyebrows with 10 coordinates; eyes with 12 coordinates; nose with 9 coordinates; mouth with 19 coordinates. All the LSTM modules used 3 × 256 structure, which is three layers of LSTM subnet, each layer is set with 256 neurons. After learning each sequence, update weights. On the VGG-S model, SGD is used to optimize the parameters. The momentum is 0.9, the weight decay is 0.004, and the initial learning rate is 0.001.

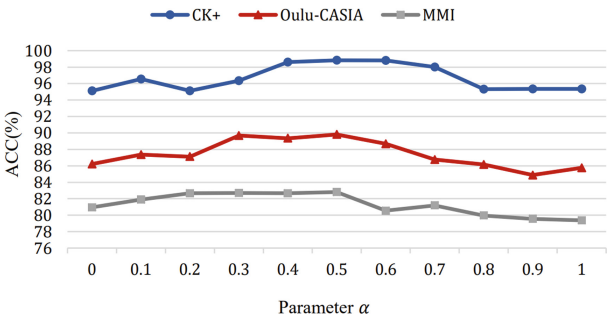


Fig. 5. The accuracy under different parameters α on three datasets.

In order to balance the information provided by the time network and the space network, we study the weight by changing the hyperparameter α of formula (16) from 0 to 1. When $\alpha = 0$, only the information provided by the spatial network is retained, when $\alpha = 1$, only the information provided by the temporal network is retained, and the change step of α is set to 0.1. Figure 5 shows the changes in the accuracy of FER on the CK+, Oulu-CASIA and MMI datasets with the change of α. In order to prevent the contingency of the experiment, the figure shows the average accuracy obtained by the 5-fold cross-validation method.

Ablation Experiments. The performance of the spatial-temporal information decision fusion network is mainly determined by the respective performance of the temporal network and the spatial network. In order to evaluate the respective functions of the two networks, we conducted ablation experiments on three datasets. Table 1 summarizes the ablation experimental results. The experimental results show that both the spatial network and the temporal network can complete the facial expression recognition task separately. However, due to the limited integrated feature information, a single network cannot achieve high recognition accuracy. After the decision fusion method, our model achieves the best performance.

Table 1. Ablation experiments on three datasets.

Method	Descriptor	CK+	Oulu-CASIA	MMI
Temporal network	BiLSTM	96.56%	85.76%	79.37%
Spatial network	VGG19-S	95.71%	86.21%	80.95%
Ours	**STDFN**	**98.83%**	**89.31%**	**82.86%**

Comparative Experiments. Table 2 shows the performance comparison of the proposed method with the SOTA method on three datasets. Most researchers chose to implement the 7-class recognition task on the CK+ dataset. The current known best performance for 7-class is PHRNN-MSCNN, which achieved 98.50% accuracy. On the

7-class classification task, our method achieved 98.83% recognition accuracy, surpassing the existing methods. Moreover, the PPDN achieved 99.30% accuracy on 6-class classification task for the CK+ dataset. We also performed a 6-class classification task by removing contempt, achieved recognition accuracy by 99.39%.

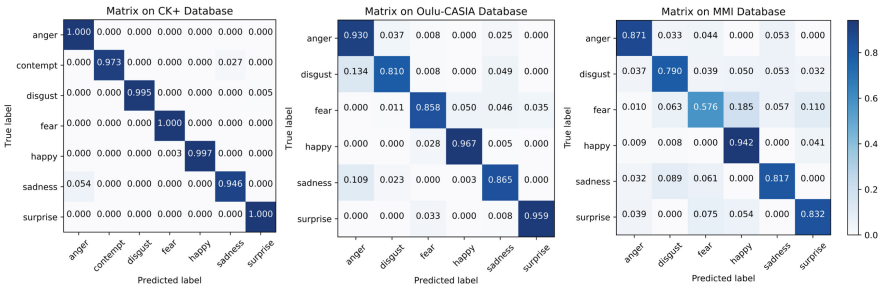
Table 2. FER accuracy (%) of various methods on CK+, Oulu-CASIA and MMI database.

Method	Descriptor	CK+	Oulu-CASIA	MMI
Klaser et al. [6]	HOG 3D	91.44	70.63	60.89
Liu et al. [7]	STM-Explet	94.19	74.59	75.12
Zhang et al. [9]	STRNN	95.40	–	–
Ding et al. [16]	FN2EN	96.80	87.71	–
Jung et al. [8]	DTAGN	97.25	81.46	70.24
Yang et al. [11]	DeRL	97.30	88.00	73.23
Hu et al. [14]	CTSLSTM	–	–	78.40
Zhang et al. [4]	PHRNN-MSCNN	98.50	86.25	81.18
Liu et al. [15]	MIC	–	–	81.29
Zhao et al. [10]	PPDN	99.30 (6)	–	–
Spatial-temporal networks	**STDFN (6-class)**	**99.39**	–	**–**
Spatial-temporal networks	**STDFN (7-class)**	**98.83**	**89.31**	**82.86**

Previously, the best performance was achieved the 88.00% accuracy by DeRL on Oulu-CASIA dataset. They trained a generating model to generate roughly neutral faces for pictures of facial expressions, and learned to recognize facial expressions by learning the residual information from the generating model. STDFN achieves the highest recognition accuracy of 89.31%, which is better than the most advanced methods at present. In addition, compared with PHRNN-MSCNN, which is based on spatial-temporal network, our model achieves a relatively satisfactory performance improvement. And our STDFN achieves 82.86% FER accuracy on the MMI dataset, which indicates that the method has a strong generalization capability. Our method has surpassed among currently known methods, both manual feature-based and spatiotemporal network-based methods.

As shown in Fig. 6 (a), on CK+, STDFN has achieved very high accuracy in recognizing the five expressions of happiness, anger, disgust, fear and surprise, which indicates that our model has been able to fully recognize these labels. However, contempt and sadness are still less effective, perhaps because the facial changes in some samples of both expressions are too slight. The confusion matrix results of STDFN in Oulu-CASIA dataset are shown in Fig. 6 (b). STDFN has high accuracy in identifying happiness, anger and surprise, and low accuracy in identifying fear, sadness and disgust. Especially for disgust, our method has a low recognition accuracy, most of which will be confused with anger, indicating that our method is still unable to accurately capture the subtle changes of facial expressions. It is worth mentioning that although the recognition accuracy of

disgust is low, our method does not appear serious confusion in expression recognition, and is generally more balanced. This shows that our method has a certain robustness. Figure 6 (c) shows the distribution of samples with classification errors in STDFN on the MMI dataset, and the two expression classes with low classification accuracy are disgust and fear. This is because some of the sample labels of these two classes in the original dataset are inaccurate.



(a) CK+ (b) Oulu-CASIA (c) MMI

Fig. 6. Confusion matrix for STDFN implements the facial expression recognition task on three datasets. (a) Confusion matrix for the CK+ dataset; (b) Confusion matrix for the Oulu-CASIA dataset; (c) Confusion matrix for the MMI dataset.

5 Conclusion and Future Works

In this paper, a FER method based on spatial-temporal decision fusion network is proposed. From the perspective of temporal information, we use BiLSTM to consider the time correlation between the frames before and after the image sequence, and integrate the local features of facial expressions through the division of facial regions. To the spatial information, VGG19 is used to obtain the shallow spatial features of the peak frames, and SENet is applied to module to learn the weight between feature channels. Finally, a decision fusion method is used to successfully combine the temporal and spatial dimension information of facial expression images. We not only implemented a spatial-temporal network-based FER method, but effectively improved the accuracy of the FER system. Finally, we discuss the parameter selection of decision fusion method, and compared our method with the current SOTA method. The experimental results show that our method has achieved an accuracy of 98.83%, 89.31% and 82.86% respectively on the most commonly used datasets CK+, Oulu-CASIA and MMI.

In future, we will focus on more extensive research on consciousness recognition and emotion recognition to further explore valuable information, and set out to develop more powerful methods to capture subtle evolution in facial expressions to further improve the accuracy of FER. We further intend to integrate physiological signals data into the FER system to realize the human emotion analysis system of multi-modal information fusion.

Acknowledgments. This work is supported by the Natural Science Foundation of Shandong Province (No. ZR2020LZH008, ZR2021MF118, ZR2019MF071), the Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO. 2021CXGC010506, NO. 2021SFGC0104).

References

1. Tayibnapis, I.R., Koo, D.Y., Choi, M.K., et al.: A novel driver fatigue monitoring using optical imaging of face on safe driving system. In: 2016 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), pp. 115–120 (2016)
2. Poria, S., Cambria, E., Bajpai, R., et al.: A review of affective computing: from unimodal analysis to multimodal fusion. Inf. Fusion **37**, 98–125 (2017)
3. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. (2020)
4. Zhang, K., Huang, Y., Du, Y., et al.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **26**(9), 4193–4203 (2017)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141. IEEE (2018)
6. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: BMVC 2008-19th British Machine Vision Conference (BMVC), vol. 275, pp. 1–10. British Machine Vision Association (2008)
7. Liu, M., Shan, S., Wang, R., et al.: Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1749–1756. IEEE (2014)
8. Jung, H., Lee, S., Yim, J., et al.: Joint fine-tuning in deep neural networks for facial expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2983–2991. IEEE (2015)
9. Zhang, T., Zheng, W., Cui, Z., et al.: Spatial–temporal recurrent neural network for emotion recognition. IEEE Trans. Cybern. **49**(3), 839–847 (2018)
10. 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
11. 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 (CVPR), pp. 2168–2177 (2018)
12. Lucey, P., Cohn, J.F., Kanade, T., et al.: 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 (CVPRW), pp. 94–101 (2010)
13. Zhao, G., Huang, X., Taini, M., et al.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)
14. Hu, M., Wang, H., Wang, X., et al.: Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks. J. Vis. Commun. Image Represent. **59**, 176–185 (2019)
15. Liu, X., Jin, L., Han, X., et al.: Mutual information regularized identity-aware facial expression recognition in compressed video. Pattern Recogn. **119**, 108105 (2021)

16. Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 118–126. IEEE (2017)
17. Valstar, M., Pantic, M.: Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proceedings of 3rd International Workshop on EMOTION (Satellite of LREC): Corpora for Research on Emotion and Affect, p. 65 (2010)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Zhou, P., Shi, W., Tian, J., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), vol. 2, pp. 207–212 (2016)