Facial Expression Recognition Based on Group Domain Random Frame Extraction

Wenjun Zhou[1], Lu Wang[1], Yibo Huang[1], Linbo Qing[1,2(✉)], Xiaohong Wu[1], and Xiaohai He[1]

[1] College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
qing_lb@scu.edu.cn
[2] Key Laboratory of Wireless Power Transmission of Ministry of Education, Chengdu 610065, China

Abstract. Modeling the dynamic variation of facial expression from a sequence of images is a key issue in facial expression recognition. However, the analysis of complete sequence temporal information requires significantly computational power. To improve the efficiency, a dynamic frame sequence convolutional network (DFSCN) is proposed in this study. In the proposed DFSCN, an expression sequence simplification method is first proposed to reduce the sequence length and takes the reduced new sequence as the input of DFSCN. An adaptive weighted feature fusion method for spatiotemporal feature learning is then put forward in DFSCN. A still frame convolutional network (SFCN) is introduced for complementing the still appearance information and the fine-tuning of DFSCN. Finally, these two models are combined together by weighted fusion to enhance the performance. Two public-available databases, CK+ and Oulu-CASIA, are used to evaluate the performance of the proposed approach. Experimental results show that the proposed method can effectively capture the dynamic process of expression sequence and the recognition performance is superior to other state-of-the-art methods.

Keywords: Facial expression recognition · Sequence simplification · Adaptive weighted feature fusion · Dynamic Frame Sequence Convolutional Network · Still Frame Convolutional Network

1 Introduction

Automatic Facial Expression Recognition (FER) has become an attractive research topic in the field of computer vision, due to its significant role in numerous applications such as medical treatment [1], security monitoring [2] and many other human-computer interaction systems. Existing researches on FER can be divided into two categories depending on the type of data: dynamic video sequence-based and static image-based. Dynamic video sequence-based approaches can effectively extract useful temporal features from consecutive frames of input, whereas static image-based methods mainly focus on spatial information from the current single image. Extensive researches have demonstrated that the performance of sequence-based methods is usually better than

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11901, pp. 467–479, 2019.
https://doi.org/10.1007/978-3-030-34120-6_38

that of image-based one [3, 19] due to better exploiting of the dynamic spatial-temporal feature of facial expression. The expression recognition based on dynamic sequence usually proceeds from original sequence or the processed facial sequence, such as STM-ExpLet [3], DTAGN [4], PHRNN-MSCNN [5], FACRN-FGRN [6] and etc. Generally speaking, the research based on facial landmarks or other indirect information is more complicated than using the original sequence, because it requires special processing, and pretreatment may affect the recognition performance of the model. Therefore, it is a meaningful essay to use only the original sequence for facial expression recognition while ensuring high recognition accuracy.

There have been many pioneers in the research of expression recognition based on original expression sequence. Jung et al. [4] proposed a deep temporal appearance network (DTAN) to extract useful temporal features and achieved satisfied recognition rate. Huang et al. [6] introduced a facial appearance convolutional recurrent network (FACRN) to combine CNNs and RNNs [10] to learn characteristics from consecutive frames. In the above methods, the sequence is input into the network with original length and these networks can hardly process all the frames in one pass because of the varied length of sequence, then the efficiency of learning dynamic change of whole expression decreased [7]. Subsequently, Zhang et al. [18] proposed a new CNN architecture which imports a frame-to-sequence model based on the last few frames of the sequence for facial expression recognition, which fixed the problem of not being able to process all frames at once. Although this method reduces the original expression sequence to analyses the facial changes and achieves a certain recognition rate, it discards the frames in front of the sequence, that is to say the temporal correlation of the whole sequence is not fully considered. Therefore, it is of great value to enable the network to efficiently process expression sequences while preserving the temporal correlation of sequences, especially in the case of long sequences.

Usually, the sequence length of dynamic expression datasets, such as CK+ [13] and Oulu-CASIA [14], are varied. Meanwhile, as shown in Fig. 1, the variety of expression includes at least three stages for all expression databases, namely initial state, transition state and peak state. In each stage, the expression changes slightly or even almost unchanged, so the similarity of expression frames in the same stage is extremely high and the information contained are also similar. Based on the above analysis, if the redundant expression frames are eliminated and then the remaining sequences are used for expression recognition, the requirement for the network to process all sequences at once can be satisfied. And the time correlation of the sequence is also preserved.

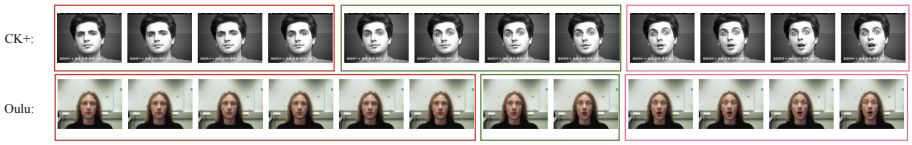


Fig. 1. Sequential expressions in two datasets. The expression of sequence can be divided into three parts: initial state (red box), transition state (green box) and peak state (pink box). (Color figure online)

In this work, we introduce a sequence simplification method to reduce the original expression sequence. Firstly, we propose a Dynamic Frame Sequence Convolutional Network (DFSCN). In our DFSCN model, we simplified the expression sequence to a fixed length and employ a convolutional structure to learn apparent characteristics for each extracted frame. Secondly, an Adaptive Weighted Feature Fusion (AWFF) algorithm is proposed in order to combine the previously obtained groups of spatial features to model the expression of the entire sequence. Considering frames still contain abundant spatial information, especially for frames with obvious expressions [19], a Still Frame Convolutional Network (SFCN) is also introduced to complement the spatial characteristics of DFSCN and fine-tune DFSCN. Finally, we use a score fusion approach to combine the DFSCN and SFCN together for final prediction. The contributions of this paper can be listed as follows:

- A deep network framework to extract temporal and spatial features of dynamic expression sequences for facial expression recognition is constructed.
- A new method to construct input sequences which can effectively shorten the length of the original sequence without losing global information of the sequence is proposed.
- A special feature fusion method is proposed to fuse the static features of different frames in the newly generated sequence, thus enhance the richness of temporal features.

This rest of this paper is structured as follows. Section 2 gives a detailed description of our DFSCN and SFCN. The performance of our proposed work compared with the state-of-art is evaluated in Sect. 3. Section 4 gives a conclusion of the whole paper.

2 Our Approach

Our proposed methodology is shown in Fig. 2. Our method consists of two kinds of networks: Dynamic Frame Sequence Convolutional Network (DFSCN) and Still Frame Convolutional Network (SFCN). Firstly, we introduce random frame extraction strategy and adaptive weighted feature fusion module in DFSCN to capture dynamic features of expression sequences. Secondly, SFCN is constructed to capture spatial features from still frames and fine-tune DFSCN. Finally, the two networks are integrated to boost the accuracy of facial expression recognition. The details of each component will be discussed in this section.

2.1 Image Pre-processing and Data Augmentation

Face Detection and Pre-processing. The expression frames in commonly used datasets, such as CK + and Oulu-CASIA, usually contain regions other than faces, which are helpless for expression recognition. Thus, we used the C++ library algorithm Dlib [15] for face detection and then cropped the detected faces. In addition, the obtained images were grayed and the size were normalized to $224 \times 224 \times 1$.

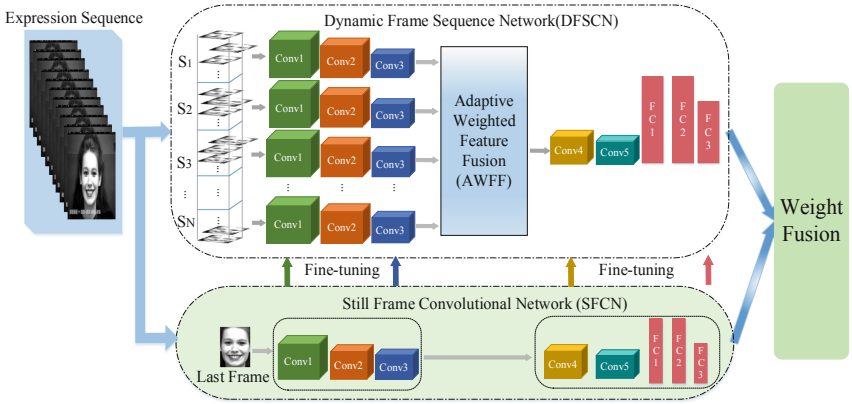


Fig. 2. Overall architecture of our method.

Data Augmentation. The existing facial expression datasets are relatively small, so the problem of overfitting is prone to occur during training. Therefore, various data augmentation techniques are required to increase the volume of data. The commonly used data enhancement methods include image rotation, noise addition and so on. Rotating image and increasing image brightness are used for data expansion in this paper. Image rotation is that the image revolves around a point and rotates at a certain angle clockwise or counterclockwise to form a new image. In this work, each image is rotated with seven angels, which can respond effectively to the change of slight. The image rotation formula is as follows:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \\ 1 \end{pmatrix} \tag{1}$$

where $P_0(\mathrm{x}_0, \mathrm{y}_0)$ is the pixel coordinate point of the original image, $P(x, y)$ is the pixel coordinate point of the rotated image, and θ is the angle of every rotation, $\theta \in \{-15°, -10°, -5°, 5°, 10°, 15°, 180°\}$. In addition, the brightness of datasets is increased by 5° and 15° respectively. Finally, we have ten times as much data as before.

2.2 Group Domain Frame Extraction

As shown previously in Fig. 1, the image sequence in public facial expression databases usually starts from a neutral face and gradually evolves into a peak expression. Thus, as shown in Fig. 1, in order to capture the information of each state and then mine the dynamic emotional changes of the whole sequence, we propose to sample a subset of frames representing the overall temporal dynamics of the sentiment sequence from the original whole sequence as the input of DFSCN. Same as [7] and [8], in this work, each expression sequence is split into N subsequences $\{S_1, S_2, \ldots, S_N\}$ of equal size, and one frame is selected randomly from each subsequence. Given the length of a

sequence is L, the newly generated clip CL is formed for representative frame extraction as follows:

$$C_i = rand(S_i), i \in \{1, 2, \ldots, N\} \tag{2}$$

$$CL = \{C_1; C_2; \ldots; C_N\} \tag{3}$$

where C_i is the extracted frame from the subsequence S_i, $rand(\bullet)$ represents the random selection of a frame from the specified sequence, CL is the newly generated sequence. After the original sequence is sampled, we employ N CNNs to process the extracted frames independently in one process. Then, N sets of feature maps are obtained, all of which contain only static information (shown in Fig. 2).

2.3 Adaptive Weighted Feature Fusion

Up to this point, the frames selected from sequences are processed independently. In order to learn how facial expressions are made up of different appearances over time, we stack the multiple groups of features with AWFF method and feed them into a 2D-CNN (Conv4) for further study. The detailed implementation process is shown in Fig. 3. When each frame in CL is sent to the three CNNs (Conv1, Conv2, Conv3) respectively, we obtained N groups of features $\{G_1, G_2, \ldots, G_N\}$. For any network, the feature map of convolution layer h can be expressed as:

$$f_h = f(w_h * x_h + b_h) \tag{4}$$

where w_h is the connection weight between the upper layer and the next layer, x_h is the initial input for the hidden layer, and b_h is the bias of the hidden layer. $f(\bullet)$ is the activation function.

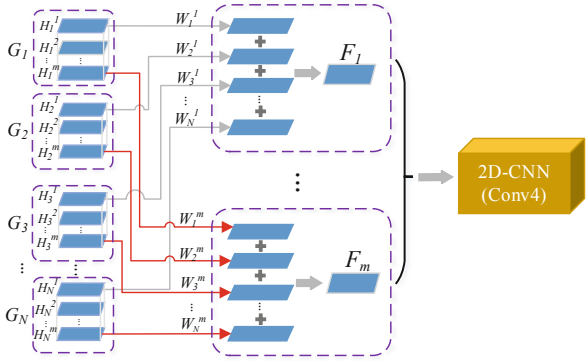


Fig. 3. Schematic diagram of spatial feature fusion.

In our AWFF model, the spatial features of the *i-th* $(i \in [1, N])$ channel include m feature maps, which are expressed as $G_i = \{H_i^1, H_i^2, \ldots, H_i^m\}$. When fusing the

N sets of features, a weight W_i^j will be assigned to the *j-th* ($j \in [1, m]$) feature map in G_i. The fusion formula is as follows:

$$F_k = W_1^k * H_1^k + W_2^k * H_2^k + \ldots + W_N^k * H_N^k, k \in \{1, \ldots, m\} \tag{5}$$

$$P = \{F_1, F_2, \ldots, F_m\} \tag{6}$$

in which F_k is the *k-th* feature fused from N group, P is the total feature output obtained after the fusion. In our work, the weight of W_i^j is updated adaptively to achieve the best fusion effect. Based on this method, we can more concisely combine the spatiotemporal information of the expression sequence.

2.4 The Detail of the Network Structure

In this paper, Our DFSCN model is designed as [Conv1(64) - Conv2(128) - Conv3 (256)] × N-AWFF - Conv4(512) - Conv5(512) - FC1(4096) - FC2(4096) -FC3(7/6), as shown in Fig. 2. The values in parentheses represent the total number of neurons used in the corresponding layer. For example, Conv2 (128) indicates that the number of convolution kernels of the second convolutional layer is 128. AWFF is designed to fuse the output of the N-channel [Conv1(64) - Conv2(128) - Conv3(256)] to learn the temporal features of the sequence. The weight information dimension in the AWFF module is $N \times m$. Besides, each convolution layer is sequentially followed by a max-pooling layer and Relu [22].

The biggest difference between DFSCN and SFCN lies in the feature processing method after the third convolution layer. For SFCN, the feature obtained after the third convolution layer is directly transmitted into the next convolution. However, for DFSCN, as shown in Fig. 2, N sets of features are obtained after N respective inputs, and then a method called adaptive weighted feature fusion is introduced to combine the features to get a new feature group and send it to Conv4.

SFCN network is designed to compensate for the spatial characteristics of DFSCN and fine-tune DFSCN as there is no relevant fine-tuning model for DFECN. The parameter settings of SFCN are similar to those of the DFSCN. We first train the SFCN network with the last frame of the original sequence, and then use parameters of Conv1, Conv2, Conv3, Conv4, Conv5, FC1, FC2, FC3 of SFCN to initialize the network layer parameters with the same name in DFSCN, respectively. Finally, the two models are fused to boost the recognition performance.

It is worth noting that the size of the convolution kernel of the five convolution layers in this paper is all 3 × 3, which is quite same with VGG [9]. There are two main reasons for this. Firstly, as the size of the convolution kernel increases, the number of parameters of the convolution kernel increases relatively, so the computational amount of convolution is bound to increase. Secondly, the network we designed must be moderate in depth and appropriate in parameters so as to avoid the problem of over-fitting caused by the small amount of data. Other than this, small receptive field is more capable of learning image features from details and distinguishable.

2.5 Model Fusion

In order to maximize the superiority of the two models, the DFSCN and SFCN are integrated by following fusion function [5]:

$$O(x) = \sum_{i=0}^{1} b_i(S_i(x) + P_i(x)) \tag{6}$$

where $P_i(x)$ $(0 < P_i(x) < 1)$ is the predicted probability of expressions in DFSCN and SFCN. $P_0(x)$ comes from the softmax layer of DFSCN while $P_1(x)$ comes from SFCN. $S_i(x)$ is sorted based on the prediction categories of expression. It can be expressed as:

$$S_i(x_1), \ldots, S_i(x_n) = Sorted(P_i(x_1), \ldots, P_i(x_n)) \tag{7}$$

where n refers to the total number of categories of expressions, $S_i(x) \in \{1, 2, \ldots, n\}$. In addition, b_i is a balance index for balancing different models. After a lot of comparative experiments, the value of b_i is set to 0.5 which can achieve the best performance as shown in Fig. 4.

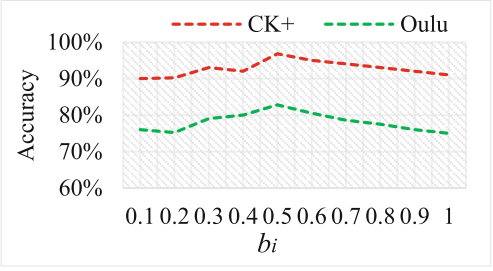


Fig. 4. Performance of our proposed work using the weighted fusion method with different values of b_i. We changed the value of b_i from 0 to 1 with interval of 0.01.

3 Experiment

In this section, we compare our models with some of the most advanced methods in facial expression recognition. This paper focuses on the emotion from neutral to peak. Thus, two widely used databases, namely CK+ and Oulu-CASIA, are used to assessing the performance of our approach. Details of our experimental results are given below.

3.1 Datasets

Description of CK+. The Extended Cohn-Kanade (CK+) database is one of the most extensively used databases for facial expression recognition. There are 593 video sequences from 123 subjects with a duration ranging from 9 to 60 frames in this database. Among these videos, 327 sequences are marked with seven typical emotional

expressions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Each expression sequence reflects the expression from neutral to emotional vertex. We adopt the most commonly used 10-fold validation method [4] to verify our experimental results.

Description of Oulu-CASIA. The Oulu-CASIA database contains 2880 expression sequences collected from 80 subjects. There are six types of emotions: anger, disgust, fear, happiness, sadness, and surprise. Similar to CK+ database, all of these sequences begins with a neutral expression and then gradually transit to the peak expression. We use 10-fold cross validation mentioned in the description of CK+.

3.2 Implementation Details

Depending on the characteristics of CK+ and Oulu-CASIA datasets, we set the value of N to 4. Figure 5 shows that more than 85% of sequences in these two datasets are between 10–30 in length. If these sequences are divided into fewer intervals, it may cause a loss of temporal features of sequences longer than 30. If the sequences are parted into more intervals, there will be more difficulties to network training. After comprehensive consideration, the sequences are divided into four groups, taking into account the sequences of different lengths. Notably, if the length of the sequence is not a multiple of $N = 4$, we pad the sequence with the last frame until the condition is satisfied.

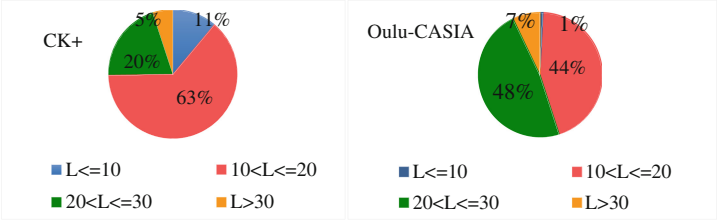


Fig. 5. Statistical analysis of sequence length in CK+ and Oulu-CASIA. L represents the length of the original sequence.

In this paper, experiments were carried out under the environment of tensorflow, a python-based deep learning framework. We firstly initialize the corresponding layer of SFCN by using the weights of VGG16 pre-trained on the ImageNet dataset [17]. When training our facial expression recognition network, we set the batch size of training to 16, the initial learning rate to 0.00001, and the training epoch to 300. For updating the network weight, we use Adam optimization algorithm [21]. Compared with the basic SGD algorithm [20], Adam can avoid local optimum and update faster. As for DFSCN, Adam optimization algorithm is also adopted. Meanwhile, we set the training batch size as 16, the initial learning rate as 0.00001, and the training epoch as 32, which means that the learning rate of 8 epochs decreased to 0.1 of the original learning rate.

It is important to note that we first train SFCN and then DFSCN. The model obtained from SFCN is not only used for classification, but for fine-tuning DFSCN. As shown in Fig. 2, the parameters obtained from the first three convolution layers of SFCN are

loaded into the four single channels of DFSCN, and then the parameters of SFCN obtained from the fourth convolution layer to the last full connection layer are loaded into the feature fusion network of DFSCN. After all the two networks are trained, the fusion approaches mentioned in Sect. 3.5 is adopted for expression classification.

3.3 Experimental Results

In the experiments, we first evaluate the performance of each network separately, then the two streams are combined together by weight fusion to achieve complementary network performance. As for the comparison experiments, our model is mainly compared with hand-crafted methods and deep learning methods [3, 4, 6, 11, 12, 16, 23, 24].

Accuracies and Analysis. Tables 1 and 2 show the recognition accuracy of our model on each database, as well as the comparisons with other algorithms. From the two tables we can see that the performances of DFSCN and SFCN alone are not comparable to many other algorithms, while the recognition accuracy after fusion is higher than most of the advanced methods. In CK+, the accuracy of our model is higher than traditional algorithm, such as STM-ExpLet [3]. Compared with the recently proposed deep learning algorithm, the final accuracy of our model is slightly lower than that of DTAGN [4], but higher than that of other methods [6, 11, 16]. In Oulu-CASIA, our recognition accuracy even exceeds all the previous researches. Meanwhile, it can be clearly found that our method is better than hand-crafted methods [3, 11, 12, 23, 24], which is mainly attributed to the feature extraction ability of our network. Our model also has significant advantages over the deep learning approaches [4, 6], which is largely related to our frame-to-sequence network connectivity and the fine-tuning approach we use.

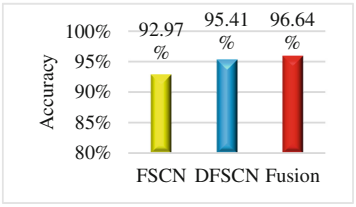
After the integrating our DFSCN and SFCN, the performance has reached the highest level, and this is shown in Fig. 6. In other words, the integration of DFSCN and SFCN improves the richness of network features, and the two channels complement each other. The combination of the two can boost the whole recognition accuracy, which proves the effectiveness of our method.

Table 1. Overall accuracy in CK+ database.

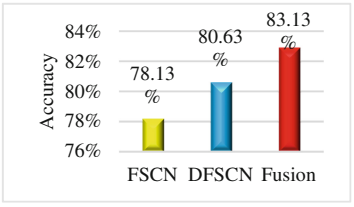
Method	Accuracy
HOG 3D [11]	91.94%
Cov3D [16]	92.30%
3DCNN [11]	85.90%
3DCNN-DAP [11]	92.40%
STM-ExpLet [3]	94.19%
DTAGN [4]	**97.25%**
FACRN-FGRN [6]	95.63%
SFCN	92.97%
DFSCN	95.41%
Fusion	**96.64%**

Table 2. Overall accuracy in Oulu database.

Method	Accuracy
HOG 3D [11]	70.63%
AdalLBP [23]	73.54%
Atlases [24]	75.52%
STM-ExpLet [3]	74.59%
3D SIFT [12]	75.83%
DTAGN [4]	81.46%
FACRN-FGRN [6]	76.50%
SFCN	78.13%
DFSCN	80.63%
Fusion	**83.13%**



(a) CK+



(b) Oulu-CASIA

Fig. 6. Comparison of accuracy of three networks (SFCN, DFSCN and Fusion) on two databases. (a) CK + . (b) Oulu-CASIA.

Confusion Matrix. Table 3 gives the confusion matrices of our model on the two datasets, respectively. In CK+, our model achieved high recognition accuracies on six emotions except Contempt (Co). In our results, Contempt (Co) is easily misdiagnosed as Sad (Sa) due to the similarity of the expression variation between sad and contempt, especially the change of mouth, as shown in Fig. 7(a). As for Oulu-CASIA database, our model performed well in recognizing Fear (Fe), Disgust (Di), Sad (Sa), and other expressions except Anger (An). The main reason for the poor recognition effect of our model on Anger (An) may be that the volunteers in this dataset have similar facial expressions when expressing Anger (An), Disgust (Di) and Sad (Sa), as shown in Fig. 7(b), which makes the algorithm of this paper not distinguish enough. Moreover, the image quality of the Oulu-CASIA is not as clear as CK+, which may be another reason for the low recognition accuracy.

Table 3. Confusion matrix of our proposed method for two databases.

(a) CK+ database

	An	Co	Di	Fe	Ha	Sa	Su
An	**96**	2	2	0	0	0	0
Co	6	**83**	0	0	0	11	0
Di	2	0	**98**	0	0	0	0
Fe	0	0	0	**96**	4	0	0
Ha	0	0	0	0	**100**	0	0
Sa	7	0	4	0	0	**89**	0
Su	0	1	0	0	0	0	**99**

(b) Oulu-CASIA database

	An	Di	Fe	Ha	Sa	Su
An	**68**	15	3	3	12	0
Di	11	**78**	1	1	6	3
Fe	1	1	**81**	7	1	7
Ha	0	0	3	**97**	0	0
Sa	15	4	0	1	**80**	0
Su	0	0	4	0	1	**95**

(a) Two types of mood samples in Ck+ (b) The samples in Oulu

Fig. 7. Sample examples of two datasets.

4 Conclusion

In this paper, we presented a new deep network structure that provides a new approach for the effectively modeling of the dynamic changes of facial expression based on the image sequences through the combination of the static characteristics of sequences. Specially, the proposed DFSCN uses simplified sequences as the network input to study the dynamic variations of expression sequences with the combination of our adaptive weighted feature fusion method. To supplement static appearance information and fine-tune DFSCN, the SFCN is proposed to extract useful spatial information from the last frame of each sequence. These two networks capture dynamic and static information, respectively, at the same time, and complement each other to improve the performance of facial expression recognition. We evaluated our two models using two public-available datasets, CK+, and Oulu-CASIA, respectively. The experimental results demonstrate that the proposed methods have achieved the same accuracy as the state-of-the-art methods.

Acknowledgments. The authors would like to thank the anonymous reviewers for their comments. This work was supported by the National Natural Science Foundation of China (No. 61871278) and the Sichuan Science and Technology Program (No. 2018HH0143).

References

1. Lucey, P., Cohn, J., Lucey, S.: Automatically detecting pain using facial actions. In: International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1–8. IEEE, Amsterdam (2009)
2. Cho, S., Kang, H.: Abnormal behavior detection using hybrid agents in crowded scenes. Pattern Recogn. Lett. **44**, 64–70 (2014)
3. Liu, M., Shan, S., Wang, R.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1749–1751. IEEE, Columbus (2014)
4. Jung, H., Lee, S., Yim, J.: Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE International Conference on Computer Vision (ICCV), pp. 2983–2991. IEEE, Santiago (2015)

5. Zhang, K., Huang, Y., Du, Y.: Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans. Image Process. **2017**, 4193–4202 (2017)
6. Huan, Z., Shang, L.: Model the dynamic evolution of facial expression from image sequences. In: Phung, D., Tseng, V., Webb, G., Ho, B., Ganji, M., Rashidi, L. (eds.) PAKDD 2018, LNCS, vol. 10938, pp. 546–557. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93037-4_43
7. Jing, L., Yang, X., Tian, Y.: Video you only look once: overall temporal convolutions for action recognition. J. Vis. Commun. Image Represent. **52**, 58–65 (2018). S10473203 18300233
8. Zolfaghari, M., Singh, K., Brox, T.: ECO: efficient convolutional network for online video understanding. In: Computer Vision 15th European Conference, pp. 1–7. ECCV, Munich (2018)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science. **2014**, 1–14 (2014)
10. Graves, A.: Generating sequences with recurrent neural networks. Computer Science. **2013**, 1–8 (2013)
11. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: Proceedings of the British Machine Vision Conference 2008, pp. 1–10. British Machine Vision Association, London (2008)
12. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: Proceedings of the 15th ACM International Conference on Multimedia, pp. 357–360. Association for Computing Machinery, Augsburg (2007)
13. Lucey, P., Cohn, J., Kanade, T.: 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, San Francisco (2010)
14. Taini, M., Zhao, G., Li, S.: Facial expression recognition from near-infrared video sequences. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE, Tampa (2008)
15. King, D.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**(3), 1755–1758 (2009)
16. Sanin, A., Sanderson, C., Harandi, M., et al.: Spatio-temporal covariance descriptors for action and gesture recognition. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 103–110. IEEE, Clearwater (2013)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 1097–1105. MIT Press, Lake Tahoe (2012)
18. Kuo, C., Lai, S., Sarkis, M.: A compact deep learning model for robust facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2121–2129. IEEE, Salt Lake City (2018)
19. Li, S., Deng, W.: Deep facial expression recognition: a survey, pp. 1–25. arXiv preprint (2018)
20. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, L., Saporta, G. (eds.) Proceedings of COMPSTAT 2010, pp. 177–186. Physica-Verlag HD, Heidelberg (2010). https://doi.org/10.1007/978-3-7908-2604-3_16
21. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR), pp. 1–15 (2015)

22. Nair, V., Hinton, G.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 807–814 (2010)
23. Zhao, G., Huang, X., Taini, M., et al.: Facial expression recognition from near-infrared videos. In: Image and Vision Computing (IVC), pp. 607–619 (2011)
24. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition using longitudinal facial expression atlases. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, pp. 631–644. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33709-3_45