Novel multi-scale deep residual attention network for facial expression recognition

Dong Liu*, Liang Jia†, Liang Fan‡, Longxi Chen†
†School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, People’s Republic of China
‡E-mail: syg_dong@163.com

Abstract: Recently, the deep convolutional neural networks (CNNs) have shown great success and for facial expression recognition (FER). These CNN-based approaches have made breakthroughs in the accuracy by using deeper networks. As the amount of data is very less, these models easily fall into the overfitting in the training. In this study, a novel multi-scale deep residual attention network (Ms-RAN) is proposed for FER. The proposed Ms-RAN is mainly based on the multi-scale residual attention unit, which consists of two different scale sub-units. Each sub-unit is composed of the convolutional layers, parametric rectified linear units (PReLUs), and the residual attention connection. By focusing on the relationship between channels and automatically learning the importance of different channel features, the proposed Ms-RAN can make the proposed model pay more attention to the most informative channel features, while suppressing those unimportant channel features. Owing to the unique design of Ms-RAN, the combination of various levels features can be enhanced in the proposed method, and valuable and different ranges of expressive information can also be provided for recognition. The experimental results demonstrate that the proposed method achieves superior performance than other state-of-the-art approaches on five databases, CK+, Oulu-CASIA, BU-3DFE, BP4D+, and MMI.

1 Introduction

Facial expression is one of the most important signals to convey emotional states and intentions for human being [1]. Recognising the expression from a facial image is a classic problem in computer vision, and has a wide range of research prospects, including intelligent medical treatment, polygraph detection, internet education, and so on. Various facial expression recognition (FER) methods and systems have been proposed in the field of computer vision and machine learning. In these works, they usually defined six basic emotions, which contain anger, disgust, fear, surprise, sadness, and happiness.

Most FER works focus on feature extraction and classifier construction, which can be classified into two categories: static image-based and image sequence-based approaches. Static image-based models [2, 3] are applicable to a single static image for FER, which include random forests, support vector machines classifiers, and Softmax classifiers. Image sequence-based approaches are applied to facial video sequences. By extracting useful temporal features from the image sequence, these methods can increase the recognition performance, such as hidden Markov model and visual machine learning. In practical application, based on these two kinds of methods, other modalities have also been used in multimodal systems to assist the FER, such as the physiological and audio channels.

In most traditional methods, they used handcrafted features or shallow learning for FER, such as the histograms of oriented gradients (HOG) [4], histograms of local binary patterns (LBP) [5], scale invariant feature transform (SIFT) [6], LBP on three orthogonal planes (LBP-TOP) [7], histograms of local phase quantisation, sparse learning [8], and non-negative matrix factorisation [9]. In addition, there are also some other spatiotemporal approaches, such as interval temporal Bayesian network [3], spatiotemporal covariance descriptors (Conv3D) [10], temporal modelling of shapes [11], and expressionlets on spatiotemporal manifold (STM-ExpLet) [12]. Due to the sparse feature representations, the performance of these methods should be improved. Since the emotion recognition competitions, like the Emotion Recognition in the Wild [13, 14] and the FER2013 [15], have been held successfully, sufficient training data have been collected from challenging real-world scenarios. The transition of FER is also implicitly promoted from lab-controlled to in-the-wild settings. In addition, with the dramatical development of the chip processing abilities and deep convolutional neural networks (CNNs), studies in various fields have begun to transfer to deep learning methods.

Due to the ability of automatically extracting useful representations from image data, deep neural networks (DNNs) and CNNs have shown excellent performance in face-related recognition tasks [16–24] in recent years. For example, Sun et al. [25] applied the region-based CNN (R-CNN) to extract features for FER. By generating high-quality region proposals, Li et al. [26] used faster R-CNN to identify facial expressions. Moreover, with the sequence image, Li et al. [27] proposed 3D CNN to capture the motion information that encoded in multiple adjacent frames. However, despite recent significant progress, FER is still a challenging task due to many difficulties, such as different subjects often display the same expression with visual appearances and diverse intensities. Moreover, there is a limit when applying the deep CNNs directly to FER databases. The major reason is that the amount of training data is too small. During the training process, the deep network is easy to fall into overfitting.

Although the DNN architectures have shown great success and superior performance in FER tasks [28–34], nevertheless, there are still some important limitations. For example, most approaches ignore the intrinsic correlations between each pair of samples and just consider each sample independently during learning, which limits the discriminative capabilities of the learned models. In addition, a DNN that has many parameters can easily fall into overfitting with a small amount of data in training. The overfitting problem will become more crucial when the training data is high dimensional. Even when regularisation techniques such as dropout [35] and batch normalisation [36] are used for network training, the results are not satisfactory.

Moreover, despite significant recent progress, FER is still a challenging problem with many difficulties, such as the different subjects often display the same expression with diverse intensities and visual appearances. For example, an expression will first appear in a subtle form and then grow into a strong display of the underlying feelings in a video stream. Second, the peak and non-
peak expressions from the same people can have significant variation in terms of attributes, such as facial wrinkles and mouth corner radius. The non-peak expressions are more commonly displayed than peak expressions. Obviously, the critical and subtle expression details from non-peak expression images will be difficult to be captured in this way, which can be hard to distinguish across expressions.

Motivated by this observation and in order to solve the problem, we propose a novel multi-scale deep residual attention network (Ms-RAN) in this paper. This model is committed to improve the feature extraction and recognition ability of the network. There are two special designs in the proposed model to improve the performance and avoid the overfitting. First, the multi-scale network structure is adopted in the proposed model, which can be applied to combine the features from multiple scale and provide very fine features for the FER. Second, the residual attention mechanism is proposed and applied in each unit. This design can improve the performance of feature extraction. Moreover, the concatenation is used to connect the features from each branch and original facial image input. The experimental results show that our proposed method can achieve better performance in feature extraction and recognition with small amount of data.

2 Related work

Recently, the methods that are based on the deep CNN have been proposed and used for FER. For example, Yang et al. [37] proposed a weighted mixture DNN to automatically extract features that are effective for FER task. Kuo et al. [38] proposed a compact frame-based FER framework for FER which achieves very competitive performance with respect to state-of-the-art methods while using less parameters. Yang et al. [39] proposed a de-expression residue learning (DeRL) method to recognise the facial expression by extracting information of the expressive component. Yu [40] proposed an ensemble of multiple deep CNNs for FER. Mollahosseini et al. [41] used three inception structures [42] in convolution to achieve the recognition of facial expression.

In addition, the VGGNets [43] and inception models [42] showed that increasing the depth of a network could significantly increase the quality of representations that it was capable of learning. By regulating the distribution of the inputs to each layer, batch normalisation can add stability to the learning process in deep networks and produced smoother optimisation surfaces [44]. Based on these works, residual network demonstrated that it was possible to learn considerably deeper and stronger networks through the use of identity-based skip connections [45]. This design not only can eliminate the degradation during the stochastic gradient descent optimisation process when training, but also can enable the data to flow across layers. In this way, the residual network can be used to counteract the internal co-variate shift and train deeper CNN efficiently. Compared with the single deep CNN, ResNet can avoid the overfitting and achieve the superior performance. Thus far, the residual network has been widely used in deep network to deal with the computer vision problem from low-level to high-level tasks, such as image classification and detection. After that, many state-of-the-art algorithms based on ResNet are proposed and achieved excellent performance in facial-related recognition tasks. Owing to the efficient residual connection, these residual-based methods become increasingly better and unceasingly refresh the recognition accuracy records.

Until now, many studies have been conducted for FER. For traditional local features, such as LBP, HOG, and SIFT, which have been applied to FER successfully and have extended in order to adapt to video. Additionally, by using conditional random fields and shape-appearance features created manually, Jain et al. [46] attempted to improve the accuracy through temporal modelling of each facial shape.

Besides the traditional methods, deep CNNs have also been widely used for FER and achieved excellent performance in recent years. Many models that based on the CNN architecture have been proposed for FER. For instance, Yu [40] applied an ensemble of multiple deep CNN model to identify the facial expression. Scovanner et al. [47] adopted three inception structures in the model network for FER. Moreover, the recent utility of generative adversarial network (GAN) [48] shows success in FER. For example, Zhou and Shi [49] used the neutral faces to synthesise facial expression images with cGANs.

3 Proposed methods

In this section, we will introduce the proposed method in detail. In the first part, the architecture of the proposed multi-scale residual attention unit is shown. The proposed model structure will be discussed and analysed in the following part.

3.1 Multi-scale residual attention unit

In this work, a multi-scale deep residual attention unit is proposed to perform feature learning and exaction. As shown in Fig. 1, the Ms-RAU consists of two sub-units, which is composed of the convolutional layers, parametric rectified linear units (PReLU), fully connection layer, and the residual attention connection. For each unit, the size of the filters is different, which are designed to exact the multi-scale feature and learn a hierarchical feature representation. Then each unit is defined as

\[
Y = \sigma(x) \cdot h(x)
\]

where \( \sigma \) represents the functions of the fully connection layer. \( x \) is described as the input of each unit. \( h \) denotes the transformation of PReLU and the convolution layers, which can be described as

\[
h(x) = f_1(g(f_2(x, w_1), w_2))
\]

here, \( f_1 \) and \( f_2 \) represent the functions of the first convolution layer and second convolution layer of each unit, respectively. Then \( g \) is considered as the PReLU layer. \( w_1 \) and \( w_2 \) are defined as the weight parameters of the first and second convolution layers in each unit, respectively.

3.2 Model architecture

As shown in Fig. 2, the proposed network consists of convolutional layers, Ms-RAUs, max-pooling layer, PReLU layers, fully connection layer, and skip connection. Given a facial expression image input, we use the first convolutional layer to extract the feature maps, which will be sent to Ms-RAUs. After that, a skip connection is used to concatenate the original facial expression input and the output from the Ms-RAUs. In this way, different levels feature maps can be combined. Lastly, we use the fully connection layer and softmax to complete the recognition of the facial expression.

4 Experimental results

4.1 Datasets

To assess our proposed model, extensive experiments have been conducted on two public facial expression databases: CK+ [50] and Oulu-CASIA [51]. These two datasets have been widely used for evaluating FER in other works. The CK+ is a dataset which has the labelled emotion number for the frame of sequences from neutral to peak states. A total of 123 subjects participated and 593 image sequences were included, with 327 of them being labelled with seven universal emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise). Fig. 3 shows the CK+ dataset examples of frontal facial image. Moreover, to further demonstrate the proposed algorithm generalises to other recognition approaches, we also evaluate the experiment performance on FER over other three popular databases, including JAFFE [52], MMI [53], BU-3DFE [54], and spontaneous expression database BP4D+ [55]. The JAFFE is a dataset consisting of grey-scale frontal facial expression images of ten Japanese women. It contains a total of 213 images including seven facial expressions. Fig. 4 shows examples of the JAFFE dataset.
4.2 Training details and parameters

During experiments, we use five landmark points to align the face region. Considering the databases do not provide the landmarks, the TSM [11] is applied for face detection and landmark localisation. After that, the aligned face region is cropped and resized to the size of $70 \times 70$ for model training. Additionally, a data augmentation method is used to generate more training data to avoid over-fitting. The data augmentation methods that are used in this paper include random crop, mirror, colour jitter, noise, rotate, and translation. With the data augmentation, the training dataset is 110 times larger than the original one.

For the network training, the proposed model is trained with ADAM. The parameters are set as $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 10^{-8}$. In addition, the mini-batch is set to 16. All variants in our network are trained with 300 epochs with an initialised learning rate $10^{-4}$. Moreover, all the experiments are performed under Ubuntu14 and Tensorflow running on a computer with a NVIDIA GTX1080 GPU.

4.3 Comparisons with state-of-the-art methods

We provide the qualitative and quantitative comparisons between the proposed method and other state-of-the-art approaches in this part. In order to examine the performance of our model fairly, the results and the average expressions recognition accuracy that are provided in this section are described as the average of the ten experimental runs. In Table 1, we illustrate the comparisons between LBP-TOP [7], HOG 3D [56], STM-Explet [12], DTAGN [57], DeRL [39], DCN [58], Kuo et al. [59], and the proposed

---

**Fig. 1** The network structure of the Ms-RAU
(a) Architecture of each sub-unit, (b) Architecture of multi-scale residual attention unit

**Fig. 2** Architecture of the proposed Ms-RAN

**Fig. 3** Examples of facial expression from CK+ dataset
model on the CK+ and Oulu-CASIA databases. In these works, we compare the proposed method with other FER models that are based on sequence and image. As shown in the experimental results, the proposed algorithm performs well for FERs on static images and achieves over 99.06 and 91.89% recognition rate on CK+ and Oulu-CASIA datasets, respectively, outperforming the compared state-of-the-art approaches. Meanwhile, we show the confusion matrix of our proposed method on the CK+ and Oulu-CASIA databases in Figs. 5 and 6, respectively, where the fear and anger show the lowest recognition rates with 95 and 82%, respectively.

In order to validate the proposed model further, a cross-database validation is conducted in this section, as shown in Tables 2 and 3. In this table, the proposed method is compared with the state-of-the-art works on JAFFE datasets. Lopes et al. [61] suggested a representative method that used a CNN in the FER. They learned data with shuffle training in order to change the order so that the method could be learned with less data. The accuracy of the basic CNN algorithm using only the static feature was 84.48%. Liu et al. [62] proposed a method to extract features of salient areas by using LBP and HOG features with gamma correction, which resulted in 90% accuracy. Goyani and Patel [63] used feature vectors by constructing the Haar wavelet of multiple levels with the face, eyes, and mouth. Finally, Kim et al. [64] proposed a hierarchical deep learning method for FER. The accuracy was 91.27%. Hamester et al. [65] proposed the MCCNN for FER and achieved 95.80% accuracy. From Table 4, we can see that our proposed model Ms-RAN achieved the best performance among the state-of-the-art works, which resulted in 96.03%.

Table 1

| Method          | CK+    | Oulu-CASIA |
|-----------------|--------|------------|
| LBP-TOP         | 88.99  | 68.13      |
| HOG 3D          | 91.44  | 70.63      |
| STM-Explet      | 94.19  | 74.59      |
| DTAGN           | 97.25  | 81.46      |
| CNN(baseline)   | 89.50  | 72.92      |
| DeRL            | 97.30  | 88.00      |
| DCN [58]        | 98.9   | —          |
| Kuo et al. [59] | —      | 91.70      |
| Ms-RAN (ours)   | 99.06  | 91.89      |

In addition, we also show the confusion matrix of our proposed method on the MMI and BU-3DFE databases in Figs. 7 and 8, respectively, where we can see the happiness and surprise expression are easy for recognition, and the fear expression is hard for recognition. Furthermore, we have also conducted a cross-database validation, as shown in Table 4. Meanwhile, the result of the confusion matrix of the proposed algorithm is shown in Table 5. The results are similar to those of the CK+ dataset.

Table 2

| Method          | Setting          | Accuracy |
|-----------------|------------------|----------|
| LBP-TOP         | sequence based   | 59.51    |
| HOG 3D          | sequence based   | 60.89    |
| STM-Explet      | sequence based   | 75.12    |
| DTAGN           | sequence based   | 70.24    |
| CNN(baseline)   | image based      | 57       |
| DeRL            | image based      | 72.23    |
| Liu et al. [60] | image based      | 78.50    |
| Ms-RAN (ours)   | image based      | 78.65    |

Table 3

| Method          | BU-3DFE | BP4D+  |
|-----------------|---------|--------|
| CNN (baseline)  | 73.2    | 76.5   |
| DeRL            | 94.17   | 74.41  |
| Ms-RAN (ours)   | 86.10   | 78.23  |
As we can see from Fig. 9, the models converge after iteration to a certain algebra. Taking the test set as an example, the following conclusions can be drawn from the analysis of Fig. 9 in our experiments. Firstly, as shown in Fig. 9, the proposed method began to converge after 25 generations of training, the convergence speed of the DeRL method was relatively slow, and the convergence speed of the CNN algorithm was slowest. Secondly, the proposed algorithm Adap-SE-ResNet can achieve higher recognition rate than other three methods after iteration to a certain extent. Therefore, it can be seen from Fig. 9 that the proposed approach can improve the recognition rate and be used for FER in complex background to a certain extent.

In summary, the proposed method outperforms the other algorithms in terms of the objective accuracy of the FER. From the experimental results, the proposed algorithm exhibits superior performance and ability to classify the facial expression than other recognition works. In future IoT systems. For example, we can control the lights in the house by recognising the facial expression. Moreover, we can make use of the facial expression to establish the health monitoring system.

Lastly, we will make short discussion about security aspects. As we all know, image recognition is important and useful for our daily life, such as the face recognition. However, since hackers have begun looking for duplicate image to trick recognition systems, there are also great risks in image recognition, same as the FER. So, we should improve the recognition rate of facial expression to prevent the cyber attacks.

As we can see from Fig. 9, the models converge after iteration to a certain algebra. Taking the test set as an example, the following conclusions can be drawn from the analysis of Fig. 9 in our experiments. Firstly, as shown in Fig. 9, the proposed method began to converge after 25 generations of training, the convergence speed of the DeRL method was relatively slow, and the convergence speed of the CNN algorithm was slowest. Secondly, the proposed algorithm Adap-SE-ResNet can achieve higher recognition rate than other three methods after iteration to a certain extent. Therefore, it can be seen from Fig. 9 that the proposed approach can improve the recognition rate and be used for FER in complex background to a certain extent.

In summary, the proposed method outperforms the other algorithms in terms of the objective accuracy of the FER. From the experimental results, the proposed algorithm exhibits superior performance and ability to classify the facial expression than other recognition works. In addition, our proposed method can be used in future IoT systems. For example, we can control the lights in the house by recognising the facial expression. Moreover, we can make use of the facial expression to establish the health monitoring system.

Lastly, we will make short discussion about security aspects. As we all know, image recognition is important and useful for our daily life, such as the face recognition. However, since hackers have begun looking for duplicate image to trick recognition systems, there are also great risks in image recognition, same as the FER. So, we should improve the recognition rate of facial expression to prevent the cyber attacks.

As we can see from Fig. 9, the models converge after iteration to a certain algebra. Taking the test set as an example, the following conclusions can be drawn from the analysis of Fig. 9 in our experiments. Firstly, as shown in Fig. 9, the proposed method began to converge after 25 generations of training, the convergence speed of the DeRL method was relatively slow, and the convergence speed of the CNN algorithm was slowest. Secondly, the proposed algorithm Adap-SE-ResNet can achieve higher recognition rate than other three methods after iteration to a certain extent. Therefore, it can be seen from Fig. 9 that the proposed approach can improve the recognition rate and be used for FER in complex background to a certain extent.

In summary, the proposed method outperforms the other algorithms in terms of the objective accuracy of the FER. From the experimental results, the proposed algorithm exhibits superior performance and ability to classify the facial expression than other recognition works. In addition, our proposed method can be used in future IoT systems. For example, we can control the lights in the house by recognising the facial expression. Moreover, we can make use of the facial expression to establish the health monitoring system.

Lastly, we will make short discussion about security aspects. As we all know, image recognition is important and useful for our daily life, such as the face recognition. However, since hackers have begun looking for duplicate image to trick recognition systems, there are also great risks in image recognition, same as the FER. So, we should improve the recognition rate of facial expression to prevent the cyber attacks.

Table 4 Comparisons between our method and the state-of-the-arts FER algorithms on the JAFFE dataset

| Method                          | Accuracy |
|--------------------------------|----------|
| CNN [62]                        | 84.48    |
| salient feature [63]             | 90.00    |
| multi-level Haar wavelet [64]    | 90.56    |
| hierarchical deep learning method [58] | 91.27 |
| MCCNN [65]                      | 95.80    |
| Ms-RAN (ours)                   | 96.03    |

Table 5 Confusion matrix of the proposed method on the JAFFE dataset

|          | AN | DI  | FE  | HA  | SA  | SU  |
|----------|----|-----|-----|-----|-----|-----|
| AN       | 96.3 | 3.0 | 0.1 | 0.0 | 0.1 | 0.0 |
| DI       | 4.3 | 93.0 | 1.6 | 0.0 | 0.1 | 0.0 |
| FE       | 0.1 | 3.8 | 89.7 | 0.0 | 3.9 | 1.7 |
| HA       | 0.0 | 0.1 | 0.0 | 98.7 | 0.1 | 0.0 |
| SA       | 2.7 | 0.5 | 9.8 | 1.6 | 80.1 | 0.0 |
| SU       | 0.0 | 0.1 | 3.5 | 2.4 | 0.0 | 93.3 |

AN: angry, DI: disgust, FE: fear, HA: happy, SA: sad, SU: surprise.
The significance of bold is represented as the accuracy of each expression.

In this study, a novel Ms-RAN is proposed for FER. Our model mainly consists of the multi-scale convolution blocks. Each multi-scale convolution block contains two different scale sub-blocks. For each sub-block, we proposed the residual attention to connect the input and output, which is firstly proposed in this study. Besides, the skip connection is applied to combine the original input facial expression image and the output form multi-scale blocks. The experimental results validate that our proposed algorithm can improve the visual feature representation and achieve high-accuracy recognition than various state-of-the-art methods on public datasets. In the future, we will continue to study and analyse the problem of FER, and expect to find better rules and strategies for superior performance.
6 Acknowledgments

This work was supported in part by the Natural Science Foundation of Shandong Province of China under grant ZR2016AM30, Social Science Planning Research Project of Shandong Province under grant 18CLY150, in part by the Shandong Soft Science Research Program under grant 2020K01B783, and in part by The Project of Shandong Province Higher Educational Science and Technology Program under grant J15L115.

7 References

[1] Tian, Y.i, Kanade, T., Cohn, J.J.: ‘Recognizing action units for facial expression analysis’, IEEE Trans. Pattern Anal. Mach. Intell., 2001, 23, (2), pp. 6–11
[2] Liu, M., Shan, S., Wang, R., et al.: ‘Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition’. Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 2014
[3] Wang, Z., Wang, S., Ji, Q.: ‘Capturing Complex spatio-temporal relations among facial muscles for facial expression recognition’. Proc. of the Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013, pp. 3422–3429
[4] Baltrušaitis, T., Mahmoud, M., Robinson, P.: ‘Cross-dataset learning and person-specific normalization for automatic action unit detection’. 2015 IEEE Int. Conf. Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015, volume 6, pp. 1–6
[5] Shan, C., Gong, S., McOwan, P.W.: ‘Facial expression recognition based on local binary patterns: A comprehensive study’. Image Vis. Comput., 2009, 27, (6), pp. 803–816
[6] Chu, W., De la Torre, F., Cohn, J.: ‘Selective transfer machine for personalized facial expression analysis’. IEEE Trans. Pattern Anal. Mach. Intell., 2016, 38, (1), pp. 519–545
[7] Zhao, G., Pietikäinen, M.: ‘Dynamic texture recognition using local binary patterns with an application to facial expressions’. IEEE Trans. Pattern Anal. Mach. Intell., 2007, 29, (6), pp. 915–928
[8] Zhong, L., Liu, Q., Yang, P., et al.: ‘Learning active facial patches for expression analysis’. 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012.
[9] Zhu, R., Flierl, M., Ruano, Q., et al.: ‘Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition’. IEEE Trans. Syst. Man and Cybern., Part B (Cybernetics), 2011, 41, (1), pp. 38–52
[10] Sanina, A., Sandino, F., Harari, Y.M., et al.: ‘Spatiotemporal covariance descriptors for action and gesture recognition’. 2013 IEEE Workshop on Applications of Computer Vision (ACCV), Clearwater Beach, FL, USA, 2013, pp. 103–110
[11] Zhu, X., Ramanan, D.: ‘Face detection, pose estimation, and landmark localization in the wild’. Proc. of the Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 2012.
[12] Liu, M., Shan, S., Wang, R., et al.: ‘Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition’. IEEE Trans. Image Process., 2015, 25, (12), pp. 5920–5932
[13] Dhall, A., Goecke, R., Joshi, J., et al.: ‘Facial expression recognition from near-infrared videos’. Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[14] Wu, F., Bao, L., Chen, Y., et al.: ‘Self-supervised adaptation of high-fidelity face models for monocular performance tracking’. Proc. of the 2019 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019.
[15] Sun, B., Li, L., Zhou, G., et al.: ‘Facial expression recognition in the wild based on multimodal texture features’. J. Electron. Imaging., 2016, 25, (6), pp. 061407
[16] Li, J., Zhang, D., Zhang, J., et al.: ‘Facial expression recognition with faster R-CNN’. Comput. Sci., 2016, 2016, pp. 1–10
[17] Li, J., Tian, W., Yi, Z., et al.: ‘Using gabor filter in 3D convolutional neural networks for human face recognition’. Proc. of the 2017 36th Chinese Control Conf. (CCC), Da Lian, People's Republic of China, 2017
[18] Villanueva, M.G., Zavala, S.R.: ‘Deep neural network architecture: application for facial expression recognition’. IEEE Latin Am. Trans., 2020, 18, (7), pp. 1311–1317
[19] Chen, A., Xing, H., Wang, F.Y.: ‘A facial expression recognition method using deep convolutional neural network based on edge computing’. IEEE Access, 2020, 8, pp. 49741–49751
[20] Wang, K., Peng, X., Yang, H., et al.: ‘Region attention network for pose and occlusion robust facial expression recognition’. IEEE Trans. Image Process., 2020, 29, pp. 4057–4069
[21] Shi, M., Xu, L.J., Chen, X.: ‘A novel facial expression intelligent recognition method using improved convolutional neural network’. IEEE Access, 2020, 8, pp. 57606–57614
[22] Bruntha, P.M., Bhanasekar, S., Sagayam, K.M., et al.: ‘A modified approach for face recognition using PRO and ABC optimization’. Int. J. Innov. Technol. Exploring Exp., 2019, 8, (7), pp. 1571–1577, ISSN:2278-3075
[23] Sagayam, K.M., Panraj, D.N., Winton, J., et al.: ‘Authentication of biometric system using fingerprint recognition with Euclidean distance and neural network classifier’. Int. J. Technol. Exploring Exp., 2019, 8, (4), pp. 766–771
[24] Goodfellow, I.J., Erhan, D., Carrier, P.L., et al.: ‘Challenges in representation learning: a report on three machine learning contests’. Int. Conf. on Neural Information Processing, Daegu, Korea, 2013, pp. 809–816
[25] Ioffe, S., Szegedy, C.: ‘Batch normalization: accelerating deep network training by reducing internal covariate shift’, arXiv preprint arXiv:1502.03167, 2015
[26] Yang, B., Cao, J., Ni, R., et al.: ‘Facial expression recognition using weighted mixture deep neural network based on double-channel facial images’. IEEE Access, 2017, 6, pp. 4630–4640
[27] Kao, C.M., Lai, S.H., Surks, M.: ‘A compact deep learning model for robust facial expression recognition’. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018
[28] Yang, H., Citcif, U., Yin, L.: ‘Facial expression recognition by De-expression residue learning’. Proc. of the 2018 IEEE Conf. on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018
[29] Yu, Z.: ‘Image based static facial expression recognition with multiple deep network learning’. Proc. of the 2015 ACM Int. Joint Conf., Okasa, Japan, 2015
[30] Maziluru, S., C. Madureira, S., Cardoso, J.: ‘Improved facial expression recognition system using convolutional neural network’. Proc. of the 2016 Portuguese Conf. on Automatic Face and Gesture Recognition, Lisbon, Portugal, 2016
[31] Simonny, K., Zisserman, A.: ‘Very deep convolutional networks for large-scale image recognition’. ICLR, 2015
[32] Santurkar, S., Tsipras, D., Ilyas, A., et al.: ‘How does batch normalization help optimization? (no, it is not about internal covariate shift)’. Conf. on Neural Information Processing Systems, Montreal, Canada, 2018
[33] He, K., Zhang, X., Ren, S., et al.: ‘Identity mappings in deep residual networks’. European Conf. Computer Vision (ECCV), Amsterdam, The Netherlands, 2016
[34] Jain, S., Hu, C., Aggarwal, J.K.: ‘Facial expression recognition with temporal modeling of shapes’. Proc. of the IEEE Int. Conf. on Computer Vision Workshops, Barcelona, Spain, 2011
[35] Scovanner, P.: ‘3-dimensional sif descriptor and its application to action recognition’. ACM Multimedia, 2007
[36] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: ‘Generative adversarial nets’. Proc. of the Int. Conf. on Neural Information Processing Systems. Montreal, Canada, 2014
[37] Zhou, Y., Shi, B.E.: ‘Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder’. IEEE computer society., 2017
[38] Lu, C., Cohn, J.F., Kanade, T.: ‘The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-speciﬁed expression’. Proc. of the Computer Vision and Pattern Recognition Workshops (CVPRW) San Francisco, CA, USA, 2010
[39] Zhao, G., Huang, X., Tian, M., et al.: ‘Facial expression recognition from near-infrared videos’. Image Vis. Comput., 2019, 79, pp. 607–615
[40] Lyons, M., Akamatsu, S., Kamachi, M., et al.: ‘Coding facial expressions with gabor wavelets’. Proc. Third IEEE Int. Conf. Automatic Face Gesture Recognition, Nara, Japan, April 2000, pp. 200–205
[41] Pantie, M., Valstar, M., Rademaker, R., et al.: ‘Web-based database for facial expression analysis’. Proc. of the Multimedia and Exposition, Amsterdam, Netherlands, 2005
[42] Yin, L., Wei, X., Sun, Y., et al.: ‘A 3D facial expression database for facial behavior research’. Proc. of the Automatic Face and Gesture Recognition, Southampton, UK, 2006
[43] Zheng, Z., Girard, J.M., Yue, W., et al.: ‘Multimodal spontaneous emotion corpus for human behavior analysis’. Proc. of the 2016 IEEE Conf. on
Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016

[56] Klaser, A., Marszałek, M., Schmid, C.: ‘A spatio-temporal descriptor based on 3d-gradients’. BMVC 2008 19th British Machine Vision Conf., Leeds, UK, 2007, pp. 271–275

[57] Jung, H., Lee, S, Yim, J., et al.: ‘Joint fine-tuning in deep neural networks for facial expression recognition’. Proc. of the 2015 IEEE Int. Conf. on Computer Vision (ICCV), Santiago, Chile, 2015

[58] Zhang, Z., Luo, P., Chen, C.L., et al.: ‘From facial expression recognition to interpersonal relation prediction’, Int. J. Comput. Vis., 2008, 126, (5), pp. 1–20

[59] Kuo, C.M., Lai, S.H., Sarkis, M.: ‘A compact deep learning model for robust facial expression to interpersonal relation prediction’. 2018 IEEE-CVF Conf. Computer Vision Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 2018, pp. 2121–2129

[60] Liu, X., Kumar, B., You, J., et al.: ‘Adaptive deep metric learning for identity-aware facial expression recognition’. 2017 IEEE Conf. Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 522–531

[61] Lopes, A.T., de Aguiar, E., de Souza, A.F., et al.: ‘Facial expression recognition with convolutional neural networks: coping with few data and the training sample order’, Pattern Recognit., 2017, 61, pp. 610–628

[62] Liu, Y., Li, Y., Ma, X., et al.: ‘Facial expression recognition with fusion features extracted from salient facial areas’, Sensors, 2017, 17, (4), p. 712

[63] Goyani, M., Patel, N.: ‘Multi-level haar wavelet based facial expression recognition using logistic regression’, Indian J. Sci. Technol., 2017, 10, p. 9

[64] Kim, J.H, Kima, B.G., Roy, P.P., et al.: ‘Efficient facial expression recognition algorithm based on hierarchical deep neural network structure’, IEEE Access, 2019, 7, pp. 41273–41285

[65] Hamester, D., Barros, P., Wermter, S.: ‘Face expression recognition with a 2-channel convolutional neural network’. 2015 Int. Joint Conf. Neural Networks (IJCNN), Killarney, Ireland, 2015