Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Brief papers

Action unit analysis enhanced facial expression recognition by deep neural network evolution

Ruicong Zhi[a,b,*], Caixia Zhou[a,b], Tingting Li[a,b], Shuai Liu[a,b], Yi Jin[c]

[a] School of Computer and Communication Engineering, University of Science and Technology Beijing, No.30, Xueyuan Road, Haidian District, Beijing 100083, PR China
[b] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, PR China
[c] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, PR China

ARTICLE INFO

Article history:
Received 21 August 2019
Revised 17 February 2020
Accepted 11 March 2020
Available online xxx

Communicated by Prof. Zidong Wang

Keywords:
Facial action coding
Emotion recognition
Emotion inference
Adaptive subsequence matching
Relationship probability

ABSTRACT

Facial expression is one of the most powerful and natural signals for human beings conveying their emotional states and intentions. Recently, facial expression recognition from facial cues based on FACS is investigated, where facial expressions can be described by a subset of AUs, and the facial expression categories could be easily extended. The goal of our work is the proposal of an action unit analysis enhanced facial expression recognition system based on evolutional deep learning approach. The main contributions of our work include the following three aspects: (1) the temporal dynamic based 3DLeNets is exploited for video analysis based AUs detection. And a general evolutionary framework is conducted for the deep neural networks optimization. (2) The correlations among AUs and the correlation between AUs and emotions are investigated, and the relationship probability model between AUs and emotions is derived by the concept of discriminative power coefficients. (3) An adaptive subsequence matching algorithm (ASMA) is adopted to measure the similarity between AUs sequences, so that to construct the inference scheme of mapping AUs to emotions. Experimental results proved that the AUs enhanced facial expression recognition system performs well comparing to existing facial expression analysis methods, and each AU has different contribution roles for different facial expressions. It is also found to be more practical than discrete facial expression recognition as most of the facial expressions can be described using a subset of AUs.

1. Introduction

Human emotion understanding is crucial for modern human-computer interaction and artificial intelligence. Facial expression is one of the most powerful and natural signals for human beings conveying their emotional states and intentions [1]. Facial expression recognition (FER) has wide applications in affective computing, and it has attracted increasing attention in the last decades. It is a great challenge to create computer systems that can automatically analyze human emotions inspired by complex social contexts.

Presently, discrete emotion categories are popular in affective analysis, and the widely used example is the six primary emotions, including anger, fear, sadness, disgust, surprise, and happiness, based on the underlying assumption that humans express the primary emotions universally regardless of cultures. While other taxonomies extend these primary emotions with attitudinal affective states, such as depression, fatigue, boredom, frustration, joy, relief etc. [2–4]. Another popular facial behavior measurement is the Facial Action Coding System (FACS) which is a sign judgment method. Facial expression categories describe facial behaviors in a global way, while facial action units depict the local variations on face. FACS is a comprehensive and anatomical system which could encode various facial movements by the combination of basic AUs (Action Units), and makes the emotion categories much wider.

Although the six discrete basic emotions is the most popular perspective for facial expression recognition, the category of discrete basic emotions is still researched by psychologists, and the number of emotion types is increasing. The affective model based on basic emotions is limited in the ability of representing the complexity and subtlety of our daily affective displays [5]. Recently, facial expression recognition from facial cues based on FACS is investigated, and there are two phases involved, that is, the facial action units are detected first, and then the emotions are inferred from the detected AUs. The AUs are considered as building blocks of facial expressions in the cascaded FER system. It is more practical

* Correspondence author at: School of Computer and Communication Engineering, University of Science and Technology Beijing, No.30, Xueyuan Road, Haidian District, Beijing 100083, PR China.
E-mail address: zhirc@ustb.edu.cn (R. Zhi).

https://doi.org/10.1016/j.neucom.2020.03.036
0925-2312/

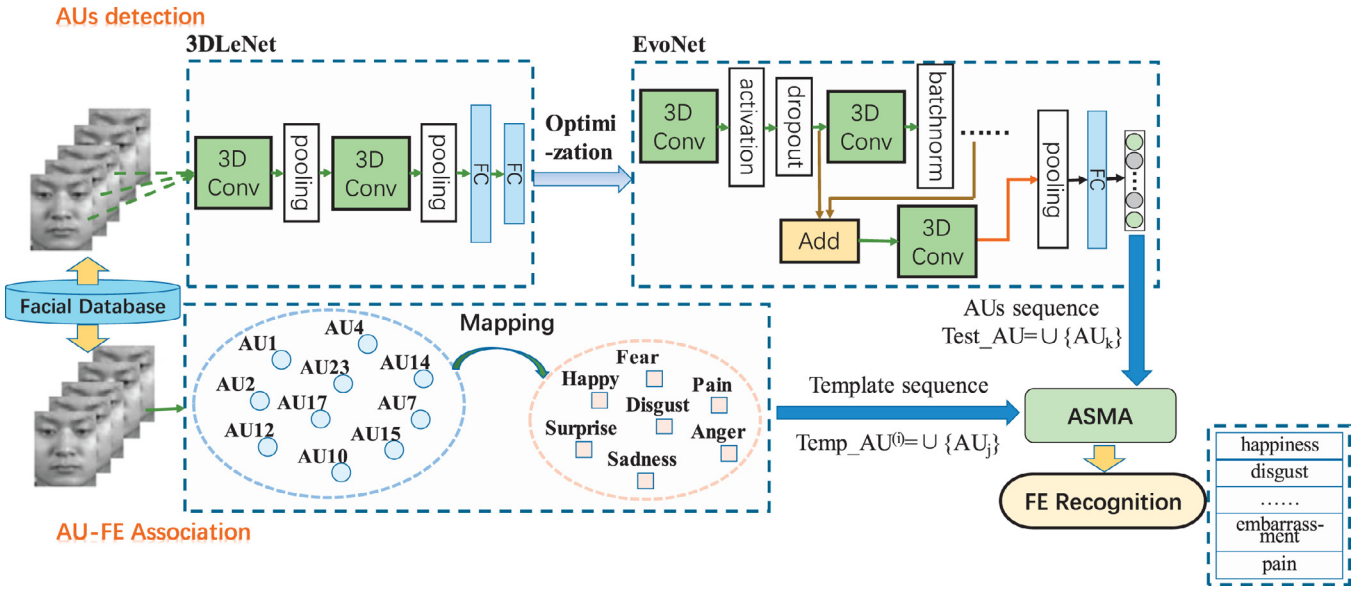


Fig. 1. The framework of our proposed method.

than discrete basic emotion analysis system, as facial expressions can be described by a subset of AUs, and the facial expression categories could be easily extended (there are around 7000 emotions in practice). Besides, the requirement for the amount of independent training data is reduced.

In this work, we present an AUs analysis enhancement based emotion inference method to map a set of AUs to one or more final emotions using a learned statistical relationship and a suitable matching measure. The main contributions of our work include the following three aspects: (1) the temporal dynamic based 3DLeNets is exploited for video analysis based AUs detection. And a general evolutionary framework is conducted for the deep neural networks optimization. (2) The correlations among AUs and the correlation between AUs and emotions are investigated, and the relationship probability model between AUs and emotions is derived by the concept of discriminative power coefficients. (3) An adaptive subsequence matching algorithm (ASMA) is adopted to measure the similarity between AUs sequences, so that to construct the inference scheme of mapping AUs to emotions.

An overview of the proposed method is shown in Fig. 1. Firstly, the AUs image sequences are input to a spatiotemporal self-learning network (called EvoNet) which is optimized automatically for the topology and hyper-parameters by evolutional scheme for 3D convolutional neural network (3DLeNet), so that to learn dynamic facial features to better represent facial characteristics. Secondly, the correlation among AUs and correlation between AUs and emotions are investigated to explore the relationship probability matrix, and the AUs template sequence for each emotion is determined based on the relationship probability matrix, which is derived by discriminative power coefficients. So that specific emotion could be represented by several AUs. Thirdly, the AUs sequences predicted by EvoNet are compared to the AUs template sequences by adaptive subsequence matching algorithm, and the emotion category is inferred by the similarity between two AUs sequences which is measured by the total probability of all the AU elements in the sequences.

The remainder of the paper is organized as follows: Section 2 reviews the related work of facial expression recognition, Section 3 introduces the evolution principle of 3DLeNets, Section 4 presents the inference scheme of AUs mapping to emotions, Section 5 illustrates the experimental results and analysis, and Section 6 provides the conclusions and future work.

2. Related work

Traditional automatic facial expression recognition system consists of several fundamental components, including face detection, feature extraction and classification. Recently, multi-view face detection and pose estimation attracted more interests for nature scenario. On the other hand, deep learning methods were reported to perform face detection task [6,7].

Numerous methods for automatic FER could be roughly divided into two categories: designed method and self-learning method. Extracting powerful features and designing effective classifiers are two key components of traditional FER system, and hand-crafted facial features are often used to represent facial expression images. Self-learning algorithms extract facial representations through a joint feature learning and classification pipeline. Facial expressions could be represented through appearance-based, geometry-based and hybrid-based aspects, and appearance-based representation is the vast applied features in both designed and self-learning methods. A great number of appearance based features are well performed, such as Gabor wavelets, Histograms of Oriented Gradients (HOG), Scale-Invariant Feature Transform (SIFT), Local Binary Pattern (LBP) and their variances (LDTP, LDP), LBP-TOP.

Feature extraction and classification are conducted separately in manually designed methods, and there is no direct guidance for the performance of FER system. Self-learning representation is a different pipeline which automatically learns features from the training data hierarchically. The most well-established self-learning representation is deep learning, which learns multi-layered hierarchical representations in low-level to high-level manner, and jointly in combination of the classifier [8]. Recently, deep neural networks (DNNs) have increasingly been leveraged to learn discriminative representations for automatic facial expression recognition.

Although vast of deep learning architectures have been proposed for specific tasks, convolutional Neural Networks (CNNs) are the most popular tool for facial expression recognition. Fasel [9] explored convolutional neural networks (CNNs) for multi-scale facial expression recognition. Jung et al. [10] trained a deep network which is based on temporal appearance features and temporal geometry features independently and fuse them through a fully connected layer by joint fine-tuning. Vielzeuf et al. [11] combined the VGG and LSTM model to conduct video emotion analysis.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

Zhang et al. [12] trained an end-to-end deep learning model with some generated expressions with different poses, which leaded to a better performance on facial expression recognition. Çuğu et al. [13] proposed a lightweight model named MicroExpNet for facial expression recognition. Their model was trained following the concept of knowledge distillation and performed well in both run time and memory requirements. Lopes et al. [14] proposed a combination of CNNs and special image pre-processing steps (C−CNNs) to recognize six facial expressions under head pose, and the accuracy was 90.96% on the BU-3DFE dataset. Tang et al. [15] trained ten facial AUs separately by fine-tuning the VGG-Face model [16]. Yang et al. [17] fused two CNNs models trained separately using the gray images and LBP images separately to improve the performance of FER. Hasani and Mahoor [18] combined the CNNs with Conditional Random Fields (CRF) to capture the spatial relation within facial images and the temporal relation between the image frames. Yang et al. [19] proposed a FER system by extracting information of the expression component through expression learning procedure, called De-expression Residue Learning (DeRL).

Moreover, facial micro-expression recognition is more challenging as it is momentary involuntary facial expression which usually occurs in high-stake situations, when people undergo but do not intend to express. Recently, facial micro-expression recognition has attracted increasing interest in computer vision, such as machine learning methods and deep learning methods. For example, Ben et al. [20] designed a maximum margin projection with tensor representation (MMPTR) algorithm to extract discriminative and geometry-preserving features from data, and the experiments were conducted on CASME facial micro-expression database. Kim et al. [21] proposed a feature representation which was learned with expression-states by using CNNs and LSTM recurrent neural networks and achieved 60.98% accuracy on CASME II.

Despite the powerful feature learning ability of deep learning, there are still several problems remained. Firstly, a large amount of training data is needed to fully extract facial features and avoid overfitting. Training deep networks with limited data may even result in poor performance due to over-fitting. Referring to the first problem, lack of large scale of facial samples for DNN training is a big problem for FER. To address the problem of insufficient expression data, data augmentation [14] and transfer learning [22–24] are two common means. Secondly, the parameters optimization of deep neural network is complex, and the direct guidance of the results is insufficient. The construction of neural networks is designed deeper and deeper to improve the ability of tackling big-data problems, which leading to an increasing number of parameters. To implement DNN optimization automatically, a number of neural network evolution algorithms have been developed over the last decade. With the popularization of deep learning, NeuroEvolution which refers to methodologies that aim at the automatic search and optimization of the NNs' parameters, has been vastly used [25]. Generally, the automatic generation of deep neural networks can be grouped into three categories: evolution of the network parameters [26], evolution of the network topology [27], and evolution of both the topology and parameters [28,29].

FER system based on AUs consists of two steps including facial action unit detection and facial expression inference, and few researches focused on the indirect AUs enhanced FER system due to the challenge of AUs detection. Emotion-specified facial expressions are not included in FACS, it is crucial to determine the inference rules of mapping AUs to emotions.

Tong et al. [30] modeled the temporal evolution and the semantic relationship among different AUs through a Dynamic Bayesian Network (DBN). The drawback is the high complexity and it cannot be used in real time systems. Khorrami et al. [31] discussed the corresponding relationships between the high-level features learned by CNNs and the facial action units (AUs). Yang et al. [32] proposed a three-layer restricted Boltzmann machine (RBM) to capture the probabilistic dependencies among facial expressions and AUs. Pumarola et al. [33] even make use of AUs to generate vivid human expressions. Kaneko and Okada [34] built an action unit-based linked data from distribution map for facial emotion recognition, and enhancing the usefulness of the data by merging them with other linked data. Salah and Aleksandar [35] proposed an action unit based FER by CNNs using 23 deep classifiers, according to the most relevant facial action units and combinations. SVRs were utilized to map AUs activations to two values, i.e. arousal and valence. Jia et al. [36] proposed a multi-layer fusion system based on AUs with a stacking framework, and the association rules were employed to mine the relationships between AUs and facial expressions. Velusamy et al. [37] proposed a Longest Common Subsequence (LCS) distance to match the AUs strings, so that to infer facial expressions from AUs sequences.

3. Evolution of 3D convolutional neural network

3.1. 3D CNNs architecture

Different from the traditional 2DCNNs, which can only input two-dimensional images for feature extraction, the 3DCNNs can learn facial features in spatio-temporal domain simultaneously from the input dynamic video sequence. The 3DCNNs is implemented by applying a 3D convolutional kernel with a depth of k, which can be thought as consisting of k 2D convolutional kernels. These 2D convolutional kernels are firstly convolved on k adjacent images to obtain feature maps, then the feature maps are weighted and summed so that to obtain one output feature map. Since the weights of the convolutional kernels in the entire cube are shared, a convolutional kernel corresponds to one type of feature.

Generally, the value at position (x, y, z) in the jth output feature map in the ith layer is calculated as follows:

$$v_{ij}^{xyz} = \sigma\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} + b_{ij}\right) \tag{1}$$

where w_{ijm}^{pqr} is the (p, q, r)th value of the filter connected to the mth output feature map in the previous layer. P_i and Q_i are the height and width of the 3D filter along the temporal dimension, and the nonlinear activation function $\sigma(x) = \max(0, x)$ is utilized in our model, named rectified linear units (ReLU).

In this research, the classic LeNet-5 [38] network with 3D convolutional layers is utilized, which is called as 3DLeNet, and the network structure is illustrated in Fig. 2. Comparing to LeNet with 2D convolution, 3D convolution can preserve temporal information of the input signals and extract the spatiotemporal features from contiguous frames, which is very important descripting the dynamic characteristics of AUs occurrence. Moreover, LeNet-5 network was well applied model with good performance on image processing tasks. The structure of LeNet is simple designed by using local connection patters and imposing constraints on the weights, and the weight sharing technique could reduce the number of free parameters, to that to reduce the capacity of the machine and reduce the gap between test error and training error. Therefore, the structure of 3DLeNet is based on conditional simple and effective LeNet-5 while the convolutional layer is updated with 3D convolution, so that to deal with dynamic feature representation problem in AUs detection task.

In the 3DLeNet structure, there are seven layers consisted in 3DLeNet except input layer, that is three convolutional layers, two subsampling layers, two fully-connected layers, and the classification results are output through the output layer. The 3D convolution operation is also shown in Fig. 2, connections of the same

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

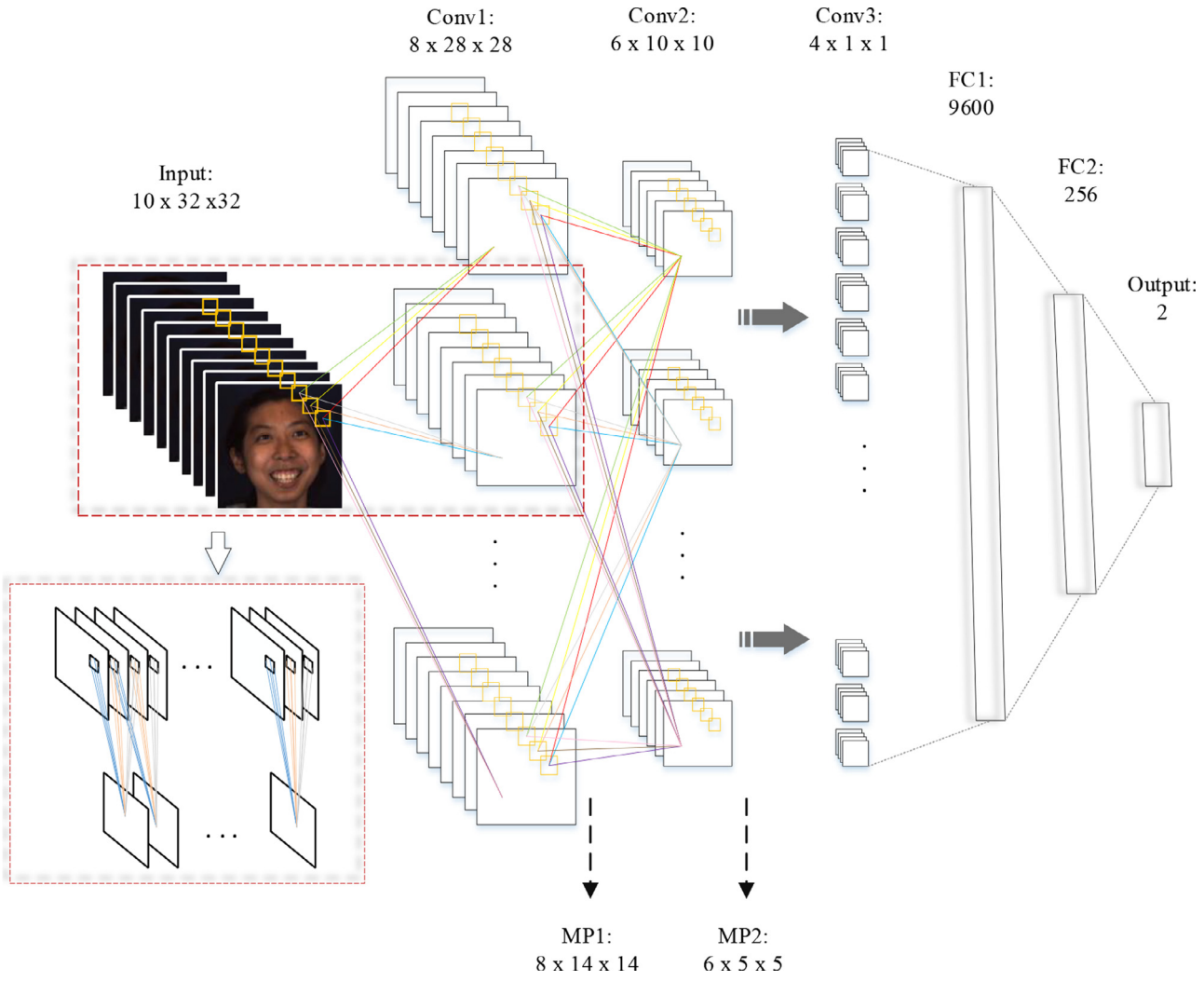


Fig. 2. Structure illustration of 3DLeNet.

color in the same layer represent weight sharing. The architecture shown in Fig. 2 is adopted in AUs detection, i.e. the output of the network represents two states of certain AU's presentation (occurrence and absence). There are several parameters that can be adjusted and optimized in the network, such as learning rate, penalty coefficient, dropout rate during training, and the number of training iterations. In addition, the convolutional kernel size (height, width, and depth) also has impact on the performance of DNN model. The detailed parameters optimization is discussed in the experimental section.

3.2. Evolution of multi-layered neural network

The Deep Neural Network (DNN) optimization is implemented based on evolutionary algorithm. The evolutionary deep neural network (EDNN) represents the structure of network by an ordered sequence of feedforward layers, and the parameters associated to the layers. Genetic Algorithms (GAs) and Dynamic Structured Grammatical Evolution (GSGE) are combined to encode the neural network structure at two different levels (i.e., DENSER [29]), corresponding to layer sequences design and parameters learning. In this research, the 3DLeNet is optimized following the DENSER evolution framework. The main steps in deep neural network evolution include: Firstly, encode the macro structure of the 3DLeNet network and their respective parameters through a user-defined Context-Free-Grammar (CFG). A population is randomly initialized where the individuals in the population are the coding of neural network. Secondly, a fitness function is defined according to the AUs detection task, and each individual is evaluated by the fitness function. Thirdly, several variation operators are conducted to generate offspring from the parent through crossover and mutation, including layer lever and parameters lever. Finally, the optimal solution is selected when the evaluation of individual is satisfied, and the solution is evolutionary deep neural network (which is called EvoNet).

The main component of evolutionary procedure is detailed as follows:

Representation: The deep neural network contains several feedforward layers with different functions, and several parameters need to be set to the layers. An innovative encoding manner is applied in neural network definition, including GA level for layer structure design and DSGE level for parameter setting. The DSGE is a form of Genetic Programming, which consists of a list of genes, one for each non-terminal symbol, and a variable length representation is used. The CFG is defined as [29]: a CFG is a tuple $G = (N, T, S, P)$, where N is a non-empty set of non-terminal symbols, T is a non-empty set of terminal symbols. S is the starting symbol, and P is the set of production rules of the form $A ::= \alpha$, with $A \in N$ and $\alpha \in (N \cup T)^*$. N and T is disjoint. Language $L(G)$ is defined as $L(G) = \{w : S \overset{*}{\Rightarrow} w, w \in T^*\}$, which is composed by all sequences of terminal symbols that can be derived from the starting symbol.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

The coding form of each individual is composed of network structure definition, and network parameters definition. For example, the valid sequence of layers can be specified as [(feature, 1,10), (classification, 1,2), (softmax, 1,1), (learning, 1,1)], where each tuple indicates the valid starting symbols, and the minimum and maximum number of times they can be used. The parameters of each layer are encoded by valid values or ranges. For example, the parameters of pooling layer include kernel size, stride, and padding, and the value of padding can be set to SAME or VALID, while stride can be set as [int, 1,1,3].

The input parameters of the evolutionary deep neural network are: the grammar describing the network structure; the maximum depth of each non-terminal symbol; the genotype (which is initially empty); the non-terminal symbol that we want to expand (the start symbol is used initially); and the current sub-tree depth (initialized to 0).

Fitness evaluation: the candidate solutions of EDNN need to be evaluated during the evolution by fitness function. As the EDNN model is applied for AUs analysis which is a typical classification task, the accuracy on the validation set is used as the fitness function.

Fitness function is used to compare the performance of individual after multiple variation operators, and it is the main clue for offspring selection during evolution. The evolution process of the neural network is terminated when the fitness function reaches the goal set by user, or the number of iteration steps achieves the upper limit which is also set by user.

Variation operators: two kinds of variation operators are involved in the evolution procedure, i.e. crossover and mutation.

Crossover is utilized to recombine two parents to generate two offspring, and there are two types of crossover applied on the GA level, i.e. one-point crossover and bitmask crossover. The one-point crossover is conducted to change layers within the same module, while bit-mask crossover is utilized to exchange modules between individuals. The module means a set of layers belonging to the same network structure, for example classification module.

Mutation is utilized to change the genetic material upon two evolution levels respectively, i.e. GA level and DSGE level. The GA level mutation deal with the structure of neural network, and the layer sequences are updated by mutation operator implemented on layers, including add layer, replicate layer, and remove layer. The DSGE level mutation treats the content of each layer, and the parameters respective to layers are changed within valid value scope.

- GA level mutation: the layer sequence design could be conducted through three operators: add layer (generates a new layer randomly while not violate the maximum number of layers), replicate layer (copies an existing layer to another position of the module), and remove layer (delete a layer from a module randomly while not violate the minimum number of layers).
- DSGE level mutation: the parameters of layers respective to each module are updated in two manners: valid value mutation (the parameter value is replaced by another valid value), or numeric mutation (integer or float value is replaced by a new generated value). For example, the parameter padding mutated from SAME to VALID, the kernel size changed from 3×3 to 4×4, and float value can be generated randomly with Gaussian perturbation.

4. Emotion recognition based on AUs estimation

4.1. Deriving relationship between AUs

The relationship between different AUs is investigated by a statistical discriminative coefficient-I (SDC-I) derived from facial database containing sufficient instances with AUs label. The SDC-I is calculated based on the label of different instances, and defined as

$$SDC_I = P(X_j|X_i) \tag{2}$$

where $P(X_j|X_i) = P(X_jX_i)/P(X_i)$. X_i and X_j denote action unit AU_i and AU_j respectively. $P(X_j|X_i)$ is the conditional probability of X_j when action unit X_i has occurred. $P(X_i)$ is the probability of X_i, and $P(X_jX_i)$ is the joint probability of when X_i and X_j occur simultaneously. Particularly, $P(X_i) = H_i/S$, $P(X_jX_i) = A_i/S$, where S represents the total number of facial images in the facial database, H_i is the number of facial images with label AU_i, and A_i is the number of facial images with both label AU_i and AU_j in the facial database.

The SDC-I is utilized to investigate the pairwise correlation of different AUs. The higher correlation between AUs is, the larger the discriminative coefficient is. So the SDC-I could reflect the relationship between AUs. The relation matrix is derived by normalization of SDC-I, and it is a basic clue for emotion inference from AUs which is described detailed later.

4.2. Deriving relationship between AUs and emotions

Similar scheme is applied to investigate the relationship between AUs and emotions. A variance of statistical discriminative coefficient-II (SDC-II) is defined on facial database containing sufficient instances with both AUs label and emotion label. The SDC-II is calculated as

$$SDC_{\mathrm{II}} = P(X_j|Y_i) \tag{3}$$

where $P(X_j|Y_i) = P(X_jY_i)/P(Y_i)$. Y_i denotes emotion type T_i, and X_j denote action unit AU_j. $P(X_j|Y_i)$ is the conditional probability of X_j when emotion Y_i occurs. $P(Y_i)$ is the probability of Y_i, and $P(X_jY_i)$ is the joint probability of when X_j and Y_i occur simultaneously. Particularly, $P(Y_i) = F_i/S$, $P(X_jY_i) = G_i/S$, where F_i is the number of facial images with label T_i, and G_i is the number of facial images with both label AU_j and T_i in the facial database.

The SDC-II could reflect whether an AU increases or decreases the probability of mapping to an emotion. The AUs-emotion relation matrix is derived from the normalized SDC-II, and the weights are associated to the emotional template selection which is crucial for inferring emotion from AUs.

The relationships between AUs and emotions are the basis of facial expression recognition. A series of AUs template sequence could be derived for each facial expression, the main procedure is: the AUs are sorted from high to low according to the discriminative coefficients between AUs and facial expressions (SDC-II), and the top-r AUs are selected to form the template sequences. r is the length of template sequence which is preset, for example, r is set to five in our work. The AUs template sequences play a key role in facial expression recognition which could identify facial expressions by AUs analysis.

4.3. Adaptive rule-based prediction for facial expression recognition

Emotion-specified facial expressions are not included in FACS, it is crucial to determine the inference rules of mapping AUs to emotions. In this section, we propose an adaptive subsequence matching algorithm (ASMA) to measure the similarity between AUs sequence of test sample and AUs template sequences of facial expressions, and make rules forming emotion inference from AUs. The ASMA algorithm is illustrated in Fig. 3. Let the AUs sequence of a test sample denote as $\text{Test\_AU} = \cup\{AU_k\},\ k \subset \{1, 2, \ldots, L\}$, where k is subscript denoting the type of AU in the test AUs sequence, L is the number of AUs classes, K is the length of AUs sequence of the test sample. The template sequence of facial expression T_i is denoted as $\text{Temp\_AU}^{(i)} = \cup\{AU_j\},\ i = 1, 2, \ldots, R,\ j \subset \{1, 2, \ldots, L\}$,

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

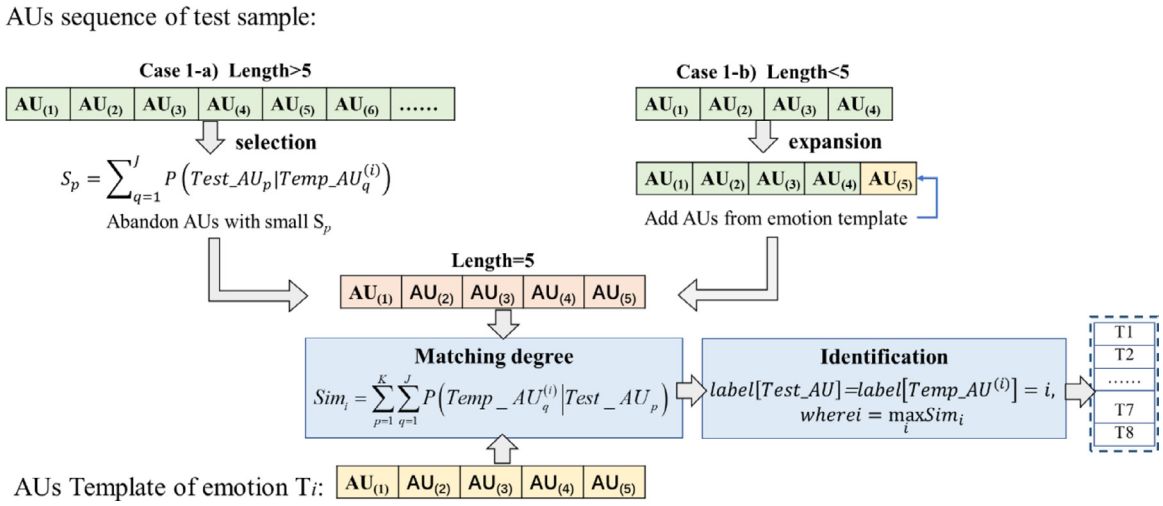


Fig. 3. Illustration of ASMA algorithm (the subscript of AU denotes the AUs involved in the sequence, not the FACS coding number).

where R is the number of facial expression classes, j is the subscript denoting the type of AU in the template sequence, J is the length of AUs template sequence of facial expression.

The main steps of the ASMA include:

Case 1: when $K \neq J$, the length of AUs sequence of test sample need to be normalized according to (a) or (b), then the target matching template could be found by the method described in case 2.

(a) if $K > J$, it means that the amount of AUs in test sample's AUs sequence is more than that in template sequence of facial expression. In that case, the redundant AUs in the test sample's AUs sequence need to be removed. The selection rule for removing AUs is as follows:

For each AU element in Test_AU $= \cup\{AU_k\}$, e.g. AU_p, calculate the correlation index between AU_p and each AU element (e.g. AU_q) in emotion template Temp_AU$^{(i)} = \cup\{AU_j\}$ using $S_p = \sum_{q=1}^{J} P(Test\_AU_p | Temp\_AU_q^{(i)})$. Therefore, there are K correlation indexes for all the AU elements of Test_AU. Sort the correlation values and find the AU element $Test\_AU_p$ corresponding to the smallest S_p, and delete the AU element AU_p, until the lengths of two AUs sequences (Test_AU and Temp_AU$^{(i)}$) are the same (i.e. $K = J$). The same procedure is repeated for all the facial expression patterns, by calculating the matching degree between the elements of test sample's AUs sequence and facial expression template AUs sequence, individually.

(b) if $K < J$, it means that the length of test sample's AUs sequence is shorter than that in template sequence of facial expression. It is necessary to expand the AUs sequence of test sample by adding AU elements. When calculating the matching degree between Test_AU and Temp_AU$^{(i)}$, randomly selected AU elements from Temp_AU$^{(i)}$ and add them to Test_AU, so that the length of Test_AU becomes the same as Temp_AU$^{(i)}$.

Case 2: when $K = J$, i.e. the amount of AUs elements in test sample's AUs sequence and template sequence of facial expression is the same. It is ready to find the target matching template for the test sample. There are two steps for the optimal solution:

(1) Compare the two AUs sequences Test_AU and Temp_AU$^{(i)}$, and summary the number of identical AUs. The facial expression template sequence which has the largest number of identical AUs with AUs sequence of test sample is the candidate optical solution of target matching template. If the optimal solution satisfying the condition is unique, then the target matching template of facial expression is determined.

(2) If the candidate optimal solution obtained in 1) is not unique, further judgment is required.

The matching degree between the test sample's AUs sequence Test_AU and the template AUs sequence Temp_AU$^{(i)}$ of the ith facial expression is defined as

$$Sim_i = \sum_{p=1}^{K}\sum_{q=1}^{J} P\left(Temp\_AU_q^{(i)} | Test\_AU_p\right) \tag{4}$$

where $Test\_AU_p$ represents the AU element in Test_AU, and $Temp\_AU_q^{(i)}$ represents the AU element in Temp_AU$^{(i)}$. The similarity between the two AUs sequences is measured by the total probability of all the AU elements in the sequences.

The optimal matching template is determined according to the matching degree derived by Eq. (4), that is, the template AUs sequence which achieves the highest matching degree with the test sample's AUs sequence is the optimal matching template. And the facial expression label of the optimal matching template is the identified label for the test sample, i.e.

$$label[Test\_AU] = label\left[Temp\_AU^{(i)}\right] = i,\ where\ i = \max_i Sim_i.$$

5. Experiments and analysis

5.1. Datasets

The experiments were conducted on the well applied facial expression databases with labels for both AUs and facial expressions. The DISFA database and BP4D database were utilized to evaluate the performance of our proposed method for AUs detection and facial expression recognition. Both DISFA and BP4D database are spontaneous facial expression databases with AUs label, and BP4D is the commonly used facial database in FERA sub-challenge for AUs occurrence detection and intensity estimation. Fig. 4 shows some examples of the AUs and facial sequences.

The DISFA (Denver Intensity of Spontaneous Facial Action) dataset [39] (http://www.engr.du.edu/mmahoor/DISFA.htm) is recorded at University of Denver. Twenty-seven adults with different ethnicity are video-recorded while they viewed video clips intended to elicit spontaneous emotion expression. Each video frame is manually coded for the presence of facial action units, also for intensities of action units on a five ordinal scale. The number of events

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

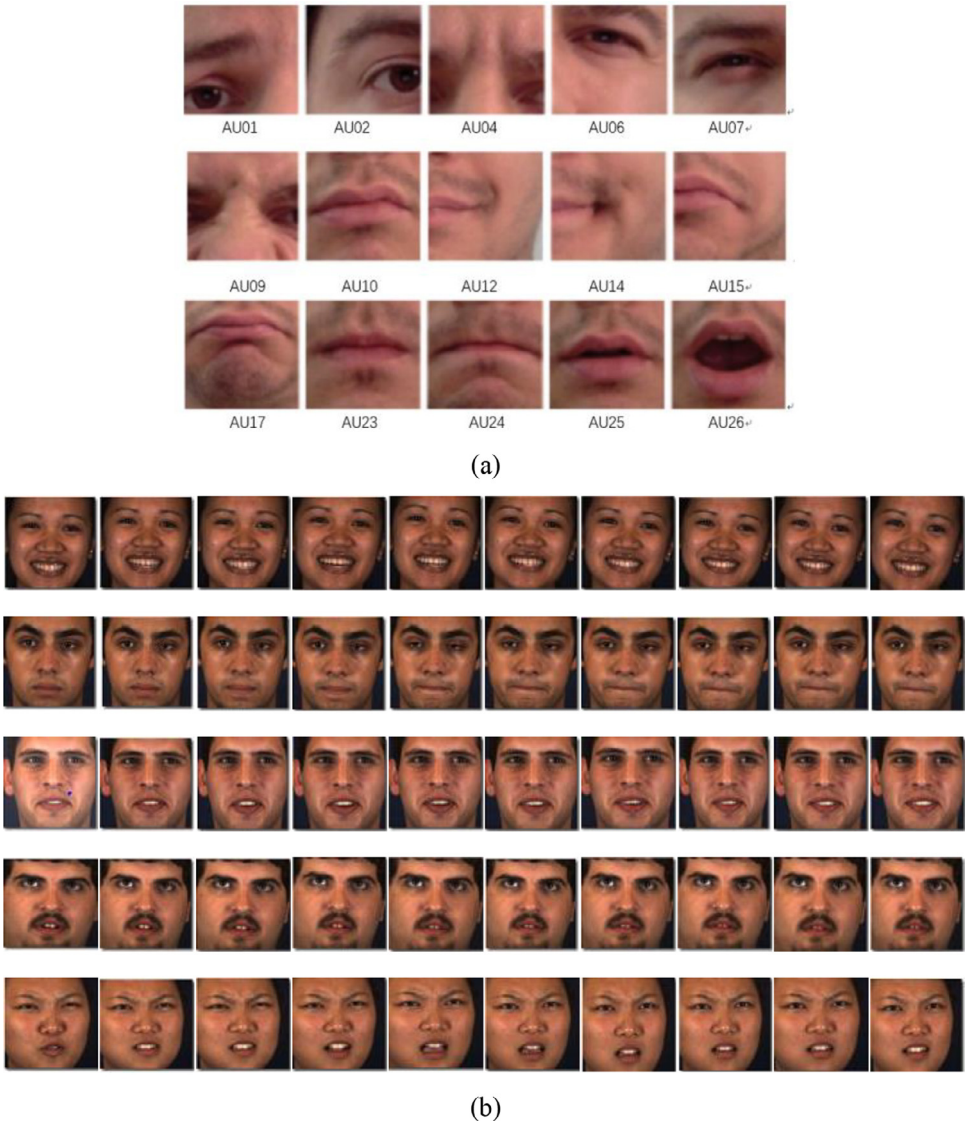


Fig. 4. Examples of facial images. (a) action units (b) image sequences.

and frames for each intensity level of each AU are reported. Event refers to the continuous occurrence of an AU from its onset (start frame) to its offset (end frame).

BP4D database [40] is a dynamic 3D video database of spontaneous facial expressions of 41 subjects (http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html). A series of effective tasks are conducted for authentic emotion induction, including face-to-face interview, social games, documentary film watching, cold pressor task, social anger induction, and experience of smell. The target emotions include happiness or amusement, sadness, surprise or startle, embarrassment, fear or nervous, physical pain, anger or upset, and disgust. Facial action units are annotated by frame-level, and the onset and offset of action units are coded.

There are 12 action units coded in DISFA database, i.e. AU1, AU2, AU4, AU5, AU6, AU9, AU12, AU15, AU17, AU20, AU25, and AU26. And the BP4D database annotated 12 action units for AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24.

Deep neural networks require a large amount of training data, and the existing facial expression databases are not sufficient to train the well-known neural network with deep architecture that achieved the most promising results in facial expression recognition tasks. In this paper, a variance of Generative Adversarial Networks which converges more stably and is easy to train was utilized, that is, Boundary Equilibrium Generative Adversarial Networks (BEGAN). The BEGAN was trained on BP4D facial dataset to generate facial samples, and the new facial samples were obtained on step 40,000. Data argumentation was conducted for AUs facial images and the dataset scale was expanded to four times larger than the original dataset. Moreover, the dataset was balanced for each AU type of positive and negative samples.

5.2. Experimental design

In our experiments, the dataset was divided into three subsets, i.e. training set, verification set and testing set. The training set and verification set were utilized for deep neural network optimization, and the performance of the deep learning model was verified on testing set.

To evaluate the generalization ability and robustness of the learning model, the person-independent scheme was utilized in the experiments, namely, the subjects of each subsets were not overlapped. For DISFA database, eight subjects were selected randomly, and their facial samples were utilized to form training set, facial samples of four subjects were used to form verification set,

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

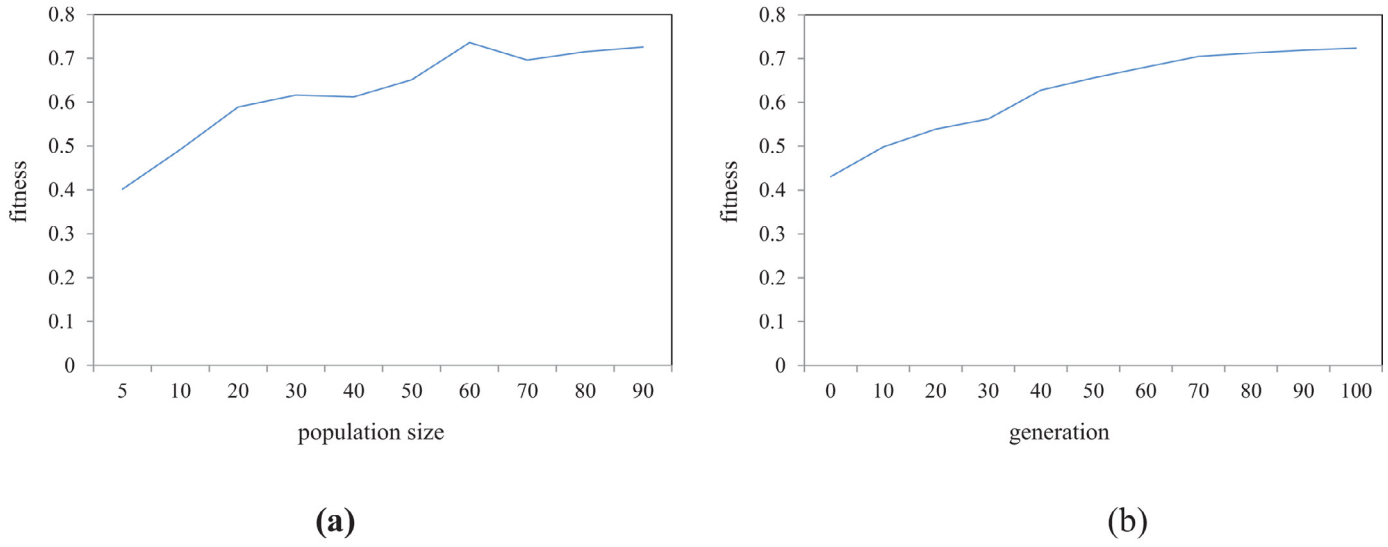


Fig. 5. Evolution of the fitness across the number of generations (a) and initial population (b).

and the remaining 15 subjects were conducted as testing. For BP4D dataset, 13 subjects out of 41 subjects were selected to obtain training set, and the verification set and testing set contained 10 subjects and 18 subjects respectively. The length of facial expression sequence was different. Therefore, 3D spline interpolation was applied to normalize the length of input sequence to 10 frames. Moreover, the facial size was normalized to 32 × 32 by taking consideration of both accuracy and efficiency.

The AUs detection is a typical binary classification task, that is, for certain action unit, it may haves one of the labels including the occurrence of AU (1) and the absence of AU (0). Two indicators could be utilized to evaluate the performance of algorithms, i.e. accuracy and F1-score. F1-score is a weighted mean of the recall and precision values, and it is better suited for imbalanced dataset.

5.3. Evolution of 3DLeNet

5.3.1. Evolutionary topology of 3DLeNet

The evolution of deep learning model was implemented based on Keras, i.e., the grammatical derivation of defined grammar was fed to a Keras model running on top of TensorFlow. Keras was specified with its modularity which could combine modules freely with low cost. The basic genes were available to form various combinations to get better network structure. Fitness was utilized to evaluate the superiority of the individual, and the bigger the fitness of individual was, the better the neural network model was. High accuracy was expected for the AU detection task, therefore the classification accuracy was chosen as the fitness function to assess the generalization and scalability ability of the neural networks.

The optimization of network during evolution is effected by generation, and it is obvious that the performance of network increases versus generation increasing. However, the training of deep neural networks is computationally expensive, and it is necessary to determine the proper generation parameter due to the hardware limitation. Moreover, the number of initial population is also investigated in the experiments. The fitness evolution of the 3DLeNet networks across the generations and initial population number were depicted in Fig. 5. The results showed that the performance of network is steadily increasing versus generation, and a change in fitness curve occurred when the initial population number was 60. The parameters were set by take consideration of the tradeoff both network performance and computational cost.

Table 1
Experimental parameters setting.

Parameter	Value	Network parameter	Value
Number of generations	100	Convolution filter size	3×5×5
Population size	60	Learning rate	0.001
Crossover rate	0.7	dropout	0.5
Mutation rate	0.3	Decay rate	0.95
Tournament size	3	Decay steps	1000
Elite size	0.01	Batch size	100

The experimental parameters used were detailed in Table 1, and the GA structure was set as: [(feature, 1, 30), (classification, 1, 10), (softmax, 1, 1)]. The 3DLeNet was trained with batches of 150, and varying learning rate policy was employed, i.e. the initial learning rate was 0.001, and it was varied along with decay steps.

The structure of 3DLeNet for AUs detection was optimized by evolutionary scheme and the best neural network which achieved the best accuracy on verification set was obtained as illustrated in Fig. 6. The figure depicted the best topology of the 3DLeNet network.

5.3.2. Performance of AUs occurrence detection

Experiments were implemented to evaluate the performance of optimized network in AUs occurrence detection. The well applied facial expression databases were utilized, i.e. DISFA and BP4D. The original 3DLeNet and the optimized network by evolution (EvoNet) were trained on the two facial databases individually. The experimental results on the testing set were compared with accuracy and F1-score.

Fig. 7 illustrated the average accuracies of AUs occurrence detection on DISFA database. It showed that the overall performance of the EvoNet outperformed 3DLeNet for all the AUs, especially that the accuracies of AU1 and AU2 classification were enhanced more than 10%. The average accuracy of all the AUs obtained by EvoNet was 87.2% while it was 84.6% for 3DLeNet. The F1-scores of 3DLeNet and EvoNet on DISFA database were recorded in Table 2. As shown in the table, the optimized EvoNet methods led to higher F1-scores than the original network (bracketed and bold numbers indicate the best performance; bold numbers indicate the second best).

The results were also compared to various AUs analysis algorithms including machine learning and deep learning methods. Zhao et al. [41] proposed deep region and multi-label Learning

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

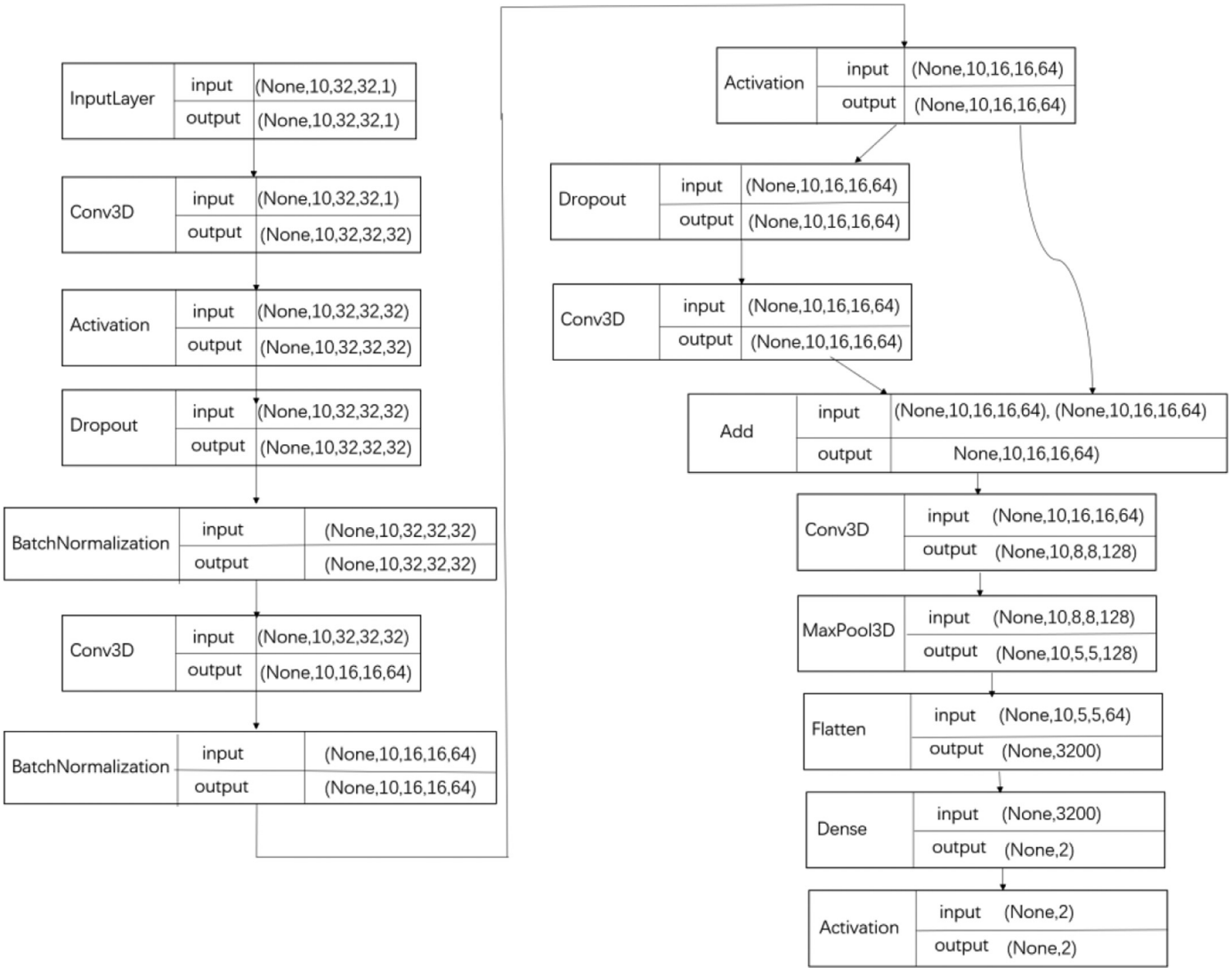


Fig. 6. Topology of the EvoNet.

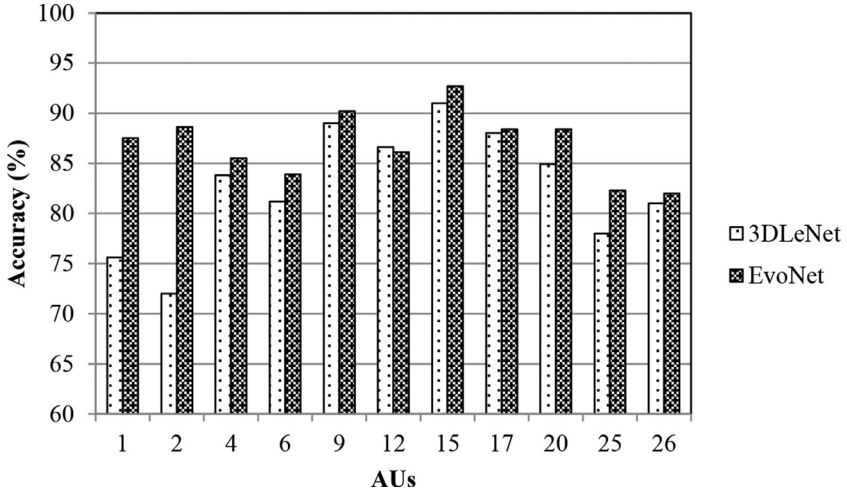


Fig. 7. Accuracy comparison between 3DLeNet and EvoNet on DISFA.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

Table 2
Comparison of algorithms on DISFA by F1-score.

AU	LSVM	APL	DRML	Alexnet	2DLeNet	3DLeNet	EvoNet
1	0.11	0.11	0.17	0.12	0.12	**0.18**	**[0.21]**
2	0.10	0.12	**0.18**	0.12	0.12	0.17	**[0.21]**
4	0.22	0.30	0.37	0.28	0.29	**0.36**	**[0.39]**
6	0.16	0.12	0.29	0.23	0.21	**0.28**	**[0.31]**
9	0.12	0.10	0.11	0.12	0.12	**0.13**	**[0.14]**
12	**0.70**	0.66	0.38	0.30	0.31	0.69	**[0.72]**
25	0.12	0.21	0.39	0.44	0.41	**0.47**	**[0.48]**
26	0.22	0.27	0.20	0.28	0.28	**0.32**	**[0.32]**
Avg	0.22	0.24	0.27	0.24	0.23	**0.33**	**[0.35]**

Table 3
Comparison of algorithms on BP4D by accuracy.

AU	LSVM	JPML	DRML	EAC	JAA	3DLeNet	EvoNet
1	0.207	0.407	0.557	0.689	**[0.747]**	0.710	**0.739**
2	0.177	0.421	0.545	0.739	**0.808**	0.784	**[0.814]**
4	0.229	0.462	0.588	0.781	**0.804**	0.757	**[0.823]**
6	0.203	0.400	0.566	0.785	**[0.789]**	0.746	**0.775**
7	0.448	0.500	0.610	0.690	**0.710**	0.685	**[0.741]**
10	0.734	0.752	0.536	0.776	**0.802**	0.784	**[0.823]**
12	0.553	0.605	0.608	0.846	**[0.854]**	0.812	**0.846**
14	0.468	0.536	0.570	0.606	**0.648**	0.616	**[0.661]**
15	0.183	0.501	0.562	0.781	0.831	**0.845**	**[0.860]**
17	0.364	0.425	0.500	0.706	**0.735**	**0.718**	**[0.747]**
23	0.192	0.519	0.539	0.810	**0.823**	0.769	**[0.829]**
24	0.117	0.532	0.530	0.824	**0.854**	0.832	**[0.876]**
Avg	0.322	0.505	0.560	0.752	**0.784**	0.755	**[0.800]**

Table 4
Comparison of algorithms on BP4D by F1-score.

AU	LSVM	JPML	DRML	EAC	JAA	3DLeNet	EvoNet
1	0.23	0.33	0.36	0.39	**[0.47]**	**0.41**	**0.47**
2	0.23	0.26	0.42	0.35	**[0.44]**	0.43	**[0.46]**
4	0.23	0.37	0.43	0.49	**0.54**	0.50	**[0.55]**
6	0.27	0.42	0.55	0.76	**[0.78]**	0.62	**0.77**
7	0.47	0.51	0.67	0.73	**0.75**	0.69	**[0.79]**
10	0.77	0.72	0.66	0.82	**0.84**	0.71	**[0.84]**
12	0.64	0.74	0.66	0.86	**[0.87]**	**0.83**	**0.85**
14	0.64	0.66	0.54	0.59	0.62	0.57	**[0.68]**
15	0.18	0.38	0.33	0.38	0.44	**0.45**	**[0.57]**
17	0.33	0.40	0.48	0.59	**0.60**	0.59	**[0.62]**
23	0.19	0.30	0.32	0.36	0.43	**0.44**	**[0.45]**
24	0.21	0.43	0.32	0.36	0.42	0.41	**[0.44]**
Avg	0.35	0.46	0.48	0.56	**0.60**	0.55	**[0.62]**

(DRML) algorithm for AUs detection, and a region layer was introduced to induce important facial regions related AUs. The results of DRML were better than other two schemes, i.e. LSVM proposed by Fan et al. [42] and APL proposed by Zhong et al. [43]. Table 2 depicted that the F1-score obtained by 3DLeNet increased for most of the AUs comparing to DRML up to 0.1. Moreover, the EvoNet yielded the best performance among various AUs detection algorithms, which enhanced the F1-score by around 0.02 for all the AUs comparing to 3DLeNet.

Both 3DLeNet and EvoNet were trained on the BP4D database using 12 AUs, and the accuracies and F1-scores were reported in Tables 3 and 4, respectively (bracketed and bold numbers indicate the best performance; bold numbers indicate the second best). The accuracy of AUs achieved by EvoNet increased obviously, especially for AU4, AU7, AU10, AU14, and AU23. There was an average enhancement of 4.5% comparing to 3DLeNet. The F1-scores for AUs acquired by EvoNet were higher than that of 3DLeNet, and the average F1-score was 0.62 and 0.55 for EvoNet and 3DLeNet individually.

The results were also compared with a number of well applied algorithms for AUs analysis on BP4D database, including LSVM [42], JPML [44], DRML [41], EAC [45], and JAA [46]. JAA was proposed by Shao et al., which utilized an adaptive attention learning module to refine the attention map of each AU adaptively, the assembled local features were integrated with face alignment features and global features for AUs detection. JAA performed better for most of the schemes, and our proposed 3DLeNet, except AU15 and AU23. The EvoNet outperformed 3DLeNet and JAA for most of the AUs, the average accuracy of EvoNet was 80.0%, which was 4.8% higher than that of EAC-net, and 1.6% higher than that of JAA. The average F1-score of EvoNet was the highest (0.62), while the F1-scores for JAA and 3DLeNet were 0.60 and 0.55, respectively. It can be seen that the optimized network through evolution was effective for AUs occurrence detection and it outperformed state-of-art researches.

Table 5
AUs templates for emotions.

Emotion No.	Action Unit					Target emotion
T1	10	12	6	7	23	Happiness
T2	4	17	7	1	15	Sadness
T3	2	12	10	6	24	Surprise
T4	12	10	6	7	14	Embarrassment
T5	4	7	6	1	14	Fear
T6	7	10	14	6	17	Physical pain
T7	23	17	14	24	7	Anger
T8	7	17	14	6	4	Disgust

5.4. Association diagram construction between AUs

This section mainly discussed the correlation among AUs, and the correlation between AUs and facial expressions. The association diagram among AUs was constructed on both DISFA and BP4D databases, while the association diagram between AUs and facial expressions was constructed on BP4D database, as AUs label was available on DISFA and both AUs label and facial expression label were available on BP4D.

Fig. 8 showed the correlation matrices on DISFA and BP4D databases which indicated element-wise probability between AUs. The pairwise correlation was calculated according to statistical discriminative coefficient (SDC-I) defined in Section 4.1. The pairwise relationship indicated that once an AU occurred, the connected AU could be deduced to also occur. For example, if we had detected the occurrence of AU1, then we could infer that AU2 also occur with high probability. Fig. 9 showed the relation matrix calculated for the eight emotions elicited in BP4D. The positive intensity (black) indicates high probability associated for an AU belonging to an emotion and the negative intensity (white) indicates high probability for an AU not belonging to an emotion. The AUs template of certain emotion was derived from the AUs-emotion relation matrix, based on the ranking results of SDC-II.

5.5. AUs enhanced facial expression recognition

In our experiments, the length of AUs template for emotion was set to 5. For each emotion, the top five AUs of highly discriminative AUs were selected to form the AUs template, and the ASMA algorithm was adopted for facial expression recognition. The entries of AUs template for eight emotions contained in BP4D database were shown in Table 5. The top five positive associations of happiness emotion composed the AUs template of happiness, i.e. AU6, AU7, AU10, AU12, and AU23. The AUs templates for emotions identified highly relevant facial actions and their association weights for various emotions, so that the emotions could be inferred by the combination of AUs flexibly. In ASMA, the AUs sequence length of test sample and template was adaptively adjusted based on the pairwise correlation between AUs as discussed in Section 5.4.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

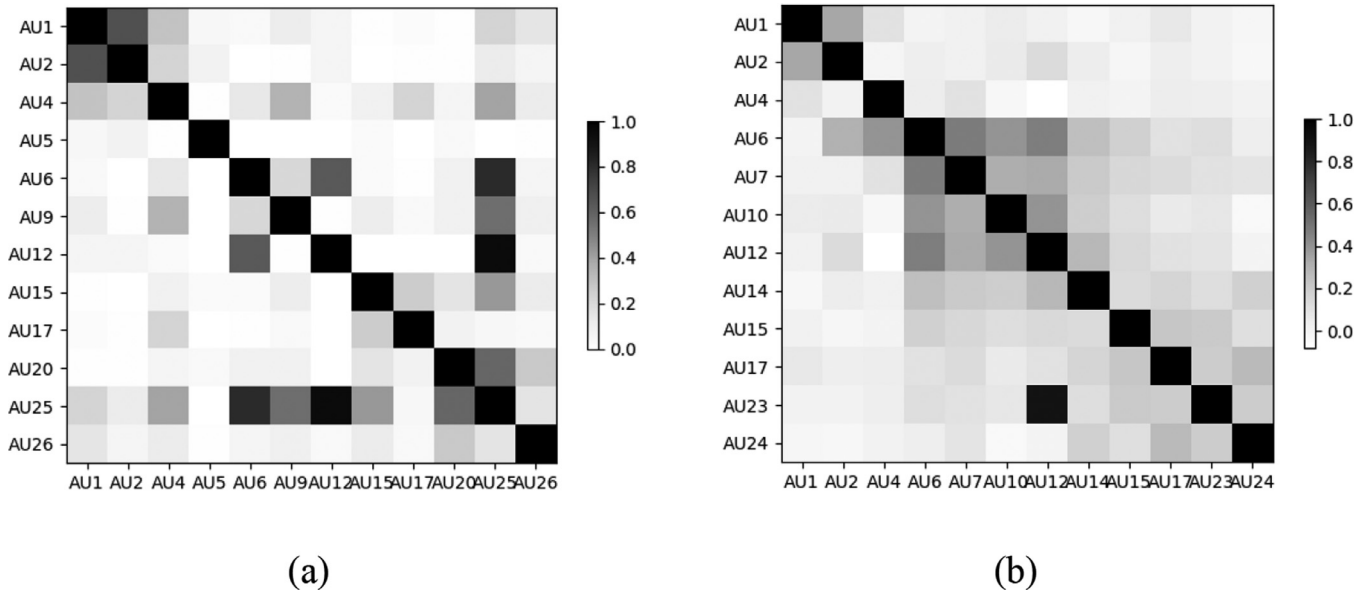


Fig. 8. Association diagrams between AUs (a) DISFA (b) BP4D.

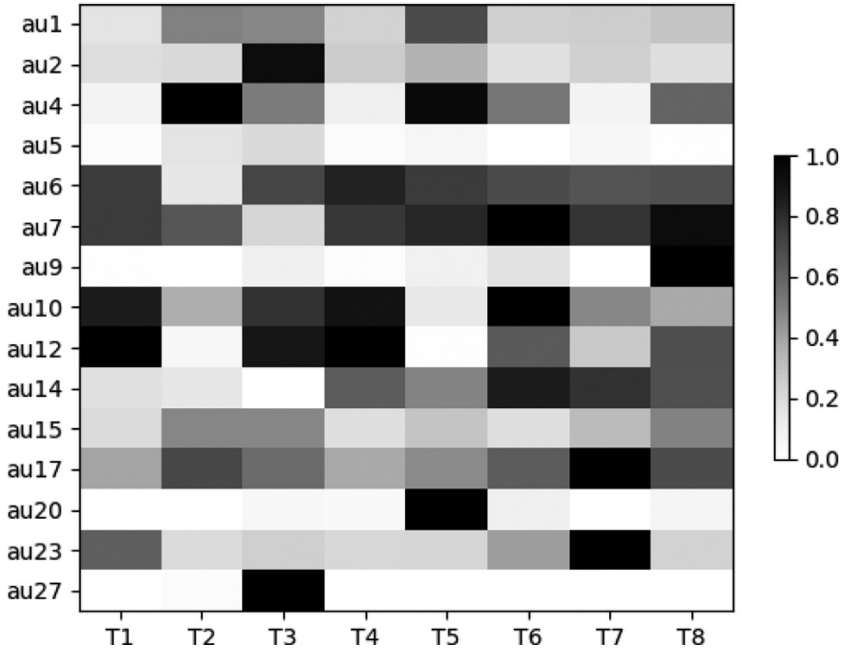


Fig. 9. Relation matrix of AUs and emotions on BP4D database.

Table 6
The association of six emotional expressions to AUs [47].

Emotional category	Primary AUs					Auxiliary AUs				
Happiness	**6**	**12**				*25*	*26*	*16*		
Sadness	**1**	**15**	**17**			**4**	**7**	*25*	*26*	
Disgust	**9**	10				**17**	*25*	*26*		
Surprise	*5*	*26*	*27*	1 + **2**						
Anger	2	4	**7**	**23**	**24**	**17**	*25*	*26*	*16*	
Fear	*20*	**1** + *5*	*5* + **7**			**4**	*5*	**7**	*25*	*26*

Table 6 was a summary of primary AUs and auxiliary AUs associated with six emotional expressions, which was investigated by Zhang and Ji [47] extending Ekman's work [48]. Primary AUs mean those AUs that can be clearly classified as or are strongly pertinent to one of the six facial expressions without ambiguities, while an auxiliary AU is the one that can be only additively combined with primary AUs to provide supplementary support to the facial expression classification [47]. There are 12 AUs well annotated in BP4D database (i.e. AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23, and AU24). Therefore, several AUs listed in Table 6 could not be detected automatically, such as AU5, AU9, AU16, AU20, AU25, and AU26. Comparing the emotional AUs templates (Table 5) and relationship between AUs and facial expressions in Table 6, it could be seen that our results were consistent with that investigated in psychology. All the primary AUs and auxiliary AUs were involved in the emotional AUs templates for happiness, sadness, surprise, and fear. The anger's primary AUs of AU2 and AU4, the disgust's primary AUs of AU10 was not involved. The consistence between our study and psychology study provided an evidence of the emotional AUs template, and it was reasonable to applied for emotional expression recognition.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036

Table 7
Confusion matrix for emotion recognition.

		Actual label							
		T1	T2	T3	T4	T5	T6	T7	T8
Predictive label	T1	**0.86**	0.00	0.08	0.09	0.00	0.00	0.00	0.00
	T2	0.00	**0.84**	0.00	0.00	0.05	0.04	0.00	0.00
	T3	0.12	0.00	**0.89**	0.00	0.06	0.00	0.03	0.00
	T4	0.00	0.05	0.00	**0.85**	0.00	0.05	0.02	0.06
	T5	0.00	0.04	0.03	0.00	**0.89**	0.00	0.00	0.00
	T6	0.00	0.07	0.00	0.00	0.00	**0.89**	0.10	0.11
	T7	0.00	0.00	0.00	0.00	0.00	0.01	**0.85**	0.00
	T8	0.02	0.00	0.00	0.06	0.00	0.01	0.00	**0.83**

Table 8
Comparison of facial expression recognition accuracies on BP4D.

Method	Overall Accuracy
Dapogny et al. [49]	0.724
Zhang et al. [40]	0.710
Dapogny et al. [51]	0.768
Danelakis et al. [50]	0.856
Zhen et al. [52]	0.817
Perveen et al. [53]	0.812
Proposed	**0.863**

The proposed AUs based facial expression recognition algorithm was conducted on BP4D database for eight emotion identification (embarrassment and pain plus six basic facial expressions were involved in the recognition system), and an average accuracy of 85% was achieved. The confusion matrix was listed in Table 7. The overall accuracies for all the eight facial expressions were higher than 80%. Table 7 indicated that happiness (T1) was easily confused with surprise (T3), embarrassment (T4) was easily confused with disgust (T8), and anger (T7) was easily confused with physical pain (T6). It could see that the entries in AUs template of emotions are partially overlapped (Table 5), such as the AU6, and AU12 were shared by T1 and T3, and AU7, AU14, AU17 were shared by T6 and T7. Therefore, the accuracies of related emotions were influenced on a degree.

Moreover, the comparison results of various facial expression recognition schemes were illustrated in Table 8. Our proposed mapping technique outperformed the rest schemes very well on facial expression recognition. The average accuracy of our proposed methods achieved 86.3%, which was higher than some of the methods by over 10% [49,40], while the best performance of the compared algorithms was 85.6% [50]. The primary advantage of the proposed method was that the relationship between AUs and facial expressions was specific, and the method did not require large look-up tables to accommodate combinations of AUs that might occur due to errors in AUs detection. The adaptive rule-based AUs to emotions mapping method could be easily applied to various emotions not limited to the basic facial expressions. And the AUs template with weight association helped to infer emotions effectively and flexibly.

6. Conclusion

This work aims to propose an action unit analysis enhanced facial expression recognition system based on evolutional deep learning approach. The 3D convolutional neural network is optimized by evolutional principles, and the facial expressions are predicted by jointly exploiting statistical probability correlation and adaptive subsequence matching algorithm.

The proposed algorithm was evaluated on facial expression databases, and the experimental results showed the effectiveness of our method. The EvoNet yielded the best performance among various AUs detection algorithms, which achieved an accuracy of up to 87.2% and 80.0% on DISFA database and BP4D database respectively. The AUs enhanced facial expression recognition problem performed well and the accuracy was higher than the compared methods by ove10%. The comparison with existing techniques showed the suitability of the proposed method, and although there was only eight facial expressions prediction in the experiments, it could be expanded to multiple facial expressions easily, and it is an effective method to infer emotions flexibly. In the future, we would extend the method to inferring more emotions by considering temporal dynamic features explored by EvoNet, and more facial databased will be utilized for performance validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Ruicong Zhi: Conceptualization, Methodology, Writing - review & editing, Supervision. **Caixia Zhou:** Data curation, Validation, Writing - review & editing. **Tingting Li:** Visualization, Investigation, Writing - original draft. **Yi Jin:** Writing - review & editing.

Acknowledgments

This work was supported by the National Research and Development Major Project [grant numbers: 2017YFD0400100], the National Natural Science Foundation of China [grant numbers: 61673052], the Fundamental Research Fund for the Central Universities of China [grant numbers:FRF-GF-19-010A, FRF-TP-18-014A2, FRF-IDRY-19-011], and the grant from Chinese Scholarship Council (CSC).

References

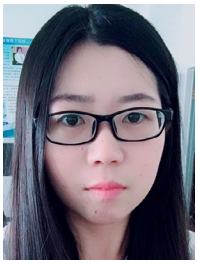
[1] P. Ekman, E.L. Rosenberg, What the Face reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), 2nd edition, Oxford University Press, 2005.
[2] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58, doi:10.1109/TPAMI.2008.52.
[3] J.M. Girard, J.F. Cohn, M.H. Mahoor, S.M. Mavadati, Z. Hammal, D.P. Rosenwald, Nonverbal social withdrawal in depression: evidence from manual and automatic analysis, Image Vis. Comput. 32 (10) (2014) 641–647, doi:10.1016/j.imavis.2013.12.007.
[4] G.M. Lucas, J. Gratch, S. Scherer, J. Boberg, G. Stratou, Towards an affective interface for assessment of psychological distress, in: Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 539–545, doi:10.1109/ACII.2015.7344622.
[5] B. Martinez, M.F. Valstar, Advances, challenges, and opportunities in automatic facial expression recognition, in: Advances in Face Detection and Facial Image Analysis, Springer, 2016, pp. 63–100, doi:10.1007/978-3-319-25958-1_4.
[6] H. Wu, K. Zhang, G. Tian, Simultaneous face detection and pose estimation using convolutiona neural network cascade, IEEE Access. 6 (2018) 49563–49575, doi:10.1109/ACCESS.2018.2869465.
[7] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: a deep multi-task learning framework for face detection, landmark, localization, pose estimation, and gender recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (1) (2019) 121–135, doi:10.1109/TPAMI.2017.2781233.
[8] R. Zhi, M. Liu, D. Zhang, A comprehensive survey on automatic facial action unit analysis, Vis. Comput. (2019) 1–27, doi:10.1007/s00371-019-01707-5.
[9] B. Fasel, Head-pose invariant facial expression recognition using convolutional neural networks, in: Proceedings of the IEEE International Conference on Multimodal Interfaces, 2002, doi:10.1109/ICMI.2002.1167051.
[10] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991, doi:10.1109/ICCV.2015.341.
[11] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the Nineteenth ACM International Conference on Multimodal Interaction, 2017, pp. 569–576, doi:10.1145/3136755.3143011.
[12] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3359–3368.

[13] İ. Çuğu, E. Şener, E. Akbşa, Microexpnet: an extremely small and fast model for expression recognition from frontal face images. (2017) 1–9. https://arxiv.org/pdf/1711.07011.pdf.
[14] A.T. Lopes, E.D. Aguiar, A.F.D. Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit 61 (2016) 610–628, doi:10.1016/j.patcog.2016.07.026.
[15] C. Tang, W. Zheng, J. Yan, Q. Li, Y. Li, T. Zhang, Z. Cui, View-independent facial action unit detection, in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition, 2017, pp. 878–882, doi:10.1109/FG.2017.113.
[16] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: Proceedings of the British Machine Vision Conference (BMVC), 2015, pp. 41.1–41.12.
[17] B. Yang, J. Cao, R. Ni, Y. Zhang, Facial expression recognition using weighted mixture deep neural network based on double-channel facial images, IEEE Access 99 (2017).
[18] B. Hasani, M.H. Mahoor, Spatiotemporal facial expression recognition using convolutional neural network and conditional random fields, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2017, pp. 790–795, doi:10.1109/FG.2017.99.
[19] H. Yang, U.A. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168–2177, doi:10.1007/978-94-017-0295-9_7.
[20] X. Ben, P. Zhang, R. Yan, M. Yang, G. Ge, Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation, Neural Computing and Applications 27 (8) (2016) 2629–2646, doi:10.1007/s00521-015-2031-8.
[21] D.H. Kim, W.J. Baddar, Y.M. Ro, Micro-expression recognition with expression-state constrained spatio-temporal feature representations, in: Proceedings of the 2016 ACM on Multimedia Conference, 2016, pp. 382–386, doi:10.1145/2964284.2967247.
[22] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade, Stacked deep convolutional auto-encodes for emotion recognition from facial expressions, in: Proceedings of the International Joint Conference on Neural Networks, 2017, pp. 1586–1593, doi:10.1109/IJCNN.2017.7966040.
[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. arXiv: 1512.03385.
[24] R. Zhi, H. Xu, M. Wan, T. Li, Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition, IEICE Trans. Inf. Syst. E102-D (5) (2019) 1054–1064.
[25] S.S. Tirumala, S. Ali, C.P. Ramesh, Evolving deep neural networks: a new prospect, in: Proceedings of the Twelfth International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2016, pp. 69–74, doi:10.1109/FSKD.2016.7603153.
[26] G. Morse, K.O. Stanley, Simple evolutionary optimization can rival stochastic gradient descent in neural networks, in: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference, 2016, pp. 477–484, doi:10.1145/2908812.2908916.
[27] K. Soltanian, F.A. Tab, F.A. Zar, I. Tsoulos, Artificial neural networks generation using grammatical evolution, in: Proceedings of the Twenty-first Iranian Conference on Electrical Engineering (ICEE), 2013, pp. 1–5, doi:10.1109/IranianCEE.2013.6599788.
[28] A.J. Turner, J.F. Miller, Cartesian genetic programming encoded artificial neural networks: a comparison using three benchmarks, in: Proceedings of the Fifteenth Annual Conference on Genetic and Evolutionary Computation, 2013, pp. 1005–1012, doi:10.1145/2463372.2463484.
[29] F. Assuncao, N. Lourenco, P. Machado, B. Ribeiro, DENSER: deep evolutionary network structured representation, Genet. Progr. Evol. Mach. 20 (1) (2019) 5–35, doi:10.1007/s10710-018-9339-y.
[30] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699, doi:10.1109/TPAMI.2007.1094.
[31] P. Khorrami, T.L. Paine, T.S. Huang, Do deep neural networks learn facial action units when doing expression recognition? in: Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW), 2015, pp. 19–27, doi:10.1109/ICCVW.2015.12.
[32] J. Yang, S. Wu, S. Wang, Q. Ji, Multiple facial action unit recognition enhanced by facial expressions, in: Proceedings of the Twenty-third International Conference on Pattern Recognition (ICPR), 2016, pp. 4078–4083, doi:10.1109/ICPR.2016.7900274.
[33] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: anatomically-aware facial animation from a single image, in: Proceedings of the Fifteenth European Conference on Computer Vision (ECCV), 2018, pp. 835–851, doi:10.1007/978-3-030-01249-6_50.
[34] K. Kaneko, Y. Okada, Action unit-based linked data for facial emotion recognition, in: Proceedings of the International Conference on Active Media, 2013, pp. 211–220, doi:10.1007/978-3-319-02750-0_22.
[35] Salah Al-Darraji, K. Berns, Aleksandar Rodić, Action unit based facial expression recognition using deep learning, in: Proceedings of the International Conference on Robotics in Alpe-adria Danube Region, 2017, pp. 413–420, doi:10.1007/978-3-319-49058-8_45.
[36] X. Jia, S. Liu, D. Powers, B. Cardiff, A multi-layer fusion-based facial expression recognition approach with optimal weighted AUs, Appl. Sci. 7 (112) (2017) 1–23, doi:10.3390/app7020112.
[37] S. Velusamy, H. Kanna, B. Anand, A. Sharma, B. Navathe, A method to infer emotions from facial action units, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, doi:10.1109/ICASSP.2011.5946910.
[38] Y. LÉcun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, doi:10.1109/5.726791.
[39] S.M. Mavadati, M.H. Mahoor, K. Bartlett, P. Trinh, J.F. Cohn, DISFA: a spontaneous facial action intensity database, IEEE Trans. Affect. Comput. 4 (2) (2013) 151–160, doi:10.1109/T-AFFC.2013.4.
[40] X. Zhang, L. Yin, J.F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, J.M. Girard, BP4D-Spontaneous: a high-resolution 3D spontaneous dynamic facial expression database, Image Vis. Comput. 32 (10) (2014) 692–706, doi:10.1016/j.imavis.2014.06.002.
[41] K. Zhao, W. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3391–3399, doi:10.1109/CVPR.2016.369.
[42] R. Fan, K. Chang, C. Hsieh, X. Wang, C. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (9) (2008) 1871–1874, doi:10.1145/1390681.1442794.
[43] L. Zhong, Q. Liu, P. Yang, J. Huang, D.N. Metaxas, Learning multi-scale active facial patches for expression analysis, IEEE Trans. Cybern. 45 (8) (2015) 1499–1510, doi:10.1109/TCYB.2014.2354351.
[44] K. Zhao, W. Chu, F. De la Torree, J.F. Cohn, H. Zhang, Joint patch and multi-label learning for facial action unit and holistic expression recognition, IEEE Trans. Image Process. 25 (8) (2016) 3931–3946, doi:10.1109/TIP.2016.2570550.
[45] W. Li, F. Abtahi, Z. Zhu, L. Yin, EAC-Net: a region-based deep enhancing and cropping approach for facial action unit detection, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2017, pp. 103–110.
[46] Z. Shao, Z. Liu, J. Cai, L. Ma, Deep adaptive attention for joint facial action unit detection and face alignment, in: Proceedings of the European Conference on Computer Vision, 2018.
[47] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 699–714.
[48] P. Ekman, W.V. Friesen, Facial Action Coding System (FACS): Manual, Consulting Psychologists Press, Palo Alto, Calif, 1978.
[49] A. Dapogny, K. Bailly, S. Dubuisson, Dynamic facial expression recognition by joint static and multi-time gap transition classification, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2015, pp. 1–6.
[50] A. Danelakis, T. Theoharis, I. Pratikakis, P. Perakis, An effective methodology for dynamic 3D facial expression retrieval, Pattern Recognit. 52 (2016) 174–185, doi:10.1016/j.patcog.2015.10.012.
[51] A. Dapogny, K. Bailly, S. Dubuisson, Pairwise conditional random forests for facial expression recognition, in: Proceedings of the International Conference on Computer Vision, 2016, pp. 3783–3791, doi:10.1109/ICCV.2015.431.
[52] Q. Zhen, D. Huang, H. Drira, B.B. Amor, Y. Wang, M. Daoudi, Magnifying subtle facial motions for effective 4D expression recognition, IEEE Trans. Affect. Comput. (2017), doi:10.1109/TAFFC.2017.2747553.
[53] N. Perveen, D. Roy, C.K. Mohan, Spontaneous expression recognition using universal attribute model, IEEE Trans. Image Process. 27 (11) (2018) 5575–5584, doi:10.1109/TIP.2018.2856373 .

Ruicong Zhi received the Ph.D. degree in signal and information processing from Beijing Jiaotong University in 2010. From 2016 to 2017, she visited the University of South Florida as a visiting scholar. She visited the Royal Institute of Technology (KTH) in 2008 as a joint Ph.D. She is currently a full professor in School of Computer and Communication Engineering, University of Science and Technology Beijing. She has published more than 60 papers, and more than twenty patents. She has been the recipient of more than ten awards, including the National Excellent Doctoral Dissertation Award nomination. Her research interests include facial and behavior analysis, artificial intelligence, and pattern recognition.

Caixia Zhou received the Bachelor degree in Anhui University of Finance and Economics in 2018. She is currently pursuing the MS degree at School of Computer and Communication Engineering, University of Science and Technology Beijing. Her research interest includes computer vision and emotion analysis.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036



Tingting Li received the MS degree at School of Computer and Communication Engineering from University of Science and Technology Beijing in 2019, and received the Bachelor degree in Computer Science from University of Science and Technology Beijing in 2016. Her research interest includes computer vision and emotion analysis.

Yi Jin was born in Heibei, China, in 1982 and received the Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University, China, in 2010. She is currently an assistant professor in the School of Computer Science and Information Technology, Beijing Jiaotong University. Her research interests include image processing, pattern recognition, computer vision and machine learning.

Please cite this article as: R. Zhi, C. Zhou and T. Li et al., Action unit analysis enhanced facial expression recognition by deep neural network evolution, Neurocomputing, https://doi.org/10.1016/j.neucom.2020.03.036