Landmark-Based Facial Feature Construction and Action Unit Intensity Prediction

Jialei Ma,1 Xiansheng Li,1 Yuanyuan Ren,1 Ran Yang,2 and Qichao Zhao2

1School of Transportation, Jilin University, Changchun 130022, China
2Beijing KINGFAR Co., Ltd., Beijing 100085, China

Correspondence should be addressed to Yuanyuan Ren; renyy@jlu.edu.cn

Received 23 November 2020; Accepted 2 March 2021; Published 12 March 2021

Academic Editor: Fu-Kwun Wang

Copyright © 2021 Jialei Ma et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Human face recognition has been widely used in many fields, including biorobots, driver fatigue monitoring, and polygraph tests. However, the end-to-end models fit by most of the existing algorithms perform poorly in interpretation because complex classifiers are constructed using facial images directly. In addition, in some of the models, dynamic characteristics of subjects as individuals are not fully considered, so dynamic information is not extracted. In order to solve these problems, this paper proposes an action unit intensity prediction model. The three-dimensional coordinates of 68 landmarks of human faces are obtained based on the convolutional experts constrained local model (CE-CLM), which enables the construction of dynamic facial features. Based on the error analysis of the CE-CLM algorithm, dimension reduction of the constructed features is performed by the principal components analysis (PCA). The radial basis function (RBF) neural network is also constructed to train the action unit prediction models. The proposed method is verified by the experiments, and the overall mean square error (MSE) of the proposed method is 0.01826. Lastly, the network construction process is optimized, so that for the same training samples, the models are fitted using fewer iterations. The number of iterations is decreased by 27 on average. In summary, this paper provides a method to rapidly construct action unit (AU) intensity prediction models and constructs automatic AU intensity estimation models for facial images.

1. Introduction

The human face is a carrier of multiple information types. It can not only directly show personal information (e.g., age, gender, and race) but also indirectly convey various emotions (e.g., pleasure, anger, sorrow, and joy). Moreover, a facial expression of emotion is one of the important types of interpersonal communication. According to Mehrabian et al. [1], in human interactions, 55% of the information is conveyed by facial expressions.

The facial action coding system (FACS) has been widely used in research on facial information. It was first developed by Ekman and Friesen in 1975 [2] and then improved in 2002 [3]. In the FACS, a unique set of basic facial muscle actions is defined and denoted as the action unit (AU). The FACS involves 44 visible AUs linked to facial muscular movements [4], including 33 independent AUs and 11 additional nonstrictly defined action descriptors (ADs) [5], which are given in Tables 1 and 2, respectively.

The FACS is comprehensive and objective. Any facial expression of emotion is generated by activating one or more groups of facial muscles. Also, each possible facial expression can be represented as a combination of different AUs. Due to the complex definition of coding rules in FACS, AU annotations can be completed only by FACS-certified coders. In order to meet the FACS coding requirements, the FACS-certified coders need to be professionally and strictly trained for at least 100 hours [4]. In addition, manual coding of facial images is a tedious and time-consuming process; for instance, even a well-trained coder needs about one hour to complete the facial coding of a one-minute video clip [6].

With the aim of solving the problem of time-consuming and inefficient facial coding, researchers have considered using computer technology for the purpose of coding.
This study aims to realize automatic AU intensity prediction by constructing facial landmark-based facial features, thereby improving the accuracy and efficiency of facial coding.

The rest of this article is organized as follows. In Section 2, the relative literature is reviewed, and the existing AU intensity estimation algorithms are analyzed. In Section 3, feature construction and dimension reduction of the AU sample library are conducted by facial feature analysis based on which a neural network model for AU intensity estimation is constructed and trained. In Section 4, the trained network model is optimized based on the experimental result to reduce training time. Lastly, in Section 5, the main conclusions are stated, and future research directions are presented.

### 2. Literature Review

Remarkable progress has been made in the field of automatic recognition of facial expressions over the past two decades [7, 8]. According to the previous studies on facial expression recognition, the two main tasks in the facial expression recognition process are facial information description and machine learning model design [9, 10]. The purpose of facial information description is to obtain a set of features from the original facial images, ensuring that similarity of features of different faces expressing the same information is more similar than that of the same faces expressing different information.

Facial landmark detection is the basic way to describe facial information. There are many well-developed algorithms for facial landmark detection. The active shape model (ASM) [11], where the range of landmarks is constrained based on Mahalanobis distance, is the earliest algorithm for landmark model construction. It not only contributes to facial landmark detection but can also apply to gesture detection and body movement detection. By adding texture information to the ASM during the shape feature statistics, the active appearance model (AAM) [12] that can locate the landmarks more accurately based on shape features has been constructed, but its real-time performance has been unsatisfactory. In contrast, the constrained local model (CLM) [13] makes a balance between accuracy and real-time performance using the advantages and disadvantages of the ASM and AAM. It achieves a higher computing efficiency by changing the global texture features in the AAM into the texture features near the landmarks of the public.

The enhancement of computing efficiency brings an increasingly complex structure of machine learning models. Mahoor et al. [14] classified combined AUs using the sparse representation (SR) classifier. Both Wang and Lien [15] and Valstar and Pantic [16] constructed the AU recognition models using the hidden Markov model (HMM). In a study by Kaltwang et al. [17], relevance vector regression was used to predict the AU intensity. The support vector machine (SVM) model was used by Mahoor et al. [18] and Valstar et al. [19] for emotion recognition. Savran et al. [4] predicted the AU intensity using the support vector regression (SVR). Recently, widely applied neural networks have become dominant in model construction. The deep neural networks were used by Liu et al. [20], Gudi et al. [21], and Zhao et al. [22] to build the AU recognition models.

Hitherto, most of the studies on automated AU analysis have focused on AU recognition, including simple AU recognition [23], combined AU recognition [24], AU-emotion coupled recognition [25], and AU dependence...
recognition [26]. However, little attention has been paid to the AU intensity prediction. Compared with the AU recognition, AU intensity prediction represents a more challenging task [27]. Due to the complexity of AU detection, including large numbers of categories, more subtle models, and minor differences between AUs, an automatic facial AU analysis still remains an open challenge in both AU recognition and AU intensity prediction [28].

3. AU Intensity Prediction Model

3.1. Facial Landmark Localization Method. Facial muscles contract and stretch when humans convey emotions through facial expressions, which causes the corresponding facial regions to change in shape or location. Therefore, facial images contain a large amount of emotional information. In this study, the emotional information contained in facial images is identified based on the changes in the locations of facial landmarks.

The convolutional experts constrained local model (CE-CLM) [29] is used to recognize the three-dimensional information of facial landmarks. By combining the advantages of the convolutional neural network and the original CLM algorithm, the CE-CLM achieves high robustness against interference factors, such as sunlight, angle, and obstruction. The input is a facial image, the human face in which is locked down by means of the human face detection algorithm. Next, three-dimensional facial alignment is conducted using the trained average face model. The exact location of each of 68 landmarks is determined using the local constraint algorithm. Lastly, the three-dimensional coordinates of all 68 facial landmarks are determined. As shown in Figure 1, the coordinate origin is the center of the camera lens. The optical axis of the camera denotes the $z$-axis, where the direction of the camera pointing to the face represents a positive direction. The direction perpendicular to the lens center represents the positive direction of the $y$-axis; the direction from the lens center horizontally to the right represents the positive direction of the $x$-axis.

3.2. Construction of Complete Facial Information Feature. After obtaining the three-dimensional coordinates of 68 facial landmarks using the CE-CLM algorithm, the complete facial information features are constructed. Points in space can differ significantly due to their differences in the origin and direction of each axis when building the coordinate system. Therefore, constructing the features based on the relationship between the points in space, the change in sample features caused by different coordinate system construction methods can be avoided.

3.2.1. Feature Construction Method. The facial features constructed based on the relations between facial landmarks can be divided into three following groups: the Euclidean distance between two points, the angle formed by a point and two other points, and the perpendicular distance from a point to the line between two other points, as shown in Figure 2. The angle formed between a point and two other points varies in a small range. Because of the principle of triangle similarity, angles cannot fully express the three-dimensional facial information. According to the calculation equation of a triangle area, the perpendicular distance from a point to the line between two other points that is shown in Figure 3 can be calculated by

$$
S_{\triangle ABC} = \frac{1}{2} BC \times AD = \frac{1}{2} CA \times BE = \frac{1}{2} AB \times CF,
$$

$$
\therefore BC : CA : AB = \frac{1}{A'D'} - \frac{1}{BE} \cdot \frac{1}{CF}.
$$

According to equation (1), this distance is equivalent to the Euclidean distance between two points, but the dimension $(C_3^{68} = 50116)$ constructed based on the perpendicular distance from one point to the line between two other points is far larger than that of the Euclidean distance between two points $(C_2^{68} = 2278)$. From the engineering perspective, the distance from a point to a straight line increases the computation complexity without introducing an effective feature. For this reason, it is not used as one of the facial features in this work. As shown in Figure 4, the facial features with 2278 dimensions are constructed based on the Euclidean distances between 68 facial landmarks.

3.2.2. Feature Construction Optimization. Considering the huge advantages of the convolutional neural networks in image processing, researchers have recently placed emphasis on AU recognition based on the prior model constructed using static images. However, the static image feature-based methods can only obtain representative geometric facial features from images, such as texture, color, and shape. They can present individual differences well but cannot focus on a certain AU of the whole group or give a summary of the features of AUs. In order to overcome the individual difference between static images, the individual sample calibration-based facial feature with movement differences is proposed. It can not only eliminate the difference between individuals but also match the AU description in the facial coding system.

In common calibration-based facial action feature construction methods, the difference between eigenvalues of facial images with and without expressions can be calculated by

$$
\Delta d_{i,j} = d_{i,j}(\text{expression}) - d_{i,j}(\text{neutral}),
$$

where $d_{i,j}$ denotes the Euclidean distance between landmarks $i$ and $j$ and $\Delta d_{i,j}$ denotes the difference between the distances from $i$ to $j$ with and without facial expressions.
However, some different $\Delta d_{i,j}$ values, such as $\Delta d_{51,8}$ and $\Delta d_{51,57}$ as shown in Figure 5, are similar due to the characteristics of facial landmark distribution, resulting in a reduced amount of feature information. For this reason, the variation rate of the features of facial images with and without expressions is used for facial feature construction in this study, and it is expressed as

$$
\Delta d_{i,j} = \frac{d_{i,j}(\text{expression}) - d_{i,j}(\text{neutral})}{d_{i,j}(\text{neutral})}.
$$

3.3. Feature Dimension Reduction. The facial features with 2278 dimensions are constructed from 68 landmarks to obtain complete facial information. These features are not valid for a particular AU from an intuitive perspective; in other words, some features are invalid for particular AUs. Therefore, feature dimension reduction is necessary for each AU. Valid features are extracted, which substantially enhances the construction efficiency.

3.3.1. Dimension Reduction Methods. Traditional feature dimension reduction methods mainly depend on the subjective understanding of the existing features. The most representative features are determined based on the judgment of several authoritative experts on each feature, which can be a practical method for a lower feature dimension. However, the workload of experts increases with the feature dimension, which makes this method impractical. Currently, there are several methods for feature dimension reduction, including the genetic algorithm, Fisher’s linear discriminant (FLD), maximum relevance and minimum redundancy (mRMR), and principal components analysis (PCA).

As a common optimization algorithm, the genetic algorithm simulates the natural survival of the fittest. This algorithm can generate and determine the most appropriate
feature based on particular rules. However, the parameters of this method and the method of new feature formation depend on manual setup, which together with a lack of theoretical basis for some certain features makes these methods unsuitable for extensive application.

The FLD represents a widely used dimension reduction method. In this method, samples are projected to a straight line, i.e., data are projected to the one-dimensional space. It divides features into valid and invalid ones, but valid and linearly correlated features cannot be removed.

The purpose of the mRMR is to minimize the correlation between selected features and maximize the correlation between the selected features and mathematical expectations. This method is the most efficient feature selection
method, but it does not consider new feature formation by means of feature combination.

3.3.2. Principal Component Analysis. The steps of the PCA are as follows: gather a dataset, select direction with the greatest change and set it as a new axis, check changes in the remaining data, and find an axis that is perpendicular to the first one and covers as many changes as possible. These steps are repeated until all the possible coordinate axes are identified. In this way, all the variables denote the axes along the rectangular coordinate system, while the covariance matrices denote diagonal matrices. Namely, each new variable is related only to itself rather than any other variable.

Essentially, it rebuilds a new coordinate system without changing the relationships between data points, which ensures the original data can be converted without any information loss. Thus, dimensions with small data variations in some directions can be removed directly without affecting data variability. In Figure 6, $XY$ represents the original coordinate system and $XY'$ represents the coordinate system after PCA conversion. However, there is no uniform standard for the thresholds on which the dimension removal procedure relies.

Considering that an eigenvalue threshold needs to be set in the PCA-based dimension reduction method, dimensions corresponding to eigenvalues that are smaller than the threshold are removed for the purpose of dimension reduction. Due to different data forms used in the PCA algorithm, there is no uniform way to set the eigenvalue threshold. For the sake of optimal dimension reduction, it is necessary to analyze the errors and set a reasonable threshold since there are errors in dynamic video acquisition and algorithm processing. The specific method is as follows.

The videos of subjects at complete rest were acquired to control the statistical error. The subjects were required not to do any facial expression during data acquisition, including blinking, so it did not take too much time to conduct the acquisition. In order to ensure the quality of statistical data, a one-minute video of each subject in the neutral state was acquired. A total of 18 clips of video data where the subjects did not blink were acquired. In this study, the videos were acquired on 30 fps with a resolution of $1280 \times 960$ using a Microsoft Kinect sensor. Finally, 18 groups of image data that contained 1800 images each were obtained.

First, the data of each image were processed and converted into features with 2278 dimensions using the aforementioned algorithm. For each frame of features, the variance corresponding to each dimension of features from the start frame to the current frame was calculated. In the end, the variances of all the dimensions were averaged to obtain the average feature variance of each frame of data, as shown in Figure 7. By using 100 frames as a step size, the information entropy of the data within every 100 frames was calculated, and the results are shown in Figure 8. As presented in Figure 8, the information entropy started to stabilize after 1500 frames, suggesting that the number of 1500 frames was the minimum requirement for error statistics in order to ensure high stability of data error. In this study, the average feature variance of 1800 frames of data was 1.16.

According to the mathematical derivation of the PCA algorithm, the eigenvalue is equal to the variance of the dimension corresponding to the coordinate after data rotation. If the variance in a certain dimension is smaller than the average variance without facial expression changes, this dimension contains less information and will not impact the overall information content if removed. Therefore, the PCA segmentation threshold is set to 1.16 in this work.

4. Model Construction

4.1. Sample-Tag Relationship. When sample data are preprocessed, the sample features after dimension reduction are obtained. Additionally, the sample data also contain the AU intensity tags of all images. These tags are calibrated by FACS-certified experts after long-time careful image observation. The FACS is an objective and comprehensive system constructed based on experimental psychological research. It aims to provide observers with an objective method for measuring facial actions. In behavioristics, the FACS is the most widely accepted measure of facial emotions. Thus, these AU intensity tags have been the most recognized result of subtle facial expressions.
There are two difficulties in constructing an automatic AU intensity prediction model for sample features and their AU tags as follows:

1. The FACS definitions of various intensities are shown in Figure 9. Levels A, B, C, D, and E represent the faint sign of action, slight but inconspicuous actions, actions that have been obviously taken, drastic actions, and actions that have reached the limit, respectively. Each intensity level involves a series of appearance changes. Thus, it can be inferred from the FACS definitions of AU intensities that there can be a nonlinear relationship between the intensity level and the range of facial actions.

2. Considering that the AU intensity tags, which represent the annotations of subject facial expressions made by experts, of the sample database are determined by the FACS, the results are still generally acceptable despite the subjectivity. Thus, the results denote the true values of the sample data output.

4.2. Radial Basis Function Neural Network. In order to solve the aforementioned problems, the radial basis function (RBF) neural network regression is used for AU intensity prediction.

The RBF is a real-value function whose value depends only on the distance from a sample to a point \( w \) in space, namely, \( g(x, w) = g(||x - w||) \). Any function \( g \) that satisfies \( g(x) = g(||x||) \) can be called an RBF. Since the Euclidean distance is the most commonly used distance measure in the RBFs, the RBF is also known as the Euclidean radial basis function.

In this study, the RBF is a Gaussian kernel function that can be expressed as \( g(x, w, \sigma) = \exp((||x - w||^2)/2\sigma^2) \), where \( x \) denotes the sample input, \( w \) denotes the center of the kernel function, and \( \sigma \) denotes the width parameter that controls the range of the radial action of the function.

The RBF neural network is a three-layer neural network consisting of the input layer, the hidden layer, and the output layer. The conversion from the input layer to the hidden layer is nonlinear, whereas that from the hidden layer to the output layer is linear. The specific structure of the RBF neural network is shown in Figure 10.

The basic idea of the RBF network is to use the RBF function as the activation function of the hidden-layer neurons so that the input vector can be mapped directly without requiring connections via weight. Once the central point of RBF is determined, the mapping relationship is also determined. The mapping from the hidden layer to the output layer is linear, which means network output represents a linear weighted sum of the hidden-layer output. The weight is the adjustable network parameter. In addition, the mapping from the network input to the network output is nonlinear, whereas the mapping from the network output to adjustable parameters is linear. The network weight can be solved directly by the system of linear equations, thus speeding up the learning process and avoiding the problem of the local minimum.

The RBF network’s features are ideal for the application considered in this study. Namely, the relationship between sample features and sample tags can be properly handled, and the correspondence between the features of training samples and their tags can be effectively regressed. Therefore, the RBF neural network regression is performed to fit the AU intensity prediction model. The mean square error (MSE) is used as the cost function during the model training and as the final evaluation indicator of model regression.

The regression error is quantized by calculating the MSE between the true and predicted values based on which model parameters are optimized to ensure model validity. The MSE is calculated by

\[
MSE = \frac{1}{n} \sum_{i=0}^{n} (y_i - \hat{y}_i)^2, \tag{4}
\]

where \( y_i \) denotes the AU intensity tag corresponding to the \( i \)th sample and \( \hat{y}_i \) denotes the predicted value of the \( i \)th sample.

5. Model Evaluation

The Bosphorus [30] and Lucey et al. (CK+) [31] databases were used for regression and evaluation of the proposed AU strength prediction model. In addition, the model training was optimized based on the experimental result in order to ensure model regression could be completed in a shorter time in the second training under the same experimental data and condition.

5.1. Database Introduction. The Bosphorus database was developed by Bosphorus University in Turkey; it included 105 subjects (60 males and 45 females), of whom 29 were professional actors/actresses, and consisted of 4652 facial images. The images were marked by FACS-certified experts as AU1, AU2, AU4, AU9, AU10, AU12, AU14, AU15, AU16, AU17, AU18, AU20, AU22, AU23, AU24, AU25, AU26, AU27, AU28, AU34, AU43, and AU44. This database is one of the AU-tag databases that contain the largest sample size.

The CK+ database was developed by Jeffrey F. Cohn and Takeo Kanade; it included 210 subjects (145 females and 65 males) and consisted of 2105 facial images that were marked by
FACS-certified experts with AU tags. However, this database was intended to present six basic emotions of subjects. Here were a number of different AU tags in this database, and the number of samples that contained a single AU was small. According to the statistics, the number of tags that were valid for each AU was smaller than 30. Considering that it would be difficult to conduct an effective model fitting with so few samples each AU tag contained, the above two databases were combined for the purpose of sample size expansion.

5.2. Dataset Construction. A sample library was constructed for each AU in this study. Take AU1 as an example. The images with AU1 tags and those of subjects without facial expressions were extracted from the two databases. Next, the aforementioned sample feature construction was performed to obtain the sample features and the corresponding tag data. Lastly, all the data with AU1 tags were used to form a complete dataset that contained the features and AU1 tags.

5.3. Evaluation Criterion. With the aim to evaluate the model regression effect more effectively, the five-fold cross-validation was introduced and used to measure the regression effect of the final model. By using this method, the dataset was randomly divided into five equal parts, of which one was used as the validation set, while the remaining parts were used as training sets. The training was repeated five times, and the MSE values of the five validation sets were averaged and taken as the model evaluation indicator.

In addition, the correlation coefficient (CORR) of sample tags and predicted values was introduced as another indicator for model regression evaluation, and it was calculated by

\[
\text{CORR}(y, \tilde{y}) = \frac{\text{COV}(y, \tilde{y})}{\sigma_y \sigma_{\tilde{y}}} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{n} (\tilde{y}_i - \bar{\tilde{y}})^2}}.
\]

(5)

5.4. Fitting Results. In this study, the AU intensities predicted by the radial basis neural network (RBNN) without feature dimension reduction, the backpropagation neural network (BPNN), and the support vector regression (SVR) algorithm [25] were compared, and results are presented in Table 3.

MSE shows the difference between the estimated values and the true values. A lower MSE value represents a more accurate prediction. In addition, a correlation coefficient closer to 1 represents a better correlation.

As shown in Table 3, the MSE of the proposed method was smaller than the MSEs of the three other methods for all AUs, and CORR values were all larger than 0.98. This indicated that there was a high correlation between predicted and true values.

5.5. Training Process Analysis. When constructing the radial basis neural network (RBNN) using the existing algorithms, the neurons with uniform step sizes were added. In order to minimize the neural network structure, the neurons were added at a step size of one. In this way, no more neurons would be added when the model converged, i.e., when the MSE stopped decreasing during the model training.

The number of neurons and the MSE of each AU in the RBNN during the training was counted. In Figure 11, the abscissa denotes the number of iterations during the training process and the ordinate denotes the MSE at the
Table 3: Result for the AU intensity estimation.

| AU   | Proposed | MSE  | BPNN | SVR  | Proposed | MSE  | BPNN | SVR  |
|------|----------|------|------|------|----------|------|------|------|
| AU01 | 0.024178 | 0.026| 0.047| 0.051| 0.992    | 0.991| 0.957| 0.960|
| AU02 | 0.009237 | 0.012| 0.066| 0.095| 0.995    | 0.997| 0.970| 0.978|
| AU04 | 0.009893 | 0.012| 0.066| 0.095| 0.995    | 0.997| 0.970| 0.978|
| AU05 | 0.027837 | 0.030| 0.134| 0.123| 0.992    | 0.989| 0.881| 0.895|
| AU06 | 0.022889 | 0.025| 0.119| 0.107| 0.989    | 0.989| 0.841| 0.859|
| AU07 | 0.012078 | 0.011| —    | —    | 0.994    | 0.994| —    | —    |
| AU09 | 0.005872 | 0.010| —    | —    | 0.998    | 0.998| —    | —    |
| AU10 | 0.024854 | 0.024| 0.036| 0.033| 0.996    | 0.988| 0.924| 0.939|
| AU12 | 0.038977 | 0.039| 0.042| 0.039| 0.985    | 0.985| 0.939| 0.930|
| AU14 | 0.016743 | 0.021| —    | —    | 0.994    | 0.988| —    | —    |
| AU15 | 0.038062 | 0.042| 0.056| 0.060| 0.981    | 0.979| 0.890| 0.892|
| AU16 | 0.015238 | 0.017| —    | —    | 0.993    | 0.992| —    | —    |
| AU17 | 0.021056 | 0.023| 0.043| 0.041| 0.990    | 0.989| 0.896| 0.923|
| AU18 | 0.018616 | 0.019| 0.064| 0.056| 0.992    | 0.992| 0.955| 0.947|
| AU20 | 0.020243 | 0.025| 0.058| 0.046| 0.987    | 0.983| 0.878| 0.913|
| AU22 | 0.009127 | 0.018| —    | —    | 0.997    | 0.997| —    | —    |
| AU26 | 0.017167 | 0.018| 0.025| 0.031| 0.988    | 0.988| 0.954| 0.976|
| AU27 | 0.008465 | 0.008| 0.097| 0.104| 0.996    | 0.996| 0.931| 0.969|
| AU28 | 0.016481 | 0.024| —    | —    | 0.994    | 0.994| —    | —    |

AU, action unit; BPNN, backpropagation neural network; CORR, correlation coefficient; MSE, mean square error; RBNN, radial basis neural network.

Figure 11: The mean square error (MSE) of each action unit (AU) during the model training.
As shown in Figure 11, the MSE decreased until it converged during all AU model training processes. The MSE showed two different local downtrends: downtrend of the concave function and downtrend of the convex function. Finally, it converged after the downtrend of the convex function.

In the proposed RBNN construction method, neurons were added at dynamic step sizes. Specifically, the MSE values of the five models were calculated during the network construction, and the concavity and convexity of the MSE were identified. If the MSE was concave, the step size of neuron addition was increased; if the MSE was convex, the step size of neuron addition was reset to one. The pseudocode (Algorithm 1) is given as follows.

A comparison of the model training times before and after algorithm improvement is given in Table 4, where it can be seen that both the number of model trainings and the model fitting time decreased after the model was constructed by means of dynamic neuron addition.

### 6. Conclusions

This paper proposed an AU intensity prediction model. First, the three-dimensional coordinates of 68 landmarks of human faces in the image set are calculated using the CE-CLM algorithm. Next, the variation rate between the Euclidean distances of images with AUs and those without facial expressions is calculated to obtain variations of facial features. Then, the variations of facial features without expressions in the videos are used to obtain the range of feature error using the CE-CLM algorithm. A reasonable threshold is determined using the PCA algorithm to eliminate the error and retain information as much as possible. Lastly, the RBNN is constructed to fit the developed model for each AU. The fitting algorithm is optimized by observing the fitting process, thereby reducing the number of fittings in the second model fitting and shortening the model training time.

In order to decrease the detection error, further optimization of the CE-CLM algorithm is necessary, which will be part of our future work. In addition, the arrangement of 68 landmarks should also be optimized to achieve a more efficient model detection.

---

**Algorithm 1:** The training algorithm of the neutral networks for AU intensity estimation.

```plaintext
(1) step = 1
(2) n = 0
(3) While (NN is not convergent)
(4) { 
(5) n = n + 1
(6) If (n < 10)
(7) [add cell and training NN]
(8) Else{
(9) If [mse (n − 9), mse(n − 8),..., mse (n)] is concave
(10) { 
(11) step = step + 1
(12) add cell and training NN
(13) }
(14) Else if [mse (n − 9), mse(n − 8), . . . , mse (n)] is convex
(15) { 
(16) step = 1
(17) add cell and training NN
(18) }}
(19) Train done
```

| AU no. | Before | After |
|--------|--------|-------|
| AU01   | 349    | 322   |
| AU02   | 288    | 261   |
| AU04   | 344    | 326   |
| AU05   | 278    | 251   |
| AU06   | 230    | 212   |
| AU07   | 1129   | 1102  |
| AU09   | 143    | 116   |
| AU10   | 181    | 150   |
| AU12   | 348    | 321   |
| AU14   | 174    | 151   |
| AU15   | 113    | 95    |
| AU16   | 219    | 192   |
| AU17   | 307    | 289   |
| AU18   | 277    | 234   |
| AU20   | 101    | 92    |
| AU22   | 88     | 69    |
| AU23   | 128    | 110   |
| AU24   | 252    | 225   |
| AU25   | 897    | 870   |
| AU26   | 448    | 430   |
| AU27   | 186    | 168   |
| AU28   | 82     | 61    |

AU, action unit.

**Table 4:** Comparison of the number of iterations before and after optimization.
Data Availability
All data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest
The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments
This work was supported by the National Key Research and Development Project of China under grant 51578262.

References
[1] A. Mehrabian, "Communication without words," Psychology Today, vol. 2, no. 4, pp. 53–55, 1968.
[2] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, CA, USA, 1978.
[3] P. Ekman and W. V. Friesen, Facial Action Coding System, the Manual, Research Nexus Division of Network Information Research Corporation, South Atlanta, GA, USA, 2002.
[4] A. Savran, B. Sankur, and M. Taha Bilge, "Regression-based intensity estimation of facial action units," Image and Vision Computing, vol. 30, no. 10, pp. 774–784, 2012.
[5] Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in Advances in Face Detection and Facial Image Analysis, pp. 63–100, Springer, Cham, Switzerland, 2016.
[6] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 974–989, 1999.
[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," IEEE Trans. Pattern Anal. Mach. Intell, vol. 31, no. 1, pp. 39–58, 2009.
[8] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.
[9] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 356–370, 2019.
[10] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in Handbook of Face Recognition, pp. 487–519, Springer, London, UK, 2011.
[11] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38–59, 1995.
[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681–685, 2001.
[13] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," Pattern Recognition, vol. 41, no. 10, pp. 3054–3067, 2008.
[14] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in Proceedings of the 2011 International Conference Automatic on Face and Gesture Recognition, pp. 336–342, Santa Barbara, CA, USA, March 2011.
[15] T.-H. Wang and J.-J. Lien, "Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation," Pattern Recognition, vol. 42, no. 5, pp. 962–977, 2009.
[16] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 1, pp. 28–43, 2012.
[17] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly sparse relevance vector machine for continuous facial behavior estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 1748–1761, 2016.
[18] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn, "A framework for automated measurement of the intensity of non-posed facial action units," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 74–80, Miami, FL, USA, June 2009.
[19] M. F. Valstar, T. Almaev, J. M. Girard et al., "FERA 2015–second facial expression recognition and analysis challenge," in Proceedings of the 2015 IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–8, Ljubljana, Slovenia, May 2015.
[20] M. Liu, S. Li, S. Shan, and X. Chen, "AU-aware deep networks for facial expression recognition," in Proceedings of the 2013 IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–6, Shanghai, China, April 2013.
[21] H. E. T. Gudi, T. M. den Uyl, and A. Maroulis, "Deep learning based FACS action unit occurrence and intensity estimation," in Proceedings of the 2015 IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 1–5, Ljubljana, Slovenia, May 2015.
[22] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3391–3399, Las Vegas, NV, USA, June 2016.
[23] Y. Li, B. Wu, B. Ghanem, Y. Zhao, H. Yao, and Q. Ji, "Facial action unit recognition under incomplete data based on multi-label learning with missing labels," Pattern Recognition, vol. 60, pp. 890–900, 2016.
[24] S. Wang, S. Wu, G. Peng, and Q. Ji, "Capturing feature and label relations simultaneously for multiple facial action unit recognition," in Proceedings of the 2015 IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, pp. 187–194, Xi’an, China, May 2013.
[25] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly sparse relevance vector machine for continuous facial behavior estimation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 9, pp. 1748–1761, 2016.
[29] T. B. Zadeh and L.-P. Morency, "Convolutional experts constrained local model for facial landmark detection," in *Proceedings of the 2017 IEEE Conference on CVPRW*, p. 2051, Honolulu, HI, USA, July 2017.

[30] A. Savran, N. Alyüz, H. Dibeklioğlu et al., "Bosphorus database for 3D face analysis," in *Lecture Notes in Computer Science*, pp. 47–56, Springer, Berlin, Germany, 2008.

[31] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade sataset (CK+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the 2010 IEEE Conference on CVPRW*, pp. 94–101, San Francisco, CA, USA, June 2010.