A Facial Expression Recognition Method Based on Dlib, RI-LBP and ResNet

Jianxing Yang¹, Jianhua Adu¹*, Hongguang Chen¹, Jie Zhang¹ and Jianfeng Tang¹

¹ School of Software Engineering, Chengdu University of Information Technology, Chengdu, Sichuan, 611230, China
*Corresponding author’s e-mail: adjh_work@163.com

Abstract. A method of facial expression recognition using a composite feature is proposed. The method combines the expanded Dlib facial feature detector, the rotation-invariant local binary pattern (RI-LBP) and the 50-layer ResNet neural network model (ResNet_50). First, the expanded Dlib was used to locate 83 feature points on the face, obtaining the Dlib feature after preprocessing and dimensionality reduction (PCA). Then, the rotation-invariant LBP feature was extracted from 8 important regions after tilt correction. Furthermore, a 50-layer ResNet neural network was used to extract the low level features from the images. Finally, the three features were combined and extreme learning machine (ELM) was used to classify the composite facial features. The experimental results on Jaffe and CK+ datasets showed that the proposed method performs better compared with other methods.

1. Introduction
The recognition and classification of facial expressions is an important field in facial recognition [1]. In the process of facial recognition, the integrity and robustness of facial features have a great impact on the recognition results [2]. Many researches have been done on the feature extraction process of facial recognition. Niu et al. [3] used PCA to extract two-dimensional features from facial images. Matthews et al. [4] proposed the active appearance model (AAM) to extract facial contours and facial feature points. Other common methods include texture feature extraction methods such as local binary pattern (LBP) [5] and Gabor transform [6], however, they are not effective in distinguishing different facial expressions. Li et al. [7] proposed a multi-direction LBP coding scheme which performs better than traditional LBP in accuracy. Wei et al. [8] combined a 2-layer feature selection algorithm and AdaBoost to reduce the curse of dimensionality of Gabor features. However, it is prone to interference by lighting, background, etc. Ping et al. [9] proposed a Feature Disentangling Machine that divided the extracted facial features into common features and expression-specific features, which demonstrated better generalization capability of selected features. In recent years, more and more deep learning models such as VGGNet, GoogleNet and ResNet were used for facial expression recognition. For example, Mollahosseini et al. [10] combined AlexNet and GoogleNet to construct a 7-layer convolutional neural network for facial expression recognition and achieved impressive results.

In this paper, a method incorporating the expanded Dlib, the rotation-invariant local binary pattern region feature (RI-LBP) and a 50-layer ResNet neural network model are proposed. The features extracted by the method are robust. The experimental results on Jaffe [11] and CK+ [12] dataset using the combined features is better than using the respective features alone.
2. Proposed Algorithm

The workflow of the proposed algorithm are shown in Figure 1. First, the input face images are preprocessed. Then a feature composed of the expanded Dlib feature, the rotation-invariant LBP features, the ResNet Feature are extracted. Finally the composite feature is sent into the extreme learning machine for classification.

![Model workflow diagram](image)

**Figure 1. Model workflow**

2.1. Image Preprocessing

The datasets used in this paper are Jaffe and CK+. First, the images are resized into the same size and converted into grayscale. In order to remove non uniform illumination of the images, illumination correction is performed. The common correction method is the gamma transformation, which is simply a power law transform:

$$g(x, y) = (f(x, y))^{\gamma}$$

where $g(x, y)$ and $f(x, y)$ represent the brightness at pixel point $(x, y)$ before and after illumination correction respectively, gamma is the adjusting coefficient. Gamma transformation can stretch the contrast of the low brightness regions, resulting in more visible detailed textures. Gamma is usually between 0 and 1, we used 0.3 in this paper.

Facial Landmark detection is a very important process in facial detection. There are originally 68 feature points in the Dlib’s facial landmark detector. There are many important details in the forehead and eye regions of human beings that can be used to classify facial expressions. Therefore, we used the 83 facial feature point model expanded on the original Dlib facial landmark detector (Figure 2). Of the 15 features added, 13 (point 69-81) are in the forehead and 2 (point 82-83) are in the center of eyes. The expanded Dlib facial landmark detector is more accurate, and able to extract more important face features.

![83 facial feature point model](image)

**Figure 2. The distribution of 83 facial feature point model**
In the dataset, due to varied shooting angles, etc., the feature points may not be accurately located, hence tilt correction is needed. Suppose that the line connecting point 82 and 83 in Figure 2 and horizontal line form an acute angle $\theta$, using the coordinates of the center points of the eyes, we can calculate $\theta$, which can be used to correct the images. Assume that $M$ is the transformation matrix, $(t_x, t_y)$ is the coordinate of the rotation point. $M$ is given by:

$$
M = \begin{bmatrix}
\cos \theta & -\sin \theta & t_x - (\cos \theta t_x - \sin \theta t_y) \\
\sin \theta & \cos \theta & t_y - (\sin \theta t_x + \cos \theta t_y)
\end{bmatrix}
$$

(2)

Given the original point $(x_i, y_i)$, the $i$-th feature point coordinates $(x_i', y_i')$ after correction are given by:

$$
\begin{bmatrix}
x_i' \\
y_i'
\end{bmatrix} = M \begin{bmatrix}
x_i \\
y_i \\
1
\end{bmatrix}
$$

(3)

Local areas such as eyes, forehead, mouth, etc. contain most of the information about facial expressions. In this paper, the left eye, right eye, left eyebrow, right eyebrow, forehead, nose, mouth and tooth area of the expression were cropped as ROI area. In both horizontal and vertical directions, 5 is used as a safe distance to avoid incomplete ROI interception. For example, the upper-left point, the width and height of the left eye ROI area is given by:

$$
\begin{cases}
x_{left\_eye} = x_{37} - 5 \\
y_{left\_eye} = \min(y_{38}, y_{39}, y_{41}, y_{42}) - 5 \\
w_{left\_eye} = x_{40} - x_{left\_eye} + 5 \\
h_{left\_eye} = \max(y_{38}, y_{39}, y_{41}, y_{42}) - y_{left\_eye} + 5
\end{cases}
$$

(4)

In order to normalize the extracted features, the tilt corrected face area detected by Dlib are resized to 64x64x1 using bilinear interpolation. The 83 facial feature points were also recalculated using bilinear interpolation for the extraction of Dlib83-feature. The $(x, y, w, h)$ of 8 ROI areas were also recalculated which are used for subsequent extraction of rotation-invariant local binary pattern features (RI-LBP Feature). For ResNet training, the input is transformed to 64x64x3 images.

2.2. Facial Feature Extraction

As mentioned, in this part we extracted 3 different facial features from the images and combine them into a composite feature for classification.

(1) Global Expanded Dlib Feature: the extracted 83 featured points provide many information for effectively distinguishing different types of expressions. A free combination of any two feature point resulted in 3043 ($C_{83}^2$) point pairs. The Euclidean distance $d_{i,j}$ between the 3043 point pairs, and the acute angle $\alpha_{i,j}$ formed by the line connecting the point pairs and the horizontal line are calculated. They are expressed in the following formula:
\[ d_{i,j} = \left( (x_i - x_j)^2 + (y_i - y_j)^2 \right)^{1/2} \]  

\[ \alpha_{i,j} = \begin{cases} \arctan \frac{y_i - y_j}{x_i - x_j}, & x_i \neq x_j \\ 1, & x_i = x_j \end{cases} \]  

where \( i \) and \( j \) represent the \( i \)-th and \( j \)-th feature point respectively, \( i, j \in \{1,83\} \). Then \( d_{i,j} \) and \( \alpha_{i,j} \) are concatenated to form a feature of size 6086×1:

\[ Dlib83\_Feature = \{d_{1,2}, \alpha_{1,2}, d_{1,3}, \alpha_{1,3}, \ldots, d_{45,46}, \alpha_{45,46}, d_{45,47}, \alpha_{45,47}, \ldots, d_{82,83}, \alpha_{82,83} \} \]

6086 is relatively large, hence Principal Component Analysis (PCA) was used to reduce the dimensions of Dlib83\_Feature to size 258×1.

(2) Local RI-LBP feature: the rotation-invariant local binary pattern (RI-LBP) is used to encode the cropped ROI regions for generating their corresponding RI-LBP descriptors. Every RI-LBP descriptor is a 36-d vector. They are cascaded to form a local RI-LBP feature in the order of left eye, right eye, left eyebrow, right eyebrow, forehead, nose, mouth and tooth.

(3) Global ResNet feature: convolutional neural networks is capable of extracting the global features of facial expressions. ResNet, proposed by He et al.[13], added the residual unit (Figure 4) to VGGNet. By directly summing the input and output information, the features are more likely to be passed on to the next residual block, which makes ResNet easier to optimize, and can gain accuracy from increased depth. In this paper, a 50-layer ResNet is used to obtain the global features.
Figure 4. ResNet residuals unit diagram.

The labeled images were put into ResNet_50 for training. When the training loss is stable and classification accuracy is optimal, each image is sent to the model and its output feature vector in the last fully-connected layer was extracted. The extracted feature has 8192 dimensions. PCA is used to reduce it to 2048 dimensions as the final ResNet Feature.

(4) Feature fusion: combining the aforementioned 3 features. Since the features are differently distributed, they are normalized using the zero-mean method respectively:

\[ z_{n}^{i} = \frac{x_{i} - \mu_{i}}{\sigma_{i}} \]  

where \( z_{n}^{i} \) represents the features after normalization, \( x_{i} \) is the feature vectors, \( \mu_{i} \) and \( \sigma_{i} \) are the mean and variance of feature vectors. After normalization, the features are concatenated in the order of Dlib83-PCA Feature, RI-LBP Feature and ResNet Feature to form the composite feature vector.

Finally, the extreme learning machine (ELM) classifier [14] was used for the classification of the combined feature vectors.

3. Experiment and Analysis

The datasets used in this paper are the Jaffe and CK+ datasets. The Jaffe dataset has 7 categories: angry, disgusted, afraid, happy, sad, surprised and neutral. The CK+ dataset has 7 categories: anger, disgust, fear, happy, sadness, surprise and contempt. Images are randomly divided into test set and training set at a ratio of 3:7. The experiments were repeated for 3 times to calculate the average accuracy. The accuracy is given by

\[ \text{accuracy} = \frac{\text{Number of correctly identified images}}{\text{Number of test images}} \times 100\% \]  

The code are written with Tensorlow 1.10 on Pycharm 2018, Windows 10 with Intel Core i7-8750H @ 2.20GHz and 16GB RAM.

Table 1 and Table 2 is the confusion matrix of the proposed method on Jaffe and CK+ datasets respectively.

Table 1. The average confusion matrix for expression recognition of Jaffe test data set.

| Expression | Anger | Disgust | Fear | Happy | Neutral | Sad | Surprised |
|------------|-------|---------|------|-------|---------|-----|-----------|
| Anger      | 100.00| 0       | 0    | 0     | 0       | 0   | 0         |
| Disgust    | 0     | 100.00  | 0    | 0     | 0       | 0   | 0         |
| Expression | Anger  | Disgust | Fear  | Happy  | Contempt | Sad  | Surprised |
|------------|--------|---------|-------|--------|----------|------|-----------|
| Anger      | 97.44  | 2.56    | 0     | 0      | 0        | 0    | 0         |
| Disgust    | 3.70   | 96.30   | 0     | 0      | 0        | 0    | 0         |
| Fear       | 0      | 4.17    | 83.33 | 0      | 0        | 8.33 | 4.17      |
| Happy      | 0      | 0       | 0     | 95.24  | 1.59     | 0    | 3.17      |
| Contempt   | 0      | 0       | 0     | 0      | 100.00   | 0    | 0         |
| Sad        | 0      | 3.70    | 0     | 0      | 0        | 96.30| 0         |
| Surprised  | 0      | 0       | 2.67  | 1.33   | 0        | 0    | 96.00     |

Table 2. The average confusion matrix for expression recognition of CK+ test data set.

A comparative experiment between the proposed method, LBP, ResNet_50 and RI-LBP were carried out. The classification accuracy of the four methods is shown in Figure 5.

As can be seen from Figure 5, compared with the other three methods, the proposed method has the highest average classification accuracy. This is due to the composite feature’s overall better and stable performance on all categories, as can be seen in Figure 6. In addition, because of tilt correction, the differences between different expressions are more prominent in the composite feature. For example, the proposed method perform much better on fear than other methods.
Table 3 lists the average inference time of all methods. The proposed method has longer inference time than methods using LBP, RI-LBP and on par with ResNet_50. The proposed method achieved better classification accuracy at the cost of slightly longer running time.

Table 3. Time comparison of the methods.

| Methods   | inference time (ms) |
|-----------|---------------------|
| LBP       | 1124                |
| ResNet_50 | 1849                |
| RI-LBP    | 1378                |
| Proposed  | 2037                |

4. Conclusions
In this paper, a facial expression recognition method based on expanded Dlib facial feature detector, RI-LBP, and ResNet_50 network is proposed. The expanded Dlib facial feature detector was used to perform tilt correction on the images and extract the ROI regions effectively. This is conducive to further differentiate the facial expressions. The use of ResNet_50 neural network model and the local RI-LBP features of the facial ROI regions accentuates the details of the facial features. Using the extreme learning machine for classification, the experimental results showed that the proposed method can significantly improve the facial expression recognition accuracy, but the computational cost should also be further reduced.

References
[1] Murray, E. M., Krause, W. H., Stafford, R. J., Bono, A. D., Meltzer, E. P., Borod, J. C. (2015) Asymmetry of Facial Expressions of Emotion., pp. 73-99
[2] Tehmina, K., Muhammad, A. S., Muhammad, M., Bilal, K., Muhammad, A. S. (2018) Emotion recognition from facial expressions using hybrid feature descriptors. Iet. Image. Processing., 12(6): 1004-1012.
[3] Niu, Z., Qiu, X. (2010). Facial expression recognition based on weighted principal component analysis and support vector machines. In: International Conference on Advanced Computer Theory & Engineering. Chengdu. pp. V3-174-V3-178.
[4] Matthews, I., Baker, S., (2004) Active Appearance Models Revisited. International Journal of Computer Vision., 60(2):135-164.
[5] Ojala, T., Pietikäinen, M., Mäenpää, T. (2002) Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. Transactions on Pattern Analysis and Machine Intelligence., 24(7): 971-987.
[6] Zhang, Z., Mu, X., Lei, G. (2011) Recognizing facial expressions based on Gabor filter selection. In: 2011 4th International Congress on Image & Signal Processing. Shanghai. pp. 1544-1548.

[7] Li, H., Li, G. (2019) Research on Facial Expression Recognition Based on LBP and Deep Learning. In: 2019 International Conference on Robots & Intelligent System (ICRIS). Haikou. pp. 94-97.

[8] Wei, Y., Sun, Z., Yan, Z. (2008) Optimal Gabor Feature for Facial Expression Recognition. Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao/journal of Computer Aided Design & Computer Graphics., 20(1): 511-520.

[9] Ping, L., Zhou, J. T., Tsang, W. H., Meng, Z., Han, S., Tong, Yan. (2014) Feature Disentangling Machine - A Novel Approach of Feature Selection and Disentangling in Facial Expression Analysis. European Conference on Computer Vision. pp. 151-166.

[10] Mollahosseini, A., Chan, D., Mahoor, M. H. (2016) Going Deeper in Facial Expression Recognition using Deep Neural Networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Lake Placid, NY. pp. 1-10.

[11] Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding facial expressions with Gabor wavelets. In: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. Nara. pp. 200-205.

[12] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I. (2010) The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops. San Francisco. pp. 94-101.

[13] He, K., Zhang, X., Ren, S., & Jian, S. (2016) Deep Residual Learning for Image Recognition. In: the IEEE Conference on Computer Vision & Pattern Recognition. Las Vegas, NV. pp. 770-778.

[14] Huang, G. B., Zhu, Q. Y., Siew, C. K. (2006) Extreme learning machine: Theory and applications. Neurocomputing., 70(1/3): 489-501.