Facial Expression Recognition based on Landmarks

Yinghong Qiu[1], Yi Wan[1]
1.School of Information Science and Engineering, Lanzhou University
Lanzhou, China
shineredqiu@163.com, wanyi@lzu.edu.cn

***Abstract*—Facial expression is presentation of people's emotion, which plays a vital role in person's daily communication. Therefore, facial expression recognition (FER) is becoming an increasingly significant task in contemporary society. Currently the majority of the proposed methods for facial expression recognition use deep convolutional neural networks (CNN) in a supervised learning fashion. In this paper, we try to answer two important questions. The first is that in human FER, what are the key features that mostly constitute an expression, if such features exist? The second question is to find a way to increase the FER accuracy. To this end, we propose a new FER framework that relies solely on facial landmarks. In this framework, in the training process, we first extract facial landmarks, and feed them into a shallow network for recognition/classification. We show that just by using 68 facial landmark points, it's possible to achieve state-of-the-art FER results, thus opening the possibility to further study human emotion cognition process. On the other hand, our framework also produce better results than typical deep CNN-based methods with fast implementation.**

Keywords-Facial expression recognition; landmarks; MLP.

I. INTRODUCTION

Facial expressions are essential in person's mutual communication, which indicate the emotion and the intention of humans. Ekman [1] shows the strong evidence from both literate and preliterate is supportive of universals in facial expressions. And Ekman and Friesen define six basic emotions based on cross-culture study [2], which shows that particular facial expressions are universally related to particular emotions. These six basic emotions are anger, disgust, fear, happiness, sadness and surprise. Then contempt also is regarded as a universal basic emotion [3].

Humans are capable of recognizing expressions and understanding emotions of others, while computer needs huge computations to discriminate different expressions from their face. Humans will benefit a lot if machines can recognize people's facial expression. For instance, if robots are prone to understand person's intention with facial expression recognition, they can offer more friendly service to human. In addition, facial expression recognition (FER) shows promising prospects in many fields such as computer interfaces, health management, autonomous driving and etc.

Automatic recognition of face emotion aims to classify facial expressions into several basic emotions such as anger, disgust, happiness, sadness and ect. And the first and preprocessing step to FER is face detection. Wang [4] proposes a face detection algorithm called Viola-Jones algorithm, and Zhang [5] introduces a face detection algorithm using multitask cascaded convolutional networks. The next step of FER is extracting features from facial expression images. There are a variety of methods to extract features of facial expression images, ranging from the traditional algorithms to the deep learning algorithms. The majority of the traditional methods mainly contain feature extraction and classification two parts. And Verma [6] first uses Gabor filters to extract features of the detected faces, then classifies facial expressions with multilayer feed forward back propagation algorithm. Traditional methods of facial expression recognition depend on handcrafted features, such as LBP [7], Gabor texture [8] and histogram of oriented gradients (HOG) [9]. But due to the complicated design procedure of the traditional features, it's hard for human to design satisfactory hand-crafted methods for facial expression recognition. Meanwhile, benefiting from rapidly increased chip processing ability, deep learning has been applied to many research fields, such as face detection and object detection. Many researchers proposed many methods based on deep architectures. And Hu [12] introduces three types of supervised blocks for shallow, intermediate and deep layers in deep CNN respectively to enhance the supervision degree for FER. A novel CNN architecture called HoloNet [14] which uses CReLU [15] instead of ReLU in lower convolutional layers and combines CReLU and residual structure in the middle layers was proposed to address emotion recognition task. Deep CNNs have achieved the state-of-the-art results in a variety of fields.

But deep CNN methods always suffer from hard-training and time-consuming drawbacks. Some researchers also use facial landmarks for facial expression recognition. Ivona in [17] proposes an approach to solve expression recognition, which is incorporating facial landmarks as a part of the classification loss function. Munasinghe in [18] also proposes a methodology to identify facial emotions using facial landmarks which is related to eyebrows and mouth while our work use the total 68 landmarks.

In our work we also use landmarks for FER task and use 68 landmarks for expression recognition. This paper we explore the potential ability of facial expression recognition method which is based on facial landmarks. In our work first we normalize our points of landmarks, divide the landmarks according to different regions, then use the MLP for expression classification. Results show that our method can achieve high

978-1-7281-1907-6/19/$31.00 ©2019 IEEE

accuracy on our test set, which probably indicates that human brain may need less data information when it comes to classifying facial expressions.

II. The Proposed Methodology

Our proposed methodology is described in this section, we use facial landmark-based vectors which describe facial expressions and use MLP to identify different facial emotions. The proposed method was trained to classify 7 expressions and those are neutral, anger, disgust, fear, happy, sad, surprise.

Due to the fact that facial landmarks changes when human make different expressions, we postulate that expression features can be represented by the landmarks changes. In section.A part, we normalize all points according one origin. And in order to capture better facial landmarks feature vectors, we choose different origins in different regions, which is described in section.B part. Figure 1 shows some changes of facial landmarks when people make different expressions.

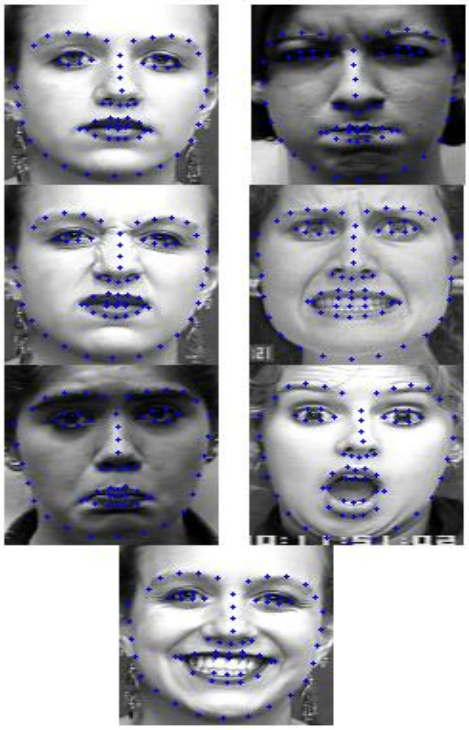

Fig. 1. 68 points of landmarks on different expressions (left to right, up to down are neutral, anger, disgust, fear, sad, surprise, happy).

A. Normalize Landmarks based on One Origin Point

Before train the MLP network, we first normalize the coordinates of 68 landmarks by subtracting the point on the nose. As shown in Figure 2, we choose the red point on the nose as our origin point, then all other points need to subtract the origin point. So we can obtain total 68 vectors being normalized, which will be sent to the MLP network.

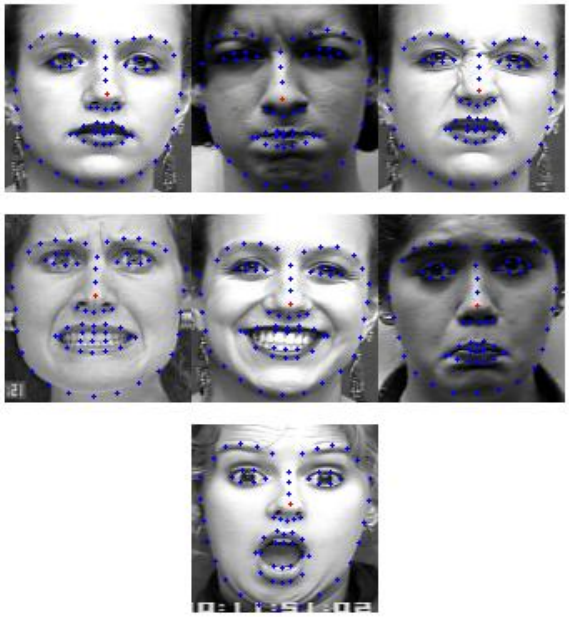

Fig. 2. Landmarks (rep point indicates the origin point, left to right, up to down are neutral, anger, disgust, fear, happy, sad, surprise)

B. Normalize Landmarks based on Multiple Origin Points

As shown in Figure 3, we divide landmarks into 4 parts, the first part are landmarks located in the left red rectangle, the second are located in the right red rectangle, the third are located in the below red rectangle, the forth are consisting of the remaining landmarks.

The coordinates of origin points of left, right, below are calculated as following:

$$C_n = \frac{(L_1 + L_2)}{2} \tag{1}$$

where C_n represents the coordinates of origin point in the nth corresponding regions, and L_1 represents the coordinates of left corner and L_2 represents the right corner. For instance, in left region, L_1 represents the the coordinates of left eye left corner, and L_2 represents the coordinates of left eye right corner.

As shown in Figure 4, the input vectors in eye and mouth regions are formulated after substract the origin point.

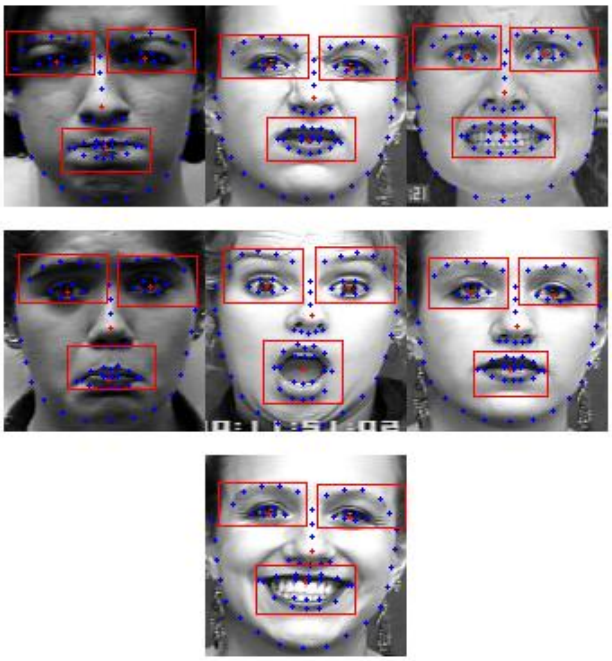

Fig. 3. Landmarks of different regions (The red point represents the origin point of different region coordinates, left to right, up to down are anger, disgust, fear, sad, surprise, neutral, happy).

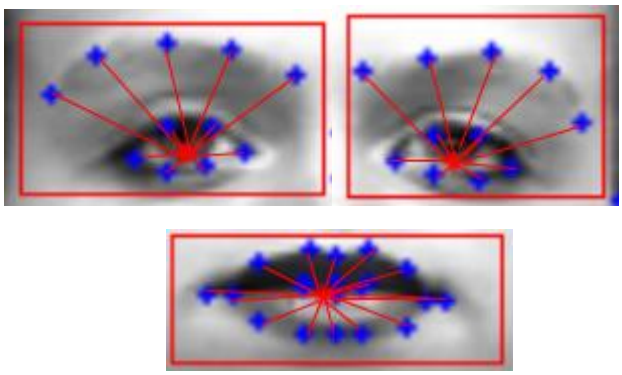

Fig. 4. Input vectors being formulated in eye region and mouth region

C. Training the Neural Network

In our work, we choose 3-hidden layers MLP in our experiment, which contains 100 neurons in the first hidden layer and 100 in the second layer, 500 in the third hidden layer. Aiming to classify expressions into 7 classes (neural, anger, disgust, fear, happy, sad, surprise), our output layer contains 7 neurons. The input layer of MLP contains 136 neurons which is related to all the coordinates of 68 landmarks. The output layer we use activation function is softmax function, and loss function is cross-entropy function.

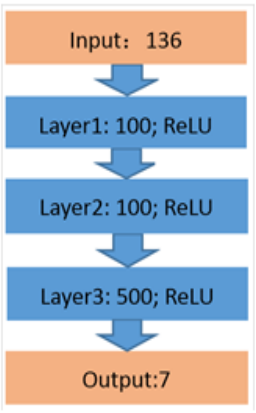

Fig. 5. Network architecture of our MLP model

III. EXPRERIMENTS

In order to demonstrate the effctiveness of our proposed methodology, we did some experiments on CK+ dataset. We choose CK+ dataset as our benchmark which has provided 68 points landmarks for each images. We also compare our method with [19], [20], and [21]. Here, because our input data has been cropped into face images, so we ignore PCA part as described in [17] in our own code. And [18] and [19] code we implemented are based on [22]. The maximum epoch we choose in our implement code of [18] and [19] is 60 and learning rate we choose is 0.01. In our code of implementing our proposed method, the learning rate we choose is 1e-4, and maximum epoch is 1000.

We evaluate the performance of proposed FER methodba sed on Pytorch framework. Two methods [20] and [21] trainin g process is performed using a Standard Tesla P100 PCIe (16 G) a NVIDIA CUDA framework 9.0 and a cuDNN library on the Linux platform. And we evaluate classification time of diff erent methods at Windows 7 with AMD Fx(tm)-8320 Eight-C ore Processor.

A. Exprimental Datasets

The Extened CohnKanade(CK+) dataset [16] contains 593 video sequences from 32 subjects, and 327 video sequences have labels of seven basic expression (anger, contempt, disgust, fear, happiness, sadness and surprise). Each video sequences show a shift from a neutral facial expression to the peak expression. We use the seven expressions, including happiness, sadness, surprise, neutral, disgust,fear and anger. For each sequence, we select the last 5 frames as corresponding expression samples and the first frame as neutral expression. This work we choose 1816
expression images (neutral-316, anger-225, disgust-280, fear-120, happy-345, sad-140, surprise-390). Here we randomly

choose 400 samples (neutral-75, anger-49, disgust-61, fear-27, happy-73, sad-33, surprise-82) as our test set.

B. Experimental Results Analysis

1) The comparison of different methods is indicated in Table I and the confusion matrix of our proposed model is shown in Table II. In our experiment, we first trained the neural network in our own implementation using our train set, then using the trained neural network to make prediction on our test samples. The learning rate we for our neural network we choose is 1e-4, the max training epoch is 2000.

Fig. 6. Seven expressions from CK+ dataset. (The first row are neutral, anger, and disgust. The second are fear, happy, sad and last row are surprise.)

TABLE I. ACCURACY ON CK+ DATASET

Method	ACC. (%)
Transfer learning[19]	91.50
VGG16[20]	90.00
Resnet18[21]	90.00
Proposed(one-origin)	87.75
Proposed(multiple-origin)	**92.00**

TABLE II. CONFUSION MATRIX ON CK+ DATASET (NEU,DN, DIS, FE, HAP, SA, SUR ARE ABBREVIATIONS OF NEUTRAL , ANGER, DISGUST, FEAR, HAPPY, SAD, SURPRISE)

	NEU	AN	DIS	FE	HAP	SA	SUR
NEU	**83%**	5%	4%	0%	1%	0%	7%
AN	10%	**78%**	12%	0%	0%	0%	0%
DIS	0%	0%	**100%**	0%	0%	0%	0%
FE	0%	0%	0%	**100%**	0%	0%	0%
HAP	0%	0%	0%	0%	**100%**	0%	0%
SA	24%	0%	0%	0%	0%	**76%**	0%
SUR	0%	0%	0%	0%	0%	0%	**100%**

TABLE III. RUNNING TIME OF DIFFERENT METHODS ON TEST SET

Method	Time. (ms)
Transfer learning[19]	458
VGG16[20]	183284
Resnet18[21]	74418
Proposed(multiple-origin)	**61**

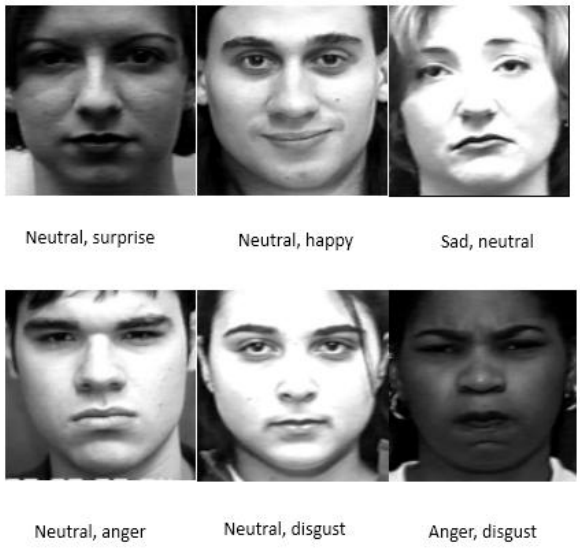

Fig. 7. Images being recognize incorrectly in test set using our proposed multi-origin method (left label below each image is correct label, right is predicted label).

As shown in Table I, our proposed (multiple-origin) model can obtain an accuracy of 92% on our test samples of CK+ dataset while our one-origin method can achieve 87.75%, which indicates that multiple origins method can capture more characteristics of landmarks. And our multiple-origin method can also achieve comparable or better performance with other deep convolutional network, which shows the the great potential of landmarks in recognizing facial expression. As shown in Table II, our proposed multiple-origin method can achieve 100% accuracy in disgust, fear, happy,surprise expressions in our test samples of CK+ dataset. But results show that neutral, anger and sad expressions are not easy to recognize.

As shown in Table III, the time for classification task using our trained model is less than other convolution network methods. Here, we only calculate classification processing time with extracted landmarks as input data for our network. The results in Table III shows that our method need can

quickly detects person's expression, which can indicate that our methods may have a promising prospect in the future.

We also explore some images in test set are recognize incorrectly are shown in Figure 8. From Figure 8, some images are also hard for human to recognize correctly, and some images are confusing. For example, the final image in Figure 7, the expression on the face is mixed with anger and disgust, which adds more difficulties to give an accurate prediction.

I. CONCLUSION

This paper we explore the potential ability for recognizing facial expression based on landmarks, which may could prove that human brain can recognizes facial expressions by using only 68 points instead of all pixels of face image. We use the landmarks provided by CK+ dataset for experiment. Our input vectors is normalized by multiple-origin or one-origin, then trained into MLP for training to classifying expressions. The result on our test samples shows that the method based on landmarks also have comparable performance with CNN based methods. But, the performance of our proposed method also is related to the precision of landmark extracting algorithm. The method we proposed may also give some inspirations to study of human expression cognition process. Sometimes natural expression
always contains multiple emotions, which is hard to recognize for human, let alone for machines. In future, we will improve the robustness and the accuracy of our proposed method in facial expression recognition.

REFERENCES

[1] Ekman P. Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique[J]. 1994.

[2] Ekman P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of personality and social psychology, 1971, 17(2): 124.

[3] Matsumoto D. More evidence for the universality of a contempt expression[J]. Motivation and Emotion, 1992, 16(4): 363-368.

[4] Wang Y Q. An analysis of the Viola-Jones face detection algorithm[J]. Image Processing On Line, 2014, 4: 128-148.

[5] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503.

[6] Verma K, Khunteta A. Facial expression recognition using Gabor filter and multi-layer artificial neural network[C]//2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC). IEEE, 2017: 1-5.

[7] Shan C, Gong S, McOwan P W. Facial expression recognition based on local binary patterns: A comprehensive study[J]. Image and vision Computing, 2009, 27(6): 803-816.

[8] Liu W F, Wang Z F. Facial expression recognition based on fusion of multiple Gabor features[C]//18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 3: 536-539.

[9] Chen, Junkai, et al. "Facial expression recognition based on facial components detection and hog features." International workshops on electrical and computer engineering subfields. 2014.

[10] Ding, Hui, Shaohua Kevin Zhou, and Rama Chellappa. "Facenet2expnet: Regularizing a deep face recognition net for expression recognition." 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

[11] Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." arXiv preprint arXiv:1804.08348 (2018).

[12] Hu, Ping, et al. "Learning supervised scoring ensemble for emotion recognition in the wild." Proceedings of the 19th ACM international conference on multimodal interaction. ACM, 2017.

[13] Zhao, Shuwen, et al. "Feature Selection Mechanism in CNNs for Facial Expression Recognition." BMVC. 2018.

[14] Yao, Anbang, et al. "HoloNet: towards robust emotion recognition in the wild." Proceedings of the 18th ACM International Conference on Multimodal Interaction. ACM, 2016.

[15] Shang, Wenling, et al. "Understanding and improving convolutional neural networks via concatenated rectified linear units." international conference on machine learning. 2016.

[16] Lucey, Patrick, et al. "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.

[17] Tautkute, Ivona, Tomasz Trzcinski, and Adam Bielski. "I know how you feel: Emotion recognition with facial landmarks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[18] Munasinghe, M. I. N. P. "Facial Expression Recognition Using Facial Landmarks and Random Forest Classifier." *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*. IEEE, 2018.

[19] Ravi, Aravind. "Pre-Trained Convolutional Neural Network Features for Facial Expression Recognition." *arXiv preprint arXiv:1812.06387* (2018).

[20] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[21] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[22] Qin, Zhenyue, and Jie Wu. "Visual Saliency Maps Can Apply to Facial Expression Recognition." *arXiv preprint arXiv:1811.04544* (2018).