Multi-Network Fusion Based on CNN for Facial Expression Recognition

Chao Li[1,2], Ning Ma[3,*] and Yalin Deng[3]
[1]Beijing Engineering Laboratory of IOT content security, School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[2]Shenzhen Key Laboratory of Data Vitalization, Research Institute in Shenzhen, Beihang University, Shenzhen 518057, China
[3]School of Computer Science and Engineering, Beihang University, Beijing, China
[*]Corresponding author

***Abstract*—We propose a method which is multi-network fusion (MNF) based on CNN to recognize facial expressions. Our experimental data adopts the ICML2013 facial expression recognition contest's dataset (FER-2013) and JAFFE dataset. Based on the classic Tang's network structure and Caffe-ImageNet structure, we perform pre-training separately to extract the optimal initialization parameters which are applied for the MNF. We adjust the MNF's parameters through fine-tuning and use L2-SVM for classification. Our experiment has achieved a high accuracy, and the result shows that the effect of the MNF is more obvious than a single network on the facial expression recognition. In this paper, we will describe the specific MNF structure and our training process, as well as the accuracy on the test set.**

***Keywords*—facial expression recognition; fusion; multi-network; CNN; SVM**

I. INTRODUCTION

Facial expression recognition (FER) technology is an emerging research area of artificial intelligence, which has broad application prospects in intelligent interpersonal interaction. At the same time, the technology is also widely used in transportation, healthcare and public safety. In recent years, FER technology has attracted more scholars' attention and become a hot topic in the field of artificial intelligence. Therefore, the research on FER technology has typically much value on theoretical research and practical applications. The technology of FER focuses on seven basic categories: neutrality, anger, disgust, fear, happiness, sadness and surprise, which is widely accepted by the public [1]. The seven expression samples in FER-2013 dataset are shown in Figure 1. The experiment on the FER-2013 facial expression dataset is challenging, because the expressions are the real expressions of people in the natural environment, rather than the human expressions making a gesture over against a camera, just like JAFFE dataset. Furthermore, the low image resolution of the dataset and different slope angel increase the difficulty of recognition. In addition, the relatively small dataset makes the training process difficult and the model prone to over-fit [1, 2]. Still, there are many researchers conduct experiments on the FER-2013 and JAFFE datasets, the two datasets have become accepted datasets.

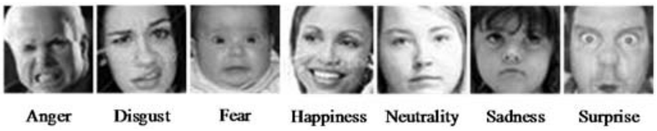

FIGURE I. SEVEN EXPRESSION SAMPLES IN FER-2013 DATASET

In recent year, CNN has become the most popular structure to solve the problem of FER. Theoretically, the more convolution layers are, the more complex features of the active mapping will be obtained [3, 4]. However, deep network is not suitable for small datasets, which can lead to inadequate training. In order to extract discriminative and convincing features of the facial expression, we adopt the technology of MNF. The method is different from the existing methods of FER which are usually trained by single CNN [3], and the performance is better than those. It is not appropriate to apply network structures directly to FER-2013 and JAFFE datasets. The main reason is that the datasets are too small, and too many network parameters make it easy to over-fit in the training process. Therefore, we use the data augmentation in our experiments, after all, collecting more facial expression data is a time-consuming project [5].

In this paper, we use the MNF based on classic Tang's network structure [6] and the Caffe-ImageNet structure [7] to train the dataset respectively. However, the original Softmax classification at the end of the Caffe-ImageNet is replaced by the L2-SVM to classify the expression features. Then we select the most competitive network model from the learning task which can represent the face expression features well as the initialization parameters of MNF, training data through fine-tuning. L2-SVM classifier is used to achieve better effect on the face expression classification. Finally, the recognition accuracy is 68.7% on the FER-2013 validation set, 70.3% on the FER-2013 test set, and 95.7% on the JAFFE dataset (the JAFFE dataset doesn’t distinguish between validation set and test set).

II. RELATED WORK

In the various kinds of classification methods which are widely used at present, the method based on CNN is outstanding. CNN can fuse feature extraction and classification in a framework, reducing the large workload of manual design features. In addition, there are many levels and parameters in deep learning models, it can obtain good effect on image and video.

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

CNN structure is widely used in the recognition and classification of facial expressions. Based on CNN, many researchers have made new changes and innovations to make FER effect better. For example, Tang, the winner of the ICML contest, trained with a CNN structure on the face expression images, and used L2-SVM instead of Softmax to classify facial expression features [6]. In our experiment, L2-SVM classification will also be used. This classification method is suitable for multi-classification tasks, and the result really has slightly improvement. The Caffe-ImageNet structure is designed to classify ImageNet datasets into 1000 categories [7], the number of final output nodes is reduced to seven in [8], but the classification accuracy of Caffe-ImageNet is generally higher than those of other CNN-based methods, such as [9, 10].

Many researchers fused different training networks to train in the latter study. For example, in [5], the extracted appearance features of the image sequence and the extracted geometric features of the temporary facial landmark were merged to enhance the FER effect. Besides, the author in [3] fused different images to extract respective features cross-transformations, and the results showed the effectiveness of the method. In [11], Kim proposed a method which constructed a hierarchical architecture of the committee with exponentially-weighted decision fusion in order to form a better committee in structural and decisional aspects. Our approach is to train two single network model first, and then extract the best network model parameters from them for the MNF's initialization parameters, which is followed with fine-tuning fusion network's parameters. What's more, the MNF structure can be dynamically adjusted, we can try different combinations of the fusion network structure as long as it has a good extraction of the face expression features.

In the process of training, due to the small amount of FER-2013 dataset and JAFFE dataset, it is easy to over-fit when using complex CNN models. In order to solve the problem, researchers in [2] randomly mirror the images, remove 3 pixels in every direction, rotate 45 degrees and magnified to 1.2 times, and finally the transformed image was crop into the size of 42 × 42. There are many predecessors in this field who also use transfer learning between different tasks. Before fine-tuning on the target dataset, the network parameters of CNN are initialized by training on the related tasks [1, 12, 13]. These methods are all better performed than training directly on the original small dataset, and the method we adopt is data augmentation.

III. OUR APPROACH

A. Dataset Preparation

Since the FER-2013 image is naturally-obtained, there may be a problem with uneven lighting, and the illumination or contrast greatly affects the accuracy of the result. Therefore, the original image usually needs to be processed to reduce the illumination effect. Histogram equalization (Hist-eq) is a good way to normalize the image grey-scale value and enhance the discrimination of brightness between the foreground and background in face images [8]. As shown in Figure 2, the contrast of the image enhanced after Hist-eq, and the image details become clearer.

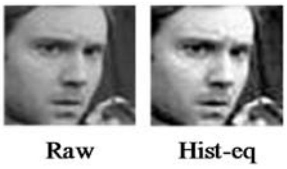

FIGURE II. COMPARISON BETWEEN THE RAW IMAGE AND THE PROCESSED IMAGE WITH HIST-EQ

The image of FER-2013 has low resolution, not aligned, that means in early training of the facial expression classification models, quantity is more important than quality to improve experiment results [1]. Moreover, stochastic disturbances through cropping images can essentially generate additional invisible training samples, thus making the network more robust [9]. In this article, images are cropped and rotated for data augmentation and promoting data diversification. We first rotate each image with fixed angle in {-10, -5, 5, 10} (the experiment proves that the rotation of 15 degrees will decrease the recognition rate, because rotating a large angle will destroy the original features of the image), and this makes the model robust to slight rotation variation of the input images. Figure 3 shows the original image and the rotated image. Then, we proceed with image normalization, decentralized (subtract the mean value), and mirroring image. Finally, we cut the original 48 × 48 face images into five 42 × 42 images (center, top left, top right, bottom left and bottom right), and we get the augmented data.

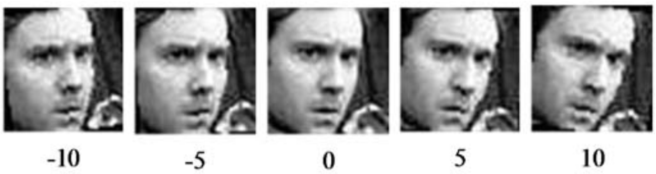

FIGURE III. THE ROTATED FACE IMAGES (THE NUMBERS UNDER THE IMAGES REPRESENT THE ROTATED ANGLES)

B. Network Architecture

Our proposed MNF structure is based on Tang's network structure [6] and Caffe-ImageNet's structure [7], both of them are based on CNN, and the two networks are chosen because they have very good effect on FER problem. Tang's structure using SVM instead of Softmax classification has improved the classification results and won the champion of FER in ICML2013 competition. Caffe-ImageNet structure is a deep level network, which has better ability to extract facial expression features. We use the augmented training set to train the network model, the validation set to tune the learned models, and the test set to test the trained model's ability of distinguishing facial expressions.

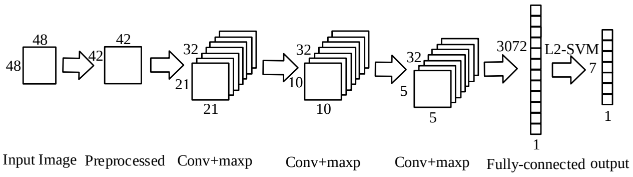

FIGURE IV. TANG'S NETWORK STRUCTURE

The first CNN structure based on Tang's structure consists of three convolutional neural network and max pooling layers, followed by a fully connected layer, and finally using L2-SVM to classify facial expressions into 7 types, as shown in Figure 4. The second network based on Caffe-ImageNet structure consists of two convolutional neural network, max pooling and LRN layers, then two convolutional neural networks, a convolutional network and max pooling layer, followed by a convolutional layer and a fully connected layer. Finally, the original classifier is changed to L2-SVM classifier, as shown in Figure 5. We select the best model parameters from each of the two trained neural network models, regarded as the initialization parameter of the MNF. Then two based network structures are fused. Similarly, L2-SVM classifier is adopted at the end. We keep the parameters before the two networks fuse unchanged. The MNF parameters after fully connected layer are fine-tuned. This process is illustrated in Figure 6. In the network models, rectified linear units (ReLU) are used between the fully connected layer and the output layer, so that the convolutional neural networks can approximate any function, which a linear combination of inputs can't.

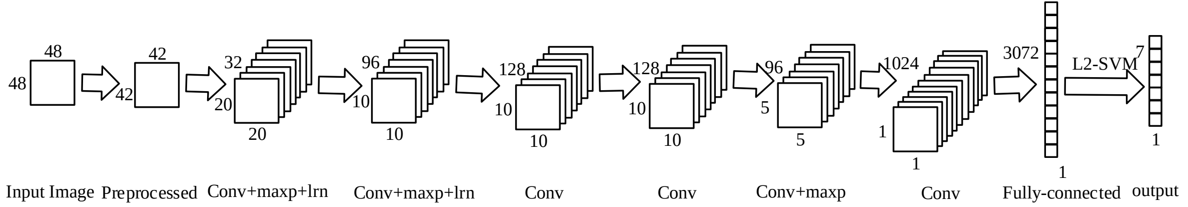

FIGURE V. CAFFE-IMAGENET STRUCTURE

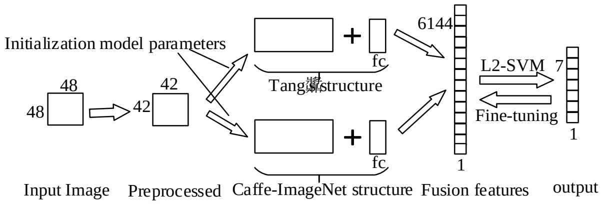

FIGURE VI. THE MNF STRUCTURE

IV. EXPERIMENT

A. Image Database and Experimental Setup

MNF experiments based on CNN are implemented on the FER-2013 dataset and JAFFE dataset. Our experiments are conducted on the NVIDIA GeForce GTX TITAN X GPU using Caffe framework.

The experiment input is a 42 × 42 size of facial expression image, which is preprocessed by clipping, mirror transformation, and reduction of the average value to normalize the image. We use different test methods to evaluate the experiment results on the two datasets. In order to improve the accuracy of classification, the performance between Softmax and L2-SVM classification is compared. It was found that the model effect will be enhanced using the L2-SVM, so we adopt L2-SVM to classify the facial expressions.

B. Experiments on the FER-2013 Database

FER-2013 dataset comes from the ICML 2013 PERL workshop's facial expression recognition contest, which includes 28709 images in training set, 3589 images in validation set, and 3589 images in test set. Each sample is a 48 × 48 face image with an expression label.

As we can see from Figure 7, the best method is Tang's, which achieves accuracy of 71.2%. We reproduce Tang's method, but we achieve accuracy of 64.2% which differs by 5% from Tang's result, maybe we have different initialization parameters. We use Tang's method as one part of the MNF structure, and find that the recognition accuracy of MNF is 70.3% which is higher than that of the individual training network. Therefore, the MNF structure is effective for the improvement on the FER accuracy.

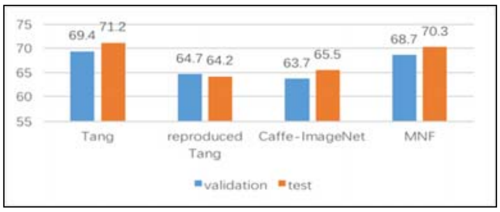

FIGURE VII. THE COMPARISON OF OUR METHOD WITH SOME CLASSIC APPROACHES ON FER-2013 DATASET (%)

TABLE I. CONFUSION MATRIX OF OUR PROPOSED METHOD ON FER-2013 VALIDATION DATASET (%)

	AN	DI	FE	HA	SA	SU	NE
AN	56.1	0.4	9.2	6	9.9	1.9	16.5
DI	19.6	60.7	0	3.6	10.7	0	5.4
FE	7.5	0	46.6	3.4	17.9	8.9	15.7
HA	1.5	0	1.1	85.9	1.5	1.6	8.5
SA	7.5	0	8.0	4.9	58.0	1.7	19.9
SU	1.5	0	3.4	4.6	2.7	84.6	3.4
NE	5.3	0	2.3	7.4	12.5	1.0	71.5

TABLE II. CONFUSION MATRIX OF OUR PROPOSED METHOD ON FER-2013 TEST DATASET (%)

	AN	DI	FE	HA	SA	SU	NE
AN	57.2	0.4	7.7	5.3	14.5	1.4	13.4
DI	23.6	63.6	3.6	1.8	3.6	0	3.6
FE	9.7	0.2	52.1	4.0	15.2	6.8	12.1
HA	1.8	0	1.5	86.9	3.0	1.9	4.9
SA	6.9	0	11.3	5.2	55.9	0.7	20.0
SU	0.5	0.2	4.8	4.8	1.7	84.9	3.1
NE	3.5	0.2	2.6	5.8	10.9	1.0	76.2

Table 1 and Table 2 are confusion matrix of our method, we can find that the accuracy is high in classifying happiness, surprise and neutrality expressions, this is because it's obvious to distinguish the three types of expression images from other types in test set, so as the human eyes. Other expressions, especially fear, have low recognition accuracy because the human eyes can't categorize them well either.

C. Experiments on the JAFFE Database

The JAFFE dataset contains 213 images (resolution: 256 × 256 pixels per image) of Japanese women's faces. There are 10 people in the expression database, and each person has seven expressions. JAFFE database is all frontal face, the face size is basically the same, the illumination is frontal light source, but the illumination intensity is different. Since the expression database is completely open and the expressions are indexed very standardly, it is now used in most studies of expression recognition.

There are no specific training and testing samples in JAFFE dataset, so we use the ten-fold cross-validation method to evaluate the recognition effect. In the experiment, the dataset is randomly divided into ten folds, and each contains about 21 face images. We take nine out of ten for training network models and the remaining one for testing in turn, and finally get an average accuracy of 95.7%. The accuracy rates of ten experiments are shown in Table 3, we can see that nine out of ten of the experimental recognition accuracies is over 90%, and three of them reach 100%. Recognition rate's comparison of our proposed method with the state-of-art methods is shown in Table 4, it shows that the MNF structure we design is more effective.

TABLE III. THE ACCURACY OF OUR EXPERIMENTS ON JAFFE DATASET

Ten times experiment on different folds of the dataset			
Fold number	***Accuracy (%)***	***Fold number***	***Accuracy (%)***
1	95.2	6	100
2	85.7	7	100
3	95.2	8	95.2
4	100	9	90.9
5	95.2	10	100

TABLE IV. THE COMPARISON OF OUR METHOD WITH SOME CLASSICAL APPROACHES ON JAFFE DATASET

Method	***Accuracy (%)***
Zhao X, et al [16]	91.0
Mlakar, et al [17]	86.7
MNF	**95.7**

V. CONCLUSION

In this paper, we introduce a new method which is MNF based on CNN to recognize facial expressions. We first train two network structures in the experiment, one based on Tang's network structure, the other based on Caffe-ImageNet, and then use L2-SVM for classification. The best network model parameters are extracted from the two previous trained networks as the initialization parameters of MNF structure, and then the MNF structure is fine-tuned. Our experiment is conducted on the FER-2013 dataset and JAFFE dataset. The results demonstrate that the recognition effect of MNF is better than that of each single network, and MNF structure is competitive with other state-of-the-art recognition networks. What's more, MNF structure is extensible, we can find possible combinations of other different models for further research.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No.61370122).

REFERENCES

[1] Winkler, Stefan, et al. "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning." ACM on International Conference on Multimodal Interaction ACM, 2015:443-449.

[2] Shin, Minchul, M. Kim, and D. S. Kwon. "Baseline CNN structure analysis for facial expression recognition." IEEE International Symposium on Robot and Human Interactive Communication IEEE, 2016:724-729.

[3] Xie, Siyue, and H. Hu. "Facial expression recognition with FRR-CNN." Electronics Letters 53.4(2017):235-237.

[4] Matthew D. Zeiler, and Rob Fergus. "Visualizing and Understanding Convolutional Networks." 8689(2013):818-833.

[5] Jung, Heechul, et al. "Joint Fine-Tuning in Deep Neural Networks for Facial Expression Recognition." IEEE International Conference on Computer Vision IEEE, 2016:2983-2991.

[6] Tang, Yichuan. "Deep Learning using Support Vector Machines." Eprint Arxiv (2013).

[7] Jia, Yangqing, et al. "Caffe: Convolutional Architecture for Fast Feature Embedding." Acm International Conference on Multimedia ACM, 2014:675-678.

[8] Devries, Terrance, K. Biswaranjan, and G. W. Taylor. "Multi-task Learning of Facial Landmarks and Expression." Computer and Robot Vision IEEE, 2014:98-103.

[9] Yu, Zhiding, and C. Zhang. "Image based Static Facial Expression Recognition with Multiple Deep Network Learning." ACM on International Conference on Multimodal Interaction ACM, 2015:435-442.

[10] Kahou, Samira Ebrahimi, et al. "Combining modality specific deep neural networks for emotion recognition in video." ACM on International Conference on Multimodal Interaction ACM, 2013:543-550.

[11] Kim, Bo Kyeong, et al. "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition." Journal on Multimodal User Interfaces 10.2(2016):1-17.

[12] Ross Girshick, et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." (2013):580-587.

[13] Yosinski, Jason, et al. "How transferable are features in deep neural networks?." International Conference on Neural Information Processing Systems MIT Press, 2014:3320-3328.

[14] Mayya, Veena, R. M. Pai, and M. M. M. Pai. "Automatic Facial Expression Recognition Using DCNN ☆." Procedia Computer Science93(2016):453-461.

[15] Mollahosseini, Ali, D. Chan, and M. H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." Applications of Computer Vision IEEE, 2016:1-10.

[16] Yadan Lv, Zhiyong Feng, and Chao Xu. "Facial expression recognition via deep learning." IETE Technical Review 32.5(2015):347-355.

[17] Lopes, André Teixeira, et al. "Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order." Pattern Recognition 61(2017):610-628.