Facial Expression Recognition Based on Multi-scale CNNs

Shuai Zhou[1,2,3], Yanyan Liang[1], Jun Wan[2,3(✉)], and Stan Z. Li[2,3]

[1] Faculty of Information Technology,
Macau University of Science and Technology, Macau, China
shuaizhou.palm@gmail.com, yyliang@must.edu.mo
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jun.wan,szli}@nlpr.ia.ac.cn
[3] University of Chinese Academy of Sciences, Beijing, China

Abstract. This paper proposes a new method for facial expression recognition, called multi-scale CNNs. It consists several sub-CNNs with different scales of input images. The sub-CNNs of multi-scale CNNs are benefited from various scaled input images to learn the optimalized parameters. After trained all these sub-CNNs separately, we can predict the facial expression of an image by extracting its features from the last fully connected layer of sub-CNNs in different scales and mapping the averaged features to the final classification probability. Multi-scale CNNs can classify facial expression more accurately than any single scale sub-CNN. On Facial Expression Recognition 2013 database, multi-scale CNNs achieved an accuracy of 71.80 % on the testing set, which is comparative to other state-of-the-art methods.

Keywords: Facial expression recognition · Multi-scale CNNs · CNN · Deep learning · Patten recognition

1 Introduction

Human facial expression, as an external representation of psychological states, plays an important role in social communication and observation. We can learn one's emotion and mental activity better by seeing his or her facial expression than just listening to words. Researches on facial expression started from nineteen century, Darwin claimed the universality of emotions [2], later on in 1971, Paul Ekman classified six facial expressions as basic: anger, disgust, fear, happiness, sadness, and surprise [3].

In the future, facial expression recognition will be a vital part of artificial intelligence. It can be applied into different crucial fields, such as human computer interaction, criminal investigation and commercial analysis. It will be of great significance to national security and economic development.

Though facial expression recognition has been researched for years, there still are several obstacles to overcome. This is a cross-discipline research, it needs the

Z. You et al. (Eds.): CCBR 2016, LNCS 9967, pp. 503–510, 2016.
DOI: 10.1007/978-3-319-46654-5_55

cooperation between computer science, social psychology and medical science. The difficulties that face detection research was facing remain in facial expression recognition. Large visual variations from human faces in the cluttered backgrounds and variational head poses and different lighting environments make this issue more and more complicated.

In recent years, convolutional neural network has been widely applied in image and video recognition [8,10], recommender systems and natural language processing, and have made great success in these fields. In image pattern recognition, most of the state of the art model were developed or constructed from convolutional neural network. In 2014, DeepID [10] convolutional neural network architecture were proposed by Tang et al. for face verification. In 2015, Kaiming He proposed deep residual networks [4] and won the 1st places in: ImageNet classification, ImageNet detection and ImageNet localisation. For face detection, a convolutional neural network cascade [8] were proposed by Li et al.

To solve facial expression and emotion recognition problem, some methods were reported in [7,9,11]. In particular, Tang [11] reported a deep CNN jointly learned with a linear support vector machine (SVM) output. This method achieved the first place on both validation and testing subset on the FER 2013 Challenge [1]. Liu et al. [9] proposed a facial expression recognition framework with 3D CNN and deformable action parts constraints in order to jointly localizing facial action parts and learning part-based representations for expression recognition. Yu and Zhang [12] proposed a method contains a face detection module based on the ensemble of three state-of-the-art face detectors, followed by a classification module with the ensemble of multiple deep convolutional neural networks (CNN). They used two schemes to learn the ensemble weights of the network responses: by minimizing the log likelihood loss, and by minimizing the hinge loss.

Considering the high quality features extracted from convolutional neutral network, we choose CNN to solve the facial expression recognition problem too. During this research, unlike using simple single CNN or even deep complex CNN, we proposed multi-scale CNNs, a method combined multiple simple CNNs with different input image size, and predicting classes together. Multi-scale CNNs can classify facial expression more accurately than any single scale sub-CNN of it. At last, the multi-scale CNNs got an accuracy of 71.80 % on FER database testing set, surpassed the result of the winner of FER Challenge 2013.

2 Multi-scale CNNs

Multi-scale CNNs are two or more CNN models which have different scales of input images. It predicts the result by fusing the features extracted from each sub-CNN, then makes the final prediction based on all the features extracted.

2.1 Training Multi-scale CNNs

To solve facial expression recognition problem, we use three sub-CNNs of different input scale rates of 1.0x, 2.0x, 2.14x. As showed in Fig. 1, we adopt similar

network architecture of winner of FER Challenge 2013. There are three convolutional layers, three pooling layers and two fully connected layers in each net. The number of output neurons in last fully connected layer is 7, because the samples of facial expression images were labeled to seven categories: angry, disgust, fear, happy, sad, surprise and neutral.

The difference between sub-CNNs is that the input image size for 1.0x sub-CNN is 42×42 pixels, 84×84 for 2.0x sub-CNN and 90×90 for 2.14x sub-CNN, because the training images, training set from Facial Expression Recognition Challenge 2013, are in shape of 48 by 48 pixels grayscale. 1.0x sub-CNN uses randomly cropped images from the raw images, and 2.0x and 2.14x sub-CNNs use resized 96 by 96 pixels images with random cropping size of 84 by 84 and 90 by 90 pixels. Also the kernel sizes of convolutional layers are different. Since the input image size is larger for 2.0x and 2.14x sub-CNNs, the first convolutional layer's kernel size is enlarged to 7×7 and kernel size of second convolutional layer is 5×5 instead of 4×4 for 1.0x sub-CNN. We apply PReLU nonlinearity function [5], after each pooling layer and second last fully-connected layer.

In the training phase, training samples were first resized to three different sizes and each sub-CNNs were trained separately with corresponding scaled

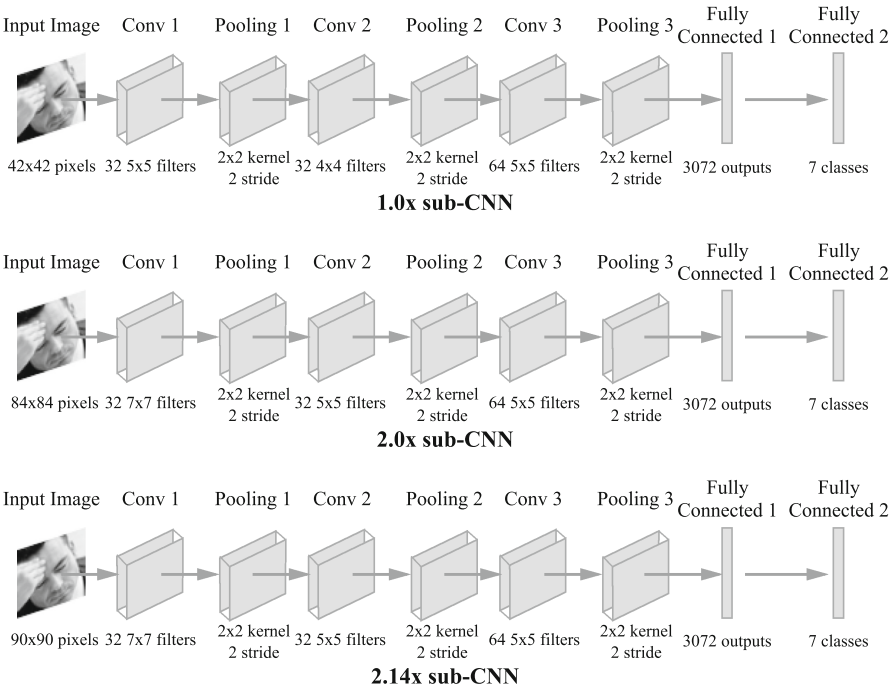

Fig. 1. Architecture of each sub-CNNs. Input image size and kernel size of partial convolutional layers are different. Please note that all these three sub-CNNs were trained separately.

training samples. In order to increase the classification accuracy, we compared the performances between softmax with loss and hinge loss as the loss function to train each sub-CNNs, and experiment of comparison between each loss function will be detailed later in Sect. 3.1. As proposed by [11], using hinge loss do enhanced the model a little.

2.2 Classification Using Multi-scale CNNs

As for predicting a test image, the image need to be resized to corresponding size and then feed into one of the different scale CNNs according to its size. After all the forward calculations of each network were done, extract features from the last fully connected layer, which should get three 7-dimensional vectors from three sub-CNNs, and we denote those feature matrixes as F_1, F_2 and F_3 for 1.0x scale, 2.0x scale and 2.14 scale CNNs. As shown in Eq. (1), we calculate the arithmetic mean on the same dimension, denoted by j, of extracted F_1, F_2 and F_3 as the averaged feature, then we got $\overline{F} : \{\overline{f_j},\ j = 1, ..., 7\}$, since we have 7 categories of facial expressions. N is 3, because there are three sub-CNNs.

$$\overline{f_j} = \frac{1}{N} \sum_{i=1}^{N} f_{ij},\ j = 1, ..., 7\ . \tag{1}$$

At last, for seven classes, $K = 7$, we use softmax function Eq. (2) to map the averaged feature to categorical probability distribution. So that each $\sigma(f)_j$ represents the probability of that class.

$$\sigma(f)_j = \frac{\exp(\overline{f_j})}{\sum_{k=1}^{K} \exp(\overline{f_k})},\ j = 1, ..., K\ . \tag{2}$$

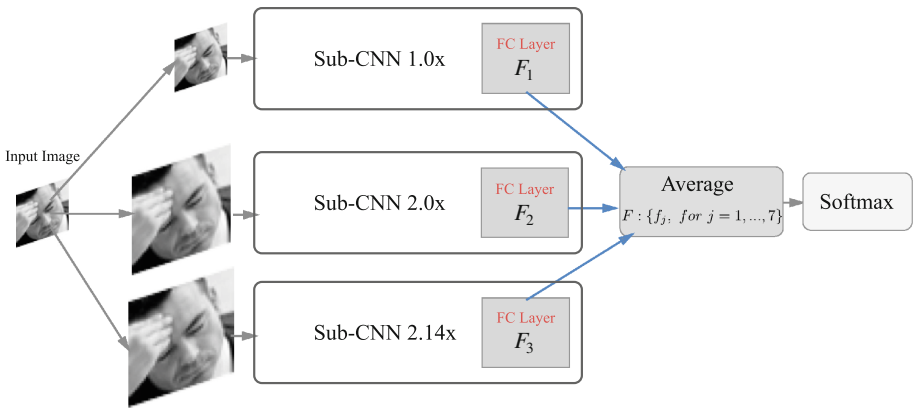

Fig. 2. Illustration of Multi-scale CNNs. Test images were first resized to different sizes then feed into corresponding sub-CNN, and features of last FC layer were extracted and combined, then using softmax to map the averaged features to classification probability.

3 Experiments on FER 2013 Database

The FER 2013 database was published for the challenge in facial expression recognition at [1]. There are 35,887 facial expression images in this database and they are in form of grayscale 48×48 pixels. The publisher claimed that all the faces have been automatically registered and the face is more or less centered at the picture. Image samples as shown in Fig. 3. The training set consists of 28,709 images, and both validation and testing sets have 3,589 images. All the images in this database were labeled to one of the seven categories: angry, disgust, fear, happy, sad, surprise and neutral. Table 1 showed the number distribution of each categories.

Fig. 3. Example images of seven classes from FER 2013 database.

Table 1. Image number distribution of each categories of FER 2013 database

	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral	Total
Training Set	3995	436	4097	7215	4830	3171	4965	28709
Validation Set	467	56	496	895	653	415	607	3589
Testing Set	491	55	528	879	594	416	626	3589
Total	4953	547	5121	8989	6077	4002	6198	35887

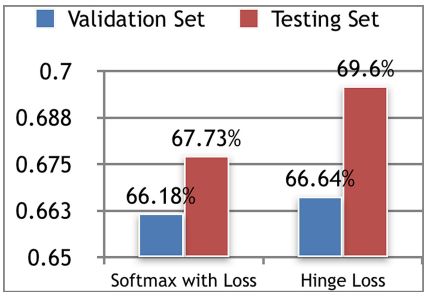

Fig. 4. Accuracy of 1.0x sub-CNN with softmax loss and hinge loss on FER 2013 database.

3.1 Comparison Between Softmax Loss and Hinge Loss

In Sect. 2.1, we tested softmax with loss and hinge loss functions to find the best loss function for solving this facial expression recognition problem. The experiment was conducted on FER 2013 using 1.0x scale CNN and only the last loss function is different. As shown in Fig. 4, by using hinge loss, the accuracy of the model do increase around one percent than softmax loss. So we choose to use the model with hinge loss in later experiments.

3.2 Performance of Sub-CNNs and Multi-scale CNNs

Figure 5 shows the accuracy of different scale sub-CNNs and the multi-scale CNNs. We can see that the accuracy of multi-scale CNNs overcomes all other single scale sub-CNNs. Various scaled input images benefit the models to learn the optimalized parameters, and multi-scale CNNs combined all of these representative features and fused them to make the final prediction together. Multi-scale CNNs do need a little bit more time compared to single sub-CNN, and with the TITAN X GPU, it costs 43 ms to predict one image while 2.0x sub-CNN only takes 18 ms. Figure 6 shows the confusion matrix of multi-scale CNNs.

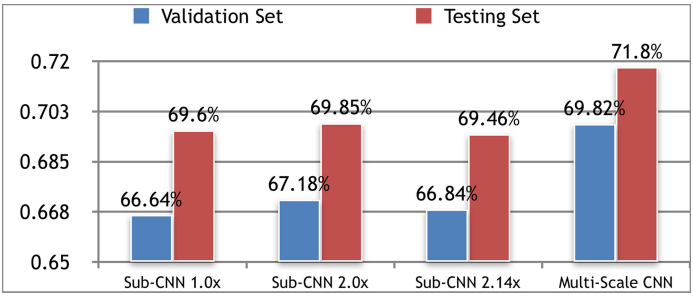

Fig. 5. Accuracy of single sub-CNNs with different scales and the multi-scale CNNs

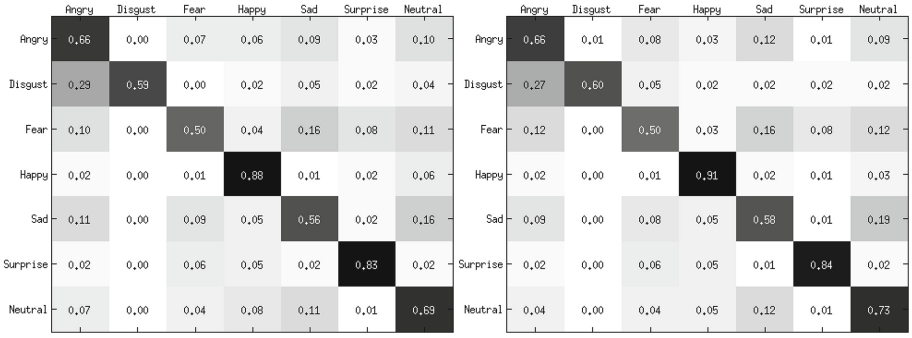

Fig. 6. Confusion matrix of multi-scale CNNs on FER 2013. (Left, validation set. Right, testing set)

3.3 Comparison with State-of-the-Art Methods

We compare our multi-scale CNNs with other state-of-the-art method, and Fig. 7 showed the performance of these methods on FER 2013 database. Our method surpassed the winner of FER 2013 Challenge on both validation and testing set of FER 2013 database. But Multiple Deep Network Learning [12] do perform a little better than us, and this may because that it used more complex data preprocessing like data perturbation and voting.

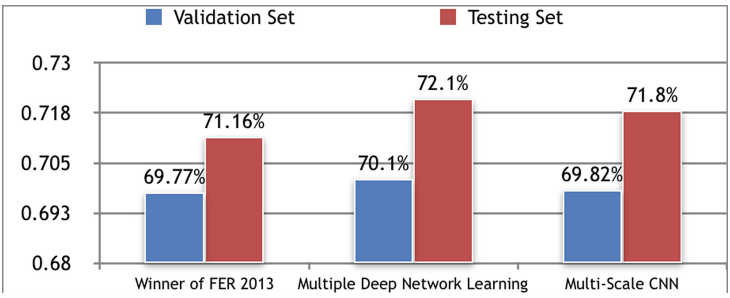

Fig. 7. The recognition accuracy of the proposed method is comparative to other methods.

4 Conclusion

In this work, we propose a new method for facial expression recognition, called multi-scale CNNs. Multi-scale CNNs consists of three sub-CNNs with different image input size and each sub-CNN is designed best to fit with the input size. After trained these sub-CNNs separately, the prediction extracted the output features of last fully connected layer and combined them by calculating the arithmetic mean on same dimension, then used softmax to map the averaged features to categorical probability of facial expression. At last, experiments on FER 2013 database showed that multi-scale CNNs is comparative to other state-of-the-art methods.

Acknowledgement. This work was supported by the National Key Research and Development Plan (Grant No.2016YFC0801002), the Chinese National Natural Science Foundation Projects ♯61473291, ♯61572501, ♯61502491, ♯61572536, Science and Technology Development Fund of Macau (No. 019/2014/A1), NVIDIA GPU donation program and AuthenMetric R&D Funds.

References

1. Challenges in representation learning: facial expression recognition challenge. https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge. Accessed 30 June 2016

2. Darwin, C., Ekman, P., Prodger, P.: The Expression of the Emotions in Man and Animals. Oxford University Press, USA (1998)
3. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. ArXiv e-prints, 12 (2015)
5. He, K., Zhang, X., Ren, S., Sun, J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
6. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
7. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 543–550. ACM (2013)
8. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)
9. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
10. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
11. Tang, Y.: Deep learning using linear support vector machines (2013). arXiv preprint: arXiv:1306.0239
12. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the ACM on International Conference on Multimodal Interaction, pp. 435–442. ACM (2015)