2018 13th IEEE International Conference on Automatic Face & Gesture Recognition

From Macro to Micro Expression Recognition: Deep Learning on Small Datasets Using Transfer Learning

Min Peng
Chongqing Institute of Green Intelligent Technology
Chinese Academy of Sciences
Chongqing, China

Zhan Wu, Zhihao Zhang, Tong Chen*
Chongqing Key Laboratory of Non-linear Circuit and Intelligent Information Processing
Southwest University
Chongqing, China
*e -mail: c_tong@swu.edu.cn

***Abstract*—This paper presents the methods used in our submission to 2018 Facial Micro-Expression Grand Challenge (MEGC). The object of the challenge is to recognize micro-expression in two provided databases, including holdout-database recognition and composite database recognition. Considering the small size of the databases, we follow a rout of transfer learning to implement convolutional neural network to recognize the micro-expression. ResNet10 pre-trained on ImageNet dataset was fine-tuned on macro-expression datasets with large size and then on the provided micro-expression datasets. Experimental results show that the method can achieve weighted average recall (WAR) of 0.561 and unweighted average recall (UAR) of 0.389 in Holdout-database Evaluation Task, and F1 Score of 0.64 in Composite Database Evaluation Task, which are much higher than what baseline methods (LBP-TOP, HOOF, HOG3D) can achieve.**

Keywords-micro expression recognition; deep learning; trasnfer learning

I. INTRODUCTION

Micro Expression is a rapid and weak facial movement that can hardly controlled by human will. It often reveals people's genuine emotion. Therefore, the recognition of micro-expression finds applications in many areas, such as emotion monitoring [1], criminal detection [2], and homeland security [3]. Due to the characteristics of micro-expression, i.e. short in duration and low in intensity of facial movement, the recognition of micro-expression has been a challenging problem.

Methods that were successfully used in recognizing macro-expression were modified for the micro-expression recognition. Yan et al. [4] combined Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) and SVM to perform the recognition. Wang et al. [5] used LBP-Six Intersection Points (LBP-SIP) to perform the recognition. Li et al. [6] used LBP-TOP, Histogram of Oriented Gradients (HOG), and Histogram of Image Gradient Orientation (HIGO) to perform the recognition.

Optical Flow was recently used as a good feature for the recognition. Liu et al.[7] used Main Directional Mean Optical-flow (MDMO) to estimate subtle facial movement for the recognition. Xu et al. [8] used Facial Dynamics Map (FDM) to characterize micro-expression for the recognition.

Patel at al [9] used features extracted from ImageNet by using ImageNet VGG f CNN and features extracted from macro-expression databases by using a designed CNN to recognize micro-expression. The features were further selected by using the genetic algorithm before inputting to classifiers. It is found that the selected features from the designed CNN is better than the features from VGG f CNN in terms of recognition rate. The recognition performance achieved by this method (features extracted from macro-expression by using CNN + classifier) is better than baseline method (LBP-TOP) but worse than the state-of-art method (STCLQ). Peng et al [10] designed a middle size neural network called Dural Temporal Scale Convolutional Neural Network (DSTCNN) for the micro-expression recognition. This is the first time that deep learning is used directly on the micro-expression datasets so that the features of micro-expression can be directly extracted. To avoid overfitting, the DSTCNN did not employ very deep architecture: it only has 4 convolutional layers and 4 pooling layers to adapt to the small databases of micro-expression. The recognition rate achieved by DSTCNN is nearly 10% higher than some state-of-art methods (STCLQ, MDMO, and FDM).

In this paper, we still employ CNN to recognize micro-expression. What makes our work different is we employ a deep (not medium) CNN architecture [11] and train it directly on the micro-expression datasets. To avoid overfitting, we follow the route of transfer learning. This transfer-learning-based method has proved to be efficient in applying deep CNN on small databases. Specifically, RestNet10 [12] trained on ImageNet [13] was fine-tuned on some public macro-expression databases, and finally fine-tuned on the CASMEII [14] and SAMM [15] databases by using apex frames. The experimental results show that the proposed method can achieve weighted average recall (WAR) of 0.561 and unweighted average recall (UAR) of 0.389 in Holdout-database Evaluation Task, and F1 Score of 0.64 in Composite Database Evaluation Task.

II. RELATED WORK

CNN is very successful in image-related recognition tasks [16]. It can extract high-level features from the raw images. However, a CNN with deep architecture has thousands of weights that need to be determined. Only very large database can be used to train deep CNN. Therefore, deep CNN cannot be directly applied on small databases.

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00103

The transfer learning transfers knowledge in one or more sources tasks to improve the learning in target task. It can be employed to solve the problem of applying deep CNN on small databases [17]. A normal route for this kind of method would be initializing the weights on general large image databases, and then fine tuning the weights on the target small databases. We will follow this route in the paper.

III. METHOD

A. CNN Architecture

Residual Networks (ResNet) was presented by He et al [18] in 2015, and achieved good performance in many recognition tasks. ResNet features a stack of residual blocks. A typical residual block is illustrated in Fig. 1, where Conv indicates convolutional layer. In each block, a shortcut connection (the rout from x to F(x) directly in Fig. 1) is used to do element-wise summarization of the input and output of the block. This design could improve the network performance with less training parameters. By stacking such blocks, the whole network is more robust to the degenerating problem. In this paper, we use ResNet10 [12] as network for the micro-expression recognition.

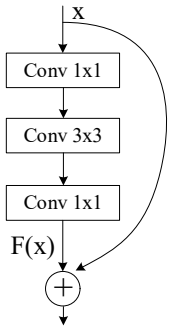

Fig. 1. A typical residual block

B. Training in Macro-expression

The ResNet10 was trained on ImageNet by using the Caffe framework. The top-1 and top-5 Single-crop error rates of ResNet on ImageNet were 36.1% and 14.8%, respectively. Since the micro-expression datasets are generally small in size, we will employ the ResNet10 by fine-tuning it to avoid the problem of small datasets.

Although the ResNet10 was trained on ImageNet, we believe that the expression recognition task is different from the other recognition task. Considering that there are many public macro-expression datasets available, we therefore trained the ResNet10 on four macro-expression datasets. They are the Extended Cohn-kanade dataset (CK+) [19], Oulu-CASIA NIR&VIS facial expression [20], Jaffe [21], and MUGFE [22].

CK+ [19] includes 593 video clips of 123 subjects. . The action units were labeled on the last frame of every image sequence. Among the 593 clips, 327 clips have emotion type labels, which are anger, contempt, disgust, fear, happiness, sadness, and surprise. Each clip starts from normal facial expression frame and ends at apex frame. We selected the last three frames from each corresponding video clip (belonging to 5 types of facial expressions in MEGC 2018) and finally got 852 images.

Oulu-CASIA NIR&VIS facial expression database [20] contains videos from 80 subjects between 23 to 58 years old. 73.8% of the subjects are males. It includes six typical expressions, i.e. happiness, sadness, surprise, anger, fear, and disgust. The whole database includes two parts, one was taken in in Oulu, consisting 50 subjects, most of whom are Finnish people. The other was taken in Beijing, consisting of 30 subjects, all of whom are Chinese people. The frame rate is 25fps and image resolution is 320*240 pixels. The videos were captured by two imaging systems, i.e. NIR (Near Infrared) and VIS (Visible light) systems. Three different illumination conditions were used when capturing the videos, i.e. normal indoor illumination, weak illumination (only computer display is on) and dark illumination (all lights are off). In our work, we selected the last three frames of each video taken by VIS system under normal indoor illumination and finally got 1200 images.

Jaffe [21] was published in 1998. The database contains 219 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each woman was asked to perform 7 types of expressions, i.e. sad, happy, angry, disgust, surprise, fear, and neutral. In our work, we selected images from the corresponding the images with right expression labels and finally got 151 images.

The MUGFE [22] database consists of 1032 video clips of 86 subjects. There are 35 women and 51 men, all of them are Caucasian origin between 20 and 35 years of age. The frame rate is 19fps, and resolution is 896×896 pixels. One part of database contains performed six basic expressions, and the other part includes laboratory induced expressions. Each clip contains 50 to 160 frames, and start from and end at normal expression frame with apex frame in between. We selected 6 to 10 frames near the apex frames from each clip and finally got 8228 images.

We selected 10431 images from four macro-expression datasets to from a new datasets. The face area of each image was segmented out by using AAM [23] and then normalized to 224*224 pixels. We randomly selected 1/10 of the new dataset as test set and the rest as training set. To avoid unbalanced problem in the training set, we employed resampling techniques and got 5000 images on each type of training set. Moreover, to increase the robustness of the network, we perform color shift (maximum value of 20), rotation (maximum degree of 10), smoothing (maximum window size of 6) with probability of 0.5 on images in the training set. We employed batch gradient descent with a momentum of 0.9 during training. The batch-size was set as 50, learning rate was initialized as 0.01, which then decreased 10 times after every 20 epochs. After 100 epochs, the ResNet10 (previously trained on ImageNet) can achieve average recognition rate of 99.35% on 5 types of macro-expressions.

C. Transfering Kownledge to CASMEII and SAMM

The MEGC 2018 uses CAMSEII [14] and SAMM [24] datasets. For the CASME II, the temporal resolution is 200 fps and the face resolution is around 280×340 pixels. 255 micro-expression sequences were selected with action units and emotions labeled. 26 participants were a mean age of 22.03 year (standard deviation (SD) of 1.6). In the SAMM, there are 159 spontaneous micro-facial movements, 32 participants (13 different ethnicity; a mean age of 33.24, SD=11.32; 16 male and 16 female

participants). Both SAMM and CASMEII provides 7 types of micro-expressions. The MEGC 2018 regrouped the micro-expressions in CASMEII according to the movements of action units and suggested 5 types of micro-expressions shown in Table I [25].

TABLE I. THE SUMMARY OF THE DISTRIBUTION OF SAMPLES FOR CASME II AND SAMM

Objective Class	CASMEII	SAMM
I	25	24
II	15	13
III	99	20
IV	26	8
V	20	3
Total (I+II+III+Iv+V)	185	68

The onset frame, apex frame, and offset frame of every micro-expression clip in both datasets are labeled. We located each apex frame and employed AAM [23] to segment the face area out from the apex frame. The face area of each apex frame was then normalize into 224*224 pixels and input to the network for recognition.

There are two tasks in MEGC 2018: Holdout-database Evaluation (HDE) and Composite database evaluation (CDE). HDE requires training on CASMEII and testing on SAMM, vice versa. CDE requires combining both CAMSEII and SAMM into a single composite database and performing Leave-One-Subject-Out cross-validation (LOSO). The recognition performance of baseline methods (LBP-TOP, HOOF, HOG3D) are provided by MEGC 2018.

For the CDE task, the results of baseline methods in CASMEII and SAMM are given separately by MEGC 2018. To compare our work with baseline results [24], we also report the LOSO results in CASMEII and SAMM separately in CDE task.

- Holdout-database Evaluation (HDE)

a) The first fold: CASMEII as training set, SAMM as test set. Data augmentation is performed firstly. The number of images in each type of micro-expression is increased to 200 by using resample techniques. The color shift (maximum value of 20), rotation (maximum degree of 8) were performed with probability of 0.5 during training process. The batch-size was set as 100, learning rate was initialized as 0.0001, which then decreased 2 times after every 10 epochs.

b) The second fold: SAMM as training set, CASMEII as test set. The data augmentation method and setting of the training method is the same as that in a).

- Composite database evaluation (CDE)

a) For SAMM dataset, 20 training-test processes are required. In each process, the training set was enlarged by using resampling techniques so that the samples in each type have the same number. Moreover, each image to the network was normalized to 240*240 pixels, and then cut into a 224*224 image by randomly cutting the four corners of the 240*240 image. The 224*224 images were then used as input images to the network. The batch-size was set as 10, learning rate was initialized as 0.001, which then decreased 10 times after every 20 epochs, and the weight decay was set as 0.05.

b) For CASMEII dataset, 26 training-test processes are required. The training and parameter settings of b) are the same as those of a).

c) For composite dataset (CASMEII+SAMM), 46 training-test processes are required. The training and parameter settings of b) are the same as those of a).

IV. RESULTS AND DISCUSSION

A. Holdout-database Evaluation (HDE)

a) The first fold: CASMEII as training set, SAMM as test set. After 50 epochs, the ResNet10 (previously trained on ImageNet and macro-expression datasets) can achieve WAR) and UAR of 0.544 and 0.44, respectively. The confusion matrix is given in Table II

TABLE II. CONFUSION MATRIX OF THE FIRST FOLD VALIDATION IN HDE

	I	II	III	IV	V
I	**91.67**	0	4.17	4.17	0
II	23.08	**7.69**	30.77	15.38	23.08
III	30	0	**50**	10	10
IV	50	0	12.5	**37.5**	0
V	0	0	66.67	0	**33.33**

b) The second fold: SAMM as training set, CASMEII as test set. The data augmentation method and setting of the training method is the same as that in a). After 50 epochs, the ResNet10 (previously trained on ImageNet and macro-expression datasets) can achieve WAR and UAR of 0.578 and 0.337, respectively. The confusion matrix is given in Table III

TABLE III. CONFUSION MATRIX OF THE SECOND FOLD VALIDATION IN HDE

	I	II	III	IV	V
I	**20**	24	56	0	0
II	0	**46.67**	53.33	0	0
III	0	2.02	**93.94**	4.04	0
IV	3.85	11.54	76.92	**7.69**	0
V	0	10	85	5	**0**

The average WAR and UAR achieved by baseline methods [24] and proposed method of both folds are summarized in Table IV.

TABLE IV. WAR AND UAR ACHIEVED BY DIFFERENT METHODS IN HDE

	Hold-database Evaluation (HDE)	
	WAR	***UAR***
LBP-TOP	0.285	0.332
HOOF	0.353	0.348
HOG3D	0.363	0.228
Our method	**0.561**	**0.389**

It is seen from Table IV that the posed method can achieve much higher WAR and UAR that those of baseline methods. This may suggest deep CNN with transfer learning is an alternative suitable way for recognize micro-expressions from small datasets.

From the confusion matrixes, we can observed that the recognition performance is worst on type II expression in the first fold validation (Table II). This is due to the training samples of type II expression in the CASMEII is the least (see Table I). Because the training samples of type IV and type V expression in the SAMM is the less

than others, we can also see that the recognition performance on type IV and V expression in the second fold validation (Table III) are worse than those of other type of expressions. The recognition performance of one system that uses the characteristics of micro-expressions learnt from another system is heavily influenced by unbalanced distribution of source system. A balanced and large enough micro-expression dataset are desperately needed to improve the recognition performance of this kind.

B. Composite database evaluation (CDE)

a) For SAMM dataset, the average accuracy and F1 score of baseline methods [24] and our method are summarized in Table V

TABLE V. RECOGNITION ACCURACY AND F1 SCORE OF DIFFERENT METHODS ON SAMM IN CDE

	Leave-One-subject-Out (LOSO)	
	Accuracy (%)	***F1 score***
LBP-TOP	44.70	0.35
HOOF	42.17	0.33
HOG3D	34.16	0.22
Our method	**70.59**	**0.54**

b) For CASMEII dataset, the average accuracy and F1 score of baseline methods [24] and our method are summarized in Table VI

TABLE VI. RECOGNITION ACCURACY AND F1 SCORE OF DIFFERENT METHODS ON CASMEII IN CDE

	Leave-One-subject-Out (LOSO)	
	Accuracy (%)	***F1 score***
LBP-TOP	67.80	0.51
HOOF	69.64	0.56
HOG3D	69.53	0.51
Our method	**75.68**	**0.65**

It is seen from Table IV and V that our method outperforms the baseline methods. For SAMM dataset, the proposed method can achieve 70.59% accuracy (25% higher than that of LBP-TOP that has best performance among the three baseline methods), and achieve a F1 score of 0.54 (0.19 higher than that of LBP-TOP). For the CASMEII, the proposed method can achieve 75.68% accuracy (16% higher than that of HOOF that has best performance among the three baseline methods), and achieve 0.65 of F1 score (0.09 higher than that of HOOF).

The proposed method can achieve better results in CASMEII dataset. This trend accords with the trends of the three baseline methods.

c) For the composite datasets (CASMEII +SAMM), the average accuracy and F1 score of our method are 74.70% and 0.64, respectively (shown in Table VII). Because the performances of the baseline methods on the composite datasets are not provided by MEGC 2018, we did not report them in Table VII.

TABLE VII. RECOGNITION ACCURACY AND F1 SCORE OF OUR METHODSON THE COMPOSITE DATASETS IN CDE

	Leave-One-subject-Out (LOSO)	
	Accuracy (%)	***F1 score***
Our method	**74.70**	**0.64**

The recognition performance of the proposed method on the composite database is better than that on SAMM and similar to that on CASMEII. This may be because CASME II contributes more suitable features to ResNet10 in this experiment.

V. CONCLUSION

In this paper, we employed a deep neural network to recognize micro-expression in small database. Transfer learning was used so that the deep ResNet10 which requires large amount of data can be used. The ResNet10 was fined tuned on four public macro-expression databases and then on the provided micro-expression databases. Only the apex frame of each micro-expression clip was input to the ResNet10 for the recognition.

We have finished two recognition tasks in MEGC 2018. The experimental results show that the proposed methods can achieve much higher WAR and UAR than those of baseline methods in HDE task, and much higher F1 score in CDE task.

ACKNOWLEDGMENT

We would like to thank the financial support from National Natural Science Foundation of China (No. 61301297).

REFERENCES

[1] S. Porter, B. L. Ten, "Reading between the lies: identifying concealed and falsified emotions in universal facial expressions," Psychological Science, 2008, 19(5):508-514.

[2] T.A. Russell, E. Chu, M.L. Phillips, "A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool," British Journal of Clinical Psychology, 2006, 45(4):579–583.

[3] S. Weinberger, "Airport security: Intent to deceive," Nature, 2010, 465(7297):412-5.

[4] G. Zhao, M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(6):915.

[5] Y. Wang, J. See, C.W. Phan, et al, "LBP with Six Intersection Points: Reducing Redundant Information in LBP-TOP for Micro-expression Recognition,"Computer Vision--Asian Conference on Computer Vision.Springer International Publishing, 2014:21–23.

[6] X. Li, X. Hong, A. Moilanen, et al, "Towards Reading Hidden Emotions: A Comparative Study of Spontaneous Micro-expression Spotting and Recognition Methods," IEEE Transactions on Affective Computing, 2017, DOI 10.1109/TAFFC.2017.2667642.

[7] Y. J. Liu, J. K. Zhang, W. J. Yan, et al, "A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition," IEEE Transactions on Affective Computing, 2016, 7(4):299-310.

[8] F. Xu, J. Zhang, J. Wang, "Micro-expression Identification and Categorization using a Facial Dynamics Map," IEEE Transactions on Affective Computing, 2017, 8(2): 254-267.

[9] D. Patel, X. Hong, and G. Zhao, "Selective Deep Features for Micro-Expression Recognition," 23rd International Conference on Pattern Recognition, 2016:2258-2263.

[10] M. Peng, C. Wang, T. Chen, et al, "Dual Temporal Scale Convolutional Neural Network for Micro-Expression Recognition," Frontiers in Psychology, 2017, 8:1745.

[11] Y. Lécun , L. Bottou, Y. Bengio, et al, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, 1998, 86(11):2278-2324.

[12] M. Simon,E. Rodner,J.Denzler, "ImageNet pre-trained models with batch normalization," arXiv:1612.01452.

[13] O. Russakovsky, J. Deng, H. Su, et al, "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, 2015, 115(3):211-252.

[14] W.J. Yan, X. Li, S.J. Wang, et al, "CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation," Plos One, 2014, 9(1):e86041.

[15] A. Davison, C. Lansley, N. Costen, K. Tan; M.H. Yap, "SAMM: A Spontaneous Micro-Facial Movement Dataset," IEEE Transactions on Affective Computing , 2016, doi: 10.1109/TAFFC.2016.2573832.

[16] C. Szegedy, W. Liu, Y. Jia, et al, "Going deeper with convolutions," arXiv:1409.4842.

[17] T. Kamishima, "Transfer Learning," Journal of Japanese Society for Artificial Intelligence, 2010, 25:572-580.

[18] K. He, X. Zhang, S. Ren, et al, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[19] P. Lucey, J. F. Cohn, T. Kanade, et al, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," Computer Vision and Pattern Recognition Workshops IEEE, 2010:94-101.

[20] G. Zhao, X. Huang, M. Taini, et al, "Facial expression recognition from near-infrared videos," Image & Vision Computing, 2011, 29(9):607-619.

[21] M. Lyons, S. Akamatsu ,M. Kamachi, et al, "Coding Facial Expressions with Gabor Wavelets," Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, 1998: 200-205.

[22] N. Aifanti, C. Papachristou, A. Delopoulos, "The MUG Facial Expression Database," Proc 11th Int. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Desenzano, Italy, April 12-14 2010.

[23] T.F. Cootes, G.J. Edwards, C.J. Taylor, "Active appearance models," European Conference on Computer Vision. Springer-Verlag, 1998:484-498.

[24] T. Sherwood, E. Ahmad, M.H. Yap, "Formulating efficient software solution for digital image processing system," Software Practice & Experience, 2016, 46(7):931-954.

[25] A. K. Davison, W. Merghani, and M. H. Yap, "Objective Classes for Micro-Facial Expression Recognition," arXiv:1708.07549.