Automatic facial expression recognition based on MobileNetV2 in Real-time

Lida Hu*, Qi Ge*

School of communication and information engineering Nanjing University of Posts and Telecommunications Nanjing, China

*Corresponding author e-mail: geqi@njupt.edu.cn, *1218012333@njupt.edu.cn.

Abstract. Facial expression recognition (FER) plays a vital role in human computer interaction and has become important filed of choice for researchers in computer vision and artificial intelligence over the last two decades. As we know, the background or non-face areas of image will seriously affect the accuracy of expression recognition. In the era of mobile networks, the demand for lightweight networks and real-time is growing. However, many expression recognition networks cannot meet the real-time requirements due to excessive parameters and calculations. In order to solve this problems, our proposed methodology combines a supervised transfer learning strategy and a joint supervision method with island loss, which is crucial for facial tasks. In addition, newly proposed Convolutional Neural Network (CNN) model, MobileNetv2, which has both accuracy and speed, is deployed in a real-time framework that enables fast and accurate real-time output. As a result, superior performance to other state-of-the-art methods is achieved in facial expression databases CK+, JAFFE and FER2013.

1. Introduction

Facial behavior is one of the most important cues for sensing human emotion and intention in people. Driven by recent advances in human-centered computing, an automatic system for accurate and reliable facial expression analysis has emerging applications such as interactive games, online/remote education, entertainment.

Algorithms for facial expression recognition usually involve two main steps: feature extraction and classification. In the first step, a numerical feature vector is extraction. These features have many forms of expression. Traditional local hand-crafted features such as LBP [1], Gabor filters [2] and HOG [3] are the most common widely studied to extract good facial expression features. But these methods with hand-crafted feature hard to achieve the competitive recognition results.

In recent year, compared to the hand-crafted feature in the past, the methods based on deep learning are more robust in automatically extracting features of facial expression. Deep convolutional neural network (DCNN) can extract more expressive features. Jung et al. [4] proposed a joint fine-tuning method for facial expression recognition to extract the appearance features from image sequences. Mollahosseini et al. [5] proposed a deeper CNN that includes two convolution layers and three Inception layers. Shin et al. [6] presented a baseline CNN structure and image preprocessing methodology to analyze which DCNN method can achieve the best performance on facial expression.
recognition. Although based on the powerful ability of DCNN to extract features, the recognition rate drops significantly when extract useless features from the background, thereby reducing the accuracy of facial expression recognition. And due to excessive parameters and calculations of DCNN, it cannot meet the requirements of real-time.

In our work, the new CNN model MobileNetv2 [7] is applied to accomplish this FER task, providing a basis for real-time. A two-stage fine-tuning strategy is used in the training process of CNN. In addition, a new supervision signal, island loss, is leveraged jointly with the softmax loss function in optimization. The proposed system was evaluated on the FER2013, CK+ and JAFFE datasets and achieved competitive results. In summary, the main contributions of this article are as follows:

1) We proposed a new method which uses two shared networks for facial expression recognition to solve the problem from the non-face area of image. The network of face extraction is used to generate face box, and the network of facial expression recognition is introduced to recognize expression.

2) A two-stage fine-tuning strategy is used in this work to solve the problem of insufficient data of facial expression datasets.

3) An extra island loss is added to the loss function under joint supervision with the softmax loss for enhancing the discriminatory power of the proposed system.

The rest of the paper is organized as follows. In Section 2, the main method is proposed. The experiments and results are presented in Section 3. Section 4 gives the conclusions.

METHODOLOGY

In this section, an overview of the proposed methodology is presented that mainly contains two parts, as illustrated in Fig. 1. Firstly, the input image is processed by a lightweight Face Extraction module [8] to detect the precise face box and then crop it. Data augmentation and Normalization is followed by the Face Extraction module, in which the cropped face images will be resized, rotated and normalized. Finally, the pre-process cropped face images are fed to CNN model, which is pre-trained on ImageNet and fine-tuned on the datasets of FER2013 and CK+ respectively. Next subsections provide process details of each module.

1.1. Face Extraction

The image received by the FER system is first processed by the FaceBoxes algorithm [8] to detect face box for eliminate the influences of background or non-face areas on expression recognition tasks. The FaceBoxes detector achieves real-time speed, and the latter aims at enriching receptive fields and discretizing anchors over different layers to handle faces of various scales. If the detector finds the faces, then the four coordinates of the rectangular region of interest (ROI) of the faces are returned. The four vertices are then used to crop the faces and, consequently, irrelevant backgrounds are deleted. Fig. 1 shows the detect results of The FaceBoxes detector on face detection benchmark datasets.

![Figure 1. Qualitative results on face detection benchmark database.](image)

1.2. The Network of Facial Expression Recognition

As shown in Fig. 2. The MobileNet V2 is employed as the CNN architecture in real-time systems of our work. As illustrated in Fig. 3. The core of the MobileNet V2 is that Inverted Residual Block which utilizes the short-cut architecture of ResNet [9] and the combination of a depthwise convolution and a 1 × 1 pointwise convolution. The characteristics of small size, running speed, and remarkable accuracy enable this FER task to maintain a favorable trade-off between speed and accuracy.
Figure 2. The proposed structure of this work, using collaborated supervision with island loss. Note that the deep features before the fully connected layer are used for calculating island loss, and those after the fully connected layer are collected for the softmax loss.

Figure 3. Bottleneck residual block includes 1*1 convolution layers and depthwise convolution layers. Figure 3 (a) shows the structure of bottleneck residual block when stride is equal to 2. Different with (a), Figure 3 (b) utilizes the short-cut architecture of ResNet and stride is equal to 1.

1.3. Joint Supervision

Inspired by the work of Cai [10], island loss increases the inter-class differences and the intra-class variations are further reduced as compared to using the center loss [11], The calculation of island loss is given in Equation (1) below:

$$L_{IL} = L_{c} + \lambda_{i} \sum_{k=1}^{N} \sum_{m=1}^{N} \left( \frac{c_{ik} - c_{im}}{\|c_{ik}\|_2 \|c_{im}\|_2} \right) + 1$$

(1)

Where $N$ is the set of expression labels; $c_{ik}$ and $c_{im}$ denote the $k^{th}$ and $m^{th}$ center with $L_2$ norm $\|c_{ik}\|_2$ and $\|c_{im}\|_2$, respectively; $(\cdot)$ represents the dot product. Specifically, the first term penalizes the distance between the sample and its corresponding center and the second term penalizes the similarity between expressions. $\lambda_{i}$ is used for balancing the two terms. By minimizing the island loss, the samples of the same expression will get closer to each other and those of different expressions will be pushed apart. Where the center loss [11] denoted as $L_{c}$ is defined in Equation (2):

$$L_{c} = \frac{1}{2} \sum_{i=1}^{N} \left( y_{i} - c_{y_{i}} \right)^{2}$$

(2)

Where $y_{i}$ is the class label of the $i^{th}$ sample; $x_{i}$ denotes the feature vector of the $i^{th}$ sample taken from the fully connected layer before the decision layer; $c_{y_{i}} \in \mathbb{R}^{d}$ denotes the center of all samples with
the same class label as $y_i$; and $m$ is the number of samples in the mini-batch. By minimizing the center loss, the samples of the same class will be pulled towards their corresponding centers and thus, the overall intra-class variations can be reduced. The overall loss function of CNN training is given by Equation (3):

$$L = L_s + \lambda L_{cl}$$

(3)

Where $L_s$ means a softmax loss; where a hyper parameter $\lambda$ of 0.001 is used to balance the two losses. The network is optimized using the stochastic gradient descent (SGD) with momentum to stabilize the update and greatly speed up the convergence.

2. Experiments and Evaluations
In this section, the details for the attained results of this work are provided. Nine groups of the dataset for training and one group for validation, and a batch size of 64 for a mini-batch. The initial learning rate for the first-stage and the second-stage fine-tuning are set to 0.01 and 0.045, respectively. Next subsections provide results and implementation details of various comparative experiments.

2.1. Effects of face extraction
We design the first comparative experiment to verify effects of face extraction, both no face extraction and face extraction experiment in FER2013, CK+ and JAFFE datasets to test recognized performance and calculate accuracy improvement. The difference between the two experiments is cropped face image will be used as input in the face extraction experiment, however the no face extraction experiment will randomly sample input image from datasets.

Table 1 illustrates that using face extraction can lead to an improvement in accuracy (2.981% for FER2013, 2.020% for CK+ and 1.384% for JAFFE). After using this face extraction, a recognize accuracy has achieve 69.267% in the FER2013 dataset, an accuracy of 88.889% and 89.189% on CK+ and JAFFE datasets can be obtained.

| Dataset  | No Face Extraction | Face Extraction | Accuracy Improvement |
|----------|--------------------|-----------------|----------------------|
| FER2013  | 66.286%            | 69.267%         | 2.981%               |
| CK+      | 86.869%            | 88.889%         | 2.020%               |
| JAFFE    | 87.805%            | 89.189%         | 1.384%               |

2.2. Effects of Joint Supervision
To further enhance the discriminatory power of the proposed framework, island loss is employed as one part of the supervision signal. This experiment is conducted to show the superiority of island loss for improving results. Comparisons are carried out regarding both the FER2013, JAFFE and CK+ datasets, and $\lambda$ and $\alpha$ for island loss are fixed to 0.001. We compare two scenarios of adopting the joint supervision and the one only using the softmax loss for supervision. The exact accuracy of the three datasets achieved in both cases and the accuracy improvement are reported in Table 2, which illustrates that using island loss as an extra supervision signal can lead to an improvement in accuracy (3.733% for FER2013, 5.011% for CK+ and 3.571% for JAFFE).
2.3. Real-time Experiment
We also design an experiment to verify the ability of the system in real time. The faces are preprocessed following the preprocessing module that without data augmentation. The data after preprocessing are then fed into the selected best-trained model to perform the classification. The computation time for classifying one single frame is evaluated and results of comparison with the literature are shown in Table 3. Table 3 indicates that our proposed framework can perform classification (with a run-time of only approximately 3.87ms/frame on average).

Table 3. The Run-time Cost Comparison against the State-of-art Methods for Real-time Facial Expression Recognition

| Method                                      | Classification Time (ms/frame) | System Arrangement       |
|---------------------------------------------|--------------------------------|--------------------------|
| IVA + HOG + Adaboost & SVM [12]             | 66.7                           | 2.4GHz CPU with no GPU   |
| LBP + SVM/Adaboost [13]                     | 227                            | Intel i3 2.2GHz CPU      |
| Boosted Deep Belief Networks (BDBN) [14]    | 210                            | 6-core 2.4GHz PC         |
| 2D-LDA + SVM [15]                           | 35.7                           | Pentium IV with 2.80GHz   |
| 68 Facial Landmarks + SVM [16]              | 83.3                           | 2.6GHz Intel Core i5 CPU |
| **Ours Proposed Method**                    | **3.87**                       | **GTX 1080 Ryzen 5 2600**|

2.4. Comparison with other literature in accuracy
As shown in Table 4, our proposed method achieves over 97.98% recognition rate, outperforming the compared state-of-the-art methods. The confusion matrix on the FER-2013 dataset is shown in Fig. 4, it provide a better understanding of FER's limitations. As expected, confusion frequently occurs between "anger", "fear", and "sadness" because they create similar motions. The confusion matrices also show that "disgust" is easily misclassified as "anger".

Table 4. Average accuracy on the CK+ database for seven expressions classification

| Method                   | Accuracy   |
|--------------------------|------------|
| LBP-TOP [17]             | 88.99%     |
| HOG 3D [3]               | 91.44%     |
| IACNN [18]               | 95.37%     |
| DTAGN [5]                | 97.25%     |
| CNN (baseline)           | 89.50%     |
| **Ours Proposed Method** | **97.98%** |
3. Conclusion
In this article, a CNN-based system of estimating basic facial expressions that utilizes a transfer learning strategy, face extraction and joint supervision. Relative to previous methods, the proposed framework not only can obtain the state-of-art accuracy on JAFFE and CK+ datasets but it also performs the classification much faster than conventional classifiers as a result of the characteristics of MobileNetv2. Although some reported studies outperform our accuracy, those methods either do not provide real-time implementation or incur a much higher run-time cost than our approach.

Acknowledgments
This work was supported in part by the Natural Science Foundation Science Foundation of China under Grant 61972213.

References
[1] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
[2] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467-476, 2002.
[3] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "Disfa: A spontaneous
facial action intensity database," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 151-160, 2013.

[4] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2983-2991.

[5] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*, 2016: IEEE, pp. 1-10.

[6] M. Shin, M. Kim, and D.-S. Kwon, "Baseline CNN structure analysis for facial expression recognition," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016: IEEE, pp. 724-729.

[7] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510-4520.

[8] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "Faceboxes: A CPU real-time face detector with high accuracy," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, 2017: IEEE, pp. 1-9.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[10] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018: IEEE, pp. 302-309.

[11] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*, 2016: Springer, pp. 499-515.

[12] R. Islam, K. Ahuja, S. Karmakar, and F. Barbhuiya, "SenTion: A framework for sensing facial expressions," *arXiv preprint arXiv:1608.04489*, 2016.

[13] S. Zhang, X. Zhao, and B. Lei, "Facial expression recognition based on local binary patterns and local fisher discriminant analysis," *WSEAS transactions on signal processing*, vol. 8, no. 1, pp. 21-31, 2012.

[14] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805-1812.

[15] F. Y. Shih, C.-F. Chuang, and P. S. Wang, "Performance comparisons of facial expression recognition in JAFFE database," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 03, pp. 445-459, 2008.

[16] P. Kumar, S. Happy, and A. Routray, "A real-time robust facial expression recognition system using HOG features," in *2016 International Conference on Computing, Analytics and Security Trends (CAST)*, 2016: IEEE, pp. 289-293.

[17] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 915-928, 2007.

[18] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017: IEEE, pp. 558-565.