Journal of Computers Vol. 31 No. 4, 2020, pp. 54-64
doi:10.3966/199115992020083104005

Recognizing Facial Emotions Using Pooling-first Bilinear CNN from a Single Image

Hao Gong, Tianyou Pei, Qiaoyu Ma, Dongmei Jiang, Teng Yu*

School of Electronic and Information Engineering, Qingdao University, Qingdao 266000, China
gong_h@outlook.com, peity126@163.com, maqiaoyu_1@163.com, Kathy.jiang@163.com,
yutenghit@foxmail.com

Received 21 December 2018; Revised 5 May 2019; Accepted 31 May 2019

Abstract. Facial expression recognition from a single image is challenging due to the subtle differences between the various expression types. Considering that facial expressions are not only expressed by the appearance of individual facial organs, but also highly correlated with the local features of these organs, we propose a new network architecture, namely the Pooling First Bilinear Convolutional Neural Network (PF-BCNN). The network uses two simplified CNNs to extract facial appearance features, then integrates them using a modified bilinear approach, which encodes the interaction of local features. In addition, we use a multi-task learning model to simultaneously train and predict two typical outputs for emotion recognition, namely action units (AU) and emotion types (ET). Experimental results show that the method achieves the most advanced performance, especially the recognition of emotion types.

Keywords: convolutional neural network, emotion recognition, facial action unit

1 Introduction

With the development of artificial intelligence technology, researchers hope that computers have a better understanding of human beings. In addition to facial recognition, speech recognition and other fields, emotion recognition becomes an important research topic. By understanding human emotions, computers can respond to human performance and communicate better with humans, which will enhance the intelligence of computers and robots.

In general, there are two typical tasks for the emotion recognition. One is based on Ekman's emotion theory 0, which divides emotions into six basic emotion types (ET) and treats the emotion recognition task as a classification task. The other one is the facial motion coding systems (FACS) [2]. It divides the facial expressions into facial action units (AU) and recognizes emotions through the relationship between AU and expressions. The proposed model in this paper is dedicated to recognizing action units and emotion types simultaneously in a single multi-task and multi-label neural network.

In recent years, CNN-based methods have achieved impressive performance on various detection and recognition applications, such as general object detection and speech recognition. Therefore, various CNN-based models have been developed to recognize the facial emotions in images or videos [11, 13-14], in which different network structures are proposed and the number of layers in the network has been increasing. Compared to the traditional learning methods, improved results have been achieved with CNN models, whereas the feature interactions among the facial parts (sense organs) are not fully exploited yet, which we think is a crucial clue for emotion recognition.

Bilinear CNN (BCNN) model has proved very effective on fine-grained visual recognition problem [21], especially on the recognition of different car models. However, the outer product two CNNs features need numerous parameters, which may lead under-fitting for emotion recognition due to the lack of training data, and it is also not suitable for real-time processing. In this paper, we propose an end-to-end and multi-task network named Pooling-first Bilinear CNN (PF-BCNN). Two light-weight CNNs are

[*] Corresponding Author

constructed to extract the features of the facial parts, followed by a bilinear structure to model the interactions among the different features. Before the bilinear calculation, a pooling operation is applied to take out the most representative features. Through the pooling first operation, the complexity of the model can be reduced. In PF-BCNN, bilinear calculation can be regarded as a feature fusion method. In the end, we use a soft-max function to process the output for emotion recognition and a sigmoid function for AU recognition.

In summary, three major contributions are made in this paper:

(1) A novel and effective model called PF-BCNN is proposed, which is designed to encode the interactions among the most useful local features (facial organs and muscles).

(2) The two outputs AUs and ETs, are trained and predicted simultaneously in a multi-task and multi-label learning network.

(3) Because of the scarcity of labelled data for training a large net, the simplified ResNet and VGGNet are applied in the CNN layers, which could be directly transferred to fine-tuning using the facial emotion datasets.

Fig. 1 shows the structure of the proposed model. Section 2 introduces the related works of facial emotion recognition. Details about training the model are discussed in section 3. Section 4 shows the experimental results and analysis. Conclusions and future work are discussed at Section 5.

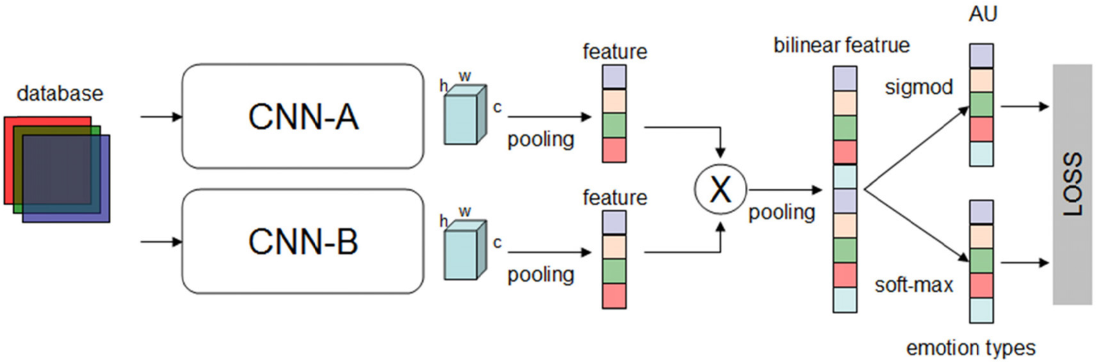


Fig. 1. Structure of the proposed model

2 Related Work

Generally, there are two tasks in the facial emotion recognition. Ekman proposed that emotions can be divided into 6 basic types 0. However, there are some flaws in dividing emotions into several types. For example, it cannot describe other emotions besides these types of emotions, nor can it describe the relationship of emotion types. FACS [2] solves this problem to some extent. FACS describes facial emotion with facial action units (AU). The relationship between AU and emotion types (ET) is given in Table 1. The methods used for emotion recognition can be divided into two categories, one is the traditional methods with hand-designed features, and the other is deep learning based approaches.

Table 1. [6] Emotion description in terms of facial action units

Emotion types	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present

In early works of emotion recognition, researchers often performed facial emotion recognition by extracting image features. Traditional hand-designed features, such as Gabor [7], Histogram of Oriented Gradients (HOG) [8], haar-like features [3], local binary patterns (LBP) [9-10], are extracted to train classifiers like Support Vector Machine (SVM), Random Forest and so on, for emotion recognition.

There are mainly two drawbacks of these methods. One is that they can extract low-level or mid-level features, but they are not able to extract high-level features with more semantic meanings. The other one is that the feature extraction step and classification step are separately designed, which make them not end-to-end trainable.

Recently, researchers have frequently used deep neural networks to recognize emotions. In particular, Convolutional Neural Networks (CNNs) have achieved good results in many detection and recognition related fields, such as object detection and speech recognition. Many deep neural CNN networks have been developed, such as the AlexNet [27], VGGNet [4], ResNet [5] and GoogLeNet [28]. Gudi et al. [11] proposed a deep convolutional neural network to participate in the FEAR2015 challenge. They pre-process the input data to detect faces and handle pose variations. Zhao et al. designed a CNN model to detect AU in [12], and they use region learning (RL) and multi-label learning (ML) to improve the model performance. Han et al. [13] proposed the Incremental Boosting Convolutional Neural Network (IB-CNN) for AU recognition and give the comparison of IB-CNN and ordinary CNN. By combining the speech information, Pons et al. [14] use multi-domain learning for emotion recognition, in which a selective sigmoid cross entropy loss function is applied. Multimodal features and temporal models are combined in [24-26]. These models outperform the traditional methods on the average accuracy, but are not applied to real-time recognition due to the complicated network structure.

To achieve real-time emotion recognition, Zhong et al. proposed a model based on the face patches in [15]. The model learns common patches across expressions and learns specific patches for individual expression. Wang et al. [19] used a Bayesian network for emotion recognition and added head motion features to improve the performance of the model. Arriaga et al. [17] designed a mini-Xception network for recognition, but the performance is not comparable to those deeper networks.

The above CNN based methods extract high-level semantic features of the individual facial parts, but the interactions among these parts have not been modeled yet, which we think is a useful way to improve the emotion recognition performance. The original bilinear models are designed for the fine-grained visual recognition [21-23], we consider it can also be used in emotion recognition, but it needs to be modified and specially-designed for real time emotion recognition.

3 Proposed Method

The overall proposed framework is shown in Fig. 1. The input of this network is an NxN face image, followed by two parallel CNNs to extract the local features. The bilinear model pools two feature vectors and combines them using an outer product to model the local feature interactions. In the end of this network, a sigmoid function and a soft-max function are applied separately to predict the AUs and ETs. Fig. 1 shows the architecture of our proposed model.

The bilinear model combines different features for recognition, which is defined as follows:

$$B=(f_A,f_B,P,CR)\,. \tag{1}$$

Here f_A and f_B are feature functions, P is a pooling function, and CR is the classification and regression functions.

3.1 Pooling-first Bilinear CNN (PF-BCNN)

The two CNN feature outputs are combined at each location using the matrix outer product in the original bilinear model. For image I and location L, the bilinear feature is as follows:

$$bilinear(l,I,f_A,f_B)=f_A(l,I)^T f_B(l,I)\,. \tag{2}$$

The pooling function P aggregates the bilinear combination of features across all locations in the image to obtain a global image representation Φ (I).

$$\Phi(I)=\sum_{l\in L} bilinear(l,I,f_A,f_B)=\sum_{l\in L} f_A(l,I)^T f_B(l,I)\ . \tag{3}$$

In case of the emotion recognition, the most representative features are from the local organ areas. Unlike the original bilinear model calculating the bilinear features of each location of the image, it is more meaningful to combine the pooled features. Base on the above analysis, in our architecture as shown in Fig. 1, two CNNs are applied parallelly to extract the image features, followed by a pooling operation. This pooling operation can guarantee that the most useful features are extracted for combination as well as to reduce the redundant features. We call it Pooling-first Bilinear CNN (PF-BCNN) which is expressed as follows

$$\Phi(I)=[\sum_{l\in L} f_A(l,I)]^T[\sum_{l\in L} f_B(l,I)] \ . \tag{4}$$

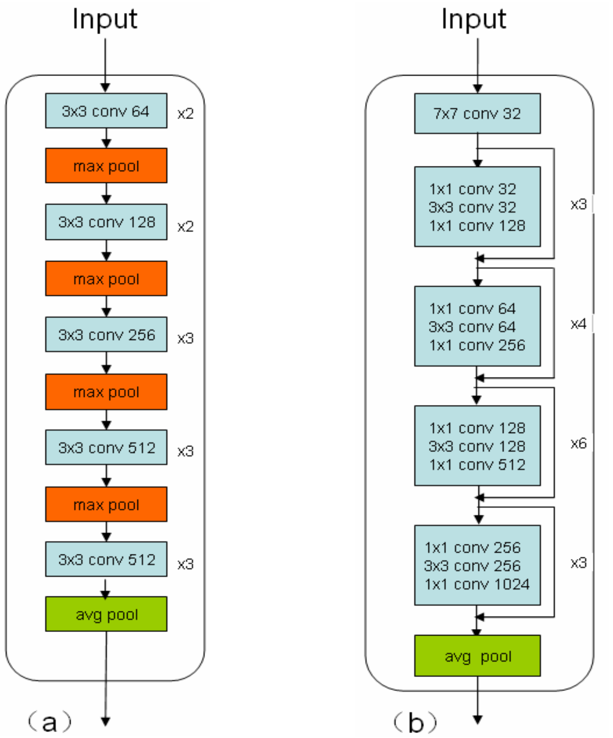
In the original bilinear model, if the shape of the CNN output array is [w, h, c], the shape of the bilinear feature array is [w, h, c^2] before pooling. It requires $w*h*c^2$ multiplication operations. In our model, it only requires c^2 multiplication operations. Since w and h are always large numbers, it will not only captures the most dominant features, but also reduce the complexity.

3.2 Simplified CNN and Fine-tuning

For the feature extraction layers, two typical CNN architectures, VGG and ResNet are considered. They have been proved useful to extract various object features for recognition. By using their pre-trained weights with ImageNet, we can directly fine-tune our model with the scarce facial emotion data.

In case of the facial emotion recognition, it needs fewer features than the ImageNet classification, so we propose two simplified versions, simplified VGG (S-VVG) and simplified Resnet (S-ResNet), as the feature extraction functions. B-VGG is a dual parallel network, each of which is simplified from the VGG16. Considering that VGG16 contains a large number of parameters in the fully-connected (fc) layer, we eliminate all the fc layers of VGG16. For comparison, we also test VGG16 with 1, 2 and 3 fc layers. The effect of deleting fc layer is given in later section.

The structure of a single net in S-ResNet is based on Res50, which can effectively avoid the disappearance of gradients using the residual structure. Res50 used on the ImageNet dataset contains many feature layers (64, 128, 256, and 512 feature layers after dimension reduction with 1x1 convolution kernel) to identify various features. We reduce the number of feature layer (32, 64, 128, and 256 feature layers after dimension reduction with 1x1 convolution kernel) in the S-ResNet. For example, for a convolution kernel of the original ResNet with a shape of [m, 3, 3, n], its shape in S-ResNet is [m/2, 3, 3, n/2]. The parameters are reduced by 75%. Fig. 2 shows the structure of S-VVG and S-ResNet.



(a) Structure of S-VGG (b) Structure of S-ResNet

Fig. 2. The structure of S-VVG and S-ResNet

We initialize the network using the weights pre-trained on ImageNet. For S-ResNet, we select the appropriate weights based on the number of feature layers in the original weight. Since the bilinear model requires two CNN networks, we tested the performance of three methods: S-VGG+S-VGG(B-V), B-ResNet+B-ResNet (B-R), and B-VGG+B-ResNet (B-R).

3.3 Multi-task and Multi-label Learning

Multitasking is a way to put multiple tasks together for learning, which can improve the generalization ability of the model. Since action units (AU) and emotion types (ET) are both representations of human emotions, we try to learn both of them at the same time. For comparison, we also use separate networks for recognition.

Action Units recognition is a multi-label task, so we use the sigmoid function to process the output data. The sigmoid function is as follows:

$$F(x) = \frac{1}{1+e^{-x}} . \qquad \textbf{(5)}$$

Where x is the output bilinear features.

When performing the AU recognition, model judges an AU appears if the corresponding output value is greater than the threshold. The threshold is selected using the verification set.

Emotion type recognition is a classification task, so we use the soft-max function to process the output to get the probability of each type. The soft-max formula is as follows:

$$\delta(Z_j) = \frac{e^{Z_j}}{\sum_{k=1}^{K} Z_k} , j = 1, 2, \ldots, K. \qquad \textbf{(6)}$$

Where Z is the output bilinear features and K is the length of it.

The model uses cross entropy as the objective function. Cross entropy is calculated as follows:

$$L = -\sum_{k=1}^{n} \sum_{i=1}^{C} t_{ki} \log(y_{ki}) . \qquad \textbf{(7)}$$

Where t_{ki} is the probability that the sample k belongs to class i and y_{ki} is the probability that the model recognizes the sample k as class i.

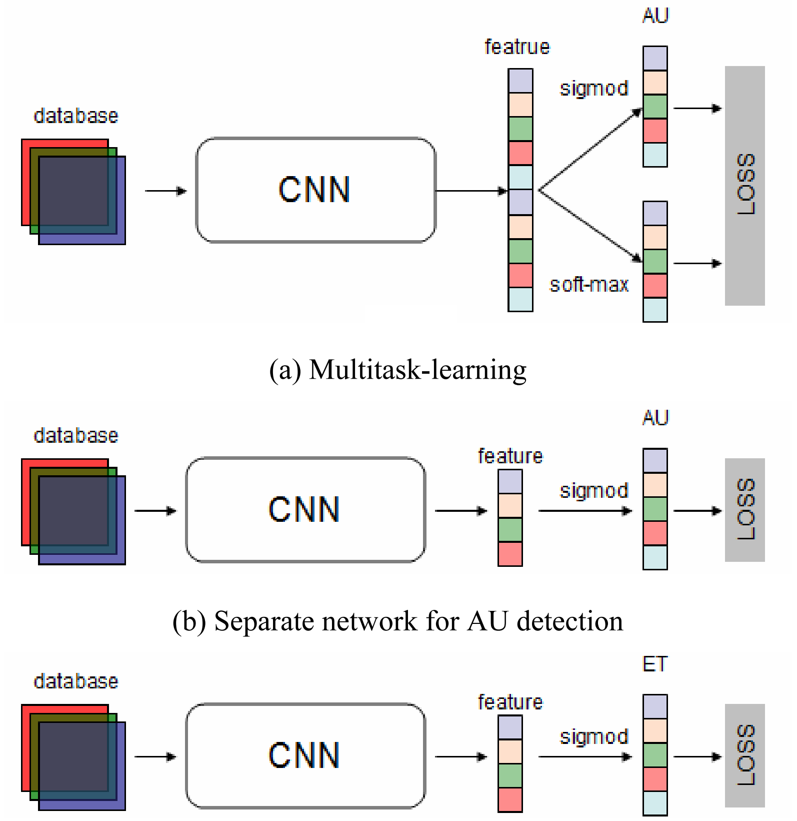
Fig. 3 shows the structures of the multi-task model and the separate models.

4 Experiments

The Extended Cohn-Kanade database (CK+) is used for training and testing, which is developed on the CK database and contains 123 subjects and 593 sequences. All sequences are from the natural facial state to the peak expression. The database uses FACS code, and each sequence corresponds to a FACS file. The FACS file contains the type and intensity of the AU in the peak frame. In the 593 sequences, 327 sequences contain emotion types. Emotion types use integer tags, which are divided into 7 types of emotions (0=neutral, 1=anger, 2=contempt, 3=disgust, 4=fear, 5=happy, 6=sadness, 7=surprise). Table 1 shows the distribution of emotions in the database. For AU recognition, we select 493 sequences as the training set, 50 sequences as the validation set, and 50 sequences as the test set. For emotion type recognition and multi-task learning, we select 277 sequences as the training set, 20 sequences as validation set, and 30 sequences as the test set.

4.1 Emotion type and AU Recognition

In our experiments, 5-fold cross validation technique is applied to evaluate the results. We use the F1 score and accuracy to evaluate the results of the AU recognition, and the ET recognition results are evaluated using the average accuracy. F1 score is a measure of our experimental accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of



(a) Multitask-learning

(b) Separate network for AU detection

(c) Separate networks for ET recognition.

Fig. 3. The structures of the multi-task model and the separate models

correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

For the S-VGG net configuration, we resize the peak frame to 224×224 as the input data. The mini-batch gradient descent method is utilized to train the model with a batch size of 4. We use the validation set to get the optimal number of steps and the threshold for judging whether an AU exists. The original VGG16 has three fully connected layers. The fully connected layer contains a large number of parameters, the number of fc layers are reduced in our experiment. We tested VGG16 with 1, 2, and 3 fc layers and observed their performance on emotion recognition compared to our S-VGG (0 fc layer). To be consistent with S-VGG, we also send the same size frame to the S-ResNet. The training steps are selected based on the validation set, too. Other parameters are generally the same as S-VGG. Fig. 4 shows the convergence of different bilinear models.

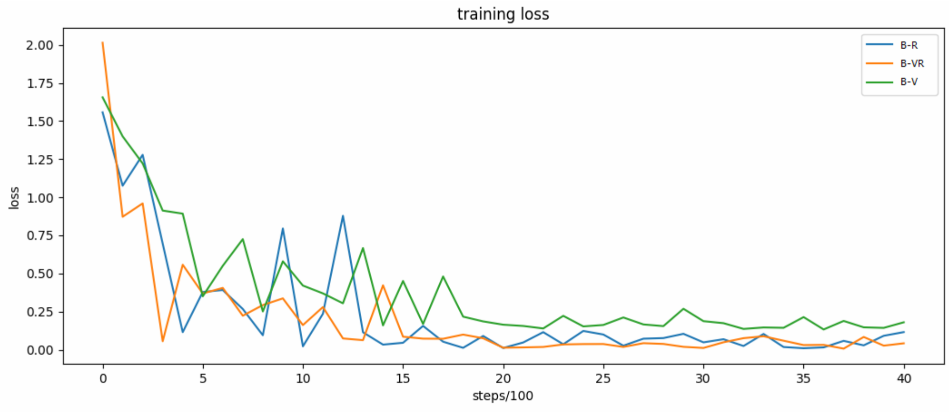


Fig. 4. Training loss of different models

Table 2 and Table 3 show the experimental comparisons of bilinear network against single CNN networks. In the experiments of separate task learning (each row), we can see that, by using the PF-BCNN model, both the B-V and B-R can improve the performance on the recognition of AU and ET. These results prove that the extracted bilinear feature is more discriminative than that of the single CNN feature on recognition of emotions. In the experiments of multi-task learning, it can be seen that both the recognition AU and ET are improved through the multi-task learning. The reason is that the recognitions of AU and ET are two close related tasks, they can benefit each other through the multi-task learning. Table 2 shows a comparison of different bilinear models. B-VR has achieved the best results on the multi-tasking learning. It means that the model of two different CNNs outperform that of two identical CNNs.

Table 2. Performance of B-V FOR emotion recognition

Task	Network	AU		Emotion type
		F1	accuracy	accuracy
AU	VGG16-1fc	0.715(0.707)	0.880(0.864)	
	VGG16-2fc	0.716(0.745)	0.874(0.856)	
	VGG16-3fc	0.717(0.723)	0.878(0.831)	
	B-V	0.723(0.732)	**0.892(0.881)**	
ET	VGG16-1fc			0.714(0.818)
	VGG16-2fc			0.821(0.909)
	VGG16-3fc			0.837(0.869)
	B-V			**0.862(0.896)**
AU & ET	VGG16-1fc	0.724(0.557)	0.758(0.762)	0.756(0.695)
	VGG16-2fc	0.728(0.685)	0.889(0.852)	0.828(0.911)
	VGG16-3fc	0.723(0.704)	0.896(0.873)	0.831(0.934)
	B-V	0.742(0.744)	**0.904(0.890)**	0.878(0.954)

Note. In parenthesis are the performance obtained on the validation set.

Table 3. Performance of B-R for emotion recognition

Task	Network	AU		Emotion type
		F1	AU accuracy	accuracy
AU	Res50	0.653(0.653)	0.878(0.829)	
	B-R	0.763(0.717)	**0.913(0.870)**	
ET	Res50			0.793(0.826)
	B-R			**0.857(0.818)**
AU & ET	Res50	0.723(0.783)	0.900(0.908)	0.793(0.869)
	B-R	0.747(0.763)	**0.921(0.905)**	**0.928(0.909)**

Note. In parenthesis are the performance obtained on the validation set.

Table 4. Performance of different Bilinear NETWORKS FOR emotion recognition using multi-task learning

Task	Network	AU		Emotion type accuracy
		F1	AU accuracy	
AU & ET	B-V	0.742(0.744)	0.904(0.890)	0.878(0.954)
	B-R	0.747(0.763)	0.921(0.905)	0.928(0.909)
	B-VR	0.763(0.739)	**0.953(0.939)**	**0.928(1.000)**

Note. In parenthesis are the performance obtained on the validation set.

4.2 Real-time Emotion Recognition System

The proposed model runs fast enough for a real-time model to recognize emotions in video sequences using a GTX1070 GPU. Fig. 5 shows the proposed real-time video emotion recognition system. We extract the face area using the Haar-like features and Adaboost method [29] in the video frame and send it to the network. Then we recognize the facial emotions and mark it on the video frame. Table 5 shows the running time of different models. It shows that the recognition calculation is fast but face detection takes a lot of time. The reason is that we only use a CPU for face detection in video sequences.

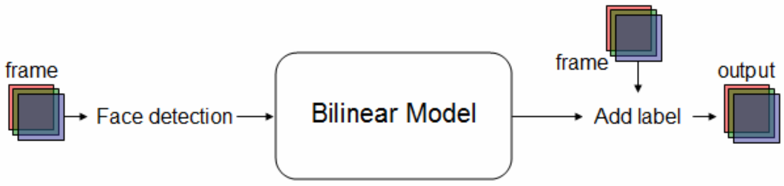


Fig. 5. Real-time recognition system

Table 5. Running time of different models. The time of face detection is 3 to 5 times that of emotion recognition

model	Running time(ms/frame)
VGG16	12.71
Res50	7.58
B-V	18.44
B-R	14.21
B-VR	16.35
Face detection	44.12

We tested the proposed model on the MMI dataset. Fig. 6 shows the sample recognition result on MMI database [18].

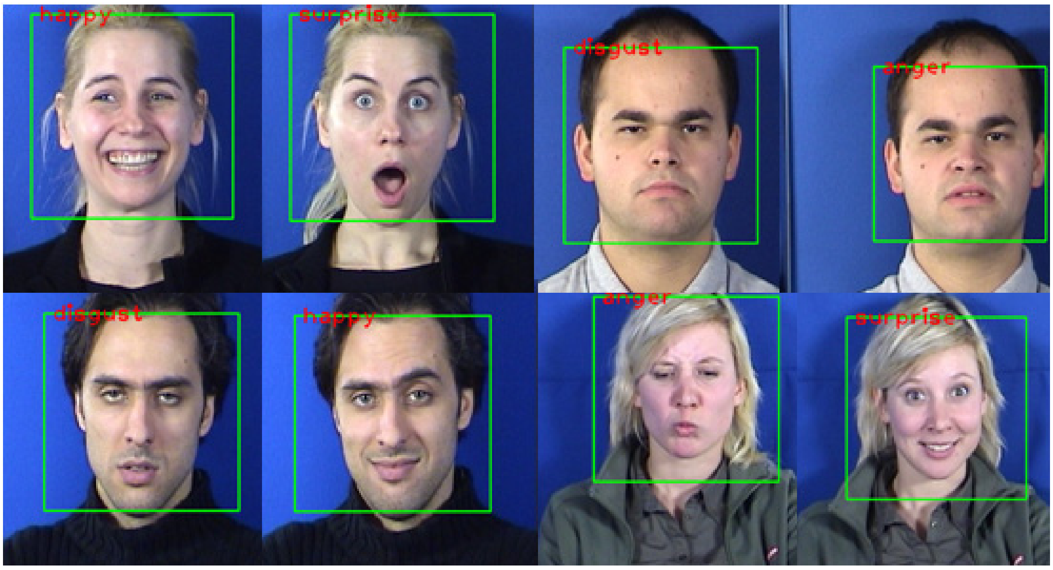


Fig. 6. Sample recognition result on MMI database

4.3 Comparison with Other Methods

In view of the good performance of the B-VR model in the test set, we chose the B-VR model for comparison. Table 6 shows the comparison of our proposed model with other emotion recognition models. We can find that our model has achieved the best performance compared to the state-of-the-art methods, especially on the recognition of emotion types. The accuracy of ET recognition is generally lower than the accuracy of AU recognition. This may be because ET is a vague description, and an emotion will be between two types. AU recognition is more accurate because AU is a description of a certain facial action. These facial actions may have distinctive features that facilitate the work of the classifier.

Table 6. Comparison of the proposed model with other methods on CK+ database

method	accuracy- AU	accuracy-ET
IB-CNN [13]	0.951	
LBP [20]	0.949	
Zhong et al. [15]		0.882
Song et al. [16]		0.897
Proposed model (B-VR)	0.953	0.928

To further verify the performance of the proposed method, we validated it on the FER-2013 dataset. We still use the B-VR structure and retrained the model on the FER2013 dataset. Because of the different resolutions, we fine-tuned the model. The results show that the proposed method achieves the highest accuracy (Table 7).

Table 7. Comparison of the proposed model with other methods on fer-2013 database

method	accuracy- AU
Jeon et al. [30]	0.707
MNF [31]	0.703
Proposed model (B-VR)	0.709

5 Conclusion

This paper presents an end-to-end real-time emotion recognition architecture using the proposed PF-BCNN model. PF-BCNN uses two independent convolutional neural networks to extract facial features. These features are then used for bilinear calculations, which describe the relationship between features. In order to reduce the amount of computation, we perform pooling operations before bilinear calculations. We found that using different network structures (VGG and ResNet) can achieve better results. By modeling the interactions among facial local features, it shows an improved performance for the recognitions of actions units and emotion types. The specially designed simplified CNN models can not only extract discriminative features, but also be applicable for real time recognition. In addition, the use of multi-task learning can improve the network performance compared to separate learnings. Since the face detection step takes up a lot of time, we will focus on speeding up the face detection by including it in a single deep neural network, as well as improving accuracy in the future work.

References

[1] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion, Journal of Personality and Social Psychology 17(2)(1971) 124.

[2] P. Ekman, W.V. Friesen, Manual for the facial Action Coding System, Consulting Psychologists Press, Palo Alto, CA, 1978.

[3] J. Whitehill, C.W. Omlin, Haar features for facs au recognition, in: Proc. 7th International Conference on Automatic Face and Gesture Recognition, 2006.

[4] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. <https://arxiv.org/abs/1409.1556>, 2014.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambada.r, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: Proc. 2010 IEEE Computer Society Conference on

Computer Vision and Pattern Recognition Workshops (CVPRW), 2010.

[7] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005.

[8] Z. Li, J. Imai, M. Kaneko, Facial-component-based bag of words and phog descriptor for facial expression recognition, in: Proc. IEEE International Conference on Systems, Man and Cybernetics, 2009.

[9] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on pattern analysis and machine intelligence 24(7)(2002) 971-987.

[10] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE transactions on pattern analysis and machine intelligence 29(6)(2007) 915-928.

[11] A. Gudi, H.E. Tasli, T. M. Den Uyl, A. Maroulis, Deep learning based facs action unit occurrence and intensity estimation, in: Proc. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015.

[12] K. Zhao, W. Chu, H. Zhang, Deep region and multi-label learning for facial action unit detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[13] S. Han, Z. Meng, A.S. Khan, Y. Tong, Incremental boosting convolutional neural network for facial action unit recognition, in: Proc. Advances in Neural Information Processing Systems, 2016.

[14] G. Pons, D. Masip, Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. <https://arxiv.org/abs/1802.06664>, 2018.

[15] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D.N. Metaxas, Learning active facial patches for expression analysis, in: Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[16] M. Song, D. Tao, Z. Liu, X. Li, M. Zhou, Image ratio features for facial expression recognition application, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40(3)(2010) 779-788.

[17] O. Arriaga, M. Valdenegro-Toro, P. Plöger, Real-time convolutional neural networks for emotion and gender classification. <https://arxiv.org/abs/1710.07557>, 2017.

[18] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proc. IEEE International Conference on Multimedia and Expo, 2005.

[19] S. Wang, Z. Liu, Y. Zhu, M. He, X. Chen, Q. Ji, Implicit video emotion tagging from audiences' facial expression, Multimedia Tools and Applications 74(13)(2015) 4679-4706.

[20] S. Han, Z. Meng, P. Liu, Y. Tong, Facial grid transformation: A novel face registration approach for improving facial action unit recognition, in: Proc. 2014 IEEE International Conference on Image Processing (ICIP), 2014.

[21] T.Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: Proc. IEEE International Conference on Computer Vision, 2015.

[22] Y. Gao, O. Beijbom, N. Zhang, T. Darrell, Compact bilinear pooling, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[23] T.Y. Lin, S. Maji, Improved bilinear pooling with cnns. <https://arxiv.org/abs/1707.06772>, 2017.

[24] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, Q. Wei, Combining multimodal features within a fusion network for emotion recognition in the wild, in: Proc. 2015 ACM on International Conference on Multimodal Interaction, 2015.

[25] S. Wang, W. Wang, J. Zhao, S. Chen, Q. Jin, S. Zhang, Y. Qin, Emotion recognition with multimodal features and temporal models, in: Proc. 19th ACM International Conference on Multimodal Interaction, 2017.

[26] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE Journal of Selected Topics in Signal Processing 11(8)(2017) 1301-1309.

[27] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proc. Advances in Neural Information Processing Systems, 2012.

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, Going deeper with convolutions, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[29] P. Viola, M. J. Jones, Robust real-time face detection, International Journal of Computer Vision 57(2)(2004) 137-154.

[30] J. Jeon, J.C. Park, Y. Jo, C. Nam, K.H. Bae, Y. Hwang, D.S. Kim, A real-time facial expression recognizer using deep neural network, in: Proc. the 10th International Conference on Ubiquitous Information Management and Communication. ACM, 2016.

[31] C. Li, M. Ning, D. Yalin, Multi-network fusion based on CNN for facial expression recognition, in: Proc. 2018 International Conference on Computer Science, Electronics and Communication Engineering (CSECE 2018), 2018.