Efficient Emotion Recognition based on Hybrid Emotion Recognition Neural Network

Yang-Yen Ou[1], Bo-Hao Su[1],Shih-Pang Tseng[2], Liu-Yi-Cheng Hsu[1], Jhing-Fa Wang[1,3], Ta-Wen Kuan[4]

[1] Department of Electrical Engineering, National Cheng Kung University.
[2] Department of Computer Science and Entertainment Technology, Tajen University.
[3]Department of Digital Multimedia Design, Ta Jen University, Pingtung, Taiwan.
[4]General Research Centers of R&D, National PingTung University of Science and Technology, Pingtung, Taiwan.
E-mail: ouyang0916@gmail.com

***Abstract*—The practical application of computer vision on robots such as emotion, age, gender recognition can improve the interactive experience between robots and users. This paper uses a webcam to capture the image as a visual system input. Then, facial image is obtained through high-performance face detect neural network. Facial landmarks is used to correct the face. After that, we input facial image into the multi-person emotion recognition system. In order to improve the accuracy of emotion recognition, a hybrid emotion recognition is proposed based on Convolutional Neural Network. Taking facial points and facial image as input, training hybrid neural network to convergence and outputting five home common emotion, neutral, happy, surprise, sad and angry. The other hand, the Microsoft Azure API is used for age and gender recognition. Finally, the experimental result shows that the accuracy of emotion recognition is as high as 86.14%. In practical applications, the system can recognize the emotions, age and gender up to thousands of people at the same time.**

Keywords— Emotion Recognition, Dilated Convolution Neural Network, Computer Vision.

I. INTRODUCTION

The Human-Machine Interface design has become an important research area to let a machine to understand human behaviors including different emotional states. Facial expressions can provide useful cues that could further improve the experience of human-machine interaction. In view of the increasing development and application of robots in various fields. As well as the rise of deep learning, big data, and the Internet of Things, emotion recognition are gaining more and more attention. The motivation of proposed research is focus on improving the image understanding about human emotion. Compared with other papers which only research on facial image[2-5] or only depends on facial landmark[5]. The emotion recognition system is proposed based on deep convolution neural network. The facial landmark points and facial local features are combined as features of deep neural network.

In the development of robots, strong computer vison is an important issue for robot system. The feedback, interaction and dialogue will be more abundant, when the robot system gets more information from the vison. In order to enhance the ability of robot vision, the age and gender recognition are integrated into our proposed system. The Microsoft Azure API is used for implement of age and gender recognition. The experiment result will focus on the evaluation of the proposed emotion recognition system.

The organization of the paper is as follows. In Section II, we described the related work for our research. Section III shows the proposed system. Experiment result is shown in Section IV. Finally, the conclusion and future work are given in Section V.

II. RELATED WORK

Emotion recognition is usually measured by facial expression as a standard. In recent years, the researchers discovered that the daily emotion is not easy to model correctly by using Ekman's FACS. Nowadays, the state-of-the-art emotion recognition approaches could be concluded in three ways, including 1) spontaneous emotion recognition, 2) emotion intensity estimation and 3) non-basis emotion analysis. The measurement algorithms of facial expression can also be classified into combined measurement, joint measurement and individual measurement, respectively [6]. Combined measurement regarded the emotion expression as a combination of action units [7-9], but the AUs combinations were limited in this method. The joint measurement was the construction of a model trained by a set of AUs for emotion identification [6, 10-12]. Although, the joint measurement is a good application on personalized system, it might face the high variance issue when the training data contained different emotions as well as users. Individual measurement attempted to estimate the emotion state of specified face regions [13-15], but some face areas may be obscure during the capturing and may cause the defect in emotion understanding system. However, the individual measurement is still adopted in the proposed system, since the accuracy can be further improved with appropriate estimation method. Emotion understanding mainly based on facial expressions and rarely focuses on emotional behaviors. Actually, human's emotion generally mixed from facial expression as well as behaviors, which results in a challenge on emotion understanding system.

Recent approaches on facial emotion recognition, which have earned a high degree of accuracy will be discussed. There are many approaches for emotion recognition, which are based on hand-engineered features such as histograms of oriented gradients features, scale-invariant feature transform descriptors, Gabor filter, or local binary patterns (LBP). Methods for emotion recognitions classify the facial image into some basic emotions like happiness, anger, and sadness, while others try to classify AUs on The face in order to describe objective characteristic of the facial expression.

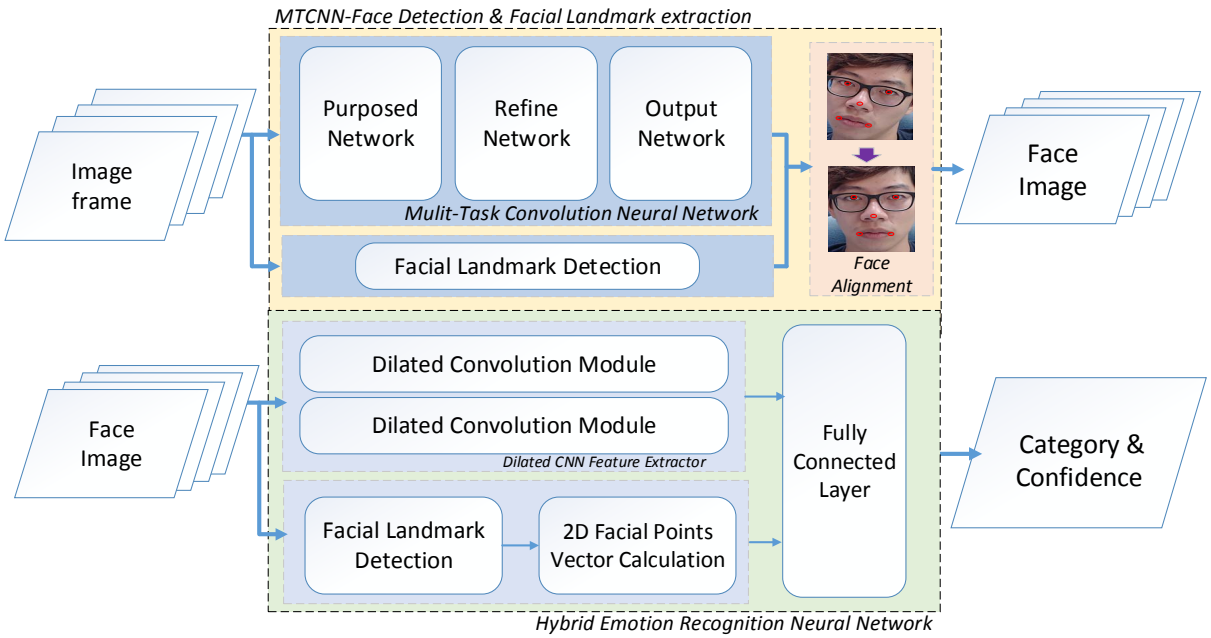

Fig 1. The emotion recognition system based on hybrid emotion recognition neural network

III. THE PROPOSED SYSTEM

A. System Overview

The face is detect by Multi-Task Convolution Neural Network, which is an efficient face detection algorithm. The hybrid emotion recognition neural network is proposed for emotion recognition. The dilated convolution neural network and facial landmark detect are contained for global and local features. A fully connected layer is used for combine the deep dilated feature extraction network and facial landmark network. The system overview is shown on Fig 1. The dilated CNN architecture and facial landmark network of Hybrid Emotion Recognition Neural Network are described below.

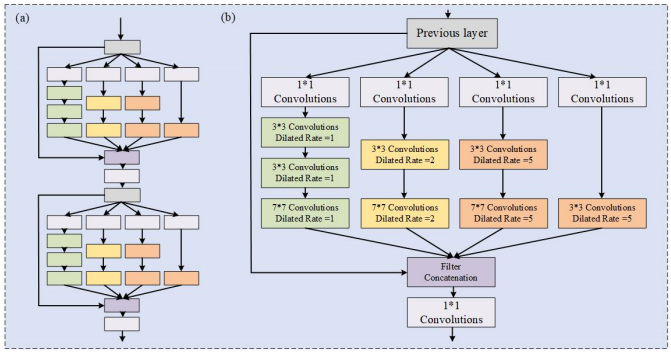

Fig 2. Dilated CNN Feature Extractor (a) the module b) the Dilated CNN architecture

B. Dilated CNN Architecture

An dilated convolution module is proposed for extraction of goal features on the face. The dilated convolution module is designed based on Inception Net and dilated convolution neural network. At the beginning, the 1*1 convolution filter is considered for reduce dimensionality on each branch. The feature map will enter the blocks of the four branches, respectively. The 3*3 convolution filter with difference dilated (1, 2, 5, 7) rate are considered for analysis of features at different scales. In the dilated convolution module, four branch of convolution are selected for facial texture analysis. The four branches are expressed as features extractor of different sizes. Finally, a 1x1 convolution filters to learn the global and local feature information of each branch.

Two architecture of dilated convolution modules are cascaded for feature extraction neural network, and global average pooling is used in the end. The dilated convolution module is shown on Fig2, and The feature extraction neural network is shown on Fig1.

C. Facial Landmark Network

The local features is select feature vectors, which are consider the distance of facial points. The dlib API is used for 68 facial landmark detection. The relative position of each point can be express, through the connection of points and points. As mentioned, we observed that in the facial expression recognition, the area of greatest change is mouth, nose and eyes. So, we divide facial vectors into four combination included eyes, noise region, mouth region and contour. The 2-D Vector and Euclidean distance, position are used for feature representation. The combination of facial regions and representation are shown on table I.

TABEL I . THE COMBINATION AND REPRESENTATION OF FACIAL REGION

Region	Facial landmark label	Representation
Mouth	49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68	2-D Vector
Nose	28, 29, 30, 31, 32, 33, 34, 35, 36	2-D Vector
Eyes	37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48	Euclidean distance and 2-D Vector
Contour	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26,27.	Position of facial landmark points.

It is not image analysis but the classification and analysis of multiple data for vectors and distances of facial landmarks. Inspired by the above application, we do not select CNN as our proposed method but use the most simple fully connected layer to build our facial landmark neural network. Facial landmark neural network, consists of one input layer, one hidden layer and one output layer, will be connected with the output of deep dilated convolution neural network and finally output five confidences. In facial landmark network, input layer has 181 nodes, one hidden layer each contain 2048 nodes and output layer outputs four emotions and their confidence values. We join the concept of drop out to avoid overfitting. The figure is as Fig 4.

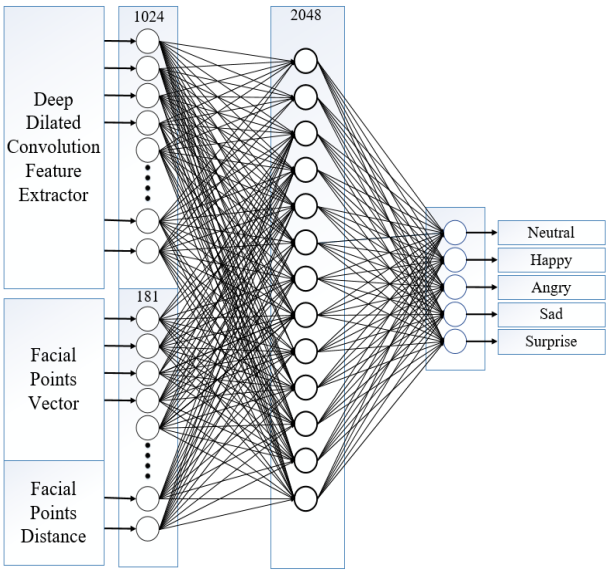

Fig 4. The Architecture of Classification model

D. *Training Phase*

The loss layer of a neural network compares the output of the network with the ground truth, i.e. processed and reference patches, respectively, for the case of image processing. It is a non-negative value, where the robustness of model increases along with the decrease of the value of loss function. Loss function is the hard core of empirical risk function as well as a significant component of structural risk function. Generally, the structural risk function of a model is consist of empirical risk term and regularization term, which can be represented as:

$$\theta^* = arg\ \min_{\theta} \frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(x_i;\theta)\big) + \lambda\Phi(\theta) \quad ,$$

where $\Phi(\theta)$ is the regularization term or penalty term, θ is the parameter od model to be learned, $f(.)$Represents the activation function and $\mathrm{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}\} \in \mathbb{R}^m$ denotes the training sample.

Multi-class Loss: In purposed emotion system, emotion recognition is a multi-class classification. We apply the most common multi-class loss function in training, Multi-class Cross Entropy. The definition of L as show in follow:

$$\mathrm{L}_{y'}(y) = -\sum_i y_i' \log(y_i),$$

where y_i' only represent the ith value in the ground-truth, y_i is the ith value of output vectors[Y_1, Y_2,…, Y_n] of softmax layer.

Regularization term: In the neural network, the problem of overfitting often occurs. We use l2 regularization to solve this issue. The Regularized network encourages small weights. In the case of small weights, some random changes in training samples do not have much impact on the neural network model. And we added a weighted square after the entire loss, through the square of the weight to reduce the impact of higher-order parameters, the formula is shown below:

$$C = C_0 + \frac{\lambda}{2n}\sum_w w^2,$$

where C_0 is the loss term, n is the number of training sample, λ is the coefficient of the regular term.

Training setting: The goal is to find the parameters of the network that minimize the average prediction log-loss after the softmax layer. Hence, Optimization is by stochastic gradient descent using mini-batches of 256 samples. The model is regularized using dropout. The coefficient of the latter was set to $5*10^{-4}$, whereas dropout was applied after the FC layers with a rate of 0.5. The learning rate was initially set to 10^{-2} and then decreased by factor of 10 when the validation set accuracy stopped increasing. The weights were initialized by random sampling from a Gaussian distribution with zero mean. Biases were initialized to zero. The training images were rescaled such that the width and height were equal to 160.

E. Age and gender Recognition

Regarding the identification of age and gender, since it is a problem of single frame recognition. It is not necessary to identify the age and gender after the recognition is completed, so we do not need to establish a new neural network for age and gender recognition to increase the burden of the system and computing resources. In view of the application mentioned above, we use the API for face recognition provided by Microsoft Azure. We can upload the detected face directly to Microsoft's server for detecting emotion, face angle, gender, age. The server will return the json file which has some information we need. Extracting the required information from json and we can get the gender and age information of the target.

IV. EXPERMENT RESULT

An emotion face dataset, FER+, is used for training and testing our proposed method. Inside and outside test are used for system evaluation. Finally, the comparison will described and discussed with difference algorithm.

A. *Dataset*

The Facial Expression Recognition plus (FER+) database consists of 4500 images of faces from FER. The task is to categorize each face based on the emotion shown in the facial expression in to one of seven categories, include angry, disgust, fear, happy, sad, surprise, and neutral. Five most common emotions, Neutral, Happy, Angry, Sad and Surprise are chosen for system evaluation.

B. *Experment Result*

The system evaluation consider the inside and outside test on FER+. The average accuracy rate for inside and outside test are 92.81% and 86.14%. An online test is

considered for evaluate the system for practical application; we asked the participant to make each expression five times. There are a total of 11 participants in experiment result of online test. The results is shown on TABLE II, and the confusion matrix of online test is shown on Table III.

TABLE II. THE EXPERIMENT RESULT OF FER+ DATASET

	Inside	**Outside**	**Online Test**
The proposed system	92.81%	86.14	85.4

The confusion matrix shows that the proposed system has well results in neutral, happy, angry. But the performance of sad and surprise emotion are not well. Sad and surprise emotion are recognized as neutral and angry, respectively. Two emotions have the same facial action on the face; for example, surprised and angry expressions have the same action with eyes open wide. The difference between sad and neutral emotion is too low.

The comparison between other emotion recognition methods and the proposed method is shown in Table IV. In the part of experiment, we can obviously find that our proposed approach improves accuracy rate 1.16% in this experiment. For the comparison with M Georgescu *et al* [16], Two reasons are discussed for improved algorithm. The image augmentation should be used on pro-processing, but this will reduce the applicability of the algorithm; the architecture of hybrid emotion recognition neural network is not deep enough for Indicates the details of the face. The method of improvement will be explained in the conclusion.

TABLE III. CONFUSION MATRIX OF EMOTION ONLINE TEST

	Neutral	Happy	Angry	Sad	Surprise
Neutral	94.5%	0.0%	0.0%	5.4%	0.0%
Happy	1.8%	98.1%	0.0%	0.0%	0.0%
Angry	0.0%	5.4%	90.9%	3.6%	0.0%
Sad	20%	5.4%	9.1%	65.4%	0.0%
Surprise	1.8%	3.6%	14.5%	1.8%	78.1%

TABLE IV COMPARISON OF DIFFERENT APPROACHES ON FER+ DATASET

	Accuracy (%)
E Barsoun *et al* [17]	84.98
M Georgescu *et al* [16]	87.76
The proposed method	**86.14**

V. CONCLUSION

A human understanding system is provided for emotion, age and gender recognition. Hybrid neural networks is proposed for emotion recognition, which is consider appearance and facial landmark points for global and local features. The experimental results have demonstrated the effectiveness of the proposed system, the recognition rate can achieve 86.14% for facial expression; the other hand, the age and gender analysis system, which are realized by Microsoft Azure API, are integrated on the human understanding system.

In order to improve the proposed algorithm, deeper convolutional network architecture should be considered for representation of facial details. Image pre-processing still reduce system performance, the problem of image augmentation will trying to be solved under the neural network architecture for improve the system.

REFERENCE

[1] P. Ekman, "Emotional and conversational nonverbal signals," in *Language, knowledge, and representation*: Springer, 2004, pp. 39-50.

[2] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern recognition,* vol. 27, no. 1, pp. 53-63, 1994.

[3] T. A. McGregor, R. L. Klatzky, C. Hamilton, and S. J. Lederman, "Haptic classification of facial identity in 2D displays: Configural versus feature-based processing," *Haptics, IEEE Transactions on,* vol. 3, no. 1, pp. 48-55, 2010.

[4] C.-T. Tu and J.-J. J. Lien, "Automatic location of facial feature points and synthesis of facial sketches using direct combined model," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on,* vol. 40, no. 4, pp. 1158-1169, 2010.

[5] K. Sandeep and A. Rajagopalan, "Human Face Detection in Cluttered Color Images Using Skin Color, Edge Information," in *ICVGIP*, 2002.

[6] M. R. Mohammad, E. Fatemizadeh, and M. H. Mahoor, "Intensity Estimation of Spontaneous Facial Action Units Based on Their Sparsity Properties," *IEEE Transactions on Cybernetics,* vol. 46, no. 3, pp. 817-826, 2016.

[7] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, 2000.

[8] C.-F. Chuang and F. Y. Shih, "Recognizing facial action units using independent component analysis and support vector machine," *Pattern recognition,* vol. 39, no. 9, pp. 1795-1798, 2006.

[9] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning Multiscale Active Facial Patches for Expression Analysis," *IEEE Transactions on Cybernetics,* vol. 45, no. 8, pp. 1499-1510, 2015.

[10] W. S. Chu, F. d. l. Torre, and J. Cohn, "Selective Transfer Machine for Personalized Facial Expression Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PP, no. 99, pp. 1-1, 2016.

[11] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-Delay Neural Network for Continuous Emotional Dimension Prediction From Facial Expression Sequences," *IEEE Transactions on Cybernetics,* vol. 46, no. 4, pp. 916-929, 2016.

[12] G. Zen, L. Porzi, E. Sangineto, E. Ricci, and N. Sebe, "Learning Personalized Models for Facial Expression Analysis and Gesture Recognition," *IEEE Transactions on Multimedia,* vol. 18, no. 4, pp. 775-788, 2016.

[13] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modeling," *IEEE Transactions on Cybernetics,* vol. 44, no. 2, pp. 161-174, 2014.

[14] 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.

[15] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing,* vol. 6, no. 1, pp. 1-12, 2015.

[16] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local Learning with Deep and Handcrafted Features for Facial Expression Recognition," *arXiv preprint arXiv:1804.10892,* 2018.

[17] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 279-283: ACM.