Fused convolutional neural network for facial expression recognition

M.K.Mohd Fitri Alif [1], A.R.Syafeeza[2*], P.Marzuki[3], A.Nur Alisa[4]

[1] Control and Mechatronics Engineering Department, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81300 Johor Bahru, Johor, Malaysia
[2,3,4] Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

[*]Corresponding e-mail: syafeeza@utem.edu.my

Keywords: Fused convolutional neural network; Facial expression recognition; Stochastic diagonal Levenberg Marquadt.

ABSTRACT – This study aims to find the optimal learning algorithm parameter, model and connection, initialization weight and normalization method using fused Convolutional Neural Network (CNN) for facial expression recognition. The best model and parameters are identified using a ten-fold cross validation method. By determining these ideal elements, a superior accuracy can potentially be achieved. CNN was utilized to a group of seven emotions from various facial expressions, namely, happy, sad, angry, surprise, disgust, fear and neutral. The four-layer CNN configuration was prepared with the JAFFE dataset, and yielded an overall accuracy of 83.72%. The outcome demonstrates that the fused CNN with the mentioned aims can generate higher accuracy with a smaller network compared to related models.

1. INTRODUCTION

One of the most important information on understanding human emotions are facial expressions. Human facial expressions are easier to be perceived as opposed to different signs. The facial expressions are one of the most complicated signal systems in our body, which consists of six muscles which can be moved independently [1]. Facial expression recognition has been utilized across numerous applications including the driver push state which distinguished facial expressions to caution the driver, the gamer response in gaming application for engineer criticisms, the advertisement industry, and the online instruction. There are seven fundamental expressions which are all inclusive among societies and countries, namely, happy, sad, surprised, angry, fear, disgust and neutral [2]. These are similar emotions that advanced facial expression scientists intend to distinguish utilizing computer vision.

In the recent years, the developments of facial expression recognition have made considerable progress. Fasel [4] achieved a state-of-the-art result in JAFFE dataset using CNN to perform facial expression recognition ensemble of five CNN layers and accuracy of 80.0%. However, the input image must be centered on a single face. Song et. al. [5] used the same number of the layer as Fasel but used a different dataset, the Extended Cohn Kanede (CK+) with an accuracy of 99.2% and uses the location of the eyes for preprocessing. The CNN model proposed by Gudi [6] uses an FER2013 dataset which has eight layers of CNN and accuracy of 65.65%. Yet, he used too many layers.

2. METHODOLOGY

The methodology for this research is divided into three sections. The first section discusses the database which is the JAFFE database, the second section discusses the applied pre-processing method and the final section discusses the CNN design in which fused CNN was used. This system ran on 2.3 GHz Intel i5-6200U processor, 12GB RAM, with Ubuntu 14.04 Linux operating system. MATLAB were used for pre-processing while GCC C compiler was used to run CNN C code.

2.1 Database

The database used in this system is Japanese Female Facial Expression (JAFFE) database [7] as shown in Figure 1 has seven total facial expressions, namely happy, sad, surprise, angry, disgust, fear and neutral. JAFFE database has 213 grayscale frontal faces pose images of ten Japanese female with a resolution of 254×256 pixels

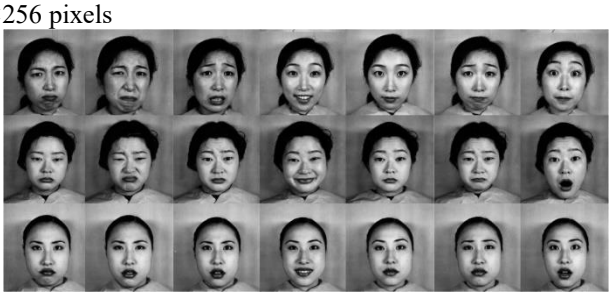

Figure 1 Sample images of JAFFE database

2.2. Pre-processing

The input to the network is expected to be in the term of facial image. However, it can be difficult for the deep network to handle high variations in the facial pose and lighting conditions. Thus, it becomes necessary to pre-process the input to make the faces more uniform. The first part is input image from the image of JAFFE database size 256×256 pixels. The second part is face detection; the algorithm used was the Viola Jones face detection algorithm. After the face was detected, the face was cropped and resized to 56×46 pixels. The image was then normalized using min-max normalization algorithm which produces pixel value within the range of -1.0 to 1.0. Output images were stored in numeric data and divided to two parts; 170 training normalized images and

© Centre for Robotics and Industrial Automation, Faculty of Electrical Engineering

43 testing normalized images.

2.3 Ten-Fold Cross Validation

The ten-fold cross-validation was used to find the best parameters and the best model of the fused CNN. The JAFFE dataset has 213 images in which 43 images were used for testing and the remaining was divided to 10 folds with one fold of validation and nine folds of training. This method will be repeated ten times, each time a different fold was selected as the validation set and the remaining sets as the training set.

2.4. CNN Design

The fused CNN design for facial expression recognition consists of four layers as shown in Figure 2. The design was inspired from LeNet-5 architecture. In the fused convolution/subsampling process that was performed at layer C1, C2, and C3, the convolution kernel was convolved with input feature map with the subsampling incorporated as a skipping operation in the convolution process. Only the size of the kernel was represented as a skipping factor between subsequent convolution in x and y directions respectively. The final layer was a full-connection layer, which is the neurons representing the seven classes. The Stochastic Diagonal Levenberg Marquadt (SDLM) learning algorithm [3] was used to accelerate convergence rate and maintain generalization ability.

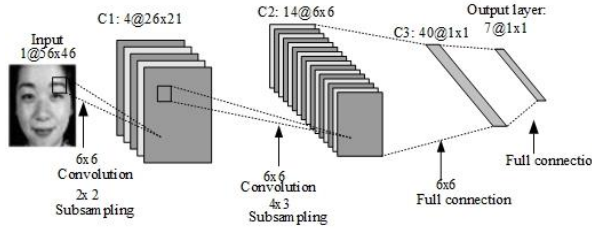

Figure 2 Proposed fused CNN design

3. RESULTS AND DISCUSSION

This section presents the result for the best parameter, model, connection and the results. These values are obtained using the 10-fold cross-validation method. There are four CNN models tested including 3-14-60, 4-14-60, 5-14-60 and 6-14-60, and these models represent the number of feature map at layer C1, C2 and C3 respectively. The forth layer is the output layer that is fixed to seven representing the seven types of emotions. SLDM learning algorithm has two critical parameters known as regularization parameter, μ and γ constant.

The two learning rate values were tested, which is 0.01 and 0.001. The best value is 0.01 since it produces the lowest Mean Square Error (MSE), at 13 epochs. There are seven regularization parameters tested which are 0.02, 0.03, 0.04, 0.05, 0.06, 0.07 and 0.08.

The best model was tested using the best learning rate constant and regularization parameter. Full connection between layer C1 and C2 was used to obtain the best model. The model 4-14-60 has been identified as the best model since it has the lowest validation error compared to the other models.

The benchmark is meant only for CNN using CPU and JAFFE databases. The proposed design has triumphed the accuracy of the state-of-the-art method by Fasel [4] and Neagoe et.al. [7] with an accuracy of 83.72% and less number of layers. The training only takes three seconds for each training session meanwhile the testing only takes less than one second. The combination of min-max normalization method with uniform weight initialization method produced the highest accuracy in this experiment

4. CONCLUSION

The proposed CNN model has proven to save computational time and burden. The four-layer fused CNN has been proposed for facial expression recognition with min-max normalization method and uniform weight initialization method, and are able to recognize facial expression with less number of CNN layer and faster converge rate with significant high accuracy, 83.72% compared to another existing result. This accuracy can be improved by adopting other methods such as maxpooling, RELU and contrast layers.

ACKNOWLEDGEMENT

This work is supported by Universiti Teknikal Malaysia Melaka (UTeM) under the grant PJP/2018/FKEKK(9D) S01622.

REFERENCES

[1] Ernst H. (1934). Evolution of Facial Musculature and Facial Expression. *Journal of Nervous and Mental Decease, 79*(109).

[2] Yu, K., Wang, Z., Zhou, L., Wang, J., Chi, Z., & D. Feng (2013). Learning realistic facial expressions from web images. *Pattern Recognition*, *46*, 2144–2155.

[3] Syafeeza A. R., Khalil-Hani, M., Liew, S. S. & Bakhteri R. (2015). Convolutional Neural Networks with Fused Layers Applied to Face Recognition. *International Journal of Computational Intelligence and Applications*, 14:1550014.

[4] Fasel, B. (2002) Mutliscale facial expression recognition using convolutional neural networks. In: *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 02), Ahmedabad, India. 1–9.*

[5] Song, I., Kim, H. J., & Jeon P. B. (2014). Deep learning for real-time robust facial expression recognition on a smartphone. In: *IEEE International Conference on Consumer Electronics,* 564–567.

[6] Gudi, A. (2015). Recognizing semantic features in faces using deep learning. Master Thesis. University of Amsterdam.

[7] Neagoe, V., Bărar, A., Sebe, N., & Robitu P. (2013). A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions. *Recent Advances in Image, Audio and Signal Processing.* 93-98.

© Centre for Robotics and Industrial Automation, Faculty of Electrical Engineering