Fused Convolutional Neural Network for Facial Expression Recognition

M.K. Mohd Fitri Alif, A.R. Syafeeza

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.

Keywords: Deep learning, emotion recognition, facial expression recognition, fused convolutional neural network, Stochastic Diagonal Levenberg Marquadt.

I. INTRODUCTION

One of the most important information on understanding human emotions is 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 applications for engineer criticisms, the advertisement industry, and the online instruction. The extraction approach of facial expression that has been utilized is either through manual inference [2] or an automatic detection approach [3]. There are seven fundamental expressions which are all inclusive among societies and countries namely, happy, sad, surprised, angry, fear, disgust and neutral [4]. These are similar emotions that advanced facial expression scientists intend to distinguish utilizing computer vision.

A promising methodology that can be utilized to address facial expression recognition issues is Convolutional Neural Network (CNN). CNN has been utilized for different applications, for example, gender recognition [5], finger-vein identification [5]–[7], face recognition [8]–[10], licence plate recognition [12] and other applications.

In [13] has introduced a state of the art CNN design; the LeNet-5 which is utilized for handwriting recognition. It comprises of seven layers, not including the input. This CNN architecture is different as subsampling layer is added after a convolution layer. The purpose of this subsampling is to compress a group of neighborhood pixels, while preserving the original information of the input image. These layers are repeated before passing the resulting feature maps to a three-layer classification task. LeNet5 architecture is shown in Fig. 1.

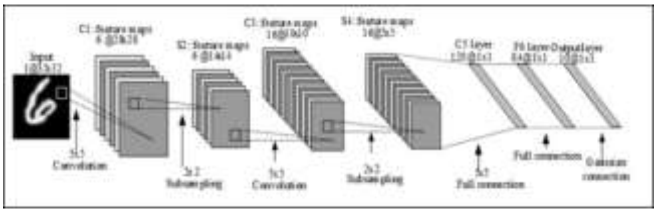

Fig. 1.LeNet-5 CNN architecture.

In the recent years, the developments of facial expression recognition have made considerable progress. In [14] 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. In [15] used the same number of the layer as in [14], 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 [16] uses an FER2013 dataset which has eight layers of CNN and accuracy of 65.65%. Yet, he used too many layers.

This paper focuses on designing the CNN model and architecture, particularly for facial expression recognition that would save computational time and burden. Ten-fold cross validation method has been used to determine the best parameters, model, connections, weights, and normalization method. It has been proven that by applying this method, the accuracy can be measured.

This paper is organized as follows: The next section introduces the CNN and theory of facial expression recognition. Section 3 describes the database used and the proposed approach. The results and analysis are shown in Section 4. Finally, Section 5 concludes this paper.

II. 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 preprocessing method and the final section discusses the CNN design in which fused CNN was used. This system

Revised Manuscript Received on September 14, 2019.

M.K. Mohd Fitri Alif, School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia, 81300 Johor Bahru, Johor, Malaysia.

A.R. Syafeeza, Faculty of Electronic Engineering and Computer Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia. Email: syafeeza@utem.edu.my

ran on 2.3 GHz Intel i5-6200U processor, 12GB RAM, with Ubuntu 14.04 Linux operating system. MATLAB was used for preprocessing, while the GCC C compiler was used to run CNN C code.

A. Dataset

The dataset used in this system is Japanese Female Facial Expression (JAFFE) dataset [17] that has seven total facial expressions, namely happy, sad, surprise, angry, disgust, fear and neutral. JAFFE dataset has 213 grayscale frontal faces pose images of ten Japanese females with a resolution of 256×256 pixels. Fig. 2 shows the sample image of JAFFE dataset.

Fig. 2.Sample images from JAFFE dataset.

B. Preprocessing

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 preprocessing step can be divided into six parts as shown in Fig. 3. The first part is input image from the image of JAFFE dataset 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 into two parts; 170 training normalized images and 43 testing normalized images.

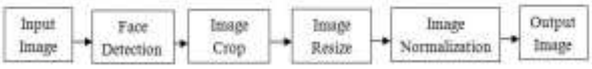

Fig. 3.Pre-processing procedures.

C. 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.

D. CNN Design

The fused CNN design for facial expression recognition consists of four layers as shown in Fig. 4. The design was inspired from LeNet-5 architecture. The fusion of convolution/subsampling process was done at CNN layers C1, C2, and C3. The convolution process was skipped by convolved convolution kernel together with input feature map and subsampling process. The seven classes represented by seven full connected layer neurons. To accelerate the convergence process and maintaining the generalization ability, the Stochastic Diagonal Levenberg Marquadt (SDLM) learning algorithm [9] was used.

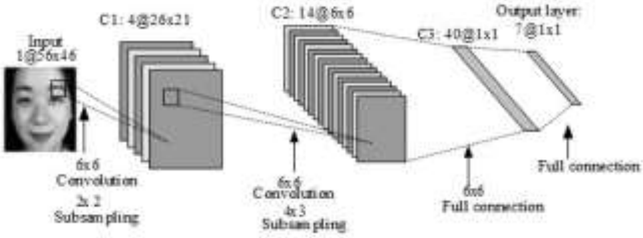

Fig. 4.Proposed fused CNN design.

The number of neurons, connections and trainable parameter was reduced due to fusion of convolution and subsampling layers. Therefore, the model complexity also reduced. This fusion approach results in four layers 4-14-60 model architecture as shown in Fig. 4. On the first layer, C1 is 4@26×21 convolutional layer that has four feature maps of size 26×21 pixels, hence producing a total number of 2184 (26×21×4) neurons, the total number of 148 (6×6×4+4) trainable weights and each 2184 neurons has 37 connections with a total number of 80808 total connections from input to layer C1. On the second layer, C2 has 14 6×6-pixel feature maps. The total number of 504 (6×6×14) neurons, the total number of 518 (6×6×14+14) trainable weights and 504 neurons has 31 connections which give the total number of 15624 (6×6×14×31) connections. Fig. 5(a) shows the normal convolution and subsampling process and Fig. 5(b) shows the fused convolution and subsampling process.

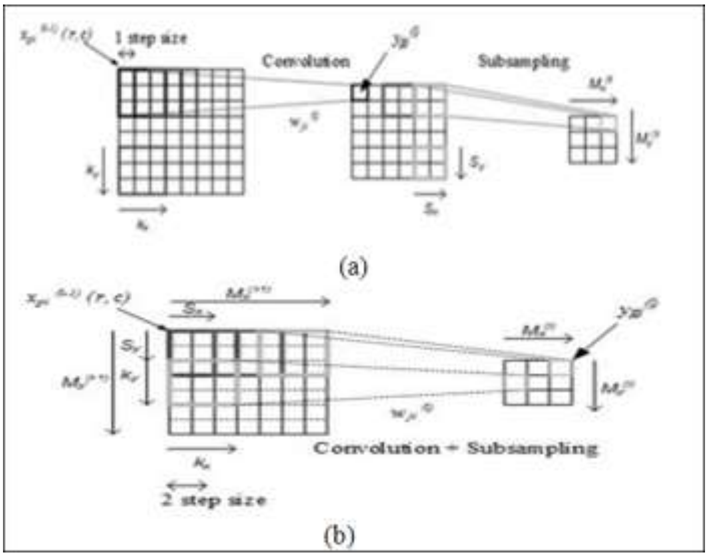

Fig. 5.(a) Normal convolution and subsampling process (b) Fused convolution and subsampling process.

DOI: 10.35940/ijitee.L1098.10812S219

III. 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 fourth layer is the output layer that is fixed to seven representing the seven types of emotions. The 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. Based on Fig. 6(a), 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 parameter is 0.02 as shown in Fig. 6(b).

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. Per Fig. 7, four models were evaluated to find 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.

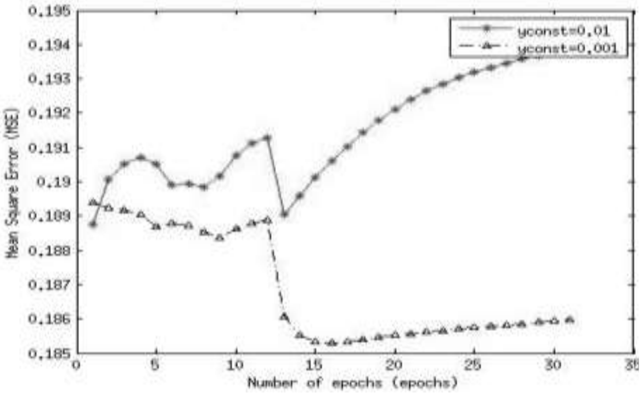

Fig. 6(a).Cross validation result for γ constant.

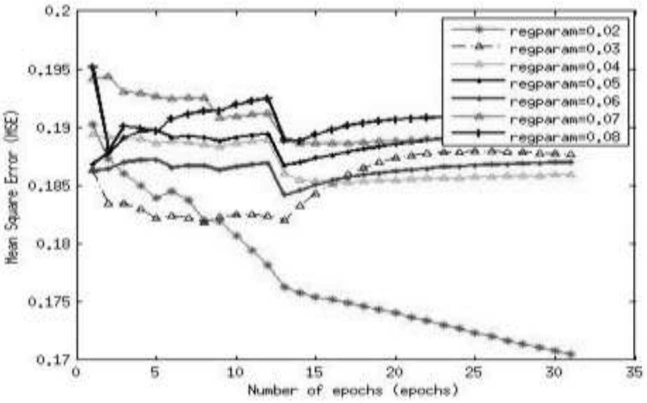

Fig. 6(b).Cross validation result for regularization parameter.

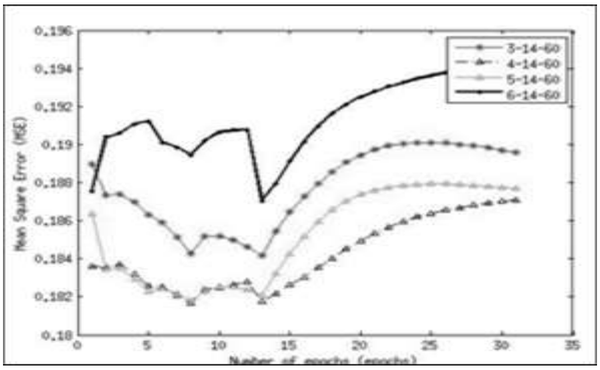

Fig. 7.Best model.

To find the best connection between layer C1 and C2, six patterns had been tested as shown in Table I. The notation "X" is indicated as the connection of feature maps between layer C1 and layer C2. The best connection between layer C1 and C2 was carried out by averaging the ten-fold cross validation of each pattern for each feature map on layer C2 start from feature map C2[0] to C2[13]. The best connection between layer C1 and C2 are shown in Table II.

Table- I: Various pattern of connection between C1 and C2

Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5	Pattern 6
X			X		X
X	X		X	X	
	X	X	X	X	
		X		X	X

Table- II: Best connection between layer C1 and C2

		C2													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
C1	0		X	X	X	X	X							X	
	1	X		X	X	X		X				X		X	X
	2	X		X				X	X	X	X	X	X	X	X
	3		X				X		X	X	X		X		X

Table III presented the result of benchmarking of the number of layers and accuracy. This 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 [14], [18] 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 as shown in Table IV. From Fig. 8, we can see that the visualization of each feature map on each layer behave as feature detector filters such as sharpening, blurring and edge detection.

Table- III: Accuracy for different combination of normalization method and weight initialization method

Normalization Method	Weight initialization method	Accuracy
Min-Max	Normal	76.74%
	Uniform	**83.72%**
	Fan In	23.26%
	Nguyen Widrow	23.26%
Z Score	Normal	23.26%
	Uniform	20.93%
	Fan In	16.28%
	Nguyen Widrow	13.95%

Table- IV: Benchmarking of model complexity

References	No. of neurons	No. of trainable parameters	No. of connections
Fasel, B (2012) [14]	2995445	237708	5304515
Neagoe, V. E. et.al. (2013) [18]	19712000	17165103	2944307200
Neagoe, V. E. et.al. (2013) [18]	396697600	12807867	39887308800
Proposed Method	**2688**	**666**	**96432**

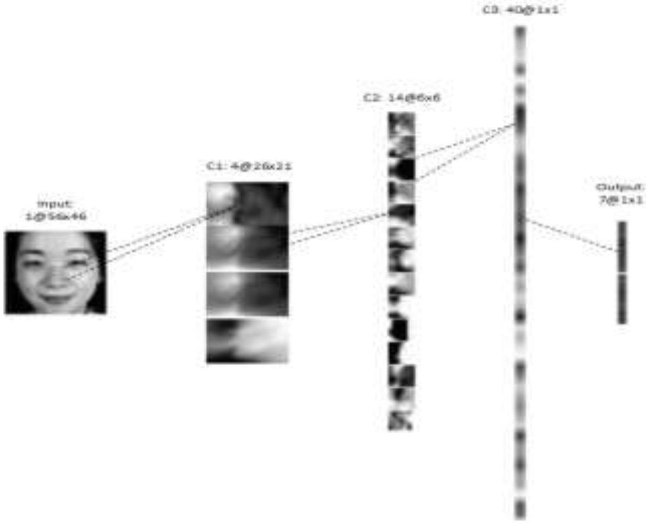

Fig. 8.Visualization of feature map for each layer.

IV. 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 the faster convergence rate with a significant high accuracy, 83.72% compared to the other existing result. This accuracy can be improved by adopting other methods such as maxpooling, RELU and contrast layers.

V. ACKNOWLEDGMENT

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

REFERENCES

1. H. Ernst, "Evolution of facial musculature and facial expression," Journal of Nervous and Mental Disease, 79, 1934, pp. 109.
2. K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," IEEE Transactions on Systems, Man and Cybernetics, Part B, 36, 2006, pp. 96–105.
3. I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," IEEE Transactions on Image Processing, 16, 2007, pp. 172–187.
4. K. Yu, Z. Wang, L. Zhuo, J. Wang, Z. Chi, and D. Feng," Learning realistic facial expressions from web images," Pattern Recognition, 46, 2013, pp. 2144–2155.
5. S. S. Liew, M. Khalil-Hani, S. Ahmad Radzi, and R. Bakhteri, "Gender classification: a convolutional neural network approach," Turkish Journal of Electrical Engineering and Computer Sciences, 24, 2016, pp. 1248–1264.
6. S. Ahmad Radzi, M. Khalil-Hani, and R. Bakhteri, "Finger-vein biometric identification using convolutional neural network," Turkish Journal of Electrical Engineering and Computer Sciences, 24, 2016, pp. 1863–1878.
7. K. Syazana-Itqan, A. R. Syafeeza, N. M. Saad, N. A. Hamid, and W. H. Mohd Saad, "A review of finger-vein biometrics identification approaches," Indian Journal Science and Technology, 9, 2016, pp. 1-8.
8. K. S. Itqan, A. R. Syafeeza, F. G. Gong, N. Mustafa, Y. C. Wong, and M. M. Ibrahim, "User identification system based on finger-vein patterns using Convolutional Neural Network," ARPN Journal Engineering and Applied Science, 11, 2016, pp. 3316–3319.
9. A. R. Syafeeza, M. Khalil-Hani, S. S. Liew, and R. Bakhteri, "Convolutional Neural Networks with fused layers applied to face recognition," International Journal of Computer Intelligence and Applied, 14, 2015, pp. 1-9.
10. A. R. Syafeeza, M. Khalil-Hani, S. S. Liew, and R. Bakhteri, "Convolutional neural network for face recognition with pose and illumination variation," International Journal Engineering and Technology, 6, 2014, pp. 44–57.
11. K. S. Itqan, A. R. Syafeeza, and M. S. Norhashimah, "A MATLAB-based convolutional neural network approach for face recognition system," Journal of Bioinformatics and Proteomics Review, 2, 2016, pp. 1-5.
12. S. A. Radzi and M. Khalil-Hani, "Character recognition of license plate number using convolutional neural network," International Visual Informatics Conference, 2011, pp. 45-55.
13. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Computation, 1, 1989, pp. 541–551.
14. B. Fasel, "Mutliscale facial expression recognition using convolutional neural networks," Indian Conference on Computer Vision, Graphics and Image Processing, 2012, pp. 1-9.
15. I. Song, H. J. Kim, and P. B. Jeon, "Deep learning for real-time robust facial expression recognition on a smartphone," IEEE International Conference on Consumer Electronics, 2014. pp. 564–567.
16. A. Gudi, Recognizing semantic features in faces using deep learning. [Online]. Available: https://arxiv.org/pdf/1512.00743.pdf.

17. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," 3rd IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205.
18. V. Neagoe, A. Bărar, N. Sebe, and P. Robitu, "A deep learning approach for subject independent emotion recognition from facial expressions," Recent Advances in Image, Audio Signal Processing, 2013, pp. 93-98.

AUTHORS PROFILE

Mohammad Fitri Alif Mohammad Kasai currently is a PhD candidate in Electrical Engineering from Universiti Teknologi Malaysia. He received B.Eng degree in Mechatronics Engineering in 2011 from Universiti Selangor and his M.Eng degree in Mechatronics and Automatics Control in 2013 from Universiti Teknologi Malaysia. His PhD research is Facial Expression Recognition and Deep Learning.

Syafeeza Ahmad Radzi received her B.Eng degree in Electrical-Electronic Engineering in 2003 and her M.Eng degree in Electrical - Electronic & Telecommunication Engineering in 2005 from Universiti Teknologi Malaysia. She also received her PhD degree in Electrical Engineering from the same university in 2014. She is currently a Senior Lecturer at the Faculty of Electronic Engineering and Computer Engineering, Universiti Teknikal Malaysia Melaka (UTeM). She has been an academician in UTeM since 2006. She dedicates herself to university teaching and conducting research. Her research interests include embedded system, pattern recognition, machine learning, deep learning, image processing and biometric.