The Improved Artificial Neural Network Based on Cosine Similarity in Facial Emotion Recognition

Kartika Candra Kirana
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
kartika.candra.ft@um.ac.id

Slamet Wibawanto
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
slamet.wibawanto.ft@um.ac.id

Nur Hidayah
Department of Guidance and Counseling
Universitas Negeri Malang
Malang, Indonesia
nur.hidayah.fip@um.ac.id

Gigih Prasetyo Cahyono
Software Engineering
Visionet Data International
Malang, Indonesia
gigih.cahyono@visionet.co.id

***Abstract*— In this study, we present the improved artificial neural network based on cosine similarity in facial emotion recognition. We apply a shifting window that employs a neural network for two concurrent processes consisting of face detection and emotional recognition. To prevent the slow and futile computations, non-face areas need to be filtered from neurons on each network layer, thus we propose the improved artificial neural network based on cosine similarity. Cosine similarity is employed to bypass the process of non-face areas in the neural network. The accuracy of the proposed method reaches 0.84, while the accuracy of the original neural network method reaches 0.74. It can be concluded that our methods work accurately. It can be concluded that our method works accurately. The proposed method is superior to the state-of-the-art algorithms**

Keywords*— *emotion recognition, neural network, cosine similarity

I. INTRODUCTION

In non-verbal communication, facial expressions present internal feelings. It also reflects the state of physical expressions. Many algorithms can be used to recognize facial emotion. One of the commonly used algorithms is the neural network. The neural network is often thought of as a robust algorithm in recognizing emotions. Several studies have shown that variations of the neural network algorithm produce better performance than other algorithms.

Kumar et al (2017) proposed the deep convolution neural network for the determine the emotion level [1]. Based on a test of the CK + dataset, the convolution neural network performs better than the SVM that has problems in multiclass classification. Y. Liu et al (2018) also proposed a novel convolution neural network using multi-pose for recognizing facial expressions in non-frontal views [2]. They show that their methods outperform than the FER methods. Furthermore, Qiuyu Li et al (2018) proposed a fused convolutional network which is applied for dividing the facial region and extracting the optical flow features in the micro-expression recognition [3]. Their result achieved competitive performance in two spontaneous micro-expression data. Also, other neural network models perform well. K.C. Kirana et all (2018) show a hierarchical Bayesian neural network for analysis the expression of neurodegenerative diseases [4]. Their methods work better than the original Bayesian framework in both multiclass classification and regression settings. Even neural networks also work well on a robot. Xiao Huang et al (2019) proposed a novel emotion rule to train a recurrent neural network in a neuromodulator system which produces the faster and more accurate methods than previous methods [5].

In a previous study [6], we also have demonstrated a reliable neural network in recognizing emotions. However, the system detects and crops the face manually. Therefore, we improvised our system by automating face search that applies a shifting window of [7]. Besides, to conserve our computational, face detection and emotional recognition that essentially extract the same object is combined into the single neural network architecture. However, if the shifting window passes through the non-face area and computes it on every neuron of a neural network like the face area, it will be the slow and futile computation. Therefore, we improvised neural networks to allow non-face areas to be filtered.

Related to the non-face area filter, the type of input needs to be considered. In this study, fisher linear discriminant (FLD) is selected as the input because this feature has a low dimension and the balance of data distribution in the form of vector. The measurement of the similarity between vectors is required in the filter of the non-face area. [8]–[10] shown that Cosine similarity measurement (CSM) has a good representation in measuring distances between vectors. Even [8] showing cosine similarity measure has an accuracy of up to 0.99 in measuring semantic vectors that have been reduced to 8 bits. Moreover [9] also shows that CSM faster than the linear scan and approximation methods.

Based on the explanation above, artificial neural networks find the best weight of each facial feature (input) iteratively[11]. However, the original neural network cannot remove non-face areas before the process is complete[6]. Thus we buried the ground between neural network functions to speed up the filter process. However, it triggers early conferences, so we immerse the calculation of the similarity between training data and test data to prevent early conferences.

The CSM measures the cosine of the angle between two non-zero vectors. In other words, CSM measures the similarity between two non-zero vectors [8]–[10]. Therefore,

we present the improved artificial neural network based on cosine similarity in facial emotion recognition. Our basic idea is cosine similarity-based filtering. If cosine similarity does not reach the expected threshold, the system will filter out the shifting window that assumed as the non-face area from the neural network. It hoped to avoid our system work slowly and in vain.

II. METHODOLOGY

In a previous study [6], we have demonstrated a reliable neural network in recognizing emotions. However, the system detects the face manually. Therefore, we improvised our system by automating face search that applies a shifting window. Every the window is shifted, three computed stages are consisting of feature extraction, face detection, and emotion recognition. Also, to conserve our computational, face detection and emotional recognition that essentially extract the same object is combined into a single neural network architecture which is named as the improved artificial neural network based on cosine similarity. The shifting window is set to 20 by 20 pixels. To address face size greater than the size of windows, the shifting window is also allowed to enlarged by 100% and then cropped and scaled to 20 by 20 pixels.

A. Feature Extraction

Fisher linear discriminant (FLD) is selected as the input of the neural network. This feature is a linear feature extraction in the form of reduction approach. It is computed by a combination of Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

PCA is an unsupervised algorithm that guarantees the distribution information. Projection matrix φ of PCA is shown in Equation 1.

$$PCA(\varphi) = \varphi^T S_T \tag{1}$$

where PCA is calculated by multiply the projection of matrix φ, the transpose of matrix φ (φ^T), and summation of between-class matrix and within-class matrix (S_t). Because the between-class matrix and within-class matrix is incorporated in S_t, PCA cannot formulate the optimal matrix projection. However, it can be handled by LDA which a supervised algorithm and it is shown by Equation 2.

$$LDA(\varphi) = \frac{\varphi^{\mathrm{T}}\,\mathrm{Sb}\,\varphi}{\varphi^{\mathrm{T}}\,\mathrm{Sw}\,\varphi}. \tag{2}$$

In contrast to PCA, LDA separate the between-class matrix S_b and within-class matrix S_w which guide the LDA being able to formulate the optimal matrix projection. However, it often encounters a small sample thus it cannot be applied directly. This problem can be solved by combining LDA and PCA.

FLD can be computed by applying Equations 1 and 2. However, both equations can be simplified into Equation 3.

$$FLD(\varphi) = PCA(\varphi)_{fisher}\Lambda \tag{3}$$

where Λ is an Eigen matrix computed by Equation 1. Whereas φ_{fisher} is the eigenvalue that has been eliminated by the value of the zero factors and sorting in descending order. Equation 3 produces an FLD vector that has a low dimension and balance of data distribution.

To project the size of the optimal window, the cosine similarity of FLD values between the current shifted window and the enlarged window is compared iteratively. If the cosine similarity of the enlarged window is larger, the shifted window size is zoomed to 100% and FLD input values are updated. Whereas the shifted window size will not change and the loop is stopped if the cosine similarity in the current shifted window is larger. The cosine similarity is calculated as follows:

$$cos\ (\theta) = \frac{\mathrm{A_i}.\overline{\mathrm{A_j}}}{\|\mathrm{A_i}\|\|\overline{\mathrm{A_j}}\|} = \frac{\sum_{i,j}^{n} \mathrm{A_i}\overline{\mathrm{A_j}}}{\sum \mathrm{A_i}^2 \sum \overline{\mathrm{A_j}}^2}, \tag{4}$$

where the cosine similarity values *cos (θ) of vector* A_i are calculated against the median vector in the training class $\overline{\mathrm{A_j}}$.

B. The Improved Artificial Neural Network based on Cosine Similarity In Facial Emotion Recognition

In this study, we present the improved artificial neural network based on cosine similarity in facial emotion recognition. The essence of the proposed approach is the using of cosine similarity to bypass the process of non-face areas in a neural network. Also, it combines face detection and emotional recognition into the single neural network architecture. The architecture of the proposed method is shown in Figure 1.

In Figure 1, assume there is an input layer X_i which is consist of a series of FLD vectors $J(\varphi)$ and one input bias, a series of hidden layer Z_j, and a series of output layer Y_k. The neurons between input layer X_i and hidden layer Z_j are fully interconnected by weight Vij, whereas neurons between hidden layer Z_j and output layer Y_k are fully interconnected by weight W_{ijk}. The output Y is divided into Y_0 as a non-face area, and $\{Y_1, ..., Y_p\}$ as face area with various emotion.

Our modification lies in the summation stage. We applied the cosine similarity which is shown in Figure 2. The cosine similarity, *cos(θ),* is represented as follows:

$$\cos(\theta) = \frac{\mathrm{wA_i}.\mathrm{A_j}}{\|\mathrm{wA_i}\|\|\mathrm{wA_j}\|} = \frac{\sum_{i,j}^{n} \mathrm{wA_i}\mathrm{A_j}}{\sum \mathrm{wA_i}^2 \sum \mathrm{wA_j}^2}. \tag{5}$$

Given the matrix of weight w and two vectors of attributes, A_i is the component of testing vector and A_j is the component of the training vector. In this stage, the weight of each vector component is considered because it is related to the basic concepts of neural networks involving the weights between neurons. The similarity ranges are {-1,1} that the in-between values indicate the similarity of training vector and testing vector. [12][13]

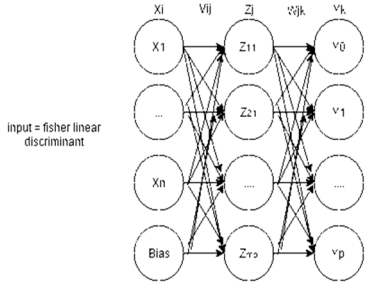

Fig. 1. An Overview of Our Proposed Methods

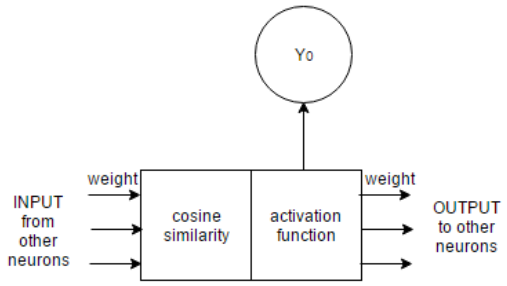

Fig. 2. The specific description of neurons of our proposed method

Based on Figure 2, the cosine similarity is fed into the activation function. Also, the activation function is allowed to jump to Y_0 without passing through neurons on the next layer. It is useful to filter the process of non-face areas in the neural network. We used the sigmoid function as the activation function σ which is shown in Equation 6.

$$\sigma(cos(\theta)) = \Sigma(1 + e^{-a\ cos(\theta)})^{-1}\ , \quad (6)$$

where the parameter determines the steep grade of the sigmoid function which is set a=1.

In the training class, error surface e is evaluated using the different of the target output T and estimated output Y as shown as follows:

$$e = T\text{-}Y, \quad (7)$$

Then the weights are adjusted simultaneously using Equation 8 [6][14].

$$w(n+1) = w(n) + \alpha x_i \sigma'(cos(\theta))\ \Sigma e_i\ w_0 \quad (8)$$

where w(n+1) is the updated weight, $w(n)$ is the previous weight, a is the steep grade of the function, x_i is input, $\sigma'(cos(\theta))$ is the derivative of activation value, Σe_i is the sum of error surface, and w_0 is the initialized weights.

III. Result and Discussion

A. Dataset

There is two basic emotion consisting of positive emotion and negative emotion. Moreover, basic emotion is also classified as neutral, happiness, sadness, anger[13], fear, surprise, and disgust [6][13] Besides, basic emotion is also classified as interest, engagement, confusion, frustration, satisfaction, hopefulness, boredom, and disappointment [15].

In this study, we implement this algorithm in real learning. However, not all expression is recorded in real learning. We only capture the interest as a positive emotion and boredom as a negative emotion. Interest expression is indicated by focusing on lecturers' presentation while bored expression is indicated by falling asleep, yawning, or not focusing on lecturer presentation.

We captured the emotion of eight UM's students using CANON EOS 700D DSLR camera and produced 550 images that the students are facing the camera. The dataset is separated into 500 training data and 50 testing data. Training data consists of 250 interest image and 250 bored images whereas the testing data consists of 25 interest image and 25 boring images. The sample of the dataset is shown in Figure 3. 3(a) shows the interest expression and 3(b) shows the bored expression.

B. Evaluation

We quantitatively evaluate the performance using accuracy shown in Equation 9.

$$accuracy = TP+TN\ /\ TP+TN+FP+FN \quad (9)$$

where TP is the really interesting face that is detected an interesting face, TN is a real boring face that is detected the bored face, FP is a real boring face or the unface that is detected an interesting face, FN is the real face of interest or the unface that is detected the bored face. The real image is classified manually by psychology organization named *'Lazuardy'*.

First, we evaluate the number of the hidden layer with variations between {5,10,15}. The evaluation of the hidden layer is shown in Table I. Based on Table I, we used ten hidden layers as the best result. The use of a little hidden layer can not distinguish emotions accurately whereas the use of many layers does not raise the performance significantly.

(a)

(b)

Fig. 3. A Sample of Dataset (a) The Sample of Interest Expression (b) The Sample of Bored Expression

Second, we not only compared the proposed method and the original neural network using accuracy, but also computation time in second. The comparison is shown in Table II. Based on Table II, our method takes a longer computation time than the original neural network method. It is caused by the multiply of the weights, the data vector train and test data in our proposed method. While the original method only computes the weights and vectors of the test data. However, the accuracy of our proposed method is significantly higher than the original neural network which is due to the comparison of training data and test data directly that can be shown in Equation 4.

Last, we evaluate the errors that occur visually. Some errors occur because of the similarity of expression, such as yawning (Figure 4(a)) and laughing (Figure 4(b)) where both expressions open the mouth. Although the eye pattern is different, the mouth area is larger than the eye area, so the influence of the mouth is greater than the eye area. This is shown in Figure 4. In further research, the weighted of critical areas is necessary to give the dominance of important areas.

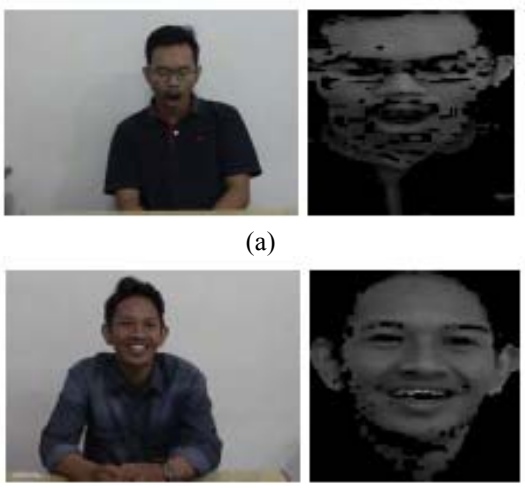

Fig. 4. The similarity of expression (a) yawning as a bored expression (b) laughing as an interest expression

TABLE I. THE EVALUATION OF HIDDEN LAYER

Number of Hidden Layer	Result				
	TP	***TN***	***FP***	***FN***	***Accuracy***
5	20	19	6	5	0.78
10	21	21	4	4	0.84
15	22	20	5	3	0.84

TABLE II. THE COMPARISON OF ALGORITHMS

Methods	Result					
	TP	***TN***	***FP***	***FN***	***Accuracy***	***Time (s)***
Original NN	18	19	6	7	0.74	2.3
Improved NN	21	21	4	4	0.84	3.1

IV. CONCLUSION

The accuracy of the proposed method reaches 0.84, while the accuracy of the original neural network method reaches 0.74. It can be concluded that our methods work accurately. However, our method works in 3.1 seconds which is longer than the original neural network method that only works in 2.3 seconds. Besides, visual results show the same expression pattern between emotions. In further research, the emphasis of important areas and the time reduction is required. Also, this method needs to be tested with a more complete type of emotion.

ACKNOWLEDGMENT

This research is fully financed by PNBP of the State University of Malang and Visionet Data International. This research is also inspired by the advice of Prof. Akira Asano

REFERENCES

[1] R. k. Kumar, G. A. R. Kumar, J. Garain, D. R. Kisku, and G. Sanyal, "Determine attention of faces through a growing level of emotion using deep Convolution Neural Network," in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, 2017, pp. 975–980.

[2] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-Channel Pose-Aware Convolution Neural Networks for Multi-View Facial Expression Recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 458–465.

[3] Q. Li, J. Yu, T. Kurihara, and S. Zhan, "Micro-expression Analysis by Fusing Deep Convolutional Neural Network and Optical Flow," in *2018 5th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2018, pp. 265–270.

[4] A. Joshi, S. Ghosh, S. Gunnery, L. Tickle-Degnen, S. Sclaroff, and M. Betke, "Context-Sensitive Prediction of Facial Expressivity Using Multimodal Hierarchical Bayesian Neural Networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 278–285.

[5] X. Huang, W. Wu, H. Qiao, and Y. Ji, "Brain-Inspired Motion Learning in Recurrent Neural Network With Emotion Modulation," *IEEE Trans. Cogn. Dev. Syst.*, vol. 10, no. 4, pp. 1153–1164, 2018.

[6] S. Wibawanto and K. C. Kirana, "Recognition of student emotion based on matrix-1 median fisher's face and backpropagation algorithm," in *2017 International Conference on Electrical Engineering and Informatics (z=ICELTICs)*, 2017, pp. 103–108.

[7] Paul Viola, "Robust Real-Time Face Detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[8] M. Karwatowski, M. Wielgosz, M. Pietron, M. Staruchowicz, and K. Wiatr, "Comparison of semantic vectors with reduced precision using the cosine similarity measure," in *2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 898–904.

[9] S. Eghbali and L. Tahvildari, "Fast Cosine Similarity Search in Binary Space with Angular Multi-Index Hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 329–342, 2019.

[10] B. Pathak and N. Lal, "Information retrieval from heterogeneous data sets using moderated IDF-cosine similarity in vector space model," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017, pp. 3793–3799.

[11] K. C. Kirana, S. Wibawanto, N. Hidayah, and G. P. Cahyono, "Improved Neural Network using Integral-RELU based Prevention Activation for Face Detection," *Int. Conf. Electr. Electron. Inf. Eng.*, 2019.

[12] K. C. Kirana, S. Wibawanto, N. Hidayah, and G. P. Cahyono, "Ant System for face detection," *2019 Int. Semin. Appl. Technol. Inf. Commun.*, 2019.

[13] K. C. Kirana, S. Wibawanto, and H. W. Herwanto, "Facial Emotion Recognition Based on Viola-Jones Algorithm in the Learning Environment," in *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 406–410.

[14] J. G. Rázuri, D. Sundgren, R. Rahmani, and A. M. Cardenas, "Automatic Emotion Recognition through Facial Expression Analysis in Merged Images Based on an Artificial Neural Network," in *2013 12th Mexican International Conference on Artificial Intelligence*, 2013, pp. 85–96.

[15] K. C. Kirana, S. Wibawanto, and H. W. Herwanto, "Emotion recognition using fisher face-based viola-jones algorithm," in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2018, vol. 2018-Octob.