Face Expressions Recognition Based on Image Processing using Convolutional Neural Network for Human Computer Interface

Nathaniel Syalomta [1, a)] Namira Fasya Rahim[2, b)] Korediantо Usman[3, c)] Nor Kumalasari Caecar Pratiwi[4, d)]

Author Affiliations
[1,2,3,4]*Telkom University, Jl. Telekomunikasi No. 1, Bojongsoang, Bandung Regency, West Java, Indonesia 40257*

Author Emails
[a)] *Corresponding author: nathanielgtg@ student.telkomuniversity.ac.id*
[b)] *namirafr@student.telkomuniversity.ac.id*
[c)]korediantousman@telkomuniversity.ac.id
[d)]nkcp@telkomuniversity.ac.id

Abstract. Feelings, communicated in various structures, are a specific relational correspondence. The emotional state helps in an independent direction, helps inventiveness, and oversees human comprehension and human-machine correspondence. In a couple of years, the need to recognize an individual's feelings is expanding, and interest in human feeling acknowledgment in different fields has been expanding, however not restricted to human-PC interfaces and metropolitan sound discernment. This study proposed a new self-constructed architecture named NNN-Net to compare it with famous AlexNet architecture. We use the same parameters, input size, optimizer, and learning rate in both architectures to find the best combinations that will perform the best result. The dataset that we use is CK+48, one of the famous datasets to study face expression recognition. We also augmented the dataset to increase the number of images for each class and balance the dataset. Furthermore, we found that our NNN-Net shows better results with an exact combination of parameters. The best accuracy result is 98.63%. at last, this study can be helpful as a foundation to classify students' expression using an online meeting platform.

INTRODUCTION

Emotions, expressed in different forms, are an inevitable part of interpersonal communication. Emotions themselves may or may observe with the naked eye [1]. The emotional state helps in decision making, assists creativity, and manages human cognition and human-machine communication. The data captured is usually related to the measurement tool, such as the video's quality, lighting, pose and size of the face on the video, and noises in a voice recording. Therefore, any indications preceding or following them can be subject to detection and recognition with the right tools. In the past few years, the need to detect a person's emotions are increasing, and interest in human emotion recognition in various fields related has been increasing, but not limited to, human-computer interfaces, animation, medicine, security, diagnostics for Autism Spectrum Disorders (ASD) in children, and urban sound perception. Several features can perform emotion recognition, such as facial expressions, speech even text. Among these features, in 1967, Mehrabian showed that 55% of emotional information was visual, 38% vocal, and 7% verbal [2]. Face changes during communication are the first signs transmitting the emotional state, which is why most researchers found this modality very interesting.

Related Works

In a previous study conducted by Claudia Primasiwi, Handayani Tjandrasa, Dini A. Navastara about Face Expression Detection Using Gabor and Haar Wavelet Features. From the study results using the Gabor and Haar Wavelet Features method and using the CK+ dataset, the accuracy obtained using the proposed features, namely Gabor and Haar wavelets, is 92%. in contrast, the highest accuracy is obtaining the Gabor and Landmark features which produce 94.8%. For the lowest accuracy, obtained by using the Haar feature, which is 70% [3]. Another study conducted by Aditya Santoso, Gunawan Ariyanto proposed hard-based deep learning for facial recognition. The study used the convolutional neural network method using a complex library and the face 94 dataset. The data used took ten male subjects, with each subject having 20 face images. Thus 17 face images will become the training dataset, and the remaining three face images will be used for the testing dataset. The data obtained measurement results of 88.57%, where the size of the image affects the level of accuracy and time of data training. The larger the size of the image trained, the longer the learning process. The use of the number of layers in the training process also affects accuracy in data testing. The more layers used, the better the results obtained [4]. Another study was also conducted by Isha Talegaonkar and friends, proposing Real Time Facial Expression Recognition using Deep Learning. From the results of the study using the FER (Facial Expression Recognition) method and the FER-2013 dataset. From these data, the results of the test accuracy are 60.12% and the validation accuracy is 89.78% [5]. In addition, another study was conducted by Miyuka Nakamura, Jiangkun Wang, Sinchhean Phea and Abderazek Ben Abdallah about Comprehensive Study of Coronavirus Disease 2019 (COVID-19) Classification based on Deep Convolution Neural Networks, this study using four CNN architecture and using COVID-19 data collection. The diagnosis accuracy of abnormal (COVID-19 and pneumonia) is 97.18 to 99.34% for the current dataset. On the other hand, the correct classification accuracy of normal/healthy lung X-ray images is 67.09 to 71.37%. It shows that there is a false positive problem. After changing the optimizer of the VGG-16 model from Adam to SGD in an additional experiment, the training model was able to binary-classify with accuracy close to other training models [6].

Convolutional Neural Network

Convolutional Neural Network (CNN) is one of the deep learning methods that designed to process images (two-dimensional) data. CNN includes a feature extraction part and a classification part. [6] The feature extraction part consist of convolutional layers, pooling layers, and activation layers. Using softmax and fully connected layer extracted features are classified. CNN is widely used in image classification and speech recognition. [7]

Convolutional Layer

Convolutional Layer is the first layer which performs convolutional operations between filters and image as input of CNN architecture. [8] The image will be convoluted with a filter to extract features from the input image that is called the feature map. [7] The optimization of convolutional layer can be reached through the optimization of filter size, stride and zero padding. [8]

Rel-U Activation

Activation layer in CNN called as RelU, which is used to performs thresholding with zero value to the pixel value in image. All over pixel values that are less than 0 on the feature map will be made 0 with this activation. [8]

Pooling Layer

The pooling layer receives output from the convolutional layer, reducing computational complexity by performing non-linear down-sampling [9]. In principle, the pooling layer consists of filters with specified size and stride that shifted throughout the feature map area [8].

Fully Connected Layer

The fully connected layer processes the feature maps of the feature extraction layers. Unlike the convolutional layer, where the neurons are connected only into specific areas on the input, the fully connected layer connects the real neurons. The fully connected layer transforms multidimensional arrays of feature maps into a one-dimensional array (flatten process), classifying data linearly [10].

Softmax Activation

Softmax Classifier is another form of Logistic Regression that can classify more than two classes. Softmax is helpful to convert the output from the last layer to its primary probability distribution [11].

SYSTEM DESIGN

Propose System Design

AlexNet

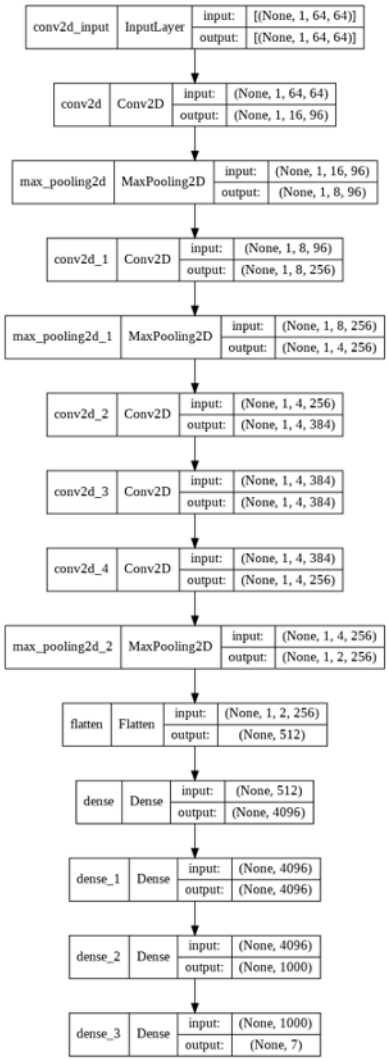

FIGURE 1. AlexNet Block Diagram

As shown in the figure, AlexNet consists of multiple sets of convolutional layers, pooling layers, and fully-connected layers stacked on top of each other. The job of the convolutional layers is to extract local features in the input images. A pooling typically follows a convolutional layer. On this layer, the image data size will reduce. Pooling layers also introduce translation invariance into the network. All the units are generally connected in the upper and final layers and thus are term fully-connected layers [12].

NNN-Net (3 Hidden-Layer)

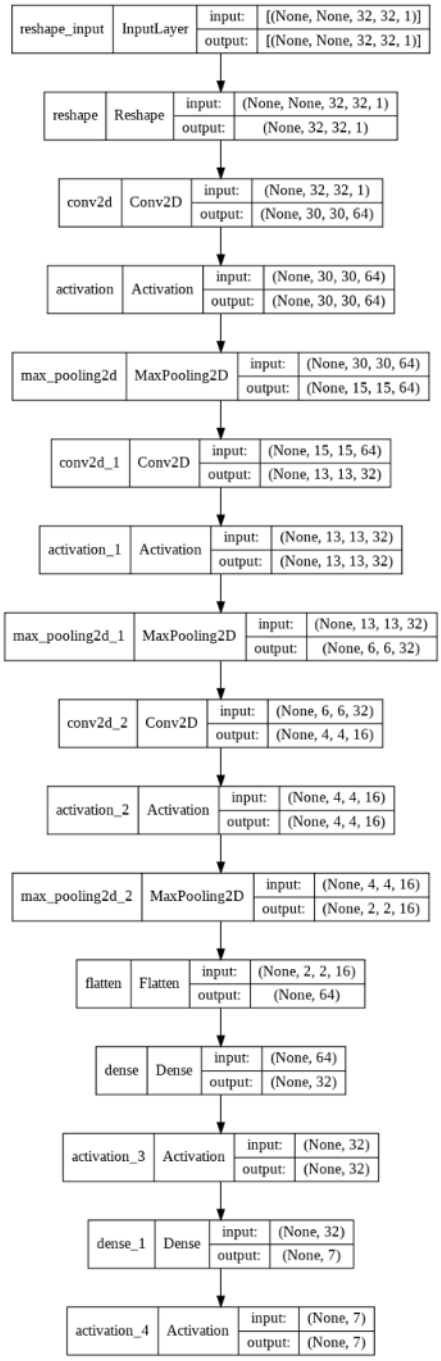

FIGURE 2. NNN-Net Block Diagram

In this study, to classify seven classes of expressions, the model consists of three hidden layers. Each hidden layer used Rel-U Activation to convert the negative value of feature maps to be zero and max-pooling, which will reduce the dimension of an image.

Evaluation Parameter

Accuracy and Confusion Matrix

TABLE 1. Confusion Matrix

Confusion Matrix	Actual A	Actual B	Actual C	Actual D	Actual E	Actual F	Actual G
Predicted A	x_{AA}	x_{AB}	x_{AC}	x_{AD}	x_{AE}	x_{AF}	x_{AG}
Predicted B	x_{BA}	x_{BB}	x_{BC}	x_{BD}	x_{BE}	x_{BF}	x_{BG}
Predicted C	x_{CA}	x_{CB}	x_{CC}	x_{CD}	x_{CE}	x_{CF}	x_{CG}
Predicted D	x_{DA}	x_{DB}	x_{DC}	x_{DD}	x_{DE}	x_{DF}	x_{DG}
Predicted E	x_{EA}	x_{EB}	x_{EC}	x_{ED}	x_{EE}	x_{EF}	x_{EG}
Predicted F	x_{FA}	x_{FB}	x_{FC}	x_{FD}	x_{FE}	x_{FF}	x_{FG}
Predicted G	x_{GA}	x_{GB}	x_{GC}	x_{GD}	x_{GE}	x_{GF}	x_{GG}
Predicted H	x_{HA}	x_{HB}	x_{HC}	x_{HD}	x_{HE}	x_{HF}	x_{HG}

The contents of the confusion matrix table are 4, namely:

1. True Positive (TP) is the number of correct predictions that an instance is negative,
2. True Negative (TN) is the number of incorrect predictions that an instance is positive,
3. False Negative (FN) the number of incorrect predictions that an instance is negative,
4. False Positive (FP) is the number of correct predictions that an instance is positive,

Accuracy is a comparison between the data that is predicted correctly with the predicted data as a whole [13]. The accuracy equation can be seen in Equation 1

$$Accuracy = \frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \quad (1)$$

Where n is number of classes, in this case the number of classes are seven.

Loss Function

Categorical Cross-Entropy (CCE) is the most commonly used loss function, loss given in Equation 2, which measures the difference between the probability distributions of one-hot encoded CNN computed class labels and ground truths [10].

$$H(p,q) = \sum_{i=1}^{n} = p(x_i) \log q(x_i), \quad Where\ \boldsymbol{x}_i \in \boldsymbol{X} \tag{2}$$

In Equation1, $q(x_i)$ and $p(x_i)$ represent the probability distributions of the one-hot encoded CNN predicted class labels and ground truths, respectively, for an input data vector $\boldsymbol{x}_i$.

DATASET DESCRIPTION

The provided CK+48 dataset consists of low-quality (48x48 pixels) labeled images. This data initially consists of 989 images (unbalance dataset). However, labeling has been done into seven classes (happy, sadness, anger, disgust, contempt, fear, and surprise).

Data Augmentation

Data augmentation is an essential tool to be implemented. Data augmentation is a technique to artificially increase the training set by adding transformations or perturbations without increasing the computational cost. Data augmentation techniques commonly used are flipping the images horizontally or vertically, crop, color jittering, scaling, and rotations in visual imagery and image classification applications. The new augmented dataset consists of 1750 images (250 for each class) [14].

According to standard practice, the dataset was split into training and validation sets. In this case, we use 75% data for training and 25% data for validation. Furthermore, training and validation images are separated randomly into disjoint sets. This challenge aims to automate the face expressions classification process to make it faster and more accurate. [15]

RESULT AND DISCUSSION

This study uses AlexNet and self-construct architecture with three hidden layers named NNN-Net as the model, with input size, optimizer, and learning rate as a parameter. In order to find the best model, we train the combinations between those parameters hierarchically, starting from training with different input sizes followed by optimizer and learning rate.

Input Size

TABLE 1. Input Image Size as Parameter for Result Comparison between NN-Net and AlexNet.

Input Size	Architecture	Accuracy	Loss
32x32	NNN-Net	0.9658	0.1487
	AlexNet	0.9636	0.1302
64x64	NNN-Net	0.9749	0.1369
	AlexNet	**0.9681**	0.1423
128x128	NNN-Net	0.9613	0.2492
	AlexNet	0.9704	0.134

The result for the classification task uses input size as a parameter listed in the figure. The best result for each input size appears in 64x64 for AlexNet with 0.9681 accuracies and 32x32 for NNN-Net with 0.9658 accuracies, considering computational time. Hence slightly different in accuracy consider as fine.

Optimizer

TABLE *3*. Input Image Size as Parameter for Result Comparison between NN-Net and AlexNet.

Input Size	Architecture	Accuracy	Loss
Adam	NNN-Net	0.9658	0.1487
	AlexNet	0.9681	0.1423
Nadam	NNN-Net	0.9681	0.1031
	AlexNet	0.9499	0.2356
SGD	NNN-Net	0.1731	1.9455
	AlexNet	0.1344	1.9459
RMSprop	NNN-Net	**0.9772**	0.081
	AlexNet	0.9636	0.158

The figure shows that we choose four optimizers: Adam, Nadam, RMSprop, and SGD. The best accuracy result on AlexNet is 0.9681 with Adam and for NNN-Net is 0.9772 with RMSprop. Overall, except for the accuracy of SGD, the average of other methods is above 0.94 to 0.97, which indicates that the results of our methods are solid. Furthermore, the overall result for Adam and RMSprop is good, proving Adam theoretically as one of the best optimizers.

Learning Rate

Input Size	Architecture	Accuracy	Loss
0.1	NNN-Net	0.1572	1.506
	AlexNet	0.1458	1.9476
0.01	NNN-Net	0.1481	1.9467
	AlexNet	0.1526	0.1582
0.001	NNN-Net	**0.9863**	0.0717
	AlexNet	0.1344	1.9464
0.0001	NNN-Net	0.9681	0.1407
	AlexNet	0.9681	0.1423

This study found that learning rate also plays an important part in accuracy. From the table, we can see that the best accuracy result for AlexNet is 0.9681 and 0.9863 for NNN-Net. This study found that smaller datasets and architectures require more total values for weight reduction, while more comprehensive datasets and deeper architectures seem to require smaller ones. Therefore, we hypothesize that complex data provides regularization, and other regularization should reduce.

Best Model

We found that the best result for combinations between parameters and architecture is NNN-Net with input size 32x32, Adam and learning rate 0.001. The graph below shows that the model is stable and reach 0.9863 in accuracy

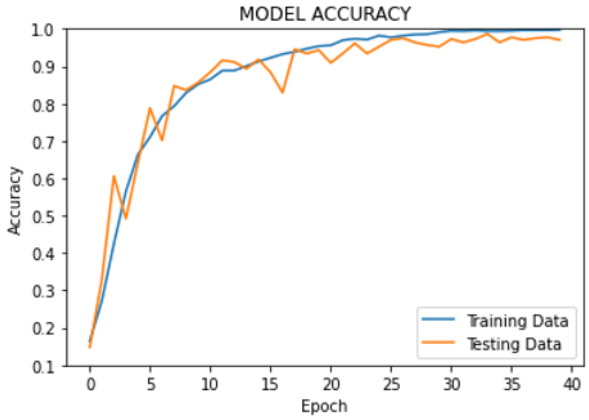

FIGURE 3. Graph of The Model Accuracy from The Best Hyperparameter Combination Scenario.

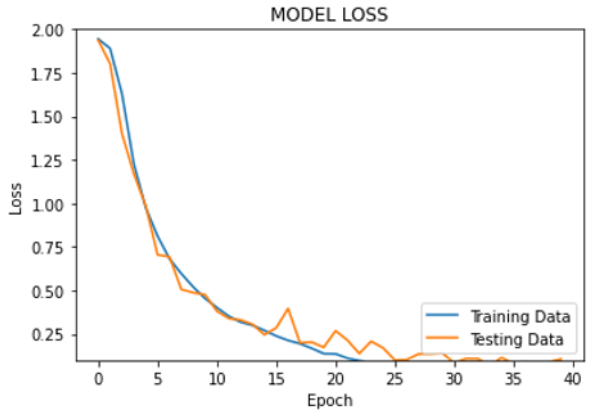

FIGURE 4. Graph of The Model Loss from The Best Hyperparameter Combination Scenario.

CONCLUSION AND FUTURE WORK

This paper applied CNN with two different architectures for facial expression recognition. Initially, we studied the fundamental structures of CNN models, and we improved the accuracy of those CNN models. Also, we calculated and compared the accuracy of each model. The two CNN models are AlexNet, and self-construct architecture with three hidden layers named NN-Net, and they examined on the same database, which is CK+48. Generally, the limited results we got are also suitable for other general situations. Finally, we found that NN-Net achieved the best overall accuracy, 0.9863. The results of our approach show that CNN can produce some beneficial results on CK+.

Since face expressions recognition is one of the critical parts of communications, the following study can continue to recognize students' expressions using online meeting platforms, for instance, zoom.

REFERENCES

1. S. Minaee, M. Minaei and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *mdpi,* vol. 21, 2021.
2. W. Mellouka and W. Handouzi, "Facial emotion recognition using deep learning: review and insights," *Procedia Computer Science,* vol. 175, no. 6, p. 690, 2020.
3. C. Primasiwi, H. Tjandrasa and D. Navastara, "Deteksi Ekspresi Wajah Menggunakan Fitur Gabor dan Haar Wavelet," vol. 7, no. 1, 2018.
4. A. Santoso and A. Gunawan, "Implementasi Deep Learning Berbasis Keras Untuk Pengenalan Wajah," vol. 18, no. 1, pp. 18-20, 2018.
5. I. Talegaonkar, K. Joshi, S. Valunj and K. Rucha, "Real Time Facial Expression Recognition using Deep Learning," in *Elsevier-SSRN*, Karvenagar, 2019.
6. M. Nakamura, J. Wang, S. Phea and A. Abdallah, "Comprehensive Study of Coronavirus Disease 2019 (COVID-19) Classification based on Deep Convolution Neural Networks," in *SHS Web of Conferences*, Aizu-Wakamatsu, 2021.
7. Y. N. Fu'adah, N. K. C. Pratiwi and N. Ibrahim, "Convolutional Neural Network (CNN) for Automatic Skin Cancer Classification System," *IOP Science,* vol. 2, no. 3, pp. 3-7, 2020.
8. Y. N. Fu'adah, N. K. C. Pratiwi, F. F. Taliningsih and S. Rizal, "Automated Classification of Alzheimer's Disease Based on MRI Image Processing using Convolutional Neural Network (CNN) with AlexNet Architecture," *IOP Science,* vol. 1844, no. 3, pp. 1-8, 2020.
9. V. Roy and S. Shukla, "Mth Order FIR Filtering for EEG Denoising Using Adaptive Recursive Least Squares Algorithm," in *IEEE*, Jabalpur, India, 2015.
10. S. Khan, H. Rahmani, S. Shah and M. Bennamoun, A Guide to Convolutional Neural Networks for Computer Vision, Morgan and Claypool Publishers, 2018.
11. J. Feng and S. Lu, "Performance Analysis of Various Activation Functions in Artificial Neural Networks," *Performance Analysis of Various Activation Functions in Artificial Neural Networks,* vol. 1237, no. 2, pp. 1-6, 2019.
12. W. Nawaz, S. Ahmed, A. Tahir and H. A. Khan, "Classification Of Breast Cancer Histology Images Using ALEXNET," in *ICIAR 2018: Image Analysis and Recognition*, vol. 10882, Springer, Cham, 2018, pp. 869-876.
13. A. Santra and C. J. Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering," *IJCSI International Journal of Computer Science,* vol. 9, no. 1, pp. 322-328, 2012.
14. P. Murugan, Implementation of Deep Convolutional Neural Network in Multi-class Categorical Image Classification, Singapore: Nanyang Technological University, 2018.