Convolutional Neural Networks Models for Facial Expression Recognition

Burhanudin Ramdhani, Esmeralda C. Djamal*, Ridwan Ilyas
Department of Informatics
Universitas Jenderal Achmad Yani
Cimahi, Indonesia
*correspondent email: esmeraldacd@yahoo.com

Abstract—**Emotion is a psychological representation of an event that arises spontaneously in a short time and can be reviewed one of them from facial expressions, which facial expressions can indicate consumer satisfaction. Facial recognition as an image can be viewed as identity, emotion, age, race, and gender. This is what makes the extraction of emotional patterns from other patterns is not easy. This research has built an image recognition system of emotion expression using Convolutional Neural Networks (CNN) by comparing two configurations using batch size 8 and 128 with two datasets that are FER-2013, self-created dataset and cross-dataset against four emotion expressions related to customer service that is happy, disappointed, angry and natural. The test results showed better results on the configuration made by researchers with batch size 8 which achieved the best results 73.98% on the dataset made by researchers and 58.25% in the FER-2013 dataset. Whereas when using batch size 128 best accuracy achieved by the previous research configuration with the FER-2013 dataset is 69.10%.**

Keywords—facial expression recognition; machine learning; deep learning; convolutional neural networks;

I. Introduction

Consumer satisfaction is one of the benchmarks of the suitability of a product or service with consumer expectations. Today many companies compete with each other to improve the quality of products and services one of them by conducting customer satisfaction surveys such as the spread of questionnaires, telephone surveys, to use the application of consumer satisfaction on the minimarket. But not every customer gives feedback to the survey. For consumers who are present at the place of service, customer satisfaction can be demonstrated by emotional expression. This is easy to do with the presence of CCTV near the service desk. But the identification and analysis of emotional expressions from the face automatically are not easy.

In general, the expression is divided into seven Happy, Shocked, Angry, Sad, Fearful, Disgusting and Natural [1]. Some emotional expressions related to customer service are Happy, Disappointed, Angry and Natural. Previous research for the facial expression recognition is using Region of Interest (ROI) and Multilayer Perceptron [2], Local Phase Quantization (LPQ) and SRC algorithm with the best accuracy 70.00% [3], Filter Gabor and Backpropagation with the best accuracy 84% [4], Active Appearance Models with the best accuracy on the angry expression that is 63.9% [5]. Other research using Local Binary Pattern get the best accuracy at 87% when combined with SVM [6]. Thus, the results of some previous studies the accuracy of each emotion are not the same, and the face is required not to use accessories such as eyeglasses [5].

The development of computing for image processing has led to the use of Deep Learning, i.e. machine learning that allows recognition with high accuracy and using images without being extracted first. Accuracy and limitations in machine learning such as Artificial Neural Networks, Support Vector Machine can be enhanced by the capabilities of these technologies but must be accompanied by high computing device specifications. The rapid development of the Graphical Processing Unit (GPU) now makes the use of Deep Learning realistic [7].

One variation of Deep Learning is Convolutional Neural Networks (CNN). CNN has a variety of configurations tailored to the dataset and the classification used. In recognition, CNN needs to be combined with classification methods such as Backpropagation [8] and Support Vector Machine (SVM) [9]. CNN can identify individuals, animals, road signs, vehicles and many other aspects of visual data [10].

Several previous research use CNN in the face detection [11], classification of face identity for authorization [12], ranging from gender identification based on face image [13], estimate age based on facial [14], detection of person movement [15], identification of vehicle type [16] to identification of plant species [17]. Earlier research identified emotional expression using CNN with two convolution layers with CK + dataset [18] and with dataset FER-2013 [19].

This research compared two CNN configurations with two dataset FER-2013 and dataset created by the researcher. Facial emotions are classified against the four classes of Happy, Disappointed, Angry and Natural.

II. Deep Learning Neural Networks

A. Convolutional Neural Networks

Deep learning has several variations: Convolution Neural Networks (CNN) and Recurrent Neural Networks (RNNs). CNN is more intended for classification of data that is not related to each other, While RNNs are usually used for related data where the past data or previous data used as a reference for the next output.

CNN is the development of an artificial neural network model that has multiple layers called Multi-Layer Perceptron (MLP). On CNN there is already a feature extraction process and

978-1-5386-5280-0/18/$31.00 ©2018 IEEE

classification process, feature extraction is obtained through a convolution process [20] and the classification process with dense layers. CNN has three-dimensional neurons of height, width, and depth which refers to the number of layers. Can be seen CNN architecture in Fig.1, CNN has two main layers: the feature extraction layer and the classification layer. The feature extraction layer serves to extract features by converting and then activate with Rectified Linear Units (ReLU) to Max Pooling. The classification layer for learning that the input is obtained from the feature extraction layer.

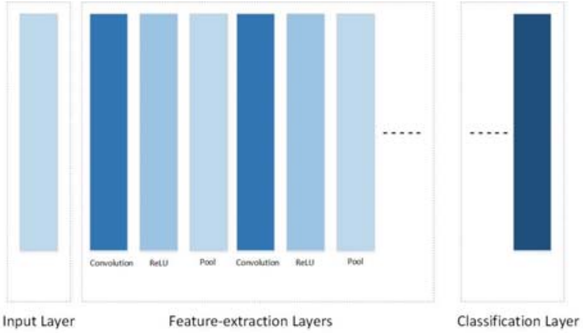


Fig. 1. Architecture of CNN

1) *Convolution*

The convolution layer is the first layer of the feature extraction layer, the operation on this convolution layer can be seen in (1).

$$FM_{(i_l,j_l)}^{(l,m_l)} = f\left(\sum_{r_l=0}^{k_h} \sum_{c_l=0}^{k_w} C_{(r_l,c_l)}^{(l,m_l)} * FM_{((r_l+i_{l-1}),(c_l+j_{l-1}))}^{(l-1)} \right) \qquad (1)$$

Where r_l and c_l are index length in the convolution with (r-l) overlap. This layer works to convert the image to different depths and extract the image by filtering (kernel), then shifting according to the Stride value. This Stride is the value that determines the filter shift (kernel) and to adjust the output dimensions of the convolution. Padding is the value that determines the number of pixels (containing the zero value) to be added on each side of the input. The convolution process can be seen in Fig.2 with (1).

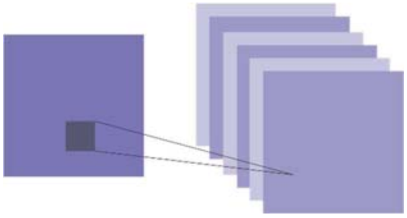


Fig. 2. Convolution Process

To calculate the size of the output of the convolution layer can be seen in (2), where W is the input height, N is the filter height, P is padding and S is the stride.

$$output = \frac{W - N + 2P}{S} \qquad (2)$$

$$\alpha + \beta = \chi.$$

2) *Activation*

The activation layer after convolution uses ReLU which has (3), x is the input value. This activation process aims to eliminate the negative value or in other words using a threshold of 0 to infinity.

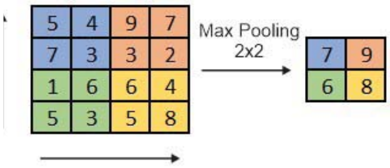
$$f(x) = \max(0.x) \qquad (3)$$

Unlike the Rectified Linear Unit (ReLU) activation, Softmax activation is used in the output layer to represent the category distribution. This Eq. of Softmax can be seen in (4), where w is the weight of the output layer.

$$y_j = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \qquad (4)$$

3) *Max Pooling*

Max Pooling layer serves to reduce the spatial size and number of parameters in the network and accelerate computing, controlling the occurrence of overfitting and generate patterns of features [21]. Max Pooling this layer can be shown in Fig.3. In this Max Pooling process, take the largest value of input layer from the result of the activation function.



5	4	9	7
7	3	3	2
1	6	6	4
5	3	5	8

7	9
6	8

Fig. 3. Max Pooling Operation

4) *Loss Function*

Loss Function or objective function or optimization score function) is one of the two parameters required. Some loss functions can be used, one of them Cross Entropy as in (5). Where D distance, Output of Softmax S and L label.

$$D(S, L) = -\sum_i L_i \log(S_i) \qquad (5)$$

5) *Dropout*

Dropout is a function to reduce overfitting during the learning process by eliminating neurons from the network layer, either the input to neurons and neurons output, the process before and after the dropout can be seen in Fig.4 [22]. The number of neurons removed according to the parameters and the temporarily removed neurons is randomly selected.

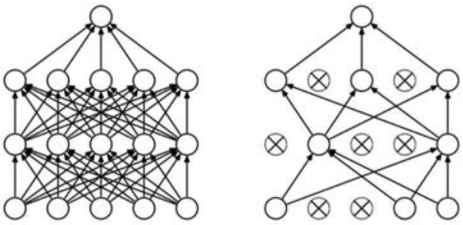


Fig. 4. Dropout Process

6) *Library*

Classification code uses Keras library CNN (Tensorflow) [23]. Keras has optimization to update the model parameters. This configuration uses Adadelta optimization [24].

III. Dataset

In this research, two datasets that are FER-2013 [19] and self-created dataset are used. The FER-2013 dataset has a 48x48 grayscale image size that contains seven expressions but later that will be trained and tested only four expressions Happy, Disappointed, Angry and Natural. The FER-2013 dataset can be seen in Fig.5. In FER-2013 dataset, there are some data that have no face image inside [19]. We removed non-face data and reconstructed images from the CSV (default) format into .png format as in Fig. 6. Non face data.

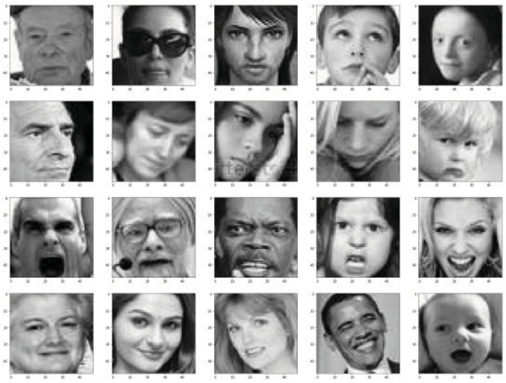


Fig. 5. Dataset FER-2013

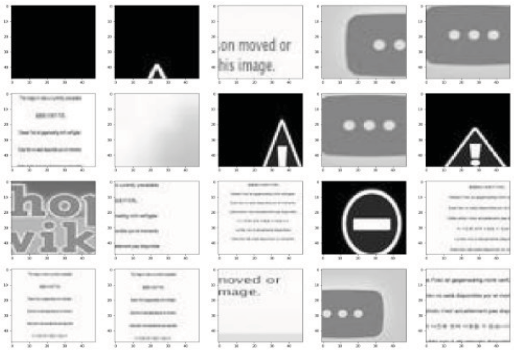


Fig. 6. The faceless image on the FER-2013 dataset

The self-created dataset consists of four expressions: Happy, Disappointed, Angry and Natural, which is the data taken from 15 objects. Each expression is taken 10 times. The self-created dataset that has been pre-processed can be seen in Fig.7.

The input of CNN is required to be consistent, hence the required pre-process. The image then goes into the pre-processing face detection using OpenCV [25] with the *haarcascade_frontalface_default.xml* classifier. After face detection then the image is a segmented only face, then resized to 48x48 and made grayscale with Luminosity Equation [26]. After going through the pre-process then entering into CNN, the first layer passed is the feature extraction then the classification layer. The process diagram can be seen in Fig. 8.

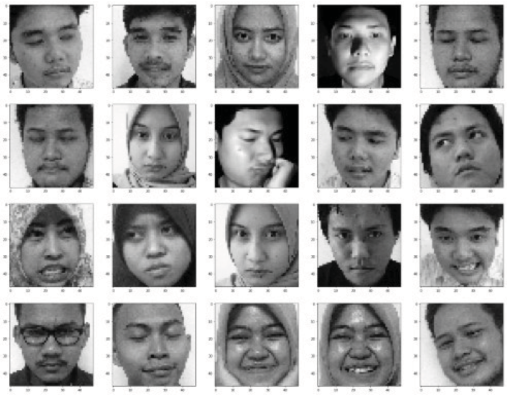


Fig. 7. The self-created dataset

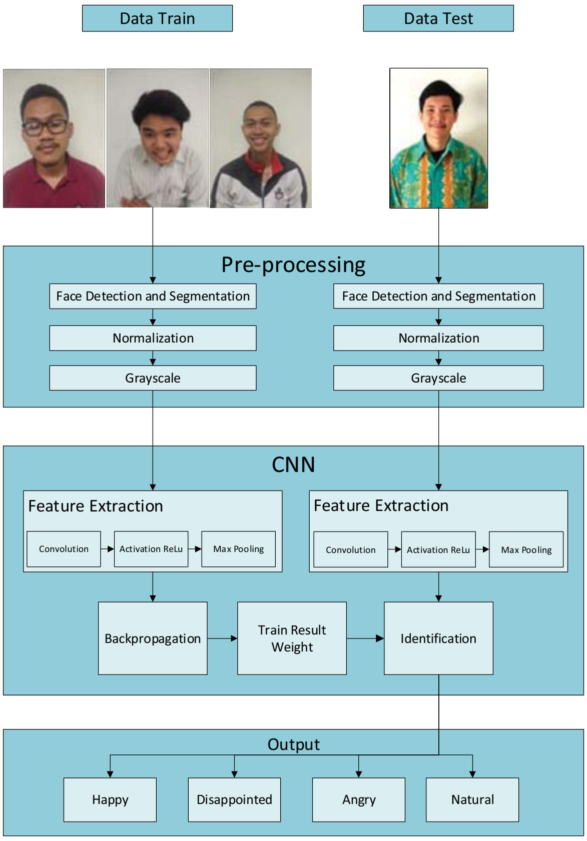


Fig. 8. Process Diagram

IV. Comparison Configurations for Facial Expression Recognitions

This research used two configurations of CNN. Each configuration was tested on two sets of data, that is self-created dataset and FER-2013 [27] and cross-dataset. The configuration created by the researcher used 9 layers in which there are 3 convolution layers, 3 layers max pooling, 2 fully connected layers, and 1 output layer. The configuration made by this researcher has 125,764 parameters. In the previous research configuration using 10 layers in which there are 3 layers convolution, 1 layer average pooling, 2 layers max pooling, 3 layers fully connected and 1 layer output. This configuration has 60,388 parameters as shown in Table I.

TABLE I. COMPARATION CONFIGURATION CNN

Layer (type)	**Proposed Config**		**Previous Research Config**	
	Kernel / Units	***Size / Dropout Probability***	***Kernel / Units***	***Size / Dropout Probability***
Convolution	32	3x3	10	5x5
Average Pooling	-	-	N/A	2x2
Max Pooling	N/A	2x2	-	-
Convolution	64	3x3	10	5x5
Max Pooling	N/A	2x2	N/A	2x2
Convolution	128	3x3	10	3x3
Max Pooling	N/A	2x2	N/A	2x2
Flattern	N/A	P = 0.25	N/A	-
Fully Connected	64	P = 0.5	256	P = 0.5
Fully Connected	4	-	128	P = 0.5
Fully Connected	-	-	4	-

V. EXPERIMENT RESULT AND DISCUSSION

In each test performed as much as 150 epoch with batch size 8 and 128. Testing the two configurations, each set of data is divided into two, namely data train and test data. The training data is taken randomly as much as 80% of the total of all data, and test data took the remaining 20%. The amount of data to be trained and tested each dataset is different. The data can be seen in Table II.

TABLE II. AMOUNT OF DATA

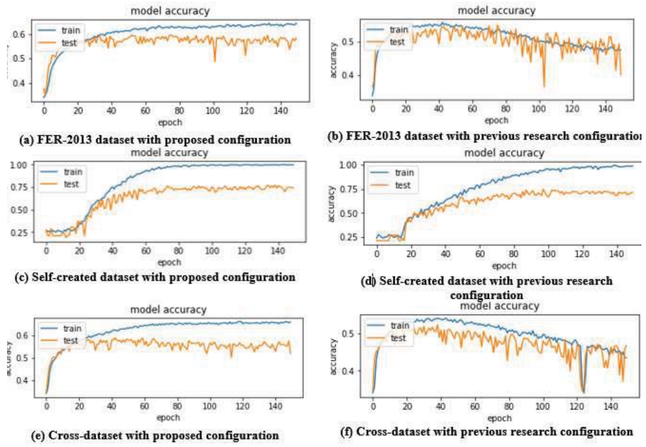
Dataset	**Train**	**Test**
FER-2013	16.746	4.187
Self-created	123.000	489.000
Cross	17.236	4.309

The first test with the batch size 8 the greatest accuracy of the test data was achieved by the dataset of researchers that is 73.98% and 71.54% used previous research datasets, both of which used configurations made by researchers. Test results with batch size 8 can be seen in Table III. The accuracy of test data and training data using the configuration created by researchers always outperform previous research configuration.

TABLE III. THE RESULT OF ACCURACY (%) WITH BATCH SIZE 8

Dataset	**Train**		**Test**	
	Proposed Config	***Previous Research Config***	***Proposed Config***	***Previous Research Config***
FER-2013	68.30	40.34	58.20	40.02
Self-created	100.00	100.00	73.98	71.54
Cross	67.76	48.36	51.98	46.78

The results of the graph using the batch size 8 can be seen in Fig. 9. When the training uses a configuration made by the researchers either using the FER-2013 dataset, the researcher dataset and cross datasets have good accuracy with the training data chart continues to rise while test data tend not fluctuate. The results using the configuration made by researchers can be seen in Fig. 9 parts a, c and e. Different when using the previous research configuration, FER-2013 dataset and cross-dataset have a fluctuating graph and tend to go down but not on a dataset of researchers who tend to rise. In this research using the average pooling parameter with batch 8 size is not suitable for large dataset.



(a) FER-2013 dataset with proposed configuration

(b) FER-2013 dataset with previous research configuratioı

(c) Self-created dataset with proposed configuration

(d) Self-created dataset with previous research configuration

(e) Cross-dataset with proposed configuration

(f) Cross-dataset with previous research configuration

Fig. 9. The result of the test graph with batch size 8

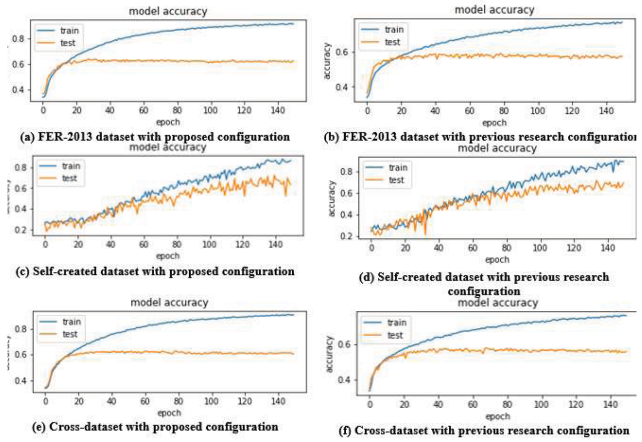
The second test with 128 batch size of the highest accuracy of the test data was achieved by the research dataset that is 69.10% using the previous research configuration. Then followed by 63,41% with a dataset of researcher using configuration made by the researcher. Test results with batch size 128 can be seen in Table IV, in that table, the lowest accuracy of the test data is 55.95% in the cross-dataset with the previous research configuration.

TABLE IV. THE RESULT OF ACCURACY (%) WITH BATCH SIZE 128

Dataset	**Train**		**Test**	
	Proposed Config	***Previous Research Config***	***Propose Config***	***Previous Research Config***
FER-2013	99.64	86.54	62.33	57.41
self-created	92.84	93.25	63.41	69.10
Cross	99.19	84.06	60.50	55.95

Graphical results using batch size 128 can be seen in Fig. 10. When the training of previous research dataset and cross-dataset using both configurations are the configurations made by researchers, and previous research configuration showed good results, the training data graphic continues up to the end of the epoch and graph of convergent test data between 55% and 60%, the graph can be seen in Fig.10 parts a, b, e, and f. The researcher dataset when tested the graph of training data and test data tend to fluctuate but keep going up to until the last epoch in Fig.10 part c and d, this is because the researcher dataset has a small amount of data making the 128 batch size less suitable for the small dataset in this test.

The best confusion matrix of each batch size test 8 and 128 can be seen in Fig. 11, in the picture a is the best confusion matrix with batch size 8 is with the dataset of the researcher using the configuration made by the researcher and part b is the best confusion matrix with batch size 128 is with the dataset of researchers using the previous research configuration.



(a) FER-2013 dataset with proposed configuration

(b) FER-2013 dataset with previous research configuratio

(c) Self-created dataset with proposed configuration

(d) Self-created dataset with previous research configuration

(e) Cross-dataset with proposed configuration

(f) Cross-dataset with previous research configuration

Fig. 10. The result of the test graph with batch size 128

In Fig. 11 the confusion matrix of section a, seven expressions of disappointed and five expressions of anger are recognized as a natural expression, and six expressions of disappointed are identified as angry expressions. Whereas in Fig. 11 part b eight expressions of disappointed are recognized as natural expressions, and 11 expressions of anger are recognized as expressions of disappointment. The happy expression of Fig. 11 always outperforms other expressions.

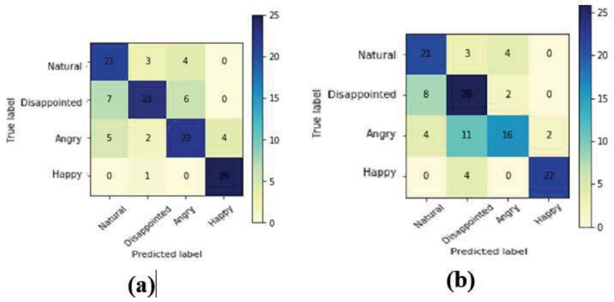


Fig. 11. Best confusion matrix with batch size 8 and 128

VI. CONCLUSION

This research gave that using batch size 8 best accuracy obtained on a dataset of researcher using researcher configuration, while the best accuracy when using batch size 128 got on set of data of researcher using previous research configuration. In each test, the configuration created by the researcher always outperformed the configuration from previous research whether it is training data and test data.

Testing of previous research datasets has never been better than self-created datasets because the previous research dataset FER-2013 has a higher complexity such as a side-facing face, not only the data of faces but also the hands, hats and other objects and different contrast.

In batch size 8, the dataset of previous research and cross-dataset using the previous research configuration, in which the accuracy graph of the training data and the test data fluctuates and tends to fall up to end of the epoch. Then the test using batch size 128, on the dataset of the researcher and cross-dataset with both configurations showed good results but not on the dataset of researchers, due to the number of sets of research data that number a bit, i.e., 612.

The test results of the accuracy of each emotional expression are not the same, but the happy expression always outperforms the other three expressions because everyone expresses happy almost the same unlike the case with other expressions.

REFERENCES

[1] [J. A. Russell and J. M. Fernandez-Dolz, *What does facial expression mean*. The Press Syndicate of The University of Cambridge, 1997.

[2] [M. Noviani and E. C. Djamal, “Identifikasi Kondisi Emosional Berdasarkan Citra Wajah Menggunakan ROI dan Multi Layer Perceptron,” in *Seminar Nasional Ipteks Jenderal Achmad Yani*, 2015, pp. 289–293.

[3] W. Zhen and Y. Zilu, “Facial Expression Recognition Based on Local Phase Quantization and Sparse Representation,” *8th International Conference Natural Computation.*, pp. 222–225, 2012.

[4] P. M. Rahardjo, “Pengenalan Ekspresi Wajah berbasis Filter Gabor dan Backpropagation Neural Network,” *ECCIS*, vol. IV, no. 1, pp. 12–17, 2010.

[5] M. S. Ratliff and E. Patterson, “Emotion Recognition using Facial Expressions with Active Appearance Models,” *Hum. Comput. Interact. (IASTED-HCI 2008)*, pp. 138–143, 2008.

[6] C. Shan, S. Gong, and P. W. McOwan, “Robust facial expression recognition using local binary patterns,” *IEEE Int. Conf. Image Process. 2005*, p. II-370, 2005.

[7] J. Schmidhuber, “Deep Learning in neural networks: An overview,” *Elsevier Neural Networks*, vol. 61, pp. 85–117, 2015.

[8] E. Zangeneh, M. Rahmati, and Y. Mohsenzadeh, “Low Resolution Face Recognition Using a Two-Branch Deep Convolutional Neural Network Architecture,” pp. 1–11, 2017.

[9] Darmatasia and M. I. Fanany, “Handwriting Recognition on Form Document Using Convolutional Neural Network and Support Vector Machines (CNN-SVM),” *2017 Fifth Int. Conf. Inf. Commun. Technol.*, vol. 0, no. c, pp. 1–6, 2017.

[10] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, “Object detection networks on convolutional feature maps,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1476–1481, 2017.

[11] S. Yang, P. Luo, C. C. Loy, and X. Tang, “From Facial Parts Responses to Face Detection: A Deep Learning Approach,” *2015 IEEE Int. Conf. Comput. Vis.*, no. 3, pp. 3676–3684, 2015.

[12] G. Ramadhan, E. C. Djamal, and T. Darmanto, “Klasifikasi Identitas Wajah Untuk Otorisasi Menggunakan Deteksi Tepi dan LVQ,” *Seminar. Nasonal. Aplikasi Teknologi. Informasi. 2016 Yogyakarta*, pp. 37–41, 2016.

[13] D. Wulansari, E. C. Djamal, and R. Ilyas, “Identifikasi Gender Berdasarkan Citra Wajah Menggunakan Deteksi Tepi dan Backpropagation,” *Seminar. Nasonal. Aplikasi Teknologi. Informasi 2017*, pp. 10–14, 2017.

[14] T. Zheng, W. Deng, and J. Hu, “Age Estimation Guided Convolutional Neural Network for Age-Invariant Face Recognition,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 503–511, 2017.

[15] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Human Behavior Untersthanding,” in *International Workshop on Human Behavior Understanding*, 2011, vol. 7065, no. November 2011, pp. 29–39.

[16] Z. Dong, M. Pei, Y. He, T. Liu, Y. Dong, and Y. Jia, “Vehicle Type Classification Using Unsupervised Convolutional Neural Network,” *2014 22nd Int. Conf. Pattern Recognit.*, vol. 11, no. 4, pp. 172–177, 2014.

[17] S. T. Hang and M. Aono, “Open world plant image identification based on convolutional neural network,” *2* 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2016.

[18] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, “Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order,” *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[19] A. Mollahosseini, B. Hassani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor, “Facial Expression Recognition from World Wild Web,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 1509–1516.

[20] A. Gibson and J. Patterson, *Deep learning*. O’Reilly Media, 2017.

[21] M. Zufar, “Convolutional Neural Networks untuk Pengenalan Wajah Secara Real - Time,” *Journal Sains dan Seni ITS*, vol. 5, no. 2, pp. 72–77, 2016.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[23] F. Chollet, *Deep Learning with Python*. 2018.

[24] S. Ruder, “An overview of gradient descent optimization algorithms,” *Inspire*, pp. 1–14, 2016.

[25] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 2001, vol. 1, p. I-511-I-518.

[26] V. Sireesha and K. Sandhyarani, “Iris Recognition Using Combined Feature Vector,” *Int. J. Adv. Res. Comput. Sci.*, vol. 7, no. 4, 2016.

[27] P. R. Dachapally, “Facial Emotion Detection Using Convolutional Neural Networks and Representational Autoencoder Units,” 2017.