Improving the Capability of Real-Time Face Masked Recognition using Cosine Distance

Devira Anggi Maharani
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
deviraanggi@students.itb.ac.id

Carmadi Machbub
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesi
carmadi@lskk.ee.itb.ac.id

Pranoto Hidaya Rusmin
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
pranoto@lskk.ee.itb.ac.id

Lenni Yulianti
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
lenniyulianti77@yahoo.com

2020 6th International Conference on Interactive Digital Media (ICIDM) | 978-1-7281-4928-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICIDM51048.2020.9339677

***Abstract*—During the pandemic, people around the world are expected to wear masks. So far, the ability to recognize the identity of a person who wears a mask is still a challenge. Face recognition is widely used in schools, hospitals, and companies as an attendance system, even as a criminal watchlist examination system. Thus, face recognition implementation is difficult to obtain the identity of people who are wearing masks, and moreover, computer systems might fail to detect the faces. This study used Haar-cascade Face detection and MobileNet while proposing the addition of the cosine distance method. This method compares the middle position of face detection results within the previous frame and the current. The proposed system can generate a person's name and identification number while wearing a mask. The system is designed to utilize multi-threading by comparing the transfer learning methods of VGG16 and Triplet Loss FaceNet for face mask recognition with an accuracy rate of 100% and 82.20%. Real-time implementation speed resulted in 4 FPS and 22 FPS and successfully added cosine distance to generate a person's ID number.**

Keywords—face mask recognition, FaceNet, MobileNet, VGG16, multi-threading, real-time, cosine distance

I. INTRODUCTION

The technology for computer vision is rising rapidly today. Many scientists are competing during the pandemic to improve touchless technologies such as face recognition. Face recognition is commonly used as an attendance scheme in schools, hospitals, and businesses. So far, it is always a struggle to be able to identify the face of a person wearing a mask.

There have been many studies on recognizing whether someone is wearing a mask or not, such as research [1] by applying Principal Component Analysis (PCA) and research [2] performing classification with YOLO V3. Several studies have been published that recognize the person's identity even if they wear a mask, as in the research in [3] which used the Convolutional Neural Network for recognition, and [4] which used SVM classifiers with face recognition feature vectors. The effect of using masks for face recognition is very sensitive [5]. The face recognition implementation is difficult to obtain the identity of a person who is wearing masks due to computer systems might fail to detect the faces because of missing some features. Several studies have developed ways so that the system can always detect objects correctly, even if occlusion occurred, as in [6], which utilized centroid tracking, [7] employed the Particle Filter, and Kalman Filter estimation algorithm [8]. In [9], centroid location was used to detect a moving target in frame t, compared to location $t-1$ with the greedy technique.

This paper proposed a method to increase the accuracy of the face mask recognition process and detection with the allocation of a person's ID number, using the cosine distance technique. In the detection process, Haar-cascade and MobileNet were implemented to obtain the face bounding box. Subsequently, the VGG16 transfer learning method and the Triplet loss FaceNet with multi-threading technique were compared in the face recognition process. The performance of various combinations of methods was compared, and the most suitable were selected.

II. TRANSFER LEARNING VGG16

A deep CNN model usually required a long time to do training, especially with an extensive dataset. One technique to reduce the time-consuming training process is to use the developed, trained model weights. According to Sinno et al. [10], knowledge transfer can improve the learning performance to avoid data labeling effort with less training image data [11]. Thus far, transfer learning is used in the regression, grouping, and classification process, as shown in Fig. 1.

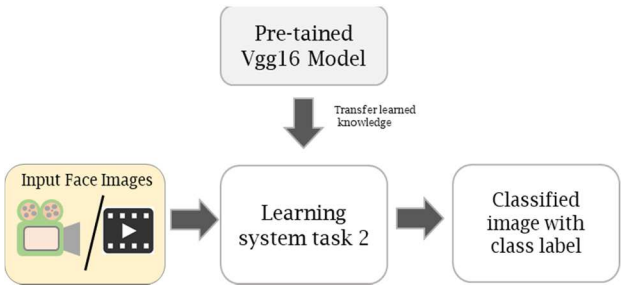

Fig. 1. Transfer learning process.

The transfer learning process takes adaptation of input domains, multitasks learning. VGG16 [12] by Karen et al.

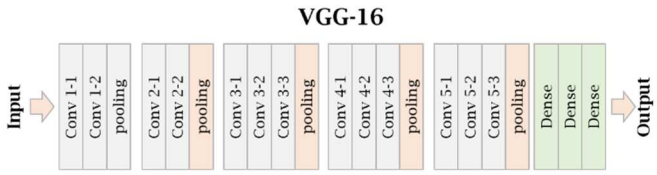

Fig. 2. VGG-16 architecture.

VGG16 uses training data from the ImageNet database [13], which can classify 1000 targets. The input size is 224x224x3 with pooling layers, 13 Conv layers, one activation layer, two drop-out layers, and three fully-connected layers.

III. Triplet Loss FaceNet

FaceNet model is a sophisticated face recognition model [14] by constructing face embedding based on triplet loss. They also experimented with [15] where the L1-distance between two face characteristics is directly optimized. FaceNet was first published in 2015 by Google researcher Schroof et al. The FaceNet model employs deep neural networks to extract face features. Input from FaceNet is an essential feature compressed into vector 128 or commonly known as embedding. FaceNet's structure, as shown in Fig. 2, where it is composed of batches as input layers and deep CNN + L2 normalization to produce face embedding and triplet loss is carried out during the training process.

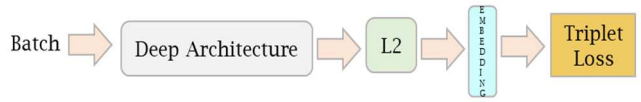

Fig. 3. Triplet loss FaceNet model.

Fig. 3 shows Triplet Loss, a learning method with anchor, positive, and negative data. Randomly, the data image will be selected as the anchor image, with the same person's image selected as positive data, a different person from the anchor will be selected as negative data. Then set with FaceNet parameter so that the positive data is closer to the anchor, and the distance is closer than the negative data.

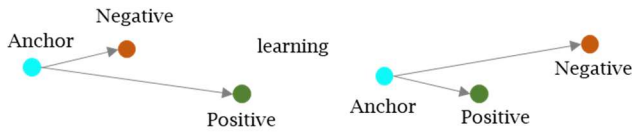

Fig. 4. Triplet loss which gets the minimum distance between the anchor and a positive.

In Fig. 4, we want the equation:

$$\|f(x_i^a) - (x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - (x_i^n)\|_2^2 \quad (1)$$

$$\forall \left(f(x_i^a), f(x_i^p), f(x_i^n)\right) \epsilon\, T \quad (2)$$

Where α is the margin between a positive and a negative pair. T is the set of all possible triplets in training and has cardinality N which makes a minimum loss.

$$L = \sum_i^N \left[\|f(x_i^a) - (x_i^p)\|_2^2 + \alpha - \|f(x_i^a) - (x_i^n)\|_2^2\right] \quad (3)$$

This triplet does not affect convergence and training due to it will be transmitted over the network. This triplet is controlling the system model to improve.

IV. The Proposed Method

In this study, Tensorflow, Scipy, and Opencv are used for calculations [16], [17], [18]. This study enhances the centroid tracker, where the center location of the bounding box (bb) face detection result in frame t will be compared with bb in frame $t-1$. Finding the closest and the most minimal distance between the center point in t and $t-1$, can use the Euclidean distance (4), Cosine distance (5), Correlation (6), and Chebyshev distance (7) which $\underline{x}$ and $\underline{y}$ are the mean values of the x and y vectors.

$$d(x,y) = \sqrt{(x-y)^2} \quad (4)$$

$$d(x,y) = 1 - \frac{x.y}{\|x\|\|y\|} \quad (5)$$

$$d(x,y) = 1 - \frac{(x-\underline{x})(y-\underline{y})}{\|x-\underline{x}\|\|y-\underline{y}\|} \quad (6)$$

$$d(x,y) = |x-y| \quad (7)$$

Fig. 5 shows the flowchart of this research's proposed methods, where the system will perform face detection using Haar-cascade [19] and MobileNet [20]. This face detection speed will be compared with the NVIDIA GEFORCE 940MX and Intel Core i5 computers. The steps for recognizing the person identity who wears masks as in Fig. 5.

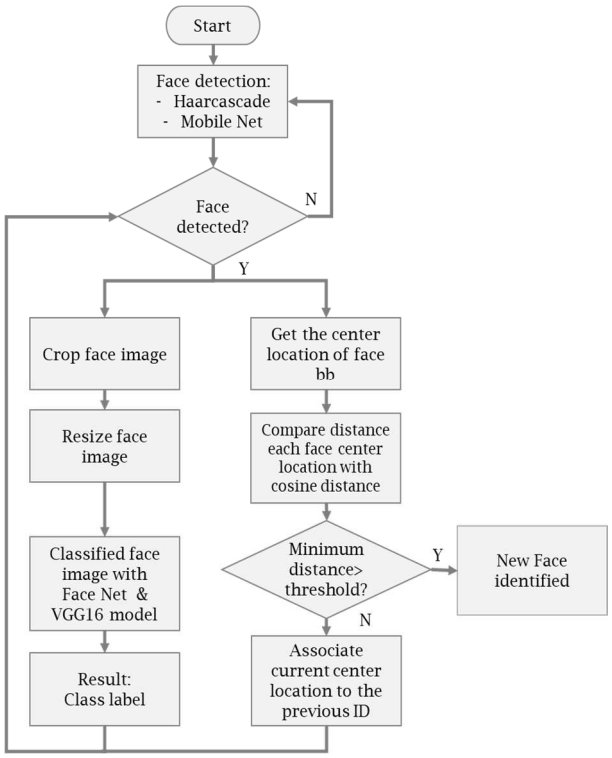

Fig. 5. Flowchart proposed architecture.

A. Preprocessing

We designed this face mask recognition implementation with three classes. We prepared 200 datasets in each class as in Fig. 6. One hundred pictures for a face mask and one hundred pictures for an unmask. In this preprocessing step,

face detection results will be reduced to 224x224 RGB images for VGG16 and 96x96 for Triplet loss FaceNet with split ratio data testing and training 20:80. Then for Triplet loss FaceNet, we compute the face image encoding into a 128-vector.

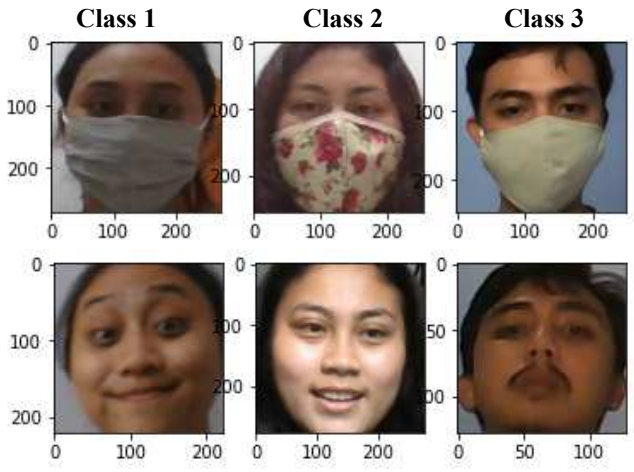

Fig. 6. Training dataset with three classes.

B. Training Process

With transfer learning, we tried to accomplish the last convolutional block of the VGG16 model. Fine-tuning starts with a trained network, then re-training using a new dataset and a very small weight with epoch 14. Later, the training model result will be stored in file.h5.

FaceNet uses 128-neuron fully connected layer as the last layer. It uses an encoding that the three face images are compared. Triplet loss The FaceNet here is trained by expecting the least triplet loss. The model is then saved file.h5 and used to recognize a person who wears a mask with calculating the encoding of the stored database image with the new face captured by the camera.

C. Centroid Tracking

Based on several papers about centroid tracking [21][6], this paper compares the most efficient, precise, and short distance method because this proposed system also considers real-time implementation.

V. Result and Discussion

There are several performances to be evaluated for the identity recognition of individuals wearing a mask, such as an accuracy rate of the training and testing process between Triplet loss FaceNet and VGG16, the FPS of multi-threaded face recognition implementation, and the impact of various centroid distances on speed computation and accuracy to assign id number.

A. Training and Testing Result

In this proposed system, we used the 80:20 ratio for training and testing data. The accuracy result can be seen in Table I.

TABLE I. Accuracy of the Training Process for Face Mask Recognition

	Accuracy	Val accuracy	Loss	Val_loss
Triplet loss FaceNet	82.20%	78.08%	1.12	1.16
VGG16	100%	100%	0.08	0.07

As in Table I, it can be concluded that the VGG16 transfer learning training process achieves a training accuracy and validity of up to 100%, while the triplet loss FaceNet reaches 82.20% and 78.08%. For Triplet loss FaceNet, this could be due to a lack of training data with epoch 14. The plot of the accuracy results is shown in Fig. 7.

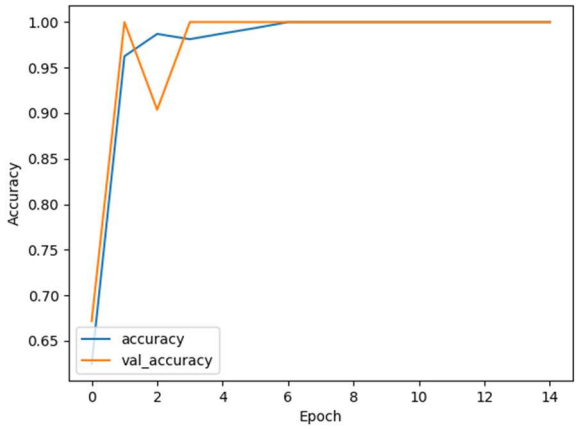

Fig. 7. Training accuracy (VGG16) on face masked recognition dataset.

Then, the results of loss training and validation of VGG16 are shown in Fig. 8.

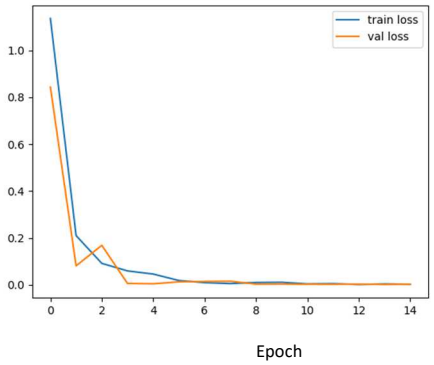

Fig. 8. Losses (VGG16) on face masked recognition dataset.

The accuracy and loss results of the triplet loss FaceNet shows in Fig. 9 and Fig. 10.

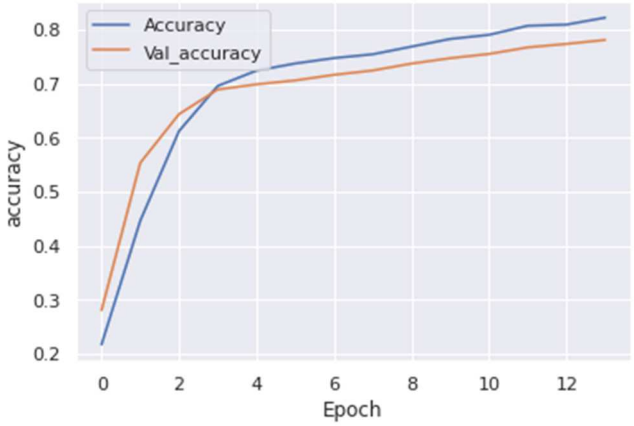

Fig. 9. Training accuracy (Triplet loss FaceNet) on face masked recognition dataset.

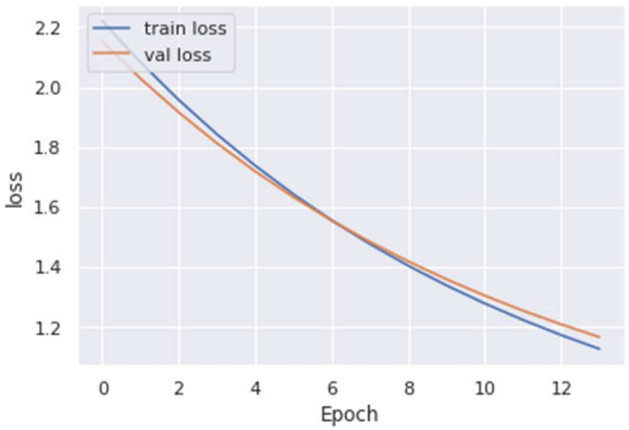

Fig. 10. Losses result (Triplet loss FaceNet) on face masked recognition dataset.

When the training process had been completed, the model was saved and used for the recognition process in .h5 format. The confusion matrix results with 20 random samples for Triplet loss FaceNet and VGG16 shows in Table II and Table III.

TABLE II. THE CONFUSION MATRIX (TRIPLET LOSS FACENET) WITH 20 RANDOM DATASETS EACH CLASS

Actual\Pred	Id 1	Id 2	Id 3
Id 1	14	4	2
Id 2	4	12	4
Id 3	6	3	15

TABLE III. THE CONFUSION MATRIX (VGG16) WITH 20 RANDOM DATASETS EACH CLASS

Actual\Pred	Id 1	Id 2	Id 3
Id 1	6	10	4
Id 2	4	16	0
Id 3	2	14	4

B. Face Recognition in Real-Time Implementation

When the training and testing process was done, the model would be stored for real-time recognition. In this study, two face detection methods had been compared. Before we recognize the person, the proposed system should proceed with the face detector first. This research attempts to add a centroid tracker by comparing the current and previous bb face position with cosine distance. The time required results show in Table V. The cosine distance calculation is faster than others.

Table IV shows the frame rate in FPS (Frames per Second). Face detection with Haar-cascade is faster than MobileNet; however, the detection process's accuracy is lower. It is a tradeoff between accuracy and frame rate. Implementation of image streaming in Fig. 11 with Queue and multi-threading shows that the system can increase the FPS, as shown in Table IV.

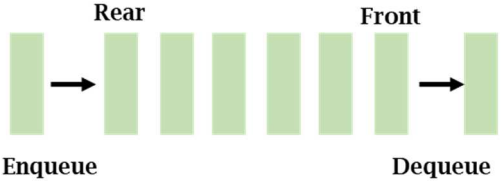

Fig. 11. FIFO (First in First Out) illustration for image stream [22].

The face detection process with the Haar-cascade method achieved 40 FPS while MobileNet reached up to 14 FPS. Table IV shows the time required for the recognition system with CPU. The computing of preprocessing of VGG16, which only changes the face images' dimensions, takes a longer time than the triplet loss FaceNet. The triplet loss FaceNet converts face images into 128-vectors, which could improve the face recognition speed.

Fig. 12 shows the detection and recognition results with the Triplet loss FaceNet + Haarcascade, which added a centroid tracker to obtain the identity number.

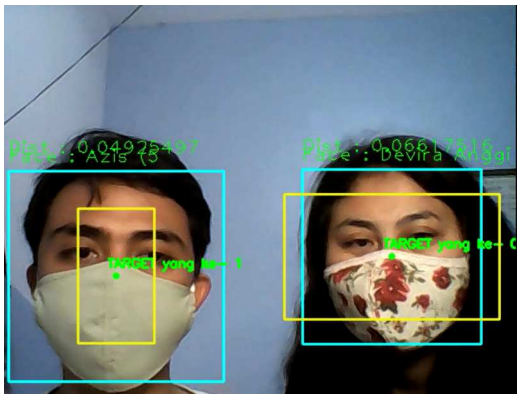

Fig. 12. The real-time face mask recognition.

TABLE IV. COMPUTATION SPEED WITH MULTI-THREADING

Face recognition method	Face detection method	Accuracy of face recognition	FPS result without multi-threading	FPS result with multi-threading
Triplet loss FaceNet	Haar-cascade	82.20%	16	22
VGG16	Haar-cascade	100%	3	4
Triplet loss FaceNet	MobilNet	82.20%	3	4
VGG16	MobileNet	100%	3	4

The mid location between the previous and current bb will be calculated and evaluated with several distance methods. The required time to assign id number shown in Table V, Fig. 13, and Fig. 14.

Fig. 13 and Fig. 14 showed the difference in distance method's accuracy to assign id number when the recognition process failed for a while. Chebyshev, Cosine, and Euclidean distances were able to identify the id number correctly. For correlation distance is one and id two on frame 128, it shows overlapping.

TABLE V. DISTANCE AND REQUIRED TIME AMONG DIFFERENCE METHOD

	Id 1	Id 2	Id 3	Time (s)
Cosine distance	3.068x 10^{-5}	8.500x 10^{-2}	3.14 x 10^{-5}	**0.0009948**
Euclidean	322.22	10.29	3.61	0.0009961
Correlation	2.00	1.11x 10^{-16}	T2.2 2 x 10^{-16}	0.0009975
Chebyshev	11.14	8.5	3	0.0009963

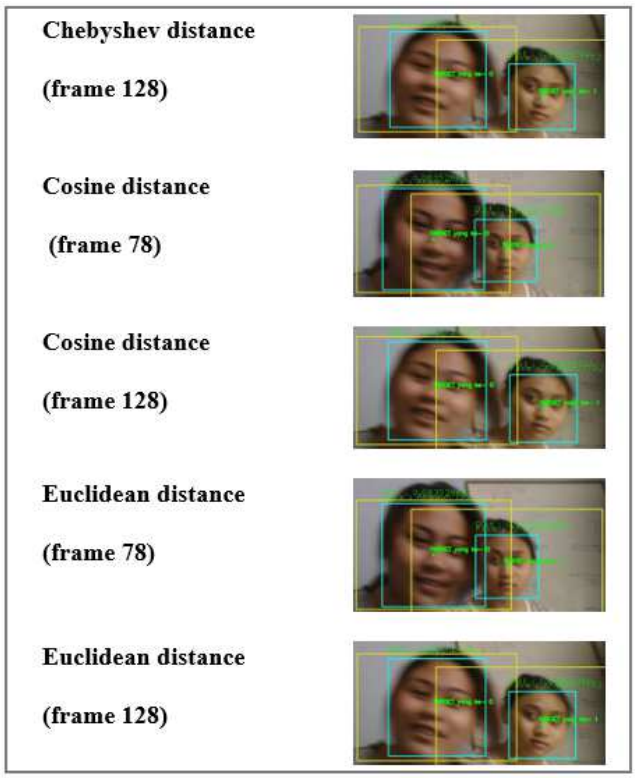

Fig. 13. The differences in distance method results.

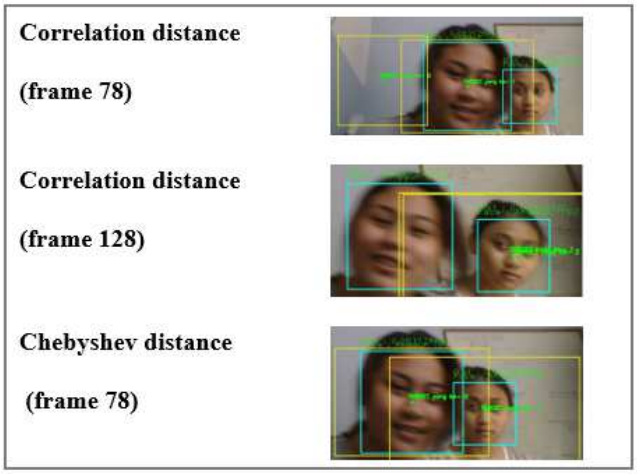

Fig. 14. The differences in distance method results.

We evaluated the face detection and recognition design with added cosine distance using two different datasets. The result is shown in Fig. 15. At the distance of 2m, the face recognition process resulted in falsely classified id two and id three due to high similarity. However, at 1.5m and 1m distances, the system successfully recognizes the face correctly. In the second dataset, with the scenario of only one person being alternately seen on the camera, the system can recognize each face correctly. In this case, cosine distance can assign the id number correctly at 2m to 1m distance.

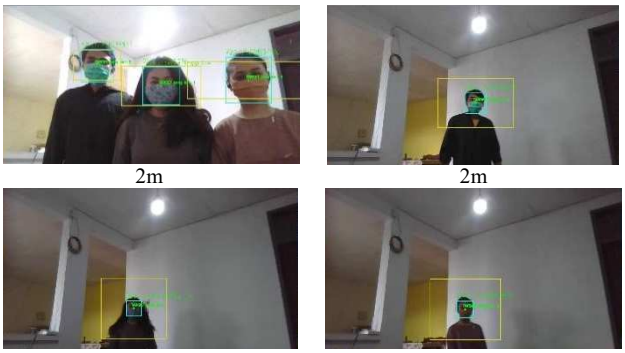

Fig. 15. Face recognition results in different distances and angles.

VI. Conclusion

This paper proposed a method to increase the accuracy of the face mask recognition process and detection. To obtain the face bounding box, Haar-cascade and MobileNet were implemented in the detection process. Subsequently, in the face recognition process, the VGG16 transfer learning method and the Triplet loss FaceNet with multi-threading techniques were compared. The evaluation of various combination was done, resulting in Haar-cascade and Triplet loss FaceNet as the most suitable for real-time implementation. To assign the ID number of a person, the cosine distance technique was used. FIFO technique was also utilized to increase the FPS result of the image stream with multi-threading. The proposed system successfully generates the name and id number even if the person is wearing a mask.

Acknowledgment

This research was funded by the Centre of Research and Community Service of the Institut Teknologi Bandung (LPPM ITB) through the *Program Penelitian dan Pengabdian Masyarakat* (Research and Community Service Program) scheme.

References

[1] M. S. Ejaz, M. R. Islam, M. Sifatullah, and A. Sarker, "Implementation of principal component analysis on masked and non-masked face recognition," in 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019, 2019.

[2] M. R. Bhuiyan, S. A. Khushbu, and M. S. Islam, "A deep learning based assistive system to classify COVID-19 face mask for human safety with YOLOv3," 2020, pp. 1–5.

[3] M. S. Ejaz and M. R. Islam, "Masked face recognition using convolutional neural network," in 2019 International Conference on Sustainable Technologies for Industry 4.0, STI 2019, 2019.

[4] I. Q. Mundial, M. S. Ul Hassan, M. I. Tiwana, W. S. Qureshi, and E. Alanazi, "Towards facial recognition problem in COVID-19 pandemic," in 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), 2020, pp. 210–214.

[5] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, "The effect of wearing a mask on face recognition performance: an exploratory study," in 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), 2020.

[6] F. M. T. R. Kinasih, C. Machbub, L. Yulianti, and A. S. Rohman, "Extending multi-object detection ability using correlative filter," in 2020 IEEE 10th International Conference on System Engineering and Technology (ICSET), 2020.

[7] D. A. Maharani, C. Machbub, L. Yulianti, and P. H. Rusmin, "Particle filter based single shot multibox detector for human moving prediction," in 2020 IEEE 10th International Conference on System Engineering and Technology (ICSET), 2020.

[8] D. A. Maharani, Carmadi Machbub, and P. H. Rusmin, "Enhancement of missing face prediction algorithm with kalman filter and DCF-CSR," in International Conference on Electrical Engineering and

Informatics (ICEEI), 2019.

[9] M. R. Sunitha, H. S. Jayanna, and Ramegowda, “Tracking multiple moving object based on combined color and centroid feature in video sequence,” in 2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014, 2015.

[10] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.

[11] I. Oztel, G. Yolcu, and C. Oz, “Performance comparison of transfer learning and training from scratch approaches for deep facial expression recognition,” in UBMK 2019 - Proceedings, 4th International Conference on Computer Science and Engineering, 2019, pp. 290–295.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–14, 2015.

[13] “‘Imagenet.’” [Online]. Available: http://www.image-net.org/.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 815–823, 2015.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[16] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016, pp. 265–283, 2016.

[17] P. Virtanen et al., “SciPy 1.0: fundamental algorithms for scientific computing in Python,” Nat. Methods, vol. 17, no. 3, pp. 261–272, 2020.

[18] G. Bradski, “The OpenCV Library,” Dr Dobbs J. Softw. Tools, vol. 25, pp. 120–125, 2000.

[19] P. Wilson and J. Fernandez, “Facial feature detection using Haar classifiers,” in Journal of Computing Sciences in Colleges, 2006, vol. 21, no. 4, pp. 127–133.

[20] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017.

[21] L. U. Ambata, I. A. P. Del Castillo, J. R. H. Jacinto, and C. M. T. Santos, “Public and private vehicle quantification and classification using vehicle detection and recognition,” in 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2019, 2019.

[22] “Queue.” [Online]. Available: https://en.wikipedia.org/wiki/Queue_(abstract_data_type).