MFRD-80K: A Dataset and Benchmark for Masked Face Recognition

Chin Poo Lee, *Member, IAENG,* and Kian Ming Lim, *Member, IAENG*

***Abstract*—Wearing face masks in public spaces has become an essential step to prevent the spread of COVID-19. This step poses some challenges to conventional face recognition due to several reasons: 1) the absence of large real-world masked face recognition dataset, and 2) the loss of some visual cues due to the occlusion by the face masks. To address these challenges, this paper presents a real-world masked face recognition dataset that consists of 80500 masked face images of 161 subjects, referred to as MFRD-80K dataset. Every subject contributes 500 masked face images, which are then partitioned into 60:20:20 for train, validation and test. Subsequently, we conduct some benchmark studies to evaluate the performance of the existing face recognition and classification methods on the MFRD-80K dataset. The methods include k-Nearest Neighbour, Multinomial Logistic Regression, Support Vector Machines, Random Forest, Multilayer Perceptron and Convolutional Neural Networks. Since the parameter settings affect the performance of each method, a grid search is performed to determine the optimal parameter settings. The empirical results demonstrate that Convolutional Neural Network achieves the highest test accuracy of 97.16% on MFRD-80K dataset.**

***Index Terms*—masked face, masked face recognition, masked face recognition dataset, machine learning, classification, CNN.**

I. Introduction

FACE recognition system is a computer vision task that aims to automatically identify an individual by the face. Face recognition is widely used in security access systems, smart payment systems, identity authentication systems, forensic investigation, attendance systems, and etc. Since the major outbreak of COVID-19 pandemic, governments have made it mandatory to wear a face mask while out in the public spaces. While wearing face masks is effective to prevent the spread of the virus, it brings some challenges as well. One of them being deteriorating the performance of the applications that involve face recognition where some parts of the face are occluded. Under the normal condition when a subject is wearing a face mask, only the eye brow and forehead if visible and useful for face recognition.

In view of this, we have collected a dataset with 80500 masked face images of 161 different subjects, referred to as MFRD-80K. Some benchmark studies are then conducted on the existing face recognition and classification methods, including k-Nearest Neighbour, Multinomial Logistic Regression, Support Vector Machines, Random Forest, Multilayer Perceptron and Convolutional Neural Networks to evaluate their performance on MFRD-80K dataset. To obtain the best performance for each method, a grid search is performed to determine the optimal parameter settings.

To this end, the main contributions of this paper are:

- A masked face dataset with 80500 images of 161 subjects was collected, referred to as Masked Face Recognition Dataset-80K (MFRD-80K). The dataset is so far the real-world masked face dataset with the highest number of masked face images. The masked face images were captured in varying backgrounds thus posing more challenges to the recognition tasks. The masked face dataset can be used for face recognition or verification purposes.
- Some benchmark studies of the existing face recognition and classification algorithms, including k-Nearest Neighbour, Multinomial Logistic Regression, Support Vector Machines, Random Forest, Multilayer Perceptron and Convolutional Neural Networks on the masked face dataset.
- A grid search to determine the optimal parameter settings based on the test accuracy on MFRD-80K dataset. The strengths of the optimal parameter settings are also discussed.

II. Related Works

This section briefly describes some publicly available real-world masked face datasets. The masked face datasets mainly serve three purposes: 1) masked face detection, 2) masked face recognition / identification, and 3) masked face verification. Masked face detection refers to classifying whether the subject is wearing a mask by locating the masked face in the image. Masked face recognition / identification aims to determine the identity of the subject based on the masked face. Masked face verification is the matching of the subject's claimed identity against the stored identity using the masked face. The following lists the existing real-world masked face datasets:

- MAsked FAces dataset (MAFA) [1] is a dataset used for masked face detection. MAFA contains 30811 images and 35806 masked faces collected from the Internet.
- Real-world Masked Face Recognition Dataset (RMFRD) [2] comprises 5000 masked face images of 525 public figures and 90000 images of the same subjects without masks. This dataset can be used for masked face recognition and masked face verification.
- Real-World Masked Face Verification Dataset contains 4015 masked face images of 426 subjects for verification purposes. The dataset is available at https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset

Manuscript received March 22, 2021; revised August 28, 2021. This work was supported by the Fundamental Research Grant Scheme of the Ministry of Higher Education under award number FRGS/1/2019/ICT02/MMU/03/7.

Chin Poo Lee, the corresponding author, is a Senior Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (phone: 606-2523172; e-mail: cplee@mmu.edu.my)

Kian Ming Lim is a Senior Lecturer in the Faculty of Information Science and Technology, Multimedia University, Melaka, 75450 Malaysia. (e-mail: kmlim@mmu.edu.my)

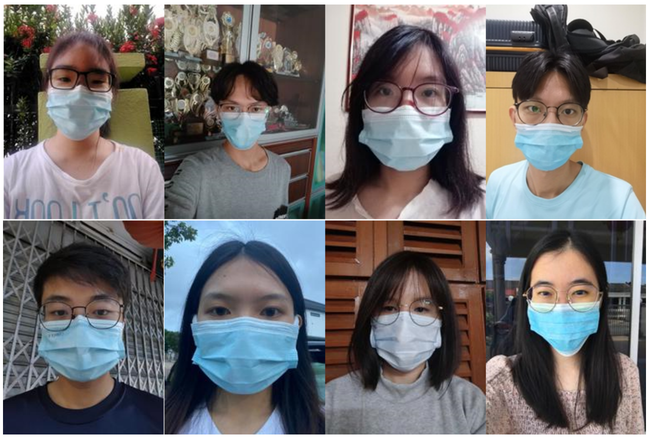

Fig. 1. Some sample images of MFRD-80K dataset.

III. MFRD-80K: Masked Face Recognition Dataset

In this paper, we present a masked face recognition dataset that contains a total of 80500 masked face images. It was collected from 161 subjects with 500 masked face images each. The images were captured from the frontal view using smartphones or webcams. All images were taken from different backgrounds either indoor or outdoor. At the time of publication, it is so far the real-world masked face dataset with the highest number of masked face images. Figure 1 shows some sample masked face images in the dataset. The comparison of the existing real-world masked face datasets is presented in Table I.

In the dataset preprocessing, all images were resized to the resolution of 192 × 256. For each subject, the masked face images were randomly partitioned into train, validation and test set at the ratio of 60:20:20. In doing so, there are a total of 48300, 16100 and 16100 images in the train, validation and test set.

The masked face recognition dataset in this work primarily aims for masked face recognition and masked face verification with masked face images. It can also be used for masked face detection when combined with publicly available unmasked face recognition datasets. The MFRD-80K is available at: https://github.com/kianming/MFRD-80K.

TABLE I
Comparison of Real-world Masked Face Datasets

Dataset	Number of Masked Face Images	Number of Unmasked Face Images	Number of Subjects
MAsked FAces dataset (MAFA)	35806	30811	-
Real-world Masked Face Recognition Dataset (RMFRD)	5000	90000	525
Real-World Masked Face Verification Dataset	4015	-	426
MFRD-80K	**80500**	-	161

IV. Benchmark Studies

Earlier on, handcrafted methods [3], [4], [5] were widely used in facial recognition. Handcrafted methods manually engineered the feature representations and then applied for classification [6], [7], [8], [9], [10]. On the other hand, learning-based methods learn the discriminative features from the input in an end-to-end manner. Later on, the learned features are used for classification [11], [12], [13].

In the benchmark studies, the performance of some existing face recognition and classification methods are evaluated on the MFRD-80K dataset. The methods include k-Nearest Neighbour [14], Multinomial Logistic Regression [15], Support Vector Machines, Random Forest, Multilayer Perceptron [16] and Convolutional Neural Networks. Since there are many parameters in each method, a grid search is leveraged to determine the optimal values for the parameters. The grid search is guided by the test accuracy obtained on the MFRD-80K dataset. The parameters and the set of values that are included in the grid search are presented in Table II. In the experiments, all images are resized to the dimension of 54 × 72 to save computational resources without compromising on the performance. Apart from that, the intensity values of all images are normalized to be within the range of [0, 1].

A. k-Nearest Neighbour

k-Nearest Neighbour (k-NN) is a supervised learning method that classifies the new data sample based on the existing data samples that are most similar to the new sample. The k-NN algorithm classifies a new data sample based on how its neighbours are classified. For the k-NN algorithm, we studied three dimension reduction techniques for feature extraction, namely Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Neighbourhood Components Analysis (NCA) [17].

PCA is an unsupervised learning method that identifies the principal components of the feature space that contribute to the most variance of the data. LDA also identifies the attributes that maximize the variance between classes, however it is a supervised learning method with known class labels. NCA similarly is a supervised learning method but it finds the feature space that is visually meaningful.

The experimental results in Table III show that the combination of PCA and k-NN yields the accuracy of 49.39% while LDA and k-NN records an accuracy of 51.95%. The results demonstrate that the combination of NCA as feature extractor and k-NN as classifier yields the highest test accuracy on the MFRD-80K dataset with 61.53%.

B. Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is a supervised classification model which is used for multi-class classification. MLR is an extension of logistic regression that supports multi-class classification. The solvers in MLR aim to find the parameter weights that minimise the cost function. Since the MFRD-80K is not a very large dataset, Limited-memory Bryoden-Fletcher-Goldfarb-Shanno (LBFGS) [18] solver records the highest accuracy of 74.90% with the shortest execution time, as in Table IV. Stochastic Average Gradient (SAG) [19] and SAGA [20] demonstrate some underfitting mainly due to they are more suitable for very large datasets.

The regularization is a technique that is commonly used to avoid overfitting in the machine learning models. In the

TABLE II
THE HYPERPARAMETER VALUES THAT ARE INCLUDED IN THE GRID SEARCH

Algorithm	Hyperparameters
k-NN	Feature extractor ∈ {Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Neighborhood Components Analysis (NCA)}

MLR	Solver ∈ {Limited-memory Bryoden-Fletcher-Goldfarb-Shanno (LBFGS), Stochastic Average Gradient (SAG), SAGA}
	Regularization ∈ {0.01, 1.0, 100}

SVM	Kernel function ∈ {Radial Basis Function (RBF), Polynomial, Linear}
	Regularization ∈ {0.01, 1.0, 100}

RF	Max depth ∈ {10, 20, 30, 40, 50}

MLP	Hidden Layers ∈ {3, 5}
	Number of Nodes in Hidden Layers ∈ {512, 256, 128, 64, 32}
	Dropout Layers ∈ {Yes, No}

CNN	Dropout value ∈ {0.2, 0.3, 0.4}
	Batch size ∈ {16, 32, 64, 128, 256}

TABLE III
THE EXPERIMENTAL RESULTS OF k-NN WITH DIFFERENT FEATURE EXTRACTORS

Feature Extraction	Accuracy
Principal Component Analysis (PCA)	49.39
Linear Discriminant Analysis (LDA)	51.95
Neighborhood Components Analysis (NCA)	**61.53**

experiments, the regularization at 1.0 shows the best generalization with the highest accuracy. From the experimental results, it is observed that the model needs at least 100 iterations to converge decently. The performance continues to improve until 100 iterations and not much improvement is observed thereafter. Therefore, the number of iterations is set to 100 with a good trade-off between performance and execution time.

TABLE IV
THE EXPERIMENTAL RESULTS OF MLR WITH DIFFERENT SOLVERS

Solver	Regularization	Accuracy	Execution Time (s)
SAG	1.0	71.16	16404.90
SAGA	1.0	74.68	19623.51
LBFGS	1.0	**74.90**	410.57

TABLE V
THE EXPERIMENTAL RESULTS OF MLR WITH DIFFERENT REGULARIZATION VALUES

Solver	Regularization	Accuracy
LBFGS	0.01	69.28
LBFGS	1.0	**74.90**
LBFGS	100	70.76

C. Support Vector Machines

Support Vector Machines (SVM) is a supervised learning method that builds the support vectors to encode the most representative similarities between data samples. The kernel function in SVM transforms the data samples into a higher-dimensional feature space so that they are easily separable. Three kernel functions are included in the experiments, namely Radial Basis Function (RBF), polynomial and linear function.

As shown in Table VI, the experimental results on MFRD-80K dataset demonstrate that linear kernel function performs well and obtains the highest accuracy of 82.58%. This is mainly because the linear function can sufficiently encode the face images with similar structure. As for regularization, the test accuracy increases from value 0.01 to 1.0 and remains stagnant since then, as in Table VII. Therefore, the regularization value of 1.0 is chosen for SVM.

TABLE VI
THE EXPERIMENTAL RESULTS OF SVM WITH DIFFERENT KERNEL FUNCTIONS

Kernel Function	Accuracy
Radial Basis Function (RBF)	74.70
Polynomial	72.75
Linear	**82.58**

TABLE VII
THE EXPERIMENTAL RESULTS OF SVM WITH DIFFERENT REGULARIZATION VALUES

Regularization	Accuracy	Execution Time (s)
0.01	80.65	864.62
1.0	**82.58**	882.18
100	**82.58**	1237.50

D. Random Forest

Random Forest (RF) algorithm decides the class label by taking the majority voting from multiple decision trees. In the experiments, we evaluate the performance of different numbers of maximum depth of the decision trees. As in Table VIII, the test accuracy increases and exhibits unspectacular improvements after the maximum depth of 50 for each decision tree. Therefore, the optimal maximum depth is set to 50 with the highest accuracy of 82.14%.

TABLE VIII
THE EXPERIMENTAL RESULTS OF RANDOM FOREST WITH DIFFERENT MAXIMUM DEPTHS

Maximum Depth	Accuracy
10	52.93
20	77.90
30	81.28
40	81.76
50	**82.14**

E. Multilayer Perceptron

Multilayer Perceptron (MLP) is a supervised multilayer neural network that contains input layer, hidden layer(s) and output layer. The input layer receives the input signals to be processed while the output layer is responsible for handling the classification task. MLP consists of one or more intermediate hidden layers as its computational engine, which is one of the differentiating advantages of MLP from other algorithms. Another distinguishing benefit is the differentiable nonlinear activation function that maps the input to the output. The hidden layers adopt Rectified Linear Unit (ReLU) activation function for more effective computation and better gradient propagation. The output layer leverages Softmax activation function to normalize the probability distributions of the classes. The loss function used is categorical cross entropy, defined as follows:

$$J(\mathbf{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right] \quad (1)$$

where $\mathbf{w}$ denotes the weights of the neurons, y_i and $\hat{y}_i$ denote the true class label and predicted class label, respectively.

The experimental results in Table IX show that adding more hidden layers to the MLP model increases the test accuracy. Not only that, adding a dropout layer after the hidden layer has also shown better generalization capability and higher test accuracy. Since the input size is moderate (54 × 72 = 3888), the MLP model with five hidden layers and the number of nodes for each hidden layer is 512, 256, 128, 64 and 32 respectively records the highest accuracy of 81.70%.

TABLE IX
THE EXPERIMENTAL RESULTS OF MLP WITH DIFFERENT ARCHITECTURES

Number of Hidden Layer	Number of Node	Dropout (Rate)	Accuracy
3	(128, 64, 32)	No	74.32
5	(512, 256, 128, 64, 32)	No	79.84
5	(512, 256, 128, 64, 32)	Yes (0.2)	**81.70**

F. Convolutional Neural Network

Convolutional Neural Network (CNN) is a deep learning algorithm with input layer, convolutional layer(s), pooling layer(s) and fully-connected layer(s) as the core building blocks. The proposed CNN consists of one input layer, three convolutional layers, three max pooling layers, and three fully-connected layers where the last fully-connected layer serves as the classification layer.

The convolutional layer involves the convolution operations where filters are multiplied with the input image to extract relevant features. The earlier convolutional layer encodes more abstract features and progressively encodes higher-level features in the subsequent convolutional layers. In the convolutional layer, the input is multiplied with a set of filters (convolution operation) to produce the feature representation known as feature map. The filters with the size of 7 × 7 are used in the first convolutional layer, while filters with the size of 3 × 3 and same padding are used in the subsequent convolutional layers. In addition, the Rectified Linear Unit (ReLU) activation function is leveraged to add nonlinearity to the convolutional layer.

The pooling layer reduces the dimension of the feature map as well as suppresses the noisy activations. The max-pooling with filter size of 2 × 2 is used in the pooling layer as it offers better de-noising and translational invariance effects. After the feature extraction by convolution and pooling layers, the feature map is then flattened and passed into the fully-connected layer for feature interpretation and classification.

The fully-connected layer learns the relation between the feature maps generated by the convolutional layers and max pooling layers to the class labels. Similarly, the Rectified Linear Unit (ReLU) activation function is adopted in the fully-connected layers to make the model less susceptible to vanishing gradient problems. Besides that, the dropout regularization is also leveraged in the fully-connected layers. The dropout regularization is a technique that works by randomly deactivating a certain portion of neurons to simulate the effects of models with different architectures. In doing so, the dropout regularization mitigates the overfitting problems caused by the same network architecture. In the experiments on MFRD-80K, the dropout rate of 0.2 yields the highest accuracy of 97.16%, as presented in Table X. The last fully-connected layer acts as the classification layer and returns the estimated probability of each class which is computed with a Softmax function. Similar to MLP, the CNN model also adopts categorical cross entropy as the loss function.

The CNN model is trained with mini batch gradient descent and achieves the highest accuracy on MFRD-80K when the batch size is 128, as shown in Table XI. A larger batch size learns slower but it results in a more steadily converged model. To optimize the gradient descent process, Adaptive Moment Estimation (Adam) optimization [21] technique is leveraged. The Adam optimization expedites the gradient descent process by calculating the learning rate adaptively based on the first and second order moments of the gradients. The number of training epochs of the CNN model is set to 100. Figure 2 illustrates the architecture of the CNN model.

TABLE X
THE EXPERIMENTAL RESULTS OF CNN WITH DIFFERENT DROPOUT RATES

Dropout Rate	Accuracy
0.2	**97.16**
0.3	95.21
0.4	95.02

TABLE XI
THE EXPERIMENTAL RESULTS OF CNN WITH DIFFERENT BATCH SIZES

Batch Size	Accuracy
16	92.71
32	95.96
64	94.49
128	**97.16**
256	96.41

Table XII presents the optimal hyperparameter settings and the results of k-kNN, MLR, SVM, RF, MLP and CNN. The experimental results demonstrate that the CNN model achieves the highest test accuracy of 97.16% among all methods in comparison.

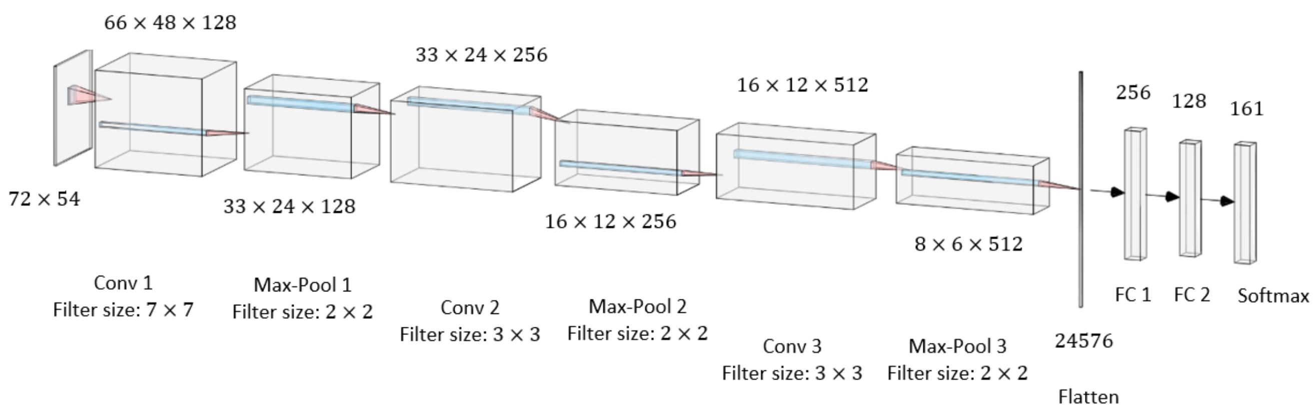

Fig. 2. The architecture of the CNN model.

TABLE XII
THE EXPERIMENTAL RESULTS OF THE METHODS ON MFRD-80K WITH THEIR OPTIMAL HYPERPARAMETER SETTINGS

Algorithm	Optimal Parameters	Test Accuracy (%)
k-NN	Feature extractor = Neighborhood Components Analysis (NCA)	61.53

MLR	Solver = Limited-memory Bryoden-Fletcher-Goldfarb-Shanno (LBFGS)
	Regularization = 1.0
		74.90

SVM	Kernel function = Linear Regularization = 1.0	82.58
RF	Max depth = 50	82.14
MLP	Hidden Layers = 5 Number of Nodes in Each Hidden Layer = 512, 256, 128, 64, 32 Dropout Rate = 0.2	81.70
CNN	Dropout value = 0.2 Batch size = 128	**97.16**

V. CONCLUSION

This paper presents a masked face recognition dataset (MFRD-80K) with a total of 80500 images of 161 subjects. The dataset is so far the dataset with the largest number of real-world masked face images. The dataset can be used for masked face recognition and verification purposes. It can also be combined with other face recognition dataset for masked face detection. The dataset is partitioned into train (60%), validation (20%) and test (20%) sets. Some benchmark studies are conducted to compare the performance of the existing face recognition and classification methods on MFRD-80K dataset, namely k-Nearest Neighbour, multinomial logistic regression, Support Vector Machines, Random Forest, Multilayer Perceptron and Convolutional Neural Networks. Since every algorithm involves some parameters, a grid search is performed to determine the optimal hyperparameter settings based on the test accuracy on the MFRD-80K dataset.

The empirical results demonstrate that the CNN model outshines the other algorithms in comparison with a test accuracy of 97.16%. The best performing CNN model on MFRD-80K comprises three convolutional layers with ReLU activation function, three max pooling layers, two fully-connected layers with dropout regularization, followed by a classification layer with Softmax function. For the k-NN model, the best result was obtained by applying NCA as the feature extractor which yields a test accuracy of 61.53%. The MLR model with LBFGS solver and regularization of 1.0 records the highest test accuracy of 74.90%. As for the SVM method, applying the linear kernel function and regularization value of 1.0 returns the highest test accuracy of 82.58% on the MFRD-80K dataset. The highest test accuracy of 81.76% is achieved in Random Forest when the maximum depth of decision trees is set to 40. The MLP model performs the best with a test accuracy of 81.70% on MFRD-80K when the architecture comprises 5 hidden layers with 512, 256, 128, 64, and 32 nodes followed by dropout layers. With the masked face recognition dataset MFRD-80K and benchmark algorithms, we look forward to more exciting and inspiring research in the near future.

ACKNOWLEDGMENTS

The authors would also like to extend our gratitude to Lim Kah Chuan, Lim Yu Jie, Luwe Yee Jia, Ngwe Jia Le, Tan Chun Keat, and the students of cohort Trimester 2 2020/2021 who took the subject TML2221 Machine Learning offered in Multimedia University for their contributions in the dataset collection.

REFERENCES

[1] S. Ge, J. Li, Q. Ye, and Z. Luo, “Detecting masked faces in the wild with lle-cnns,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2682–2690.

[2] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei *et al.*, “Masked face recognition dataset and application,” *arXiv preprint arXiv:2003.09093*, 2020.

[3] S. Sumpeno, M. Hariadi, and M. H. Purnomo, “Facial emotional expressions of life-like character based on text classifier and fuzzy logic,” *IAENG International Journal of Computer Science*, vol. 38, no. 2, pp. 122–133, 2011.

[4] Y. Kristian, H. Takahashi, E. Purnama, I. Ketut, K. Yoshimoto, E. I. Setiawan, E. Hanindito, and M. H. Purnomo, “A novel approach on infant facial pain classification using multi stage classifier and geometrical-textural features combination.” *IAENG International Journal of Computer Science*, vol. 44, no. 1, pp. 112–121, 2017.

[5] M. Monwar, S. Rezaei, and K. Prkachin, "Eigenimage based pain expression recognition," *IAENG International Journal of Applied Mathematics*, vol. 36, no. 2, pp. 1–6, 2007.

[6] Z. Sufyanu, F. S. Mohamad, A. A. Yusuf, and M. B. Mamat, "Enhanced face recognition using discrete cosine transform." *Engineering Letters*, vol. 24, no. 1, pp. 52–61, 2016.

[7] M. Fachrurrozi, A. Wijaya, M. N. Rachmatullah *et al.*, "New optimization technique to extract facial features." *IAENG International Journal of Computer Science*, vol. 45, no. 4, pp. 523–530, 2018.

[8] C. P. Lee, A. Tan, and K. Lim, "Review on vision-based gait recognition: Representations, classification schemes and datasets," *American Journal of Applied Sciences*, vol. 14, no. 2, pp. 252–266, 2017.

[9] K. M. Lim, A. W. Tan, and S. C. Tan, "A four dukkha state-space model for hand tracking," *Neurocomputing*, vol. 267, pp. 311–319, 2017.

[10] J. N. Mogan, C. P. Lee, K. M. Lim, and A. W. Tan, "Gait recognition using binarized statistical image features and histograms of oriented gradients," in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*. IEEE, 2017, pp. 1–6.

[11] C. Xing, J.-S. Wang, and B.-w. Zheng, "Hybrid face recognition method based on Gabor wavelet transform and VGG convolutional neural network with improved pooling strategy." *IAENG International Journal of Computer Science*, vol. 48, no. 2, pp. 413–427, 2021.

[12] Y. S. Tan, K. M. Lim, C. Tee, C. P. Lee, and C. Y. Low, "Convolutional neural network with spatial pyramid pooling for hand gesture recognition," *Neural Computing and Applications*, vol. 33, no. 10, pp. 5339–5351, 2021.

[13] Y. S. Tan, K. M. Lim, and C. P. Lee, "Hand gesture recognition via enhanced densely connected convolutional neural network," *Expert Systems with Applications*, vol. 175, p. 114797, 2021.

[14] A. Wirdiani, P. Hridayami, A. Widiari, K. Rismawan, P. Candradinata, and I. Jayantha, "Face identification based on k-nearest neighbor," *Scientific Journal of Informatics*, vol. 6, no. 2, pp. 150–159, 2019.

[15] S. Ongkittikul, J. Suwatcharakulthorn, K. Chutisowan, and K. Ratanarangsank, "Convolutional multinomial logistic regression for face recognition," in *2020 8th International Electrical Engineering Congress (iEECON)*. IEEE, 2020, pp. 1–4.

[16] P. Latha, L. Ganesan, and S. Annadurai, "Face recognition using neural networks," *Signal Processing: An International Journal (SPIJ)*, vol. 3, no. 5, pp. 153–160, 2009.

[17] J. Goldberger, G. E. Hinton, S. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," *Advances in Neural Information Processing Systems*, vol. 17, 2004.

[18] D. R. S. Saputro and P. Widyaningsih, "Limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method for the parameter estimation on geographically weighted ordinal logistic regression model (GWOLR)," in *AIP Conference Proceedings*, vol. 1868, no. 1. AIP Publishing LLC, 2017, p. 040009.

[19] N. L. Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," *arXiv preprint arXiv:1202.6258*, 2012.

[20] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.