Comparison of Face Classification with Single and Multi-model base on CNN

2020 15th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP) | 978-1-6654-1554-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/iSAI-NLP51646.2020.9376825

Sarin Watcharabutsarakham
Image Processing and Understanding Research Team
National Electronics and Computer Technology Center (NECTEC)
Pathum Thani, Thailand
sarin.wat@nectec.or.th

Supphachoke Suntiwichaya[1], Chanchai Junlouchai[2]
Leveraging Technology Solutions Section
National Electronics and Computer Technology Center (NECTEC)
Pathum Thani, Thailand
supphachoke.suntiwichaya@nectec.or.th[1], chanchai.junlouchai@nectec.or.th[2]

Apichon Kitvimorat
Image Processing and Understanding Research Team
National Electronics and Computer Technology Center (NECTEC)
Pathum Thani, Thailand
apichon.kitvimonrat@nectec.or.th

***Abstract*— Since the coronavirus disease 2019 (COVID-19) outbreak has spread across the country, our research applies to remind the people to wear a face mask when we go outside because a facial image detection and classification method will be used to authentication and authorization. This paper has shown that our created models based on CNN can detect the face mask-wearing, glasses-wearing, and gender with comparison two models. We training model with mix public datasets such as WIDER FACE, AFW, and MAFA. Moreover, we use VGG-Face to pre-train the model for the advance detection rate.**

Keywords*— convolutional neural network, CNN, pre-training, classification, face classification

I. INTRODUCTION

The current image classification techniques are used in extensive applications including, security features, face recognition, face verification, traffic identification, medical diagnosis, and other fields. The idea of image classification can be solved by different approaches[12]. Over the past few years, the increasing security is growing attention toward authentication based on voice, fingerprint, face and others. Face recognition is very interesting because it is useful for a wide number of real-time applications, such as surveillance, security systems, and access control. There are many works that were done in recent years, with many different approaches that have been proposed for a facial image. One topic, recognition methods, is trying to know the person and re-identify them [2-8]. The research problem in verification or re-identification[3] proposed deep discriminative representation learning (DDRL) for the unconstrained person re-identification. Moreover, [6] except for face recognition, such methods are often incapable of handling the open-set scenario of identifying new persons not included in the current set of known persons. Next, research problems from the source of the image such as [4] use to matching facial images captured from different sensors or sources with NIR-VIS to improve face recognition and [14] have additional depth images in the training data captured using depth cameras such as Kinect. In particular, we extract visual features and depth features from the RGB images and depth images. The next topic, geometry-feature-based methods, is to try to identify the position and relation between a part of the face such as eyes, nose, mouth, and shape, or size of the regions [9-10]. Once, COVID-19 was spread, the public health experts recommended universal mask-wearing, and some cities ordered residents to wear them under penalty of fine or imprisonment. In this real situation, the application may not undergo facial processes because the analysis algorithm to the face part has lost information. Anyway, COVID related application of computer vision, this one on detecting whether or not a person is wearing a face mask [9]. The research problem about detecting an object on the face such as [9] detect faces with occlusions is a challenging task due to two main reasons: the absence of large datasets of masked faces, and the absence of facial cues from the masked regions and [10] proposed method based on active appearance model (AAM) to remove eyeglasses from face images.

The purpose of this work is to propose a facial image classification method which consists of glasses and mask over the faces. So, the method appears suitable to be also applied in combination with this approach. It will focus on some parts of the face, such as eyes and mouth. Face detection is a critical technology, due to being applied in many fields such as authentication and authorization. In order to allow go to various locations, But due to the flu situation, general people have to wear a mask when they go outside. In this manner, many missing facial can be primarily recovered and exist in technology, not enough. Therefore need to create a new model based on the original model that is already effective facial detection technology. In this paper, we propose a method for facial classification based on features extracted with convolutional neural networks (CNN) and using the advantage of a pre-trained model in similar works.

II. RELATED THEORY

Recently, Almost of research is using deep learning because it becomes the first order of the state of art recognition algorithms. Especially, convolutional neural networks (CNN) have shown great potential in computation tasks. CNN use in the tasks of object recognition, tracking, classification, and face recognition. It has shown an excellent capability to solve complex problems and image classification tasks, which are impossible for human computation. The CNN creates a new image classification models which are much faster and more accurate than ever before, and they were applied in several objects [12-15]. Current CNN-based face detection methods divided into two categories. In the first category, CNN is used as a feature extractor in a traditional face detection framework to improve performance[17]. The second category, CNN, refers to the focus point on the face method, regards face detection as a particular case of generic object detection, and solves it using

978-1-6654-1554-5/20/$31.00 ©2020 IEEE

CNN-based object detection algorithm relying on identifying faces in the image [16].

VGG-16 network, proposed by K. Simonyan and A. Zisserman from the University of Oxford, is a convolutional neural network architecture. It is compounded with 16 layers. Each layer consists of convolutional layers maximum pooling, or max pooling layers, activation layers, and connected layers with fully. VGG-16 is convolution network for classification and detection. Next, The VGG-Face CNN descriptors are implemented by the University of Oxford-based on the VGG-Very-Deep-16 CNN architecture that is trained on over 2 million celebrity images as described in [5]. Other technologies were applied in the face images such as Facenet, OpenFace, and DeepFace. Even, DeepFace was developed by Facebook for face recognition models as a good alternative to VGG-Face.

III. METHODOLOGY

The algorithms used in this research based on the CNN architecture. We applied VGG-Face, the models are trained better and are able to identify different levels of image representation. The model will give the characteristic of face image such as glasses-wearing, a mask-wearing, or gender.

TABLE I. TRAINING DATASET

Group	Data Types	#no. Images
1	Glasses wearing	8,900
2	No Glasses wearing	8,900
3	Face Mask	2,200
4	No Mask wearing	2,200
5	Man	27,000
6	Woman	27,000

A. Data Collection

The public datasets used in the paper are WIDER FACE [19], Annotated Face in-the-Wild (AFW) [18], and MAFA [9]. To demonstrate their proposed method achieves state-of-the-art results. We mixed the dataset from MAFA, WIDER FACE, and AFW for training. Moreover, Deep learning needs to be trained with a huge training data set to achieve satisfactory performance. The public datasets usually contain limited images. Training dataset, the positive images are generally fewer than negative images. So, data augmentation is better to boost the performance and a widely used compensate in deep learning [20]. So, augmented data will be making many images to achieve data balancing. Data augmentation is widely used

a) The face data

b) The face mask data

Fig. 1. Example of Training dataset

TABLE II. PROPOSED CNN ARCHITECTURE

Layer type	Parameters
Input Layer	224x224 RGB image
Convolution	#64 224x224
Convolution	#64 224x224
Max Pooling	#64 2x2
Convolution	#128 112x112
Convolution	#128 112x112
Max Pooling	#128 2x2
Convolution	#256 56x56
Convolution	#256 56x56
Convolution	#256 56x56
Max Pooling	#256 2x2
Convolution	#512 28x28
Convolution	#512 28x28
Convolution	#512 28x28
Max Pooling	#512 2x2
Convolution	#512 14x14
Convolution	#512 14x14
Convolution	#512 14x14
Max Pooling	#512 2x2
Avg. Pooling	#512 1x1
Flatten	#512 1x1
Dense (Relu)	#4096 1x1
Dense (Relu)	#4096 1x1

to compensate for deep learning. Deep learning needs to be trained with a huge training data set to achieve satisfactory performance. We use the data augmentation to boost the

performance when training the deep network. We divided the image training set into six sets in our experimental detail in Table I.

The training set, we crop the input images into 224 × 224 pixels and horizontally flip them around the y-axis. To achieve data balancing, we sample the same number of positive images and negative images in each classes when start the training process.

B. Network architecture and Training

The overall framework is shown in Fig. 1. The framework is composed of two steps of neural networks. Firstly, pre-trained, this model start with the smaller networks converged and then used as initializations for the larger and deeper networks. The goal is to find the parameters of the network that minimize the average prediction loss value after the softmax layer. We applied VGG-16 and used VGG-Face structure model in pre-training step. Secondly, our layer, the classifier will have 2 convolution layers, 1 max pooling layers, 1 flattening layer and finally an output layer with Adam optimizer. The 33,605,442 parameters are trained in our proposed model. Our implementation is based on the python program with the NVIDIA CuDNN libraries to accelerate training. All our experiments were carried on NVIDIA 2070 RTX GPUs with 6GB of onboard memory.

IV. Experiments and Results

To allow for a direct comparison to previous work, while our CNNs are trained on the mixed dataset in the section before. We design two experimental groups to measure the accuracy rate, whether one model or multi-model is better for face classification with our model structure and unbalanced data. To obtain all the training data's average responses, we put all training data into the fine-turned VGG-face. Setting X= [x1, x2, x3, . . . , xn] denotes the image of training data, n is the number of training data. This method for fine-tuning the pre-trained VGG-face.

Fig. 2. Hierarachy Classification

A. First Experimental

The experiment is to try to classify the face with a hierarchy structure, as shown in Fig 2. The dataset is consists of 4 groups of face images.

-The image of 10,000 faces with a face mask or glasses and 10,000 regular faces.

-The image of 8,900 faces with glasses and 8,900 faces without glasses.

-The image of 2,200 faces with a face mask and 2,200 faces without a mask.

-The image of 27000 man faces and 27000 woman faces. We start with the pre-train of VGG-face and add three of the last layers to parameters training. This classifier will have one flattening layer, one max pooling layer, two convolution layers, and finally, an output layer with Adam optimizer. In the following, we improved the last three layers in the proposed models for increased accuracy. The input image will classify gender with first and compute the confidence value in $Score_i$. Next, the face image will classify mask-wearing and compute the confidence value in $Score_j$. Last, the face image will classify glasses-wearing and compute the confidence value in $Score_k$. Finally, the accuracy rate of testing is shown in Table III.

Condition for group classification
If found face in the image
1) If the value of $Score_i < 0.5$ and the value of $Score_j < 0.5$ and the value of $Score_k < 0.5$
classified in a confused group
2) If the value of man > 0.9
classified in a man group
or the value of woman > 0.9
classified in a woman group
or classified in a none group
3) If value of mask > 0.8
classified in a mask group
4) If value of Glass > 0.8
classified in a glass group

Fig. 3. Pseudo code

When we use the 4 models to predict the face, we select the condition and threshold by the value of data training. First, If the confidence value of gender, mask, and glasses was less than 0.5, the images were rejected into the group. Other criteria to classify groups by the confidence value are shown in Fig 4. To evaluation, we classified the face images into 6 classes as Table 1 (glasses-wearing, mask-wearing, man, woman). We use 4 trained models to classify like a hierarchy structure. In the first experiment, we trained four models with 96,000 face images to classify into four groups. The evaluation is shown that the models get an average 94.99% accuracy rate.

TABLE III. Result of Models

	Accuracy Rate (%)	
Model	**Training**	**Testing**
M1	99.44	98.58
M2	98.54	91.12
M3	97.59	96.13
M4	97.18	94.12

M1 is a trained model for classification of the face image that wears a face mask or glasses and a normal face. M2 is a trained model for face image classification to a face mask wearing and non-face mask-wearing. M3 is a trained model for face image classification to glasses and no glasses. M4 is a trained model for gender classification by face images. The accuracy rate of the experiment is shown in Table III.

B. *Second Experimemtal*

The experiment uses a trained model to classify the face into four classes, such as glasses-wearing (S1), mask-wearing (S2), man (S3), and woman (S4). The dataset uses the same as section A. We use to train with the pre-train of VGG-Face. The accuracy rate of training is 85.5%.

.

TABLE IV. CONFUSION METRIC OF RESULT

Actual	Predicted			
	S1	S2	S3	S4
S1	0.75	0.18	0.05	0.02
S2	0.01	0.90	0.01	0.08
S3	0	0	0.95	0.05
S4	0	0.10	0.10	0.80

We use our proposed architecture, explained in section A, for classification face images by only single model. In the second experiment, we trained a model with 6,000 face images to classify into four groups with balance data in every group. The evaluation is shown that the model gets an average 85.5% accuracy rate. We use the ReLu function to solve the vanishing gradient problem because it is suitable for balance data. This function does not have an asymptotic upper and lower bound. Thus, the earliest layer can receive the last layers' errors to adjust all weights between layers. By contrast, a traditional activation function like sigmoid is restricted between 0 and 1, so the errors become small for the first hidden layer. This scenario will lead to a poorly trained neural network. The experimental results showed that the model trained with many images would give an advantage. Although we use the image augmentation to make data balancing, the real face images are the best for training.

V. CONCLUSION

The comparison of single and multi-model are shown that the dataset in this experiment ware cross the group in some case so the multi-model will be loss performance to predict in the face image by our proposed model architecture was shown that the face images could give information for deep learning network to predict the objective setting. The result of this paper shows that the multi-model get high accuracy, but on the other side that we have to conscious is the space in memory of the computer. If we use a single model to classify, it uses a resource less than the multi-model. Fig 5 shows a comparison of the accuracy rate in each group. Moreover, face image was applied in many areas because pre-train VGG-Face helps the suitable tuning parameters by million face images. This advantage model structure can be helpful for researchers to create a new model of about-face images. Notably, our research uses to detect the face in the coronavirus disease 2019 (COVID-19) outbreak event to remind the people to wear a face mask when going outside or inside the public places. The compare results were shown in Fig. 4. We use a multi-model in section A and a single model in section B to classify face into four groups as a normal face, glasses-wearing, mask-wearing, and gender. The accuracy rate for classifies normal face (C1) is 75% with a single-model and 98% with multi-models. The accuracy rate for classifies the glasses-wearing group (C2) is 90% with a single-model and 91% with multi-models. The accuracy rate for classify mask-wearing group (C3) is 95% with a single-model and 96% with multi-models. The accuracy rate for classifying gender (C4) is 80% with a single-model and 94% with multi-models. In future work, the facial will classify a particular person because deep learning makes amazing in recognition applications.

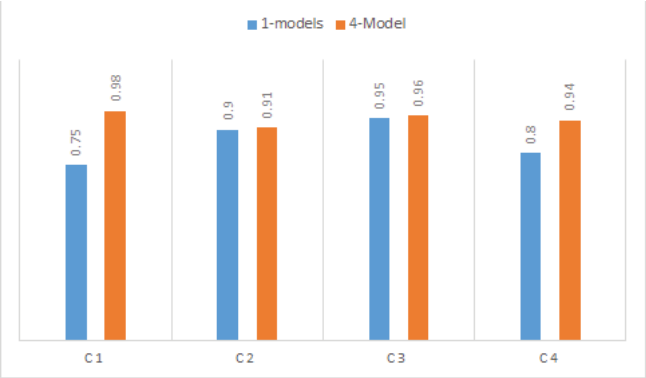

Fig. 4. The result of prediction with single and multi-model

ACKNOWLEDGMENT

We are thankful for public data distributors, which all use in this paper. Moreover, we thank our team in Artificial Intelligent Research Team which provided insight and expertise that greatly assisted the research, although the infrastructure may not be enough with all of the resources for this paper.

REFERENCES

[1] X. Mengyu, T. Zhenmin, Y. Zhenmin, Y. Lingxiang, L. Lingxiang and X. Jingsong ,"Deep Learning for Person Reidentification Using Support Vector Machines," Adv. Multim. 2017, doi:10.1155/2017/9874345.

[2] G. Artur , K. Marcin and, P. Norbert," Face re-identification in thermal infrared spectrum based on ThermalFaceNet neural network," International Microwave and Radar Conference (MIKON), May 14-17, 2018, doi: 10.23919/MIKON.2018.8405170

[3] J. Yu, D. Ko, H. Moon and M. Jeon, "Deep Discriminative Representation Learning for Face Verification and Person Re-Identification on Unconstrained Condition," 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, 2018, pp. 1658-166z2., doi: 10.1109/ICIP.2018.8451494

[4] C. Peng, N. Wang, J. Li and X. Gao, "Re-Ranking High-Dimensional Deep Local Representation for NIR-VIS Face Recognition," in IEEE Transactions on Image Processing, vol. 28, no. 9, pp. 4553-4565, Sep, 2019, doi: 10.1109/TIP.2019.2912360

[5] O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep Face Recognition," British Machine Vision Conference,2015.

[6] V. Madhawa, S. Imesha,K. Pasindu,V. Jayan and P. Indika, " Open Set Person Re-identification Framework on Closed Set Re-Id Systems," International Conference on Signal and Image Processing (ICSIP)," AUG 4-6, 2017, doi: 10.1109/SIPROCESS.2017.8124507

[7] S. Kanchanapreechakorn and W. Kusakunniran, "Robust human re-identification using mean shape analysis of face images," TENCON 2017 - 2017 IEEE Region 10 Conference, Penang, 2017, pp. 901-905., doi: 10.1109/TENCON.2017.8227986

[8] G. Agus and W. H. Dwi, "Key Frame Extraction with Face Biometric Features in Multi-shot Human Re-identification System," International Conference on Advanced Computer Science and

information Systems (ICACSIS), OCT 12-13, 2019, doi: 10.1109/ICACSIS47736.2019.8979799

[9] S. Ge, J. Li, Q. Ye and Z. Luo, "Detecting Masked Faces in the Wild with LLE-CNNs," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 426-4, doi: 10.1109/CVPR.2017.53

[10] Y. Wang, J. Jang, L. Tsai and K. Fan, "Improvement of Face Recognition by Eyeglass Removal," 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Darmstadt, 2010, pp. 228-23, doi: 10.1109/IIHMSP.2010.64

[11] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220

[12] M. Jogin, Mohana, M. S. Madhulika, G. D. Divya, R. K. Meghana and S. Apoorva, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 2319-2323, doi: 10.1109/RTEICT42901.2018.9012507.

[13] X. Xu, W. Li, and D. Xu, "Distance Metric Learning Using Privileged Information for Face Verification and Person Re-Identification," IEEE Transactions on Neural Networks and Learning Systems, vol. 26 , 2015, pp. 3150 - 3162.

[14] Y. Chen, and A. Baskurt, Person re-identification in images with deep learning, 2018.

[15] C.W. Ngo, and et al.,"Deep Learning for Person Reidentification Using Support Vector Machines," Advances in Multimedia, 2017, doi: doi.org/10.1155/2017/9874345.

[16] A. Jourabloo, M. Ye, X. Liu and L. Ren, "Pose-Invariant Face Alignment with a Single CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3219-3228, doi: 10.1109/ICCV.2017.347.

[17] M. Nakada, H. Wang and D. Terzopoulos, "AcFR: Active Face Recognition Using Convolutional Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, 2017, pp. 35-40, doi: 10.1109/CVPRW.2017.11.

[18] X. Zhu and D. Ramanan. "Face detection, pose estimation,and landmark localization in the wild," In IEEE Conference on Computer Vision and Pattern Recognition, pages 2879-2886, June 2012.

[19] Y. Ping L. Chen, C. Loy and, X. Tang,"WIDER FACE: A Face Detection Benchmark," 2015, CoRR abs/1511.06523

[20] Mei Wang and Weihong Deng, "Deep Face Recognition: A Survey",School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China,2019.