Face Recognition Using Light-Convolutional Neural Networks Based on Modified VGG16 Model

Anugrah Bintang Perdana, Adhi Prahara
Informatics Department
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
nugabp@gmail.com, adhi.prahara@tif.uad.ac.id

***Abstract*—Identification and verification systems nowadays utilize biometric technology such as face recognition, retina scan, and fingerprint mapping to recognize a person identity. For the face recognition technology, the traditional methods usually consist of four stages: face detection, face alignment, feature extraction, and classification. Recently, deep learning gains popularity in the face recognition task due to its performance that outperform the traditional methods and its simplicity that combines features extraction and classification in a single architecture. Deep learning especially Convolutional Neural Networks (CNN) has been successfully implemented in the face recognition applications. There are many variations of CNN based models that successfully improved the face recognition performance. However, the majority of the models have very deep layers and trained with large scale face image dataset that need a lot of computational resources. In this research, a light-CNN based on modified VGG16 model is proposed to recognize face with a limited dataset. The proposed light-CNN is compact yet produces good performances with 94.4% accuracy.**

Keywords—light-CNN, face recognition, deep learning

I. INTRODUCTION

Understanding how the human visual system (HVS) processes visual information involves building models that would account for the human-level performance on a multitude of tasks [1]. In the intelligence systems model, biometric technology recently becomes popular. Biometric technology performs authentication and identification data from the unique characteristics of persons such as face, retina, and fingerprint. The conventional face recognition pipeline consists of four stages: face detection, face alignment, feature extraction and classification [2]. The most important stage in the face recognition model is feature extraction. Hand-crafted features such as Eigenfaces [3], Fisherfaces [4] and Local Binary Patterns (LBP) [2], [5], [6] have achieved good performance in the face recognition task. However, the performance degrades if the features applied on unconstrained environment such as complex background, illumination, various pose, and occlusion [2].

There are several methods that can be implemented to recognize objects and deep learning is the one that commonly used. Deep learning simply uses a network with each input layers of neuron is connected to every output neuron in the next layer. Deep learning, in particular Convolutional Neural Networks (CNN) is a validated image representation and classification technique for image analysis and applications [7]. Deep learning does not need complicated method to extract the features, especially when using CNN. Datasets trained by CNN will give variety of results that depends on the architecture and the datasets. CNN has received immense success in multiple applications such as natural language processing, object classification, and image segmentation [8]. CNN provides state-of-the-art results in several computer vision problems. CNN has a large number of parameter that requires a large number of training sample which becomes the limiting factor for a small sample size problem [8].

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [9] was held to evaluate the algorithms for object recognition and classification in a large scale. There are a few popular object recognition and classification algorithms such as CNN algorithm which won the competition for the first time in 2012 with 8 layers called AlexNet [10]. Lately, many variations of CNN model able to significantly improve the performance of object detection and classification for example VGGNet [11]. VGGNet consists of 16 convolution layers that became the runner-up in 2014. The brief review of some successful CNN models for face recognition problems are presented in the following paragraph.

Schroff *et al.* [12] from Google, Inc. proposes a system called FaceNet. FaceNet learns to map a face into Euclidean space where the distance can be directly used for face recognition. They use Deep Convolutional Network in the feature extraction process. The experiment shows result 99.63% and 95.12% accuracy from LFW and Youtube Faces dataset. Taigman *et al.* [13] from Facebook proposes DeepFace model. From the process of face recognition namely face detection, face alignment, feature representation, and face recognition, they optimize the face alignment and feature representation steps using 3D face model. The experiment conducted on 4 million faces with more than 4,000 labels and shows 97.35% accuracy in LFW dataset. Following by its predecessor which using deep neural networks for object recognition, Sun *et al.* [14] motivated to apply deep neural networks on face recognition problems. The work based on VGGNet and GoogLeNet to make them suitable for face recognition. The result achieves 99.53% accuracy on LFW face recognition dataset and 96.0% of LFW rank-1 face identification accuracy.

However, aside from the great achievement of the previous deep learning models, deep neural networks need more resources for the computation. There are few articles that discuss about light CNN [15], [16]. Wu *et al.* proposes new activation function called Max-Feature-Map (MFM) to be used in their proposed light-CNN model. By reducing the parameters and accelerate the computational process, light CNN framework able to learn a robust face representation on noisy labeled dataset [15]. Zheng and Zu proposes normalized light-CNN using 11 hidden layers implemented in LFW and achieve 98.46% of face verification accuracy [16].

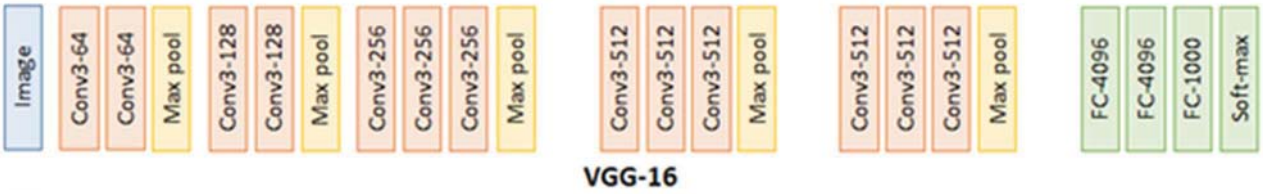


Fig. 1. VGG16 architecture. (modified from https://medium.com/coinmonks/paper-review-of-vggnet-1st-runner-up-of-ilsvlc-2014-image-classification-d02355543a11).

Face recognition generally used for identification and verification. There are several facial features that can be used to recognize persons. This paper has a goal to build light CNN architecture that can be implemented with limited data based on existing architecture namely VGG16. The implementation of the proposed CNN model will be explained in the next section that organized as follow: Section II presents the proposed light-CNN architectures, Section III presents the result and discussion, and finally the conclusion of this work is presented in Section IV.

II. LIGHT-CNN FOR FACE RECOGNITION

In this research, the proposed light-CNN architecture is built based on VGG16 which suitable for limited dataset. Figure 1 shows the baseline of VGG16 architecture for face recognition. VGG16 which design for large scale classification has quite deep layers with several small convolution layers with various number of kernel followed by max pooling.

A. Light-CNN Architecture

The proposed architecture which based on VGG16 model can be seen in Table I. From VGG16 architecture which illustrated in Figure 1, one layer from the 64 filters convolutional layers is removed and the 256 filters and 512 filters convolutional layers are completely removed. Also the size of fully connected layers is changed. By removing some layers from VGG16 model, the architecture becomes light and compact.

TABLE I. THE PROPOSED LIGHT-CNN ARCHITECTURE

Layer Type	Kernel	Size	Stride
Conv1	64	3 x 3	1 x 1
Max Pooling	-	2 x 2	1 x 1
Conv2-1	128	3 x 3	1 x 1
Conv2-2	128	3 x 3	1 x 1
Conv2-3	128	3 x 3	1 x 1
Max Pooling	-	2 x 2	1 x 1
FC	512	-	-
FC	30	-	-
Soft-max	1	-	-

The proposed architecture uses 120 x 120 pixels as the size of the input image and has only two types of convolutional layers which followed by max pooling. Each convolutional layer followed by rectified linear unit (ReLU) activation function. The first convolutional layer has 64 filters with 3 x 3 filter size followed by max pooling layer with 2 x 2 filter size. The second convolutional layers have 128 filters with 3 x 3 filter size followed by max pooling layer with 2 x 2 filter size. In the final layers, there are two fully connected layers which have 512 neurons and 30 neurons for classification into 30 labels using soft-max.

The general procedure of face recognition system proposed in this research is shown in Figure 2. From Figure 2, the face recognition process is divided into two phases namely training and testing. In the training phase, dataset of labeled face image is resized into 120 x 120 pixels then used to train the networks using the proposed architecture. The trained model will be used in the face recognition process in the testing phase. In the testing phase, face detection method is applied to the frame grabbed from video camera or webcam. Every face which found in the frame will be cropped and resize to fit the input model. After applying the face recognition procedure using the previously trained model, a threshold is applied to discard the face label which has confidence score below the threshold and pass the face label which has confidence score above the threshold. This procedure will reduce the false classification by the model. The face label then shown in the screen as the result of face recognition system.

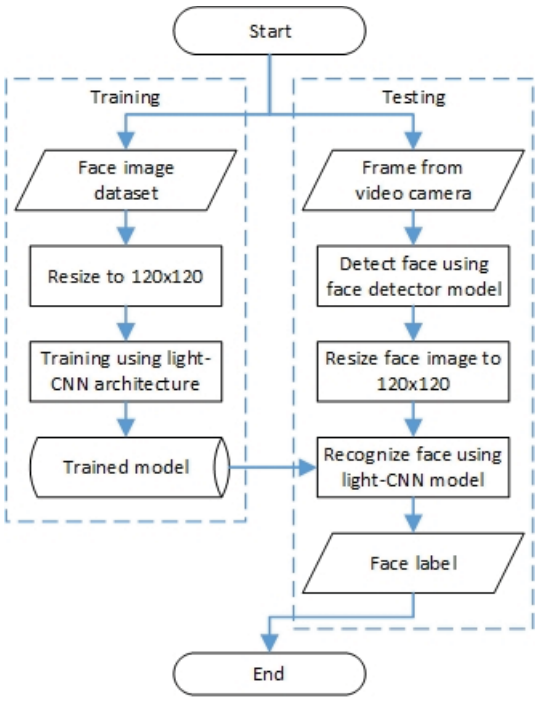


Fig. 2. The general procedure of the proposed face recognition system.

B. Perfomance Measurement

In order to evaluate the performance of the proposed model, we use confusion matrix. From the confusion matrix,

accuracy, recall, precision, and F1 measure are calculated using (1) – (4) respectively.

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

$$recall = \frac{TP}{TP+FN} \quad (2)$$

$$precision = \frac{TP}{TP+FP} \quad (3)$$

$$F1\ measure = 2 \times \frac{precision \times recall}{precision+recall} \quad (4)$$

TP is true positive, TN is true negative, FP is false positive and FN is false negative.

III. Result and Discussion

The proposed method is built using Python 3.6 with additional deep learning library Keras, image processing library OpenCV, matplotlib, and scikit-learn library. The method runs on computer with processor Intel Core-i7, NVidia GTX 1070, and 16GB of RAM while the training model uses Google Colab GPU with NVIDIA Tesla K80 and memory 16 GB.

A. Dataset

The dataset consists of 30 labels of face image in RGB format. Some face labels are acquired from ROSE-Youtu Face Liveness Detection Database which used in [17] and some are taken manually with camera. Every face that found in the video will be cropped automatically using face detection method from OpenCV library. The dataset consists of 7,250 face images which then divided into 5,075 face image for training and 2,175 face images for validation. A new set of 484 face images is used for testing. Some samples of the dataset are shown in Figure 3. As shown in Figure 3, each face image has various size and background and has not follow certain pattern.

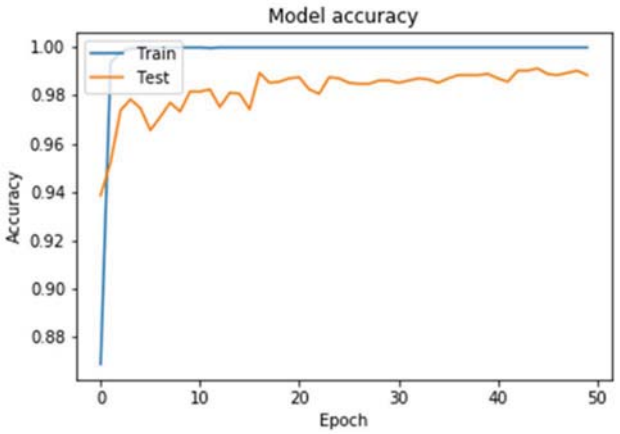
Fig. 3. Sample of face images used in this research.

B. Perfomance Evaluation

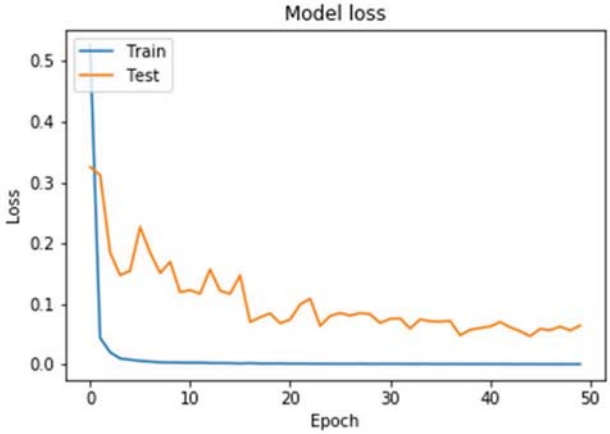
The dataset has 30 labels and consists of 7,250 face images with 5,075 images for training and 2,175 images for validation. Training is performed for 50 epochs using Stochastic Gradient Descent (SGD) optimizer. The accuracy and loss from the model are monitored and shown in Figure 4(a) and Figure 4(b) respectively. From Figure 4, the training accuracy is quite stable in 0.98 and 0.1 loss after approximately 10 epochs and does not show any overfitting case. The trained model then applied in the test dataset and achieves score 0.944 or 94.4% of accuracy. From the 484 test data, the model can correctly predict 457 data. The proposed model also achieves recall score 95%, precision score 93%, and F1 measure score 94%.

Before reaching the final architecture in Table I, we conduct a test for two architectures based on VGG16. In the first trial architecture, a 32 filters of convolutional layer are added before the 64 filters of convolutional layer from Table I. In the second trial architecture, the difference between architecture in Table I is only the input size that changed to 96 x 96 pixels. The first and the second trial architecture achieve accuracy score 92.9% and 93.3% respectively.

To compare the result with the baseline, VGG16 architecture is used to train the training dataset and test the result on the same test data as the proposed model. The baseline architecture achieves 77.8% accuracy which correctly classify 370 data from 484 test data. Table II shows the comparison of accuracy from VGG16 and the proposed method.



(a) Training accuracy



(b) Training loss

Fig. 4. The graph of training and validation of the proposed model

From Table II, the accuracy of the proposed light-CNN model is better than the rest of the modified models. Compare to the first trial and VGG16 model, the proposed architecture is more compact and light. This prove that light-CNN model is superior in accuracy and training time when the dataset is limited and the number of category is small.

TABLE II. Comparison Between VGG16 and The Proposed Method

Architecture	Accuracy
VGG16	77.8%
First trial	92.9%
Second trial	93.3%
Proposed	94.4%

IV. Conclusion

Face recognition in shallow learning pipeline consists of four stages, and the most important step is the feature extraction step. One of the advantage of deep learning is deep learning does not need a complicated method to extract the features. However, the deep convolutional networks need more resources for the computation while the light convolutional networks do not spend too much time for training the model. Our experiment shows that shallow network such as light-CNN also produces high accuracy which is 94.4% and performs better in limited dataset and small number of labels.

Acknowledgment

This research is supported by LPPM Universitas Ahmad Dahlan with research grant no. PF-135/SP3/LPPM-UAD/IV/2019.

References

[1] J. Kubilius, S. Bracci, and H. P. Op de Beeck, "Deep Neural Networks as a Computational Model for Human Shape Sensitivity," *PLOS Comput. Biol.*, vol. 12, no. 4, p. e1004896, Apr. 2016.

[2] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, "Extended local binary patterns for face recognition," *Inf. Sci. (Ny).*, vol. 358–359, pp. 56–72, Sep. 2016.

[3] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cogn. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[5] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face Recognition with Local Binary Patterns," Springer, Berlin, Heidelberg, 2004, pp. 469–481.

[6] T. Huynh, R. Min, and J.-L. Dugelay, "An Efficient LBP-Based Descriptor for Facial Depth Images Applied to Gender Recognition Using RGB-D Face Data," in *Computer Vision - ACCV 2012 Workshops: ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I*, J.-I. Park and J. Kim, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 133–145.

[7] G. Carneiro, Y. Zheng, F. Xing, and L. Yang, "Review of Deep Learning Methods in Mammography, Cardiovascular, and Microscopy Image Analysis," Springer, Cham, 2017, pp. 11–32.

[8] R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning Structure and Strength of CNN Filters for Small Sample Size Training," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9349–9358.

[9] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., pp. 1097–1105, 2012.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[14] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face Recognition with Very Deep Neural Networks," Feb. 2015.

[15] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[16] H. H. Zheng and Y. X. Zu, "A Normalized Light CNN for Face Recognition," *J. Phys. Conf. Ser.*, vol. 1087, no. 6, p. 062015, Sep. 2018.

[17] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised Domain Adaptation for Face Anti-Spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 7, pp. 1794–1809, Jul. 2018.

.