Gender Recognizer based on Human Face using CNN and Bottleneck Transformer Encoder

Adri Priadana
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
priadana3202@mail.ulsan.ac.kr

Muhamad Dwisnanto Putro
Department of Electrical Engineering
Universitas Sam Ratulangi
Manado, Indonesia
dwisnantoputro@unsrat.ac.id

Jinsu An
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
jinsu5023@islab.ulsan.ac.kr

Duy-Linh Nguyen
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
ndlinh301@mail.ulsan.ac.kr

Xuan-Thuy Vo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
xthuy@islab.ulsan.ac.kr

Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
acejo@ulsan.ac.kr

***Abstract*—Several applications, such as Human-Robot interactions and offline advertising platforms, perform gender recognition based on a human face to profile their audience. These applications demand gender recognition that can operate in real-time on a low-cost or CPU device. This work proposes a gender recognizer based on a human face using a Convolutional Neural Network (CNN) and Bottleneck Transformer Encoder (BTE) that renders low parameters and operation. BTE is offered to support the primary CNN feature extractor in learning the global representation of the feature maps efficiently. This work uses three face gender datasets benchmark, namely UTKFace, Labeled Faces in the Wild (LFW), and Adience, to train and validate the proposed network. The proposed network achieves competitive accuracy compared to the state-of-the-art network. The recognizer can operate in real-time on a CPU with 150 frames per second.**

***Index Terms*—Bottleneck Transformer, Convolutional Neural Network (CNN), Efficient Architecture, Gender Recognition, Self-Attention**

I. Introduction

As a unique soft biometric property, gender is an essential part that can be used in profiling users. Several applications utilize gender by recognizing it based on a human face to support their purpose [1]. In Human-Robot interactions [2], [3], robots can use gender property to characterize their interlocutors to personalize and make conversations more pleasant. Foggia et al. [2] demonstrated that people prefer conversations with robots that can personalize their interlocutors according to gender by using appropriate pronouns and honorifics. In advertising applications such as digital signage, gender recognition is used as a basis to provide relevant advertising content in real-time [4].

Gender recognition through a human face is initiated by detecting the audience's face via a camera and then classifying it to the gender class. In robotics or digital signage applications utilizing a low-cost device, the gender recognition process is demanded to perform in real-time properly while the audiences face them. Therefore, an efficient gender recognition mechanism with low computation is required. Hence, robotics or digital signage platform that uses a low-cost device, or at least a CPU device, can operate it properly.

Convolutional Neural Networks (CNNs) have recently demonstrated substantial effectiveness in recognition tasks. To construct a recognition system, particularly for gender recognition based on human faces, many researchers have created various CNN designs. Hamdi and Moussaoui [5] proposed a manually-designed CNN architecture to predict facial gender, gaining 89.97% based on accuracy on the UTKFace [6] dataset. HyperFace-ResNet [7], improved ResNet-101 [8] architecture by fusing the lower and deeper layers using an element-wise addition operation, is offered to predict facial gender. This presented architecture gained 94% based on accuracy on the LFW [9] datasets. Greco et al. [4] applied MobileNetV2 architecture to develop a gender recognizer performing in real-time on a CPU device supporting a digital signage platform. Other works [10]–[12] designed more efficient CNN architectures for gender recognition based on human face generating fewer parameters and achieved great accuracy on UTKFace [6], LFW [9], and Adience [13] datasets. These architectures can perform fast in a CPU device. Employing CNN architecture with fewer parameters will improve efficiency and lead the recognizer to operate more quickly, especially on low-cost or CPU devices.

Considering the efficiency and computation, this work proposes a lightweight CNN architecture combined with Bottleneck Transformer Encoder (BTE) to perform gender recognition based on a human face. This architecture generates a few parameters and low operation, appropriate to implement on a

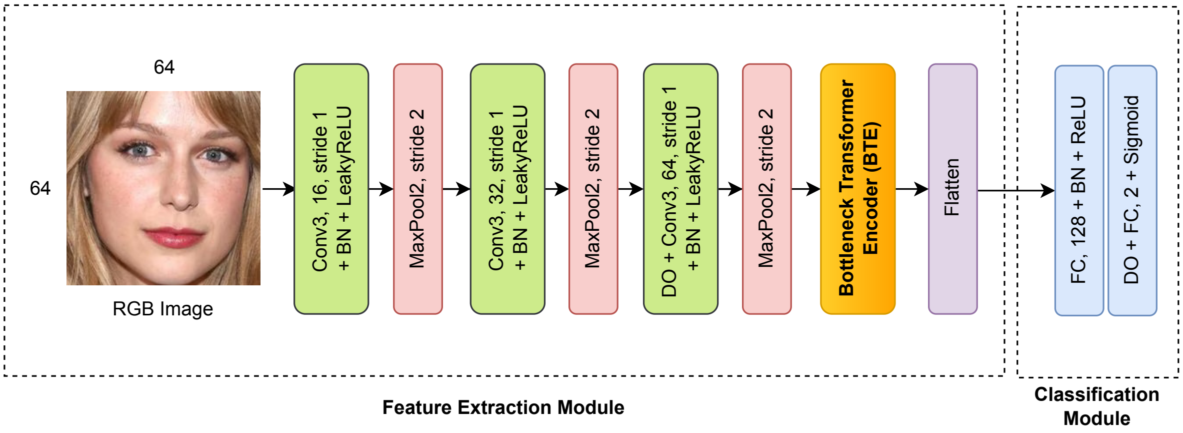

64
64
RGB Image
Conv3, 16, stride 1 + BN + LeakyReLU
MaxPool2, stride 2
Conv3, 32, stride 1 + BN + LeakyReLU
MaxPool2, stride 2
DO + Conv3, 64, stride 1 + BN + LeakyReLU
MaxPool2, stride 2
Bottleneck Transformer Encoder (BTE)
Flatten
FC, 128 + BN + ReLU
DO + FC, 2 + Sigmoid
Feature Extraction Module
Classification Module

Fig. 1. The proposed architecture of the face gender recognition based on a human face.

CPU device. In this work, the contribution outlines as follows:

1) A gender recognition based on a human face with only 555,770 parameters and 24.8 MFLOPs achieves very competitive accuracy and efficiency compared to other architectures on three datasets, UTKFace [6], Adience Benchmark [13], and Labeled Faces in the Wild (LFW) [9].
2) A Bottleneck Transformer Encoder (BTE) leads the architecture not only able to capture the local representation but also the global representation by efficiently calculating the relationship between each spatial information.
3) The proposed gender recognizer based on a human face can operate in real-time on a CPU with a speed of 131.85 frames per second, outperforming the other gender recognizer based on a human face.

II. Proposed Architecture

The proposed gender recognition architecture has a feature extraction and classification modules shown in Fig. 1. It produces only 555,770 parameters.

A. Feature Extraction Module

The feature extraction module serves as an instrument to extract facial features from the input face area image, containing three of 3×3 convolution layers with a growing number of channel from 16 to 64. This mechanism aims to extract more information on the higher level feature of the facial image. This module employs Batch Normalization (BN) [14] and Leaky Rectified Linear Unit (LeakyReLU) activation in every convolution layer to bargain with the gradient problem. A dropout (DO) mechanism [15] is applied before the last convolution layer to prevent overfitting issues. Three times of 2×2 max-pooling operations with strides two are used after every convolution layer as a downsampling mechanism. Involving several convolution layers captures only the local representation of the facial feature. Therefore, we propose a Bottleneck Transformer Encoder (BTE) to grasp the global representation of the feature maps. In this architecture, BTE is placed after the last of the max-pooling operation and before the flatten operation.

B. Bottleneck Transformer Encoder (BTE)

Modern networks concentrate on improving the self-attention mechanism oriented to the success of the Vision Transformer (ViT) [16] technique. However, it brings high operations (multiplication and addition), leading to inadequate implementation for vision tasks on low-cost devices. This work proposes a Bottleneck Transformer Encoder (BTE) to alleviate the mentioned above problem. Similar to the general Transformer Encoder, this function will transform a tensor input $\mathbf{X}$ of shape $H \times W \times C$ into a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) by applying a 1×1 convolution operation with Rectified Linear Unit (ReLU). However, a reduction channel r and multi-head mechanism with $NumHead$ are applied in this work to produce a slimmer tensor $H \times W \times ((\frac{C}{NumHead})/r)$, followed by reshaping operation to produce a $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrix with shape $HW \times ((\frac{C}{NumHead})/r)$. After that, we compute these matrices using scaled dot-product attention shown in Fig. 2, which is described as follows:

$$SDPA\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \qquad (1)$$

where d_k is a scaling factor to control the softmax temperature and T is a transpose matrix operation. To reform the output matrix shape into $H \times W \times ((\frac{C}{NumHead})/r)$, the reshaping operation is also performed. Then, a concatenation operation is used to merge all output from all heads, followed by 1×1 convolution operation with a dropout (DO) and Rectified Linear Unit (ReLU) activation layers to restore the number of channels according to the input tensor $\mathbf{X}$ with shape $H \times W \times C$ that represents a bottleneck mechanism. Once again, a linear projection using 1×1 convolution operation with a LeakyReLU activation is performed, integrating with a residual connection [8] and Layer Normalization (LN) [17] around the multi-head attention and last convolution layer.

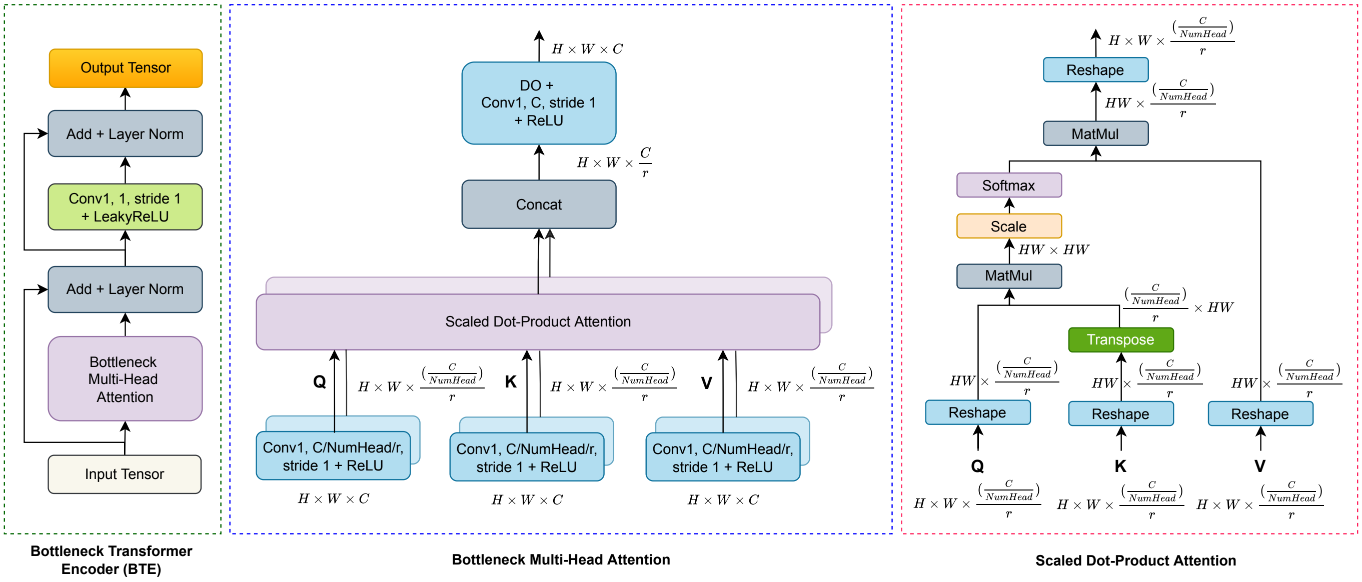

Output Tensor
Add + Layer Norm
Conv1, 1, stride 1 + LeakyReLU
Add + Layer Norm
Bottleneck Multi-Head Attention
Input Tensor
Bottleneck Transformer Encoder (BTE)
$H \times W \times C$
DO + Conv1, C, stride 1 + ReLU
$H \times W \times \frac{C}{r}$
Concat
Scaled Dot-Product Attention
Q
K
V
$H \times W \times \frac{(\frac{C}{NumHead})}{r}$
Conv1, C/NumHead/r, stride 1 + ReLU
$H \times W \times C$
Bottleneck Multi-Head Attention
Reshape
$HW \times \frac{(\frac{C}{NumHead})}{r}$
MatMul
Softmax
Scale
$HW \times HW$
MatMul
$\frac{(\frac{C}{NumHead})}{r} \times HW$
Transpose
Reshape
Reshape
Reshape
Q
K
V
Scaled Dot-Product Attention

Fig. 2. The proposed Bottleneck Transformer Encoder (BTE).

C. Classification Module

The second stage of the proposed gender recognition architecture is a classification module that consists of two fully connected (FC) layers with 128 and 2 units, respectively. The first FC uses batch normalization and ReLU (Rectified Linear Unit) activation to avoid gradient issues. The second FC applies a dropout mechanism [15] to prevent overfitting problems and Sigmoid activation to generate the input into the prediction decision. This classification module uses to compute the probability of each gender class. It directs in discriminating whether the face indicates a male or female.

III. Implementation Configuration

In this experiment, UTKFace, LFW, and Adience datasets are used to train the proposed architecture. This work applies 10^{-2} initial learning rate with a batch size of 512 on 300 epochs. This work uses Tensorflow-Keras as a framework and an Nvidia GeForce GTX 1080Ti with 11GB GPUs as an accelerator. This training applies an Adam optimizer to optimize the weight on the Binary Cross-Entropy loss. A reducing learning rate mechanism with a decrease of 75% is also performed when accuracy does not improve every 20 epochs. We calculate the speed of the proposed architecture on the Intel Core i7-9750H CPU@2.6 GHz with 20GB RAM.

IV. Experimental Results

A. Evaluation on Datasets

1) UTKFace: There are more than 23,000 facial images on this dataset labeled in gender, age, and ethnicity, covering many variations such as expression, pose, age, illumination, etc. The proposed architecture is trained on 70% of this dataset, while the remains for testing sets. The proposed architecture achieves 92,78% in validation accuracy on this dataset with only 555,770 parameters. Table I shows that the accuracy outperforms the other state-of-the-art architectures, which differed by 0.12% from the second-best. The proposed architecture also generates fewer parameters compared to the

TABLE I
Evaluation results on UTKFace, LFW, and Adience datsets

Architecture	Number of Parameters	Validation Accuracy (%)
Evaluation on UTKFace		
Hamdi & Moussaoui [5]	530,034	89.97
EfficientNetV2B1 [18]	77,54,679	90.31
Krishnan et al. (VGG-19) ([19]	143,667,240	91.50
Krishnan et al. (ResNet-50) [19]	25,636,712	91.60
Krishnan et al. (VGG16) [19]	138,357,544	91.90
Shahzeb et al. [20]	14,715,201	91.94
Savchenko (Modified MobileNet) [21]	3,491,521	91.95
SufiaNet [11]	226,574	92.05
MPConvNet [10]	659,650	92.32
MudaNet [12]	674,760	92.66
Proposed	**555,770**	**92.78**
Evaluation on LFW		
Althnian et al. [22]	15,473,190	72.50
Rouhsedaghat et al. [23]	16,900	94.63
SufiaNet [11]	226,574	95.66
MudaNet [12]	674,760	96.22
MPConvNet [10]	659,650	96.30
Greco et al. [24]	3,538,984	**98.73**
Proposed	**555,770**	**96.50**
Evaluation on Adience		
Althnian et al. [22]	15,473,190	83.30
Greco et al. [24]	3,538,984	84.48
SufiaNet [11]	226,574	84.60
MudaNet [12]	674,760	84.85
Saha et al. [25]	3,190,913	84.94
MPConvNet [10]	659,650	85.67
Opu et al. [26]	210,050	85.77
Proposed	**555,770**	**85.86**

TABLE II

MODEL ANALYSIS OF THE PROPOSED ARCHITECTURE ON UTKFACE DATASET

Settings	Number of Parameters	MFLOPs	Validation Accuracy (%)
The Backbone without BTE	549,218	23.92	92.46
The Backbone with BTE	555,770	24.80	92.78

second-best.

2) LFW: This dataset includes more than 13,000 face images, divided into testing (30%) and training (70%). This dataset has two labels, females and males, with a significant imbalance between females (23%) and males (77%). Table I describes that the proposed architecture achieves competitive validation accuracy with 96,50% on this dataset. It becomes second-best compared to the other state-of-the-art architectures. However, the proposed architecture generates far fewer parameters.

3) Adience: This dataset contains more than 26,000 face images labeled in gender and age, encompassing several variations such as age, pose, noise, lighting, appearance, etc. Some pre-processing is established on this dataset in this experiment, such as eradicating data with missing values and generating 17,492 face images. Following the two other datasets mentioned before, this data set is also split into 30% and 70% as a testing and training set, respectively. As a result, the proposed architecture gains a validation accuracy of 85.86%, shown in Table I. This result surpasses the other state-of-the-art architectures, which differed by 0.09% from the second-best.

B. Ablation Study

1) Model Analysis: This analysis is established on the UTKFace dataset to investigate the impact of the proposed BTE on recognition accuracy in this work. Table II reveals that the proposed BTE can improve the validation accuracy by 0.32%. With this increased performance, the proposed BTE only generates 6,552 and 0.88 more additional parameters and MFLOPs, respectively.

2) Channel Reduction Analysis: This work arranges this analysis to examine the optimal channel reduction value on the proposed BTE concerning the recognition performance. Table III shows that using a channel reduction value of one does not enhance performance. The proposed BTE with a channel reduction value of eight produces the highest validation accuracy in this work.

C. Runtime Efficiency

The proposed architecture, with only 555,770 parameters and 24.8 MFLOPs, is principally developed to perform on a CPU device in real-time scenarios. As a result, the proposed architecture can operate 345.69 FPS for gender recognition and 150.49 FPS when integrated with an efficient face detector, dubbed LWFCPU [27], as shown in Table IV. The proposed gender recognizer based on a human face outperforms the other public and light architecture used in this case. Fig. 3 demonstrates the gender recognition result of the proposed gender recognizer based on a human face on a CPU. The male and female faces are indicated with fuchsia and yellow bounding box, respectively.

TABLE III

CHANNEL REDUCTION ANALYSIS OF THE PROPOSED BTE ON UTKFACE DATASET

Value of Reduction r	Number of Parameters	MFLOPs	Validation Accuracy (%)
-	570,274	27.18	92.49
2	561,986	25.72	92.53
4	557,842	25.09	92.57
6	556,288	24.87	92.63
8	**555,770**	**24.80**	**92.78**
10	555,252	24.73	92.60

- indicates that the BTE is performed without channel reduction r

TABLE IV

RUNTIME EFFICIENCY ON A CPU CONFIGURATION

Architecture	Number of Parameters	MFLOPs	GR (FPS)	FD + GR (FPS)
VGG16 + BN	39,782,722	2,290	42.42	36.35
ResNet50V2	23,568,898	571	57.15	46.67
MudaNet [12]	674,760	69	58.57	48.07
MobileNetV2	2,260,546	50	118.56	80.43
SufiaNet [11]	226,574	22	127.50	84.90
MPConvNet [10]	659,650	67	265.07	131.85
Proposed	**555,770**	**25**	**345.69**	**150.49**

GR indicates the Gender Recognition

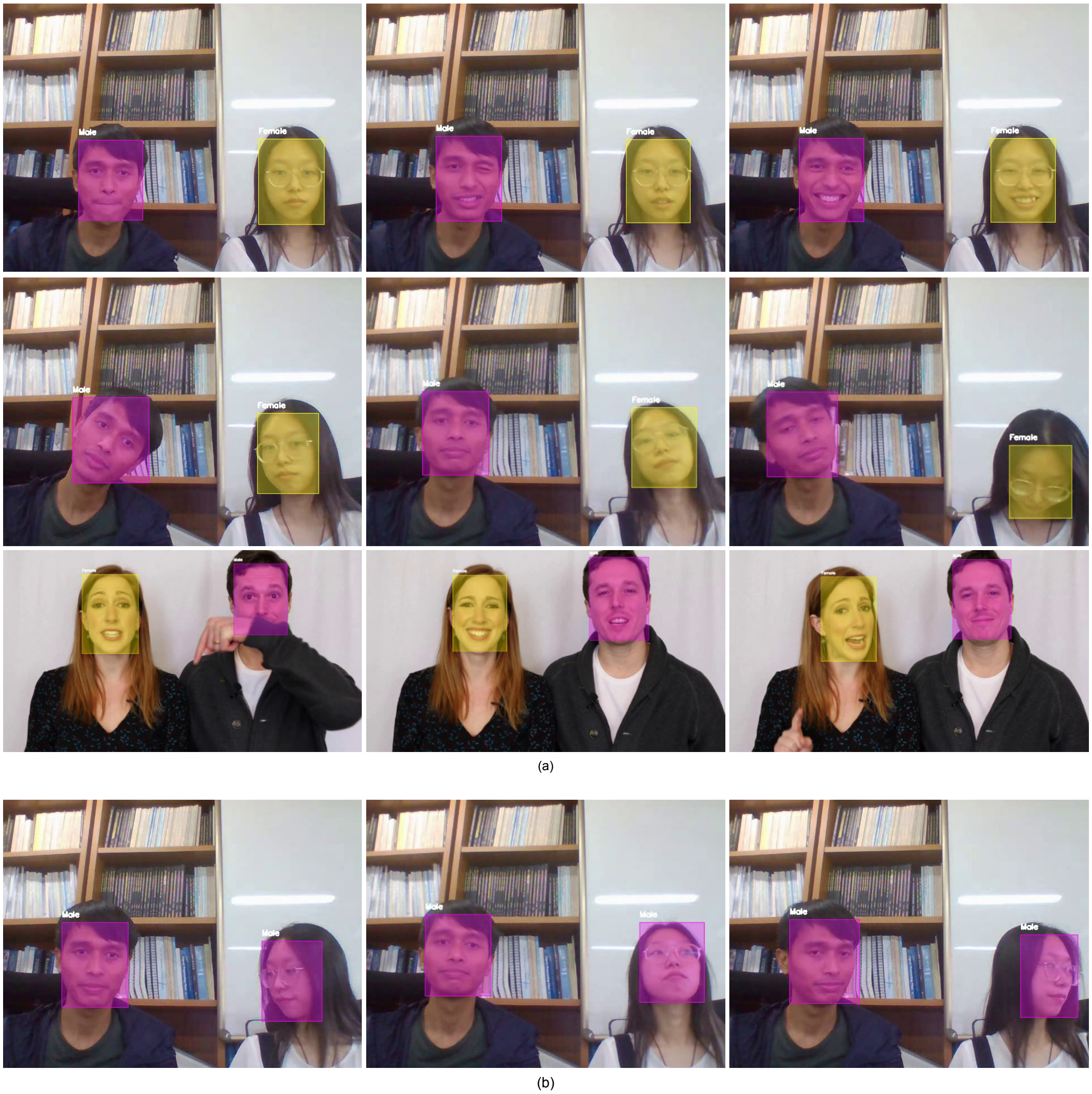
FD + GR indicates the Gender Recognition integrated with Face Detection

D. Limitations

The UTKFace dataset used to train the proposed architecture for the gender recognizer in this work covers variations in the pose. However, it does not have face instances with extreme yaw and pitch pose. It pushes the recognizer resulting in an incorrect prediction in some cases of a face with extreme yaw and pitch pose, as sown in Fig. 3 (b). In this case, a female face is recognized as a male face.

V. CONCLUSION

This work proposes a gender recognizer based on a human face. It offers a light CNN as a feature extractor supported by a Bottleneck Transformer Encoder (BTE) to learn the global representation of the feature maps efficiently. This encoder will enhance the quality of the feature map outcoming from the feature extractor. The proposed network achieves competitive accuracy compared to the state-of-the-art network on three face gender dataset benchmarks, namely UTKFace, LFW, and Adience. The proposed gender recognizer based on a human face consisting of the proposed gender recognition integrated with a face detector can operate in real-time on the CPU with 150.49 frames per second. In future work, other methods will be explored to solve the limitation in this work and to increase the accuracy and speed of the gender recognizer through a human face.



(a)

(b)

Fig. 3. The correct recognition result (a) and the incorrect recognition results (b) of the proposed gender recognizer based on a human face.

Acknowledgment

This result was supported by ”Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

References

[1] P. Foggia, A. Greco, A. Saggese, and M. Vento, “Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition,” *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105651, 2023.

[2] P. Foggia, A. Greco, G. Percannella, M. Vento, and V. Vigilante, “A system for gender recognition on mobile robots,” in *Proceedings of the 2nd international conference on applications of intelligent systems*, 2019, pp. 1–6.

[3] M. A. Uddin, M. S. Hossain, R. K. Pathan, and M. Biswas, “Gender recognition from human voice using multi-layer architecture,” in *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2020, pp. 1–7.

[4] A. Greco, A. Saggese, and M. Vento, “Digital signage by real-time gender recognition from face images,” in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. IEEE, 2020, pp. 309–313.

[5] S. Hamdi and A. Moussaoui, “Comparative study between machine and deep learning methods for age, gender and ethnicity identification,” in *2020 4th International Symposium on Informatics and its Applications (ISIA)*. IEEE, 2020, pp. 1–6.

[6] Z. Zhang, Y. Song, and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[7] R. Ranjan, V. M. Patel, and R. Chellappa, “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose

estimation, and gender recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

[9] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[10] A. Priadana, M. D. Putro, and K.-H. Jo, "An efficient face gender detector on a cpu with multi-perspective convolution," in *2022 13th Asian Control Conference (ASCC)*. IEEE, 2022, pp. 453–458.

[11] A. Priadana, M. D. Putro, C. Jeong, and K.-H. Jo, "A fast real-time face gender detector on cpu using superficial network with attention modules," in *2022 International Workshop on Intelligent Systems (IWIS)*. IEEE, 2022, pp. 1–6.

[12] A. Priadana, M. D. Putro, X.-T. Vo, and K.-H. Jo, "A facial gender detector on cpu using multi-dilated convolution with attention modules," in *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2022, pp. 190–195.

[13] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Transactions on information forensics and security*, vol. 9, no. 12, pp. 2170–2179, 2014.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] P. Vidyarthi, S. Dhavale, and S. Kumar, "Gender and age estimation using transfer learning with multi-tasking approach," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*. IEEE, 2022, pp. 1–5.

[19] A. Krishnan, A. Almadan, and A. Rattani, "Understanding fairness of gender classification algorithms across gender-race groups," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2020, pp. 1028–1035.

[20] M. Shahzeb, S. Dhavale, D. Srikanth, and S. Kumar, "Dcnn-based transfer learning approaches for gender recognition," in *International Conference on Data Management, Analytics & Innovation*. Springer, 2023, pp. 357–365.

[21] A. V. Savchenko, "Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output convnet," *PeerJ Computer Science*, vol. 5, p. e197, 2019.

[22] A. Althnian, N. Aloboud, N. Alkharashi, F. Alduwaish, M. Alrshoud, and H. Kurdi, "Face gender recognition in the wild: an extensive performance comparison of deep-learned, hand-crafted, and fused features with deep and traditional models," *Applied Sciences*, vol. 11, no. 1, p. 89, 2020.

[23] M. Rouhsedaghat, Y. Wang, X. Ge, S. Hu, S. You, and C.-C. J. Kuo, "Facehop: A light-weight low-resolution face gender classification method," in *International Conference on Pattern Recognition*. Springer, 2021, pp. 169–183.

[24] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, vol. 8, pp. 130 771–130 781, 2020.

[25] A. Saha, S. N. Kumar, and P. Nithyakani, "Age and gender prediction using adaptive gamma correction and convolutional neural network," in *2023 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2023, pp. 1–5.

[26] M. N. I. Opu, T. K. Koly, A. Das, and A. Dey, "A lightweight deep convolutional neural network model for real-time age and gender prediction," in *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*. IEEE, 2020, pp. 1–6.

[27] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot," in *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, 2020, pp. 94–99.