# Age Group Recognizer based on Human Face Supporting Smart Digital Advertising Platforms

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

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

2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE) | 979-8-3503-9971-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIE51358.2023.10227959

***Abstract*—Smart digital advertising platforms have been widely employed in public areas in big cities. An age group recognizer is indispensable to support these platforms in providing relevant advertisements for each audience. These platforms also demand a recognizer that can run properly at the minimum on a CPU device to degrade the budget during system procurement. This study proposes an age group recognizer based on human faces to predict the age group of the audience's face using an efficient architecture containing a light backbone. This work offers a residual mini multi-level module integrating two grouped convolution layers with diverse frequency levels to extract exclusive facial features maintained by residual operation. In order to improve the feature map's quality, a deep lite attention module is proposed, consisting of the deep channel and lite spatial attention part. The architecture generates few parameters with cheap operation and achieves competitive performance on the benchmark datasets. In addition, the architecture integrated with face detection, as a recognizer, can perform fast on a CPU configuration in real-time with 144 frames per second.**

***Index Terms*—smart digital advertising, real-time recognizer, face age group recognizer**

## I. Introduction

As a unique demographic feature, facial attributes like age are essential components and widely used in digital marketing, intelligent commerce [1], and digital advertising platforms [2]. This information can be used to optimize the advertising process by segmenting the audience. A digital advertising platform can apply age recognition to perform audience profiling to provide relevant ads. Therefore, the advertisements can be more targeted to reach potential customers. Usually, performing age recognition of a person upon facial image, instead of predicting an exact age [3], tends to estimate an age range or group [4]–[7]. People in a particular age group tend to have similar needs or interests. Moreover, recognizing ages based on groups can minimize incorrect prediction because the target class will be less than trying to predict the exact age.

Nowadays, due to the exceptional performance of the Convolutional Neural Network (CNN), most works applied this method to perform human face-based age recognition. Chen et al. [8] modified AlexNet as a baseline and presented Attribute-Region Association Network (ARAN) to recognize age from a face. The architecture generates more than 400 million parameters. Another researcher, Li et al. [9] offered BridgeNet based on CNN architecture that consists of local regressors that learn aware continuity weights, generating 120 million parameters. Shin et al. [10] adopted VGG16 architecture as an encoder and designed a moving window regression algorithm to perform facial age estimation.

Recently, many works have focused on designing a lightweight CNN architecture to consider efficient computation. An integrated CNN architecture [11] is proposed, combined with deep distinguishable random forest techniques to estimate age based on a face. The architecture generates 14 million parameters. Wang et al. [3] offered a novel architecture that included a fusion network and an attention module to determine the subjects' ages and dynamically find and organize age-specific patches. The architecture also generates about 14 million parameters. Another work [12] offered an efficient CNN architecture that consists of two perspective convolution branches and a novel attention module. The architecture only generates 459,347 parameters.

In the practical implementation of digital advertising platforms, a human face-based age recognizer is also demanded to operate at the minimum on CPU devices to decrease implementation costs [13], [14]. These platforms require an efficient CNN architecture to perform age recognition on a CPU device, especially can run fast in real-time while maintaining accuracy.

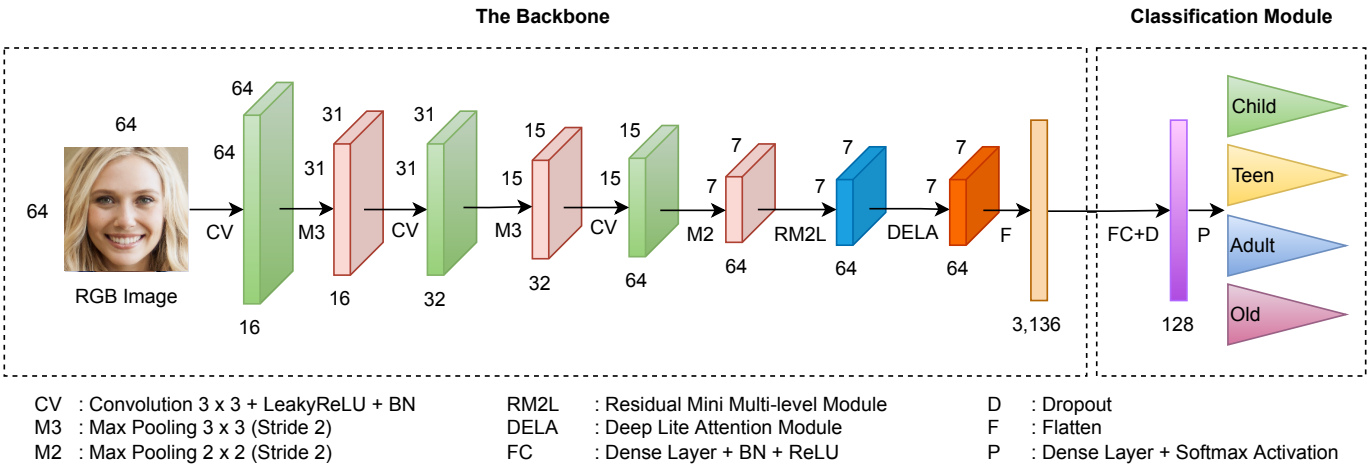


Fig. 1. The proposed face-based age group recognition architecture of the recognizer.

Based on the issue, this work proposes a face-based age group recognizer with a scant parameter. This recognizer is suitable for implementation in real-time on a CPU device.

A human face-based age group recognizer (AggerCPU) proposed a lightweight CNN architecture with a very inexpensive operation. A pair of novel modules dubbed the Residual Mini Multi-level (RM2L) convolutional module improved from [15], and the Deep Lite Attention (DELA) module, are developed to enhance the CNN architecture employed by the recognizer. RM2L gradually refines the distinctive features guarded by residual mechanisms. This module generates fewer parameters than the standard convolution. The DELA is employed to escalate and confirm the essential features established on the channel and spatial maps. Therefore, the recognizer can perform more onward and faster to recognize age based on a face. This work presents the main contributions summarized as follows:

1) A face-age group recognizer presents a lightweight CNN architecture with low computation to recognize age based on a face applied to support smart digital advertising. It achieves very competitive mean absolute error (MAE) and accuracy on two datasets, UTKFace [16] and FG-NET [17], compared with other CNN architectures.
2) A novel efficient backbone with a Residual Mini Multi-level module (RM2L) is proposed that rapidly extracts the exclusive facial features, producing low parameters and very inexpensive computation. It incorporates two grouped convolution blocks with various frequency levels to improve the diversity of the feature map, maintained by residual operation.
3) A deep lite attention module (DELA) is offered as an enhancement tactic to catch the essential features based on channel and spatial maps. It efficiently encourages the feature map grade from the high-level features, improving the recognizer performance.

## II. Proposed Architecture

The proposed face-based age group recognition architecture of the proposed recognizer consists of an efficient backbone and classification module, as shown in Fig. 1. This architecture generates 486,822 parameters.

### A. The Efficient Backbone

The backbone module employs a sequential convolution layer to extract facial features from a face. This scheme performs three times of convolution layers with the same $3 \times 3$ filter size with a small number of channels and grows from 16 to 32 and then 64. This scheme makes the architecture produce a few parameters and slight computation. This architecture applies Leaky Rectified Linear Unit activation, commonly called Leaky ReLU, in every convolution layer followed by a batch normalization (BN) to deal with the gradient issue [18]. Three max-pooling operations are also used to downsample the feature map. This architecture applies two times of $3 \times 3$ and one time of $2 \times 2$ sizes with strides 2. The lack of applying a few convolution layers cause the network over shallow. As a response, we propose a Residual Mini Multi-level (RM2L) module as an additional efficient extractor and a Deep Lite Attention (DELA) module to improve the feature map's quality. These modules are put after the final of the max-pooling operation and before the flatten operation.

### B. The Residual Mini Multi-level Module (RM2L)

To improve the capability to extract the facial feature on the backbone module, we propose a Residual Mini Multi-level module (RM2L). This module locates after the last max-pooling layer. Motivated by [15], this module intuitively combines two feature maps with distinct frequency levels, as displayed in Fig. 2. Firstly, it divides the input feature map $\mathbf{X}$ into two pieces $[\mathbf{X_1}, \mathbf{X_2}]$. The first part $\mathbf{X_1}$ is applied a convolution with $3 \times 3$ filter size to acquire low-level features, represented as follows:

$$F_l(\mathbf{X_1}) = LReLU(BN(C3(\mathbf{X_1}))), \quad (1)$$

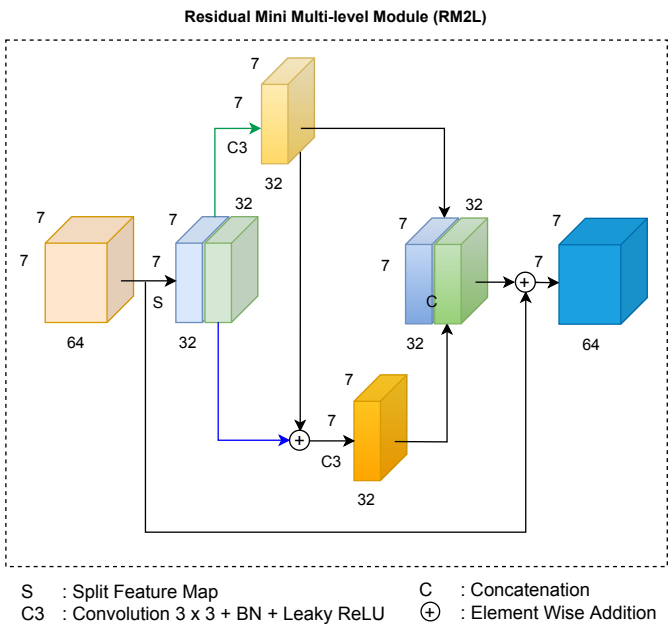


Fig. 2. The proposed Residual Mini Multi-level module.

where $C3$ is a convolution layer with $3 \times 3$ filter sizes, $BN$ is batch normalization, and $LReLU$ is Leaky ReLU activation. Another level feature is created by applying a convolution layer with $3 \times 3$ filter size to the aggregation of low-level features generated on the first part by element-wise addition with the second part $\mathbf{X_2}$, which is illustrated as follows:

$$F_h(\mathbf{X}) = LReLU(BN(C3((F_l(\mathbf{X_1})) + \mathbf{X_2}))), \quad (2)$$

Different from [15], RM2L does not apply a convolution layer after concatenating the two feature maps $F_l$ and $F_h$ to minimize the number of parameters. As a reserve, if the module unsuccessful produces improvement to extract the facial features, this module applies shortcut connections as a residual mapping [19] to construct the output, which is described as follows:

$$RM2L(\mathbf{X}) = \mathbf{X} + (F_l(\mathbf{X_1}) \oplus F_h(\mathbf{X})), \quad (3)$$

where $\oplus$ is a concatenation operation.

### C. The Deep Lite Attention Module (DELA)

The proposed architecture employs a shallow backbone that makes it weak to distill the specific features. This architecture needs an attention module as a booster before feeding the feature map into the classification module. A Deep Lite Attention module (DELA) is proposed to catch essential facial features. This module consists of two parts, Deep Channel Attention (DCA) and Lite Spatial Attention (LSA), paired sequentially, as shown in Fig. 3.

Inspired by [12], modified from [20], DCA inserts a depthwise convolution layer with $3 \times 3$ filter size before the pooling operation to give a chance for the individual channel to sharpen and deepen learning without being influenced by the others, that makes this part called deep channel attention. Unlike on [12], this module only applies global average-pooling and one convolution layer with $1 \times 1$ filter size before executing a channel-wise multiplication in the last stage to reduce the computation and number of parameters. This layer is attended by sigmoid activation to compute the independent attention weights following [20]. The proposed deep channel attention module is represented as follows:

$$DCA(\mathbf{X}) = \mathbf{X} * \sigma(C1(GAE(DC3(\mathbf{X})))), \quad (4)$$

where $\mathbf{X}$ indicates as an input of the proposed deep channel attention module, $DC3$ is a depthwise convolution with $3 \times 3$ filter sizes, $GAE$ is a global average-pooling operation each channel, $C1$ is a convolution layer with $1 \times 1$ filter sizes and $\sigma$ is a sigmoid activation.

The disadvantage of applying only a channel attention module is that this scheme neglects the essence of spatial information. Therefore, a lite spatial attention (LSA) module is proposed. Motivated by the spatial attention module on [21], this module uses a concatenation of global max-pooling and global average-pooling operations across the channel to aggregate spatial features map. Different from [21] that applies a convolution operation with a $7 \times 7$ filter size to produce a 2D spatial attention map, this module uses a $1 \times 1$ filter size to reduce the number of parameters and computation, followed by sigmoid activation and spatial-wise multiplication with the input feature map. The proposed lite spatial attention module is illustrated as follows:

$$LSA(\mathbf{X}) = \mathbf{X} * \sigma(C1(GMA(\mathbf{X}) \oplus GAA(\mathbf{X}))), \quad (5)$$

where $\mathbf{X}$ indicates as an input of the proposed lite spatial attention module. $GMA$ and $GAA$ are global max-pooling and global average-pooling operations across the channel. $C1$ is a convolution operation with $1 \times 1$ filter sizes and $\sigma$ is a sigmoid activation.

### D. The Classification Module

In the last stage, facial features generated by the backbone module will be fed to the classification module that will compute the probability of each age group class. This module guides in determining whether the face represents a child, teen, adult, or old. This module consists of two fully connected layers with 128 and 4 units, respectively. The first layer uses batch normalization, ReLU (Rectified Linear Unit) activation, and dropout mechanism to prevent gradient and overfitting problems. The second layer uses Softmax activation to generate the input into the prediction decision.

### E. The Face Detector

The proposed recognizer consists of two primary operations, face detection and face-based age recognition, integrated to perform face-based age recognition from an image. Face detection is employed to capture the face region as the Region of Interest (RoI) from an image before feeding it to the face-based age recognition network. An efficient face detection operation is demanded to support the face-based age recognizer

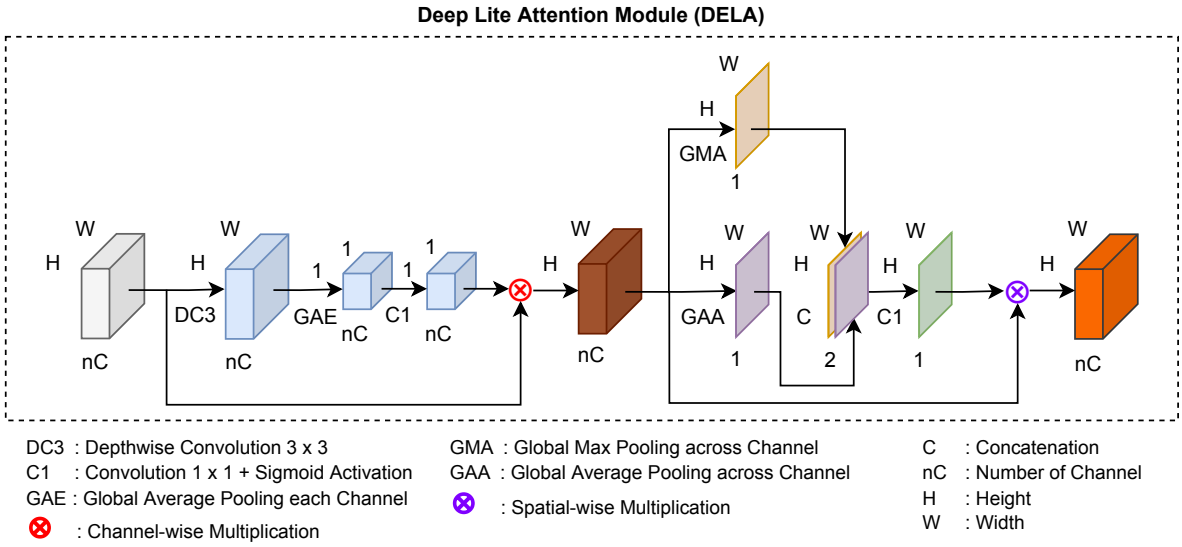


Fig. 3. The proposed Deep Lite Attention Module (DELA).

performing fast in real-time cases. Therefore, this recognizer employs an efficient face detector named LWFCPU [19] that only generates a number of parameters. The face region, the outcome of the face detector, will be cropped and resized to a precise size suitable for the face-based age recognition architecture input.

## III. Implementation Setup

In this work, the proposed architecture is trained on UTK-Face and LFW datasets by using $10^{-3}$ initial learning rate on an Nvidia GeForce GTX 1080Ti with 11GB GPUs with a batch size of 256 on 300 epochs by using Tensorflow and Keras framework. If the accuracy does not enhance every 20 epochs, the learning rate will be cut as much as 75%. This experiment uses the Adam optimizer to optimize the weight on the Categorical Cross-Entropy loss. To examine the speed in frame per second or FPS of the proposed architecture and the recognizer, we employ an Intel Core i7-9750H CPU@2.6 GHz with 20GB RAM.

## IV. Experimental Results

This part explains the dataset evaluation, ablation study, runtime efficiency, and attention modules comparison. In this experiment, three datasets are used to measure the performance evaluation of age group prediction.

### A. Evaluation on Datasets

*1) UTKFace:* This dataset accommodates 23,708 facial images with age variations ranging from 0 to 116. It contains case variations such as pose, illumination, expression, etc. Two settings of this dataset are used for evaluation. For the first regulation, i.e., *Setting I*, complying with previous research [22], [23], the dataset is split into two parts, training (80%) and testing (20%) set. To measure the proposed architecture on this dataset, the Mean Absolute Error (MAE) of the testing set is used for this setting. Table I shows that for Setting I, the proposed architecture with not more than 500K parameters achieves the second-best performance with 4.38 MAE, which differs only by 0.01 from the best one [24]. However, the proposed architecture is much more efficient than [24] based on the number of parameters.

The second regulation, i.e., *Setting II*, is proposed and used for our recognizer. This dataset is separated into a training

TABLE I
The Evaluation Results on UTKFace Dataset with Setting I.

| Architectures | Baseline | Number of Parameters (Million) | MAE↓ |
|---|---|---|---|
| OR-CNN [25] | Manually-designed | - | 5.76 |
| CORAL [22] | ResNet34 | 21 | 5.47 |
| Savchenko [26] | MobileNetV1 | 3.5 | 5.44 |
| LRTI [27] | ResNet34 | 21 | 4.55 |
| Berg et al. [23] | ResNet50 | 24 | 5.14 |
| FCRN [28] | ResNet50 | 24 | 4.47 |
| 2PDG [12] | Manually-designed | 0.46 | 4.44 |
| MWR [24] | VGG16 | 40 | **4.37** |
| **Proposed** | **Manually-designed** | **0.49** | **4.38** |

TABLE II
The Evaluation Results on UTKFace Dataset with Setting II.

| Architectures | Number of Parameters (Million) | VA (%) |
|---|---|---|
| ResNet34 [19] | 21.1 | 87.13 |
| ResNet50 [19] | 23.6 | 88.43 |
| SquezeeNet [29] + BN | 0.74 | 88.47 |
| MobileNetV2 [30] | 2.26 | 90.08 |
| VGG11 [31] + BN | 34.4 | 90.12 |
| 2PDG [12] | 0.46 | 90.12 |
| VGG16 [31] + BN | 39.8 | 90.34 |
| VGG13 [31] + BN | 34.5 | 90.42 |
| **Proposed** | **0.49** | **90.90** |

TABLE III
THE EVALUATION RESULTS ON FGNET DATASET.

| Architectures | Baseline | Number of Parameters (Million) | MAE↓ |
|---|---|---|---|
| LSDML [32] | ResNet101 | 44 | 3.92 |
| ARAN [8] | VGG16 | 414 | 3.79 |
| M-LSDML [32] | ResNet101 | 44 | 3.74 |
| DLDLF [11] | VGG16 | 14 | 3.71 |
| DRF [11] | VGG16 | 14 | 3.41 |
| DAG-VGG16 [33] | VGG16 | 24 | 3.08 |
| DAG-GoogleNet [33] | GoogLeNet | 131 | 3.05 |
| ADPF [3] | Manually-designed | 14 | 2.86 |
| 2PDG [12] | Manually-designed | 0.46 | 2.75 |
| MSFCL [1] | Manually-designed | 15 | 2.71 |
| BridgeNet [9] | Manually-designed | 120 | **2.56** |
| MWR [24] | VGG16 | 40 | **2.23** |
| **Proposed** | **Manually-designed** | **0.49** | **2.71** |

TABLE IV
THE ABLATION STUDY ON UTKFACE DATASET WITH SETTING II.

| M2L | Residual on M2L | DCA | LSA | Number of Parameters | VA (%) |
|---|---|---|---|---|---|
| | | | | 463,268 | 90.08 |
| ✓ | | | | 482,020 | 90.21 |
| ✓ | ✓ | | | 482,020 | 90.25 |
| ✓ | ✓ | ✓ | | 486,820 | 90.60 |
| ✓ | ✓ | ✓ | ✓ | 486,822 | 90.90 |

(90%) and validation (10%) set. Regarding class targets, they can be set according to the purpose of the system. In this scenario, this dataset is divided into four age groups, 0-11 as the child group, 12-17 as the teen group, 18-60 as the adult group, and 61-116 as the old group. To measure the proposed architecture on this dataset, Validation Accuracy (VA) is used for this setting. Table II shows that for Setting II, the proposed architecture achieves 90.90% VA, which outperforms the other lightweight and deep CNN architectures.

*2) FGNET:* The dataset accommodates 1,002 facial images from 82 subjects. The dataset contains case variations such as pose, expression, and illumination. Complying previous studies [11], [32], this dataset applies the k-fold cross-validation and leave-one-person-out (LOPO) approaches. One subject's facial images from each fold are used for testing, while the others are used for training. This evaluation procedure executes 82 fold representing 82 subjects. Due to the different number of instances from each person in the dataset, the number of training and testing sets is various for each fold. This evaluation calculates the results based on the average values according to the MAE metric. Table III shows that the proposed architecture achieves the third-best performance with 2.71 MAE, which differs by 0.48 and 0.15 from the best [24] and second-best [9], respectively. Nonetheless, the proposed CNN architecture produces parameters distant below both.

### *B. Ablation Study*

This section elucidates the investigation of the performance of each proposed module. Firstly, each module is pulled one by one from the proposed architecture on the UTKFace dataset with Setting II. Then calculate its performance compared based on the number of parameters to show the impact of the existence of each module. Table IV shows the result reports of this investigation, which is based on the VA metric. The report shows that employing the mini multi-level module and adding a residual connection can enhance classification validation accuracy by 0.13% and 0.04%, respectively. Further, putting the deep channel and lite spatial attention module can also improve classification validation accuracy by 0.35% and 0.30%, respectively.

TABLE V
THE RUNTIME EFFICIENCY ON THE SAME CPU CONFIGURATION.

| Architectures | Number of Parameters | MFLOPs | AGR (FPS) | AGR + FD (FPS) |
|---|---|---|---|---|
| VGG16 [31] + BN | 39,782,722 | 2,290 | 42.20 | 36.25 |
| VGG13 [31] + BN | 34,476,100 | 1,610 | 50.97 | 42.65 |
| VGG11 [31] + BN | 34,421,892 | 1,270 | 55.12 | 45.37 |
| ResNet50 [19] | 23,595,908 | 633 | 56.02 | 46.02 |
| ResNet34 [19] | 21,105,284 | 217 | 80.70 | 61.28 |
| MobileNetV2 [30] | 2,263,108 | 50.1 | 121.00 | 81.79 |
| SquezeeNet [29] + BN | 736,340 | 83.3 | 230.67 | 122.19 |
| 2PDG [12] | 459,476 | 64.5 | 316.23 | 144.00 |
| **Proposed** | **486,822** | **40.6** | **330.66** | **144.49** |

AGR indicates the Age Group Recognition
AGR + FD indicates the Age Group Recognition integrated with Face Detection

### *C. Runtime Efficiency*

The practical application emphasizes a recognizer to perform in real-time on CPU configuration to cut the budget during system procurement. The proposed architecture can work efficiently in real-time on a CPU by employing only 486,822 parameters and 40.6 MFLOPs. It can perform 330 frames per second in classifying the age group of the human face and 144 frames per second in identifying the age group based on the human face integrated with face detection, as shown in Table V. The proposed recognizer becomes the fastest compared to other competitors. Fig. 4 shows the recognition result of the proposed recognizer, in which the red, blue, yellow, and green bounding box represents an old's, an adult's, a teen's, and a child's face, respectively. This recognizer will not save the audience's face images after performing face-based age group recognition to keep the anonymity and privacy of the audience.

### *D. Limitations*

The Agger-CPU recognizer is trained on the UTKFace dataset that covers pose variation. However, the dataset does not contain many instances for every pose variation, especially faces with extreme yaw pose. The dataset also does not include an example of a face in an occluded case. Consequently, in some cases, the recognizer provides an inaccurate prediction in predicting the age group of an occluded face and a face with extreme yaw pose and its variation, illustrated in Fig. 4 (b).

### *E. Attention Modules Comparison*

DELA consisting of DCA and LSA is also compared with other commonly used attention strategies such as Convolutional Block Attention module (CBAM) [21], Bottleneck


(a)


(b)

Fig. 4. The correct (a) and incorrect (b) prediction result of the Agger-CPU.

TABLE VI
THE COMPARISONS OF DIFFERENT ATTENTION MODULES APPLIED ON THE PROPOSED ARCHITECTURE ON UTKFACE DATASET WITH SETTING II.

| Attention Modules | Number of Parameters | MFLOPs | AGR (FPS) | AGR + FD (FPS) | VA (%) |
|---|---|---|---|---|---|
| BAM [34] | 489,481 | 41.2 | 320.14 | 141.92 | 90.51 |
| SE [20] | 482,532 | 40.5 | 342.64 | 147.29 | 90.55 |
| CBAM [21] | 482,618 | 40.5 | 335.86 | 145.76 | 90.77 |
| **DELA (ours)** | **486,822** | **40.6** | **330.66** | **144.49** | **90.90** |

AGR indicates the Age Group Recognition
AGR + FD indicates the Age Group Recognition integrated with Face Detection

Attention module (BAM) [34] and Squeeze-and-Excitation (SE) [20]. In the proposed architecture, the attention module is located following the Residual Mini Multi-level module on the backbone to perform a fair comparison. Table VI shows that the proposed architecture with DELA is superior, compared to the proposed architecture with BAM, SE, or CBAM, based on the validation accuracy, which differs by 0.39%, 0.35%, and 0.13%, respectively. According to the speed, the proposed architecture with DELA is not far behind the best and second-best, with a difference of only three and one frame per second, respectively, when integrated with face detection.

## V. CONCLUSION

This paper proposes a human face-based age group recognizer using a lightweight architecture with the cheap operation. This study provides an efficient backbone with Residual Mini Multi-level (RM2L) and Deep Lite Attention (DELA) modules. The offered modules help the network to rapidly extract exclusive facial features and improve its quality with few parameters and less computation cost. The architecture achieved competitive performance on the benchmark datasets compared to other competitors. In addition, the recognizer integrated with face detection, as a recognizer, can perform 144 frames per second to recognize the face-based age group

in real-time on a CPU device. Moreover, the proposed architecture with DELA also achieves the best accuracy compared to the proposed architecture with BAM, SE, or CBAM. In the future study, other methods, such as the transformer, will be addressed to improve the recognizer performance. Another dataset will also be explored to overcome the limitation of the prediction result. It is also possible to perform face detection and face-based age group recognition with a single efficient architecture.

## 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] M. Xia, X. Zhang, L. Weng, Y. Xu *et al.*, "Multi-stage feature constraints learning for age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2417–2428, 2020.

[2] S. Suman and S. Urolagin, "Age gender and sentiment analysis to select relevant advertisements for a user using cnn," in *Data Intelligence and Cognitive Informatics*. Springer, 2022, pp. 543–557.

[3] H. Wang, V. Sanchez, and C.-T. Li, "Improving face-based age estimation with attention-based dynamic patch fusion," *IEEE Transactions on Image Processing*, 2022.

[4] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[5] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li, "Age group and gender estimation in the wild with deep ror architecture," *IEEE Access*, vol. 5, pp. 22 492–22 503, 2017.

[6] M. T. B. Iqbal, M. Shoyaib, B. Ryu, M. Abdullah-Al-Wadud, and O. Chae, "Directional age-primitive pattern (dapp) for human age group recognition and age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2505–2517, 2017.

[7] A.-T. Mai, D.-H. Nguyen, and T.-T. Dang, "Real-time age-group and accurate age prediction with bagging and transfer learning," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2021, pp. 27–32.

[8] Y. Chen, S. He, Z. Tan, C. Han, G. Han, and J. Qin, "Age estimation via attribute-region association," *Neurocomputing*, vol. 367, pp. 346–356, 2019.

[9] W. Li, J. Lu, J. Feng, C. Xu, J. Zhou, and Q. Tian, "Bridgenet: A continuity-aware probabilistic network for age estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1145–1154.

[10] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 760–18 769.

[11] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. Yuille, "Deep differentiable random forests for age estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 404–419, 2019.

[12] A. Priadana, M. D. Putro, X.-T. Vo, and K.-H. Jo, "An efficient face-based age group detector on a cpu using two perspective convolution with attention modules," in *2022 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*. IEEE, 2022, pp. 1–6.

[13] K. Mishima, T. Sakurada, and Y. Hagiwara, "Low-cost managed digital signage system with signage device using small-sized and low-cost information device," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2017, pp. 573–575.

[14] 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.

[15] M. D. Putro, A. Priadana, D.-L. Nguyen, and K.-H. Jo, "A faster real-time face detector support smart digital advertising on low-cost computing device," in *2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2022, pp. 171–178.

[16] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4352–4360.

[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[18] 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.

[19] 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.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[22] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.

[23] A. Berg, M. Oskarsson, and M. O'Connor, "Deep ordinal regression with label diversity," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 2740–2747.

[24] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 739–18 748.

[25] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 4920–4928.

[26] 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.

[27] M. M. Badr, R. M. Elbasiony, and A. M. Sarhan, "Lrti: landmark ratios with task importance toward accurate age estimation using deep neural networks," *Neural Computing and Applications*, vol. 34, no. 12, pp. 9647–9659, 2022.

[28] G. Chen, J. Peng, L. Wang, H. Yuan, and Y. Huang, "Feature constraint reinforcement based age estimation," *Multimedia Tools and Applications*, pp. 1–22, 2022.

[29] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] H. Liu, J. Lu, J. Feng, and J. Zhou, "Label-sensitive deep metric learning for facial age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 2, pp. 292–305, 2017.

[33] S. Taheri and Ö. Toygar, "On the use of dag-cnn architecture for age estimation with multi-stage features fusion," *Neurocomputing*, vol. 329, pp. 300–310, 2019.

[34] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.