2022 IEEE/ASME International Conference on
Advanced Intelligent Mechatronics (AIM)

A Faster Real-time Face Detector Support Smart Digital Advertising on Low-cost Computing Device

Muhamad Dwisnanto Putro[1], Adri Priadana[2], Duy-Linh Nguyen[3], and Kang-Hyun Jo[4]

***Abstract*— Smart digital advertising requires face detection as the initial stage to recognize the person's attributes by localizing human facial areas. This technology tends to operate with CPU-based systems. The Deep Convolutional Neural Network approach has demonstrated excellent accuracy for face detection work. However, this architecture involves heavy computations and parameters because it uses many filter operations. It causes deep architecture to slow down the detector speed. Moreover, a practical application entails using a detector that can operate in real-time. The recent CPU-based face detectors operate slowly in an integrated system. This study proposes a faster face detector to predict the human face area using efficient architecture robustly. The architecture consists of a light backbone to discriminate distinctive features and a four detection module to predict multiple faces. In order to bridge the three prediction layers, it implements a high-level transition module with a cheap operation. It also offers a new light attentive block to highlight typical facial features at each detection module efficiently. As a result, this detector achieves excellent performance and outperforms other low computing detectors. The proposed detector can fast operate at 112 frames per second on a Core I5 CPU and at 11 frames per second on a Lattepanda device, faster than other competitors.**

I. INTRODUCTION

Nowadays, the overall development of digital content creates a new digital marketing environment. Artificial intelligence also encourages the birth of smart digital advertising that can provide more benefits in the digital marketing process. Smart digital advertising allows promoted content to be displayed dynamically according to the audience, which makes it more targeted [1]. This technology provides an effective mechanism that only shows relevant promotions for targeted consumers that are achieved by personalizing the audience. The demographic information, such as gender and age, is essential information for personalized advertising targeting [2], [3]. The data was obtained by recognizing through the face of the audience, which requires face detection as an initial process [4], [5].

In recent years, Deep Convolutional Neural Networks (DCNNs) have become very popular because it is robust and provide exceptional accuracy for various computer vision works, such as object detection and classification [6]. DCNNs have provided many breakthroughs, especially in face detection works. In [7], a DCNNs model was designed to detect faces that produced high-performance detection. The DCNNs are capable of learning high-level features from faces effectively. Moreover, it succeeded in performing a face verification task and achieved a high accuracy [8]. The common tendency in developing the DCNNs model is to design more visceral and more complex DCNNs models to achieve higher accuracy [9], [10]. However, improving accuracy does not necessarily create lighter and faster networks in many real-world implementations, particularly in supporting real-time cases. The detection and recognition tasks need to be sufficiently performed, especially on low-cost computing devices.

Commonly, designing cheap operation and lightweight CNN model can generate high-speed performance of detection and recognition task [11]. Many previous works tried to make CNN models with lighter weight and low operation for faster processes. In [12], a CNN architecture was proposed to detect faces that produced rapid detection with a light network. It can be able to apply in low-cost computing devices. In another work, [13], the CNN model with a lightweight and deep approach is proposed to perform face detection on a low-cost computing device such as Jetson TX2. It is clear from the literature that most of the face detection work gives excellent results by utilizing the DCNNs. However, these works have not been tested further to provide more satisfactory results. Moreover, face detection is only an initial process to support further work on real-time gender and age recognition on low-cost computing devices.

As an initial process, high-speed face detection is needed to support further work applying the DCNNs. As seen at work [14], gender recognition is performed after the face detection process. Gender recognition is performed to support smart digital advertising such as digital signage. This kind of system requires a low-cost device [15] to reduce budget expenditures. It encourages this system to have an efficient face detector as an initial process of an integrated system. Therefore, it requires a much lighter and cheaper operation to be integrated with further work that implements the CNN and applies to low-cost computing devices.

This work presents a faster face detector based on the vision approach (ACETRON) using a very lightweight and cheap operation. Therefore, it supports integration with further work and application on low-cost computing devices in real-time. Two novel modules, namely Mini Multi-level

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the government (MSIT). (No.2020R1A2C200897212).

M.D. Putro, A. Priadana, D.-L. Nguyen, and K.-H. Jo* are with the Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan, Korea.

*Corresponding author: acejo@ulsan.ac.kr

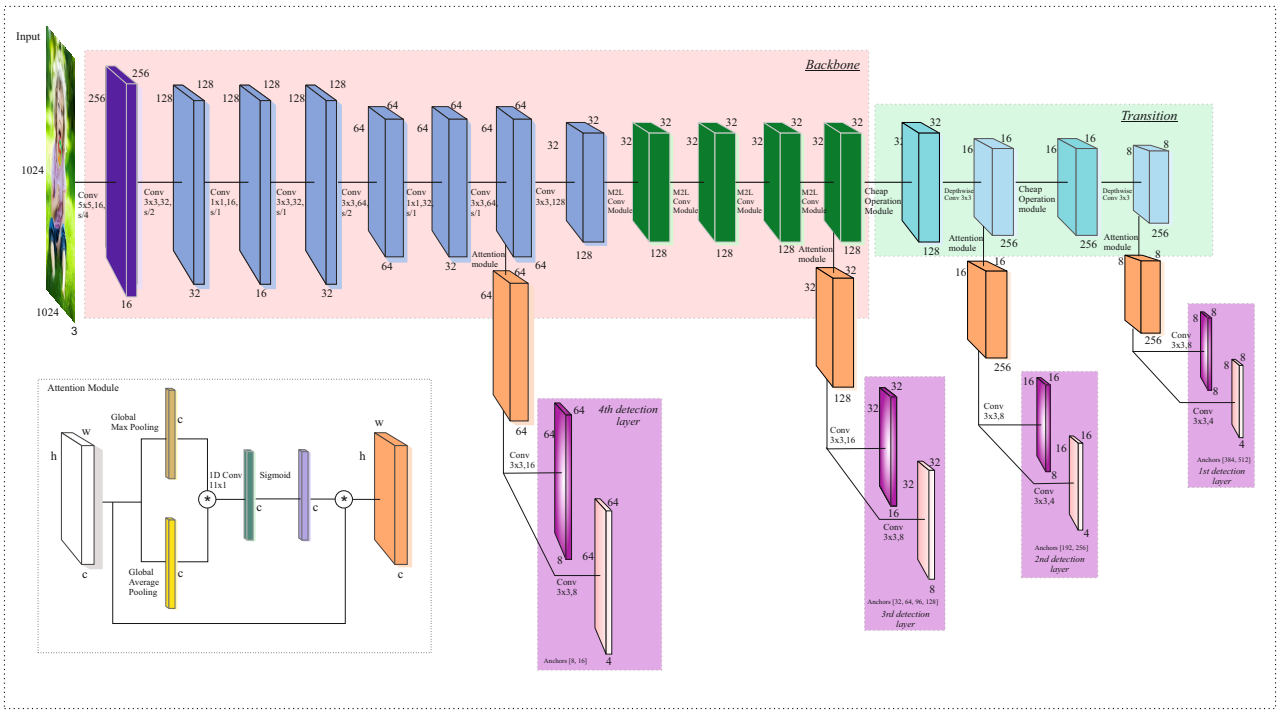

Fig. 1. The proposed architecture of the detector. It uses an efficient backbone module to extract human facial features rapidly. The anchor-based technique is applied in the multi-level detection modules to predict multiple faces based on multi-scale variants. Best viewed in color.

(M2L) convolutional module and Light Attentive module (LIHAT), are designed to improve the DCNNs architecture used by the detector. The M2L gradually extracts the exclusive features with fewer parameters than the standard convolution. The LIHAT is used to escalate and ensure the critical features based on the channel maps. Therefore, the detector can perform more advanced and faster to predict the object's class. This paper offers the main contributions as follows:

1) A novel fast face detector proposes a lightweight CNN architecture with a cheap operation and computation implemented to support smart digital advertising. It emphasizes the efficiency and efficacy of the CNN-based model, which achieves fast real-time speed on a CPU and an edge device.
2) A new efficient backbone module is introduced that rapidly extracts distinctive facial features using Mini Multi-level (M2L), generating few parameters and low-cost computation. This module combines two grouped convolution blocks with different frequency levels to enrich the feature information.
3) A Light Attentive module (LIHAT) is offered as a single enhancement module to capture the essential features according to the channel map from input features. It efficiently boosts the feature map quality at the four-level of the backbone, increasing the detector prediction performance.

II. Related Works

A CNN-based face detection is one of the works that shows outstanding progress, especially in performance. In [16], TinaFace, a backbone based on DCNNs, was proposed to perform face detection. The backbone surpassed most of the other recent more extensive models in detecting faces. In another work [17], the DCNNs combined with YOLOv5 were designed to build a face detector, namely YOLO5Face. Both works were designed to perform quickly on GPU. However, they slowly run when they perform on the CPU.

These days, face detectors specially designed for CPU or low-cost computing devices arise to respond to the shortage described in the previous paragraph. In [18], a Light and Fast Face Detector (LFFD) was built and successfully performed face detection on low-cost computing devices such as Raspberry Pi. The LFFD proposes a model with fewer filters that drives the face detector to run fast. Faceboxes [19] successfully performs face detection on a CPU in real-time. The network architecture used rapidly digest convolution layers (RDCL) and multiple-scale convolution layers (MSCL) to extract the important element and enrich the receptive field.

FlashNet [20], a lightweight network model with few parameters, was designed to detect faces. The detector can detect faces in real-time and achieves high running speed on the CPU. Later, FCPU [21] successfully performs real-time face detection not only on the CPU but also on several low-cost devices such as Lattepanda and Rasberry Pi. In this article, the proposed model constructs the CNN approach

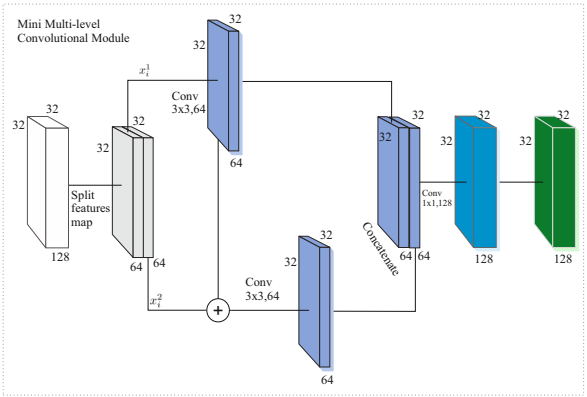

Fig. 2. Mini Multi-level convolutional block is a light extractor feature that combines different receptive frequencies.

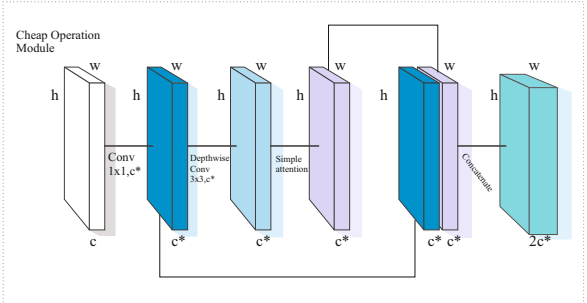

Fig. 3. High-level transition module using cheap operation block.

by using a backbone module that produces more lightweight with a cheaper operation. The detector focused on increasing running speed while maintaining the average precision.

III. Proposed Architecture

The proposed face detector consists of four components: backbone, transition, light attentive module, and multi-scale detection. Fig. 1 shows that it applies four detection layers by assigning various scales anchors to predict the bounding box's location, size, and class.

A. Efficient backbone

A CNN-based architecture generally employs feature extraction to discriminate essential features that support prediction performance. It plays an essential role in capturing important information from an input image so that the proposed network does not dismiss the performance of each sub-module. The detector utilizes a sequential convolutional block with a small number of channel layers. It causes the detector to produce a few parameters, light computation, and low memory usage. In order to support this capability, the proposed detector applies a shrinking block at the beginning of the stage, which emphasizes the efficient shrinkage of the feature map. This strategy can reduce overhead computation while avoiding heavy parameter models. Therefore, it operates a 5 $\times$ 5 kernel size followed by 3 $\times$ 3 convolution

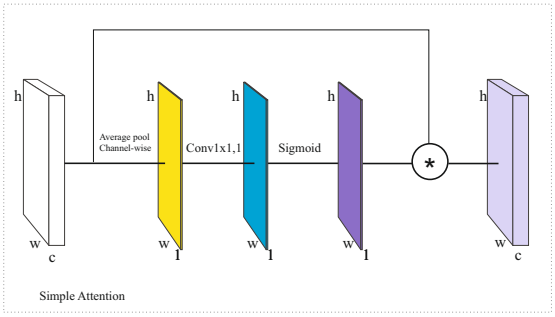

Fig. 4. Simple spatial attention module to enhance valuable elements in cheap operation module.

layers to shrink the feature map size. This architecture uses a large stride at the initial phase to drastically decrease the size of the feature map without compromising the quality of the extracted features.

In order to increase the selectivity on low-level layers, it applies a bottleneck technique using sequential 1 $\times$ 1 and 3 $\times$ 3 convolution, which compresses the channels layer at the beginning of the block. This approach benefits the model by producing a few parameters. The end of the shrinking block generates 32 times smaller feature maps than input images by providing feature information on 128 channels. It allows the detector to absorb little memory usage for extracting process with keeping its performance to filter the interest information.

Furthermore, it implements a stem module that plays a role in comprehensively extracting features by discriminating against facial elements from the background. The need for a real-time system demands that the stem module offers an efficient feature extractor. However, this does not eliminate the performance of the discriminator, so it encourages the detector to be able to distinguish facial and background features accurately. The Mini Multi-level (M2L) convolutional block intuitively fuses two feature maps with different frequency levels of input features (x_i), as shown in Fig. 2. It delivers different level features and combines them to enrich the information.

This structure adopts [22], increasing efficiency by applying only two convolutional levels. The initial process splits two channel-based feature map inputs $[x_i^1, x_i^2]$ to reduce the computational complexity in subsequent operations. Each extractor will get input with a smaller channel layer, so this approach also reduces the number of trainable parameters. The first segment (x_i^1) is applied 3 $\times$ 3 convolutions to obtain low-level features as illustrated:

$$F_l(x_i) = ReLU(BN(W_u(x_i^1))), \quad (1)$$

where it ignores bias and then sequentially applies Batch Normalization (BN) and Rectified Linear Unit activation $(ReLU)$ to prevent vanishing gradient problems. Another level feature is generated by applying a 3 $\times$ 3 convolution filter to the aggregation of low-level and second segment

(x_i^2) can be described as:

$$F_h(x_i) = ReLU(BN(W_v(F_l(x_i) + x_i^2))). \quad (2)$$

M2L output is obtained by combining the two feature maps with different levels to enrich the feature frequency by applying the concatenate operation $(\oplus)$. In addition, 1×1 convolution is applied in the following process to reconstruct every single spatial pixel, as can be illustrated as follows:

$$M2L(x_i) = ReLU(BN(W_z(F_h(x_i) \oplus F_l(x_i)))). \quad (3)$$

Mini Multi-level convolutional block gradually employs updated filter operation at different frequency levels. It emphasizes multi-scale convolutions considering both deeper semantics. Additionally, it increases the diversity of the receptive area while preventing saturation of sensitivity of the deeper convolutions. Instead of using only one M2L block, the proposed detector sequentially applies four modules to improve prediction accuracy on medium and high-level features. It reduces computational and parameter usage and applies grouped convolution to all 3×3 filters on the M2L convolutional block.

B. Multi-scale detection

Face detectors generally utilize the detection layer to predict the face area by estimating its coordinates and size boxes. Several works [17], [19], [21] have proven that multi-layer can improve predictive performance. Each prediction layer will be responsible for specific object sizes. It handles the inconsistency issue of the rigid receptive field, which has difficulty accommodating different facial scales. Therefore, the ACETRON involves this approach to enhance localization precision. It avoids excessive computation and memory, slowing down the real-time speed. Thus, the proposed detector adopts a pyramidal feature hierarchy that eliminates extra convolution and up-sampling techniques. It applies four detection levels to predict faces of varying sizes. Two layers are applied to the last two levels of the backbone, while the other is applied to the interval stage of the high-level transition module.

Each prediction layer employs a head convolutional block that applies 3×3 updated kernels to generate the regression (x, y, w, h) and classification (face or none) predictions. The center coordinates and the size of the bounding boxes are predicted in the regression layer, while the classification score provides the probabilities of the two classes. The proposed detector applies multi anchors to help initialize the size of the predicted box. This assignment applies different anchors sizes to each prediction level. It helps detectors focus on obtaining specific features according to facial dimensions and establish the consistency of each detection layer. Based on this strategy, it installs anchor sizes of 384 and 512 to predict large faces on the first level, 256 and 192 for medium faces. Tiny and small faces were predicted on the third and fourth layers by applying [128, 96, 64, 32] and [16, 8], respectively.

C. High-level Transition Module

This module has a role in bridging the three prediction layers by implementing a cheap operation module. It also transforms the dimensions of the feature map to support a multi-scale detection approach. Ghost module [23] and Depthwise convolution [24] were adopted to avoid computation overhead on low-cost devices. Fig. 3 shows that it employs 1×1 convolution (F_{1x1}) at the beginning of the process to extract channel-based from the (x_i) input features on a single spatial. Then, it applies depthwise convolution and simple attention, respectively. This module can be illustrated as:

$$Tr(x_i) = s_{att}(W_{dw}(F_{1x1}(x_i))) \oplus F_{1x1}(x_i), \quad (4)$$

where W_{dw} is the updated single filter and s_{att} is the simple attention that helps improve the feature map produced by the linear filter. In order to obtain essential information from spatial context information, it applies average pooling to summarize the mean of each spatial pixel. The simple attention module can be described as:

$$s_{att}(x_i) = x_i * \delta(W_{1x1}(Avg_{pool}(x_i))), \quad (5)$$

where 1×1 convolutional serves to scale the spatial representation, then generates the probability of each pixel by applying the sigmoid activation (δ). Fig. 4 shows that spatial attention updates the input feature using a channel-based representation to select the pixels of interest and provide better specific features on the high-level features map.

D. Light attentive module

The proposed architecture uses an enhancement module to boost the potency of crucial features from the input map. Because the lightweight backbone is weak to refine the specific features, it should be improved by inserting a special block before the head of a detection. A Light Attentive module (LIHAT) is proposed as an attention module that can capture essential facial features with low computation and memory complexity. It globally highlights long dependencies according to each map that can select rich channels without ignoring valuable information. It also ensures that every relationship between facial components is represented, which can be illustrated as:

$$Att(x_i) = x_i * \delta(W_{1d}(Avg_{pool}(x_i) * Max_{pool}(x_i))). \quad (6)$$

It applies associative aggregation of average Avg_{pool} and max-pooling Max_{pool} to summarize the channel representation. This combination obtains a definite summary of the feature map to help increase precision in selecting interest features. In addition, it uses 1D convolution to reconstruct spatial information efficiently. It then applies sigmoid activation (δ) to generate a weighted score that will be used to update the information map on input features (x_i), as shown in Fig. 1. The LIHAT module enhances the quality of the information map for adaptive feature refinement. It

allows the detector to operate fast at low computations and significantly save parameters.

IV. Training and testing configuration

A learning module requires a set of data to model the characteristics of the features. The WIDER face dataset was selected as a training dataset that provides varied instance information and contains many challenges. This dataset covers multiple faces with different gender, ages, expressions, scales, poses, illuminances, and occlusions. The benchmark contains 32,203 images, of which 12,800 images are used for the training phase. The augmentation techniques are employed to enrich knowledge and prevent overfitting issues. The method applies random cropping, scale transformation, color distortion, and horizontal flipping, which adopts a work [19]. The cropping result is resized to a high resolution of 1024×1024 used for the input dimensions of the training model.

The entire training and evaluation phases are simulated using the PyTorch framework. The ACETRON network is trained in end-to-end mode applying a mini-batch of 32 that divides the dataset into small partitions. It applies the Mean Squared Error [25] and the Softmax loss [21] as the regression and classification loss, respectively. This loss compares all anchor predictions with ground-truth through IoU (Intersection over Union), generating positive and negative boxes. Then the whole neuron performs updating the weights based on this error score. Furthermore, the training process defines random weights initialization at all kernels in the beginning iteration. The backpropagation process updates the neuron weights by employing the Stochastic Gradient Descent (SGD). Additionally, the optimizer uses the regularization decay of $5 \cdot 10^{-4}$ to optimize updating weights. The proposed detector is trained by applying a gradual learning rate: 10^{-3} learning rate at 300 epochs, 100 epochs at a 10^{-4} learning rate, 50 epochs at a 10^{-5} learning rate, and 20 epochs at a 10^{-6} learning rate. The evaluation phase instructs an anchor matching method by setting 0.5 IoU. It uses a GTX1080Ti only in the training phase to increase computation speed. The detector speed was tested on process evaluation with low-cost computing devices, such as PC Intel Core I5-6600 CPU @3.30 GHz with 8 GB RAM and Lattepanda Intel Cherry Trail Z8300 Quad Core CPU @1.4 GHz with 4 GB RAM. In the inference stage, all detection in the whole prediction layers is combined and then utilized Non-maximum Suppression (NMS) of 0.5 to select the best box as the final prediction.

V. Experimental Results

This section explains an ablation experiment and the evaluation of benchmarks, including Annotated Faces in the Wild (AFW), PASCAL face, and Face Detection Data Sets and Benchmarks (FDDB). It also describes the runtime performance compared to other detectors and the implementation detector operating on low-cost computing devices.

TABLE I
Ablation study of each proposed module.

Proposed Module	Experiment				
	1	2	3	4	5
Backbone	√	√	√	√	√
M2L module	√	√	√	√	
Multi-scale detection	√	√	√		
Cheap operation module	√	√			
LIHAT module	√				
TPR on FDDB (%)	97.52	96.77	97.02	92.11	90.50
Parameters	489,630	488,174	534,764	583,148	368,108
GFLOPS	0.23	0.23	0.24	0.22	0.16
FPS on VGA-resolution	112.06	123.56	135.22	151.84	210.10

A. Ablative study

This subsection is examined each proposed module by replacing a one-by-one module, which is then measured in its performance and efficiency. This approach will show the weaknesses and strengths of each proposed module. It applies the same training setting, besides particular module changes. The FDDB dataset is used as an evaluation dataset to assess the True Positive Rate (TPR) at 1,000 false positives as detector accuracy. TABLE I shows that it is evaluated by substituting and removing each module. Then, the performance, parameters, computational complexity, and speed are analyzed comprehensively. Firstly, the experiment removes all the light attention modules on each branch detection. It decreases the accuracy by 0.75%, but this investigation did not significantly impact parameters and GFLOPS. This module can increase the ability of detection without obstructing the detector's efficiency. Secondly, the cheap operation module is changed with 1×1 convolution. It increases 46.6K parameters and TPR by 0.25%. Thirdly, the experiment only uses one detection layer at the network end. It inserts all anchors only on a detection layer. Although the examination boosted the speed by 16.62 FPS, it declined TPR and added parameters by 4.91% and 48K, respectively. The last experiment eliminates the whole M2L block so that only the shrinking block remains. It shows that M2L can increase the performance by 1.61%, but it also adds the parameters by 215K. In addition, this module decreases the processing rate by 58.26 FPS.

B. Evaluation on dataset

The evaluation of the proposed detector is examined on the AFW, PASCAL FACE, and FDDB datasets by comparing the performance with other competitors.

1) AFW dataset: This dataset contains 205 images with 473 faces that capture from Flickr images. The challenges provide various ages, glasses, skin colors, and expressions with different backgrounds. The proposed detector achieves 99.45% of average precision (AP), as shown in Fig. 5. It outperforms the CPU-based detectors, including FCPU [21] and FaceBoxes [19]. Moreover, the commercial detectors (Face++ and Picasa) obtain low accuracy. The qualitative result shows that ACETRON can accurately locate the face, as illustrated in Fig. 8 (a). The weakness of the detector predicts

TABLE II

EFFICIENCY COMPARISON WITH OTHER FACE DETECTORS. SPEED RESULTS ARE TESTED ON VGA-RESOLUTION.

Detector	Number of Parameter	GFLOPS	TPR on FDDB (%)	PC Core i5 (FPS)	Lattepanda (FPS)
FaceBoxes [19]	654,178	0.15	96.48	70.23	6.25
DCFPN [26]	1,019,834	0.54	95.38	72.57	7.19
LFFD [18]	1,964,656	8.17	97.31	10.91	0.73
RetinaFace-Mobile [27]	426,610	0.76	97.25	21.11	1.51
FlashNet [20]	151,786	0.18	97.33	75.64	6.27
FCPU [21]	989,832	0.20	97.00	90.24	8.76
ACETRON	**489.630**	**0.23**	**97.52**	**112.06**	**11.12**

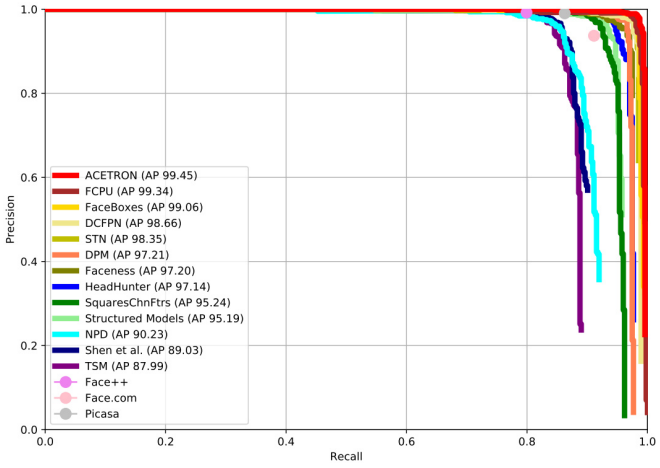

Fig. 5. Evaluation results on AFW dataset.

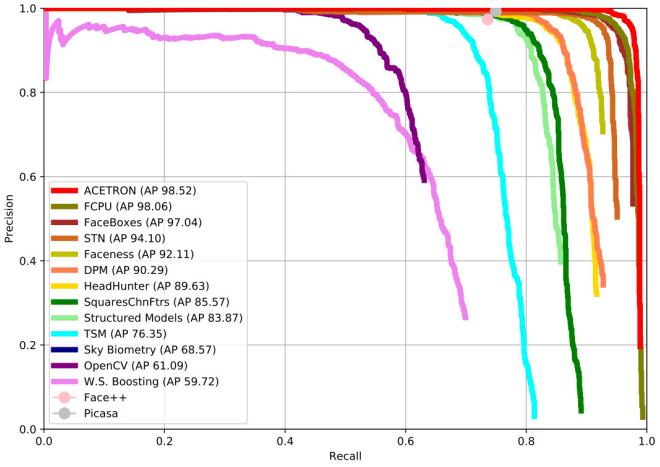

Fig. 6. Evaluation results on PASCAL FACE dataset.

background objects and dog faces as human faces. However, the multi-pose and illuminance variation challenges can be overcome by this detector.

2) PASCAL FACE dataset: This dataset consists of 851 images that contain 1,335 faces. A set of faces is obtained from the subset of the PASCAL person dataset. The variety of background supplies challenges the indoor and outdoor environment to generate variations in lighting. Fig. 6 shows

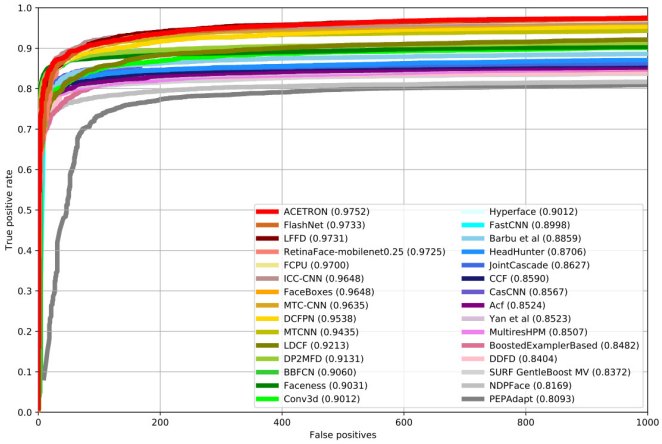

Fig. 7. Evaluation results on FDDB dataset based on true positive rate at 1000 false positives.

that the proposed detector achieves an AP of 98.52%. It outperformed FCPU and FaceBoxes by 0.46% and 1.48%, respectively. The qualitative results in Fig. 8 (b) show that ACETRON can detect faces in variation scale, pose, and occlusion. The proposed detector falsely predicts objects with textures and colors similar to faces.

3) FDDB dataset: This dataset contains 5,171 faces annotated in 2,845 images collected from Yahoo websites. It provides many challenges, including occlusions, multi-poses, illuminance, and low image resolutions. Fig. 7 shows that the proposed detectors are examined on discrete criteria with a true positive rate metric (TPR) at 1,000 false positives. The performance of ACETRON is superior than LFFD [18], Retinaface-mobile[27], FlashNet [20], FCPU, and FaceBoxes. Even the detector performance is 0.19% different from FlashNet. Fig. 8 (c) illustrates that a hand and background object traps the detector to predict faces. This problem does not weaken the detector's performance in detecting various sizes, poses, expressions, genders, and occluded faces.

C. Runtime Performance on CPU devices

A vision-based detector observes patterns from objects through visual input. It has the same way with eyes that can interact with the brain to decide perception. This system

Fig. 8. Qualitative results of the proposed detector on AFW (a), PASCAL FACE (b), FDDB datasets (c), and a video in real-world application (d).

requires a short processing time to synchronize it with the action. A practical application also emphasizes a vision method to operate in real-time. A digital advertising system automatically offers consumer interest with fast processing times. Integration with an advanced module can reduce the speed of the overall system. Therefore, the need for a fast initial detector is an appropriate solution without compromising detection accuracy. In addition, the issue of speed is often associated with its implementation on CPU-based devices, which are generally used for smart digital advertising technologies.

In the testing stage, input lives stream video taken from a webcam to examine the real scenarios performance and real-time speed of the detector at 1,000 frames on VGA resolution. Fig. 8 (d) shows that the detector can detect human faces using accessories (masks and glasses). Even a partially occluded face challenge does not restrict the detector from locating its location. On the other case, some faces are not detected in certain positions due to disappearing facial features such as eyes, eyebrows, and forehead. Nevertheless, this detector is still feasible to apply to smart digital advertising technology because most human faces are detected accurately in real-time.

The proposed detector generates 490K parameters, less than the FCPU as the fastest competitor. The implementation results show that the detector achieves a speed of 112.06 frames per second on a CPU Intel Core i5, as shown in TABLE II. This speed is superior to the FCPU of 21.82 FPS. In addition, other results also show that the proposed detector is faster than the FlashNet detector that outperforms by 0.19% TPR on the FDDB dataset. The proposed detector also achieves the fastest detector on Lattepanda devices. This device exploits a clock speed of 1.4GHz, classified as a low-cost computing device. The competitors run slowly in this device, so they tend to depend on expensive devices. The ACETRON speed shows that this efficient detector can

be operated on low-cost devices. The proposed architecture implements fewer operations and layers, resulting in low computational and memory complexity. Additionally, it keeps the precision of detection that accurately localizes the multi-scale face area.

VI. CONCLUSIONS

This paper presents a fast face detection using the lightweight CNN architecture that works in real-time on the CPU devices. The ACETRON detector employs several efficient modules that generate few parameters and computation complexity. The proposed architecture consists of four main modules: an efficient backbone, transition, lightweight attention, and multi-scale detection. Light backbone and attentive modules help the network to distinguish distinctive features rapidly and highlight the specific essential elements. These modules use few operations and layers that produce less computation cost and support the proposed detector to quickly operate on low-cost computing devices. As a result, the ACETRON achieves excellent performance compared with CPU-based detectors on the benchmark datasets. The network is fastest than other models and can operate at real-time speeds of 112 FPS on a CPU and 11 FPS on a Lattepanda device. The real scenario results show that the detector can detect multiple faces with various challenges of scales, positions, and occlusions. In future work, the transformer module will be explored to enhance the small feature face. Besides, a specific evaluation can be conducted on the animal faces dataset to examine the detector in a challenging environment.

REFERENCES

[1] W. U. Wickramaarachchi, W. Weerasinghe, and R. Rathnayaka, "Influence of smart interactive advertising based on age and gender: A case study from sri lanka," in *International Conference of Reliable Information and Communication Technology*. Springer, 2019, pp. 869–880.

[2] M. A. Moreno-Armendáriz, H. Calvo, C. A. Duchanoy, A. Lara-Cázares, E. Ramos-Diaz, and V. L. Morales-Flores, "Deep-learning-based adaptive advertising with augmented reality," *Sensors*, vol. 22, no. 1, p. 63, 2022.

[3] A. Priadana, M. R. Maarif, and M. Habibi, "Gender prediction for instagram user profiling using deep learning," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*. IEEE, 2020, pp. 432–436.

[4] S. Mittal and V. S. Rajput, "Gender and age based census system for metropolitan cities," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2020, pp. 1094–1097.

[5] 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.

[6] K. Mohan, A. Seal, O. Krejcar, and A. Yazidi, "Facial expression recognition using local gravitational force descriptor-based deep convolution neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2020.

[7] S. Zhang, C. Chi, Z. Lei, and S. Z. Li, "Refineface: Refinement neural network for high performance face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4008–4020, 2020.

[8] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[9] 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)*, 2016, pp. 770–778.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[11] S.-C. Hsia, S.-H. Wang, and C.-Y. Chang, "Convolution neural network with low operation flops and high accuracy for image recognition," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1309–1319, 2021.

[12] X. Li, S. Lai, and X. Qian, "Dbcface: Towards pure convolutional neural network face detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[13] G. Mu, D. Huang, G. Hu, J. Sun, and Y. Wang, "Led3d: A lightweight and efficient deep approach to recognizing low-quality 3d faces," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5766–5775.

[14] 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.

[15] 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)*, 2017, pp. 573–575.

[16] Y. Zhu, H. Cai, S. Zhang, C. Wang, and Y. Xiong, "Tinaface: Strong but simple baseline for face detection," *arXiv preprint arXiv:2011.13183*, 2020.

[17] D. Qi, W. Tan, Q. Yao, and J. Liu, "Yolo5face: why reinventing a face detector," *arXiv preprint arXiv:2105.12931*, 2021.

[18] Y. He, D. Xu, L. Wu, M. Jian, S. Xiang, and C. Pan, "Lffd: A light and fast face detector for edge devices," *arXiv preprint arXiv:1904.10633*, 2019.

[19] S. Zhang, X. Wang, Z. Lei, and S. Z. Li, "Faceboxes: A cpu real-time and accurate unconstrained face detector," *Neurocomputing*, vol. 364, pp. 297–309, 2019.

[20] Y. Ge, Q. Wang, B. Sheng, and W. Yang, "Flashnet: A real-time anchor-free face detector," in *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2020, pp. 441–446.

[21] M. D. Putro, L. Kurnianggoro, and K.-H. Jo, "High performance and efficient real-time face detector on central processing unit based on convolutional neural network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 7, pp. 4449–4457, 2021.

[22] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.

[23] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1577–1586.

[24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, cite arxiv:1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861

[25] C. Dewi, R.-C. Chen, Y.-T. Liu, X. Jiang, and K. D. Hartomo, "Yolo v4 for advanced traffic sign recognition with synthetic training data generated by various gan," *IEEE Access*, vol. 9, pp. 97 228–97 242, 2021.

[26] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z. Li, "Detecting face with densely connected face proposal network," *Neurocomputing*, vol. 284, pp. 119–127, 2018.

[27] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.