© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

A Single Shot Multi-Head Gender, Age, and Landmarks Detection using Shared Convolution Features

Gulraiz Khan, Kevin Pimbblet, Kenneth Wertheim, Waqas Ahmed
Data Science AI and Modelling Centre (DAIM). University of Hull, Hull, UK

***Abstract*—Considering the face as a vital and most informative portion of the human body, it reflects different high-level information about an individual. This high-level information includes Age, Gender, and Emotion. Facial muscles' shape and movement can be the best descriptors for the automatic extraction of these high-level facial features. Detection of these high-level features has applications in different areas including entertainment, surveillance, multimedia, and educational training. However, with the varying nature of these features, it becomes difficult to capture one class with the variability of other classes. This article presents a lightweight heterogeneous neural network with one shared backbone and three network heads to predict multiple face features: landmarks, age, and gender. The proposed system (MultiHeadCNN) captures these high-level facial features in the wild with extreme face pose, occlusions, and lightening conditions. The system is capable of predicting one type of feature with different variability of other types: predicting gender for different age groups and vice versa. The system is tested on comprehensive (UTKFace) and complex (Adience) datasets with varying age, gender, pose, and lightening conditions. The experiment shows promising results in terms of accuracy, with results for age and gender detection on the UTKFace and Adience datasets being 99.9%, 99.7%, 90.3%, and 61.7%, respectively. Furthermore, the parallel inference speed is 20 frames per second.**

***Index Terms*—Convolution Neural Network, Convolution Feature Sharing, ResNet, Backbone Network**

I. Introduction

Given the high complexity and variance in human face images, it is considered a challenging task to forecast age and gender from only facial photos. These variances can result from a variety of environmental constraints, including variable lighting, occlusions, positions, facial emotions, and ethnicities. Age and gender detection models should behave consistently in varying these parameters to be used in real-world applications.

A number of applications that benefit from accurate age and gender prediction have motivated different researchers to perform research and development in the area despite obstacles. These applications include behavior patterns understanding that eventually enables an in-depth understanding of societies. Detection of these automatic high-level facial features can help in authorizations for different access control systems in line with facial recognition. It can also help in crowd behavioral understanding and management and for other surveillance purposes. Different eCommerce businesses can also employ age and gender detection to improve customized user experience. The gaming and multimedia industry can use these features to refine the 3D profiling of objects and avatars.

In the past few decades, there has been a plethora of work proposed for age and gender prediction from face images. Earlier attempts in this field relied merely on hand-crafted features mainly extracted from facial images, followed by a machine learning classifier for finding the age and gender label. These works that used hand-crafted features were restricted in their ability to capture the complex variants in the face, such as variable illuminations, poses, sizes, and occlusion [1]–[3]. However, with the notable success of deep learning networks in different computer vision tasks over the past few years, now the work on age and gender prediction has shifted focus towards using convolution neural networks. Deep neural networks have shown favorable outcomes in learning hierarchical features directly from the input image, allowing them to apprehend these variations effectively based on data [4]–[8].

In this paper, we propose MultiHeadCNN, a unique heterogeneous deep learning framework to predict facial landmarks (face localized keypoints that represent the contour of eyes, nose, mouth, and chin), age, and gender from a single neural network with multiple branches. We designed this framework by considering the fact that face features can jointly predict age, gender, and face keypoints, as wrinkles can directly relate to age, whereas a beard or mustache can assist in distinguishing between genders, reducing the likelihood of a child. Considering face features as common predictive for three target classifications of interest to us, we decided that ResNet-18 is a suitable backbone feature extraction network. Figure 1 shows the backbone feature maps of a few input samples, it can be seen that majorly feature focus is on skin area with different intensity levels on different face parts. Major contributions of the proposed system are as follows,

- We have proposed a heterogeneous multi-purpose learning network for predicting age, gender, and face points.
- We utilized the concept of using a backbone shared feature network for performing multiple classification tasks.
- The proposed network consists of three very lightweight dedicated branches for age, gender, and landmarks detection.
- Overall the system achieves promising accuracy for all three targets without compromising the efficiency. These results are obtained by using a single residual network

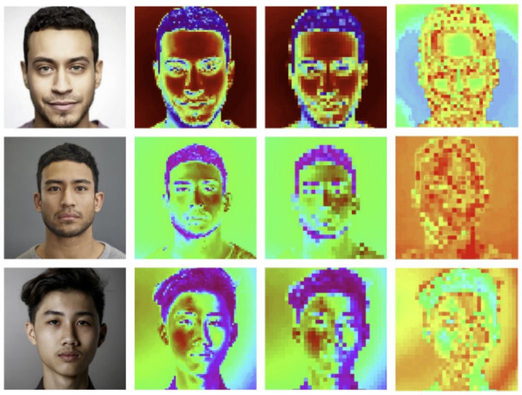


Fig. 1: Examples of features on input photos for different layers

for feature extraction and then lightweight classification heads.

The overall paper consists of the following sections. Section 2 discusses the details of the literature for predicting age, gender, and facial feature points. In section 3, a comprehensive methodology of the proposed system is provided. Section 4 summarises the experimentation with the description of datasets that we used for training and testing and results. Finally, the paper is concluded with future work in section 5.

II. Literature Review

The face is considered the most descriptive part of the human body. It is widely used in AI systems to detect facial attributes such as a person's age, gender, and emotional state. In this section, different classic age and gender detection systems [1]–[3] and deep learning based systems are discussed [4]–[8]. Classical approaches can be further divided into different categories including hand-crafted features [10], [13] based and appearance based models [11], [12].

Handcraft features are the popular traditional approach for age and gender prediction because of easy implementation and computational complexity. Whereas, they lag in accuracy and are vulnerable to face variations. Bio-inspired hand-craft features are used in [13] by using Gabor filter followed by a Support Vector Machine Classifier for predicting age. Similarly, multiple face descriptor features (Local Binary Pattern-LBP and Binarized Statistical Features) are used in [14] for age estimation.

Similarly, multiple researchers have employed hand-crafted features for gender classification. Jabid et al. [15] proposed a novel feature detection system by extracting Local Directional Pattern (LDP) to make gender detection direction invariant. This system computes edge response values in multiple directions to provide a more consistent feature set in the presence of noise. They used a Support Vector Machine to classify gender. Another research group [16] worked on the discriminative face features using Interlaced Derivative Patterns (IDP). Their experimental results showed improved performance than LBP and LDP. Few researchers fused multiple features for gender detection [17], [18]. E. Tapia et al. [17] combined multiple hand-crafted features including LBP to improve the accuracy. Their system employed four types of feature selection methods to find redundancy and relevance among different features.

With the advent of deep learning, there has been a shift of attention towards age and gender detection. Convolution Neural Networks have revolutionized the traditional way of learning and prediction. In a study by Benkaddour et al. [19], a convolution neural network with shallow layers is proposed to predict age and gender from a single network. Their proposed model is significant because it demonstrates the fact that we can achieve good accuracy with a simple shallow layered network. They tested this system on the Adiance dataset and attained promising results.

Another research group in [20], proposed a heterogeneous network by combining a convolution neural network with Extreme Learning Machine (ELM) in a hierarchical fashion. CNN is used for feature extraction from a plain image followed by ELM to classify these features into meaningful outputs. By combining the strengths of CNN with ELM, their proposed hybrid structure was able to achieve significant results in age and gender classification tasks on MORPHII and Adience datasets.

Ozbulak et al. [21] proposed the transfer learning based network for age and gender detection. Two pre-trained models on face: AlexNet and VGG are employed for fine-tuning on Adience dataset for age and gender prediction in the wild. This fine-tuned network can be effectively used for age and gender prediction, even for multiple uncontrolled environments with variable conditions in pose, size, illumination, and occlusion. Lapuschkin et al. [22] provided a comparison of different architectures in terms of predicting age and gender precisely. The authors assess the robustness of systems in variable environmental conditions based on test set swapping.

Few research groups [23], [24] employed attention mechanisms for age and gender classification. Rodriguez et al. [23] proposed a feed-forward attention network to focus more on specific face areas with more informative features for age and gender detection. Their proposed model learns to find the most descriptive portion from the face image and extract attention features followed by classification layers. Similarly, an improved version of the attention network combined with residual learning is proposed in [24]. This system utilized the attention network as a backbone feature extractor and these backbone features are then passed to two separate branches for gender and age detection.

Optimizing accuracy and efficiency is typically necessary to achieve near-real-time target feature extraction. To fulfill this need, ensuring accuracy without compromising speed is the aim of this research, enabling the concurrent processing of all target classes at the same time. Because our suggested network incorporates a moderately complex backbone network, it has an advantage in reaching optimal results. Our network, in contrast to shallow architectures like AlexNet or LeNet, uses a residual mechanism with shortcut connections to lessen the problem of vanishing gradients, which is a problem with VGG networks. Furthermore, because the attention network is inherently complex and may result in decreased efficiency, we decided against integrating it.

III. Proposed Methodology

This section elaborates on the overall proposed framework for predicting age and gender. We structured our network to perform multi-task learning, where a single backbone convolution neural network is learned on one common area of the face with three branch networks for the purpose of detecting facial feature points, age, and gender. The following figure 2 shows the basic building block of the backbone network. The prior location of facial feature points can help in gender and age prediction [25] as they contain information of muscle shape and movements. Along with the backbone features, we passed landmarks to the gender and age detection branches separately. Augmenting the feature embedding of the age and gender classifier with the predicted landmark points, the precision of age and gender prediction can be further enhanced. The entire process flow is depicted in figure 3, where input is routed through a backbone network to acquire shared facial features, and then it is passed through separate branches to forecast landmarks, age, and gender.

A. Residual Neural Network

As described earlier, we have employed a residual neural network as a backbone for the purpose of feature sharing with all three network heads. The residual learning can be paired with a simple feed-forward neural network to produce better results.

The paradigm of Machine Vision has changed significantly since the introduction of Convolution Neural Networks (CNNs), but when CNN architectures go further deeper, they run into a barrier called the vanishing gradient problem. This problem results from gradients reducing in deeper layers due to constant multiplication during back-propagation. As a result, this problem restricts deep network ability to be trained successfully and can lead performance to stagnate or even decline.

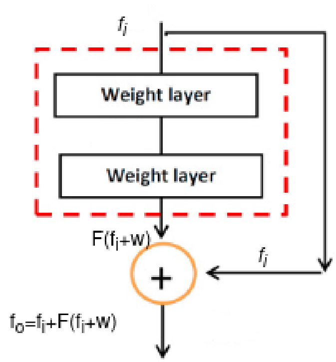


Fig. 2: ResNet Block: The basic building block of a Residual Neural Network (ResNet). This basic unit is used repeatedly throughout the network architecture.

The residual learning approach provides a novel solution to the vanishing gradient problem in a deep network by proposing the shortcut connection. A very basic residual learning structure is explained by the following equation 1.

$$f_o = f_i + F(f_i, W_i) \tag{1}$$

In equation 1, f_i represents the input feature set, f_o shows output features and W_i represents the weight matrix. The function F takes input features along with the weight matrix to produce features that are directly combined with input to produce the final output. Figure 2 shows the basic block of the residual neural network.

We built a backbone network by combining several residual blocks, which feeds features into three branches. The detailed specification of our complete network is provided in table I, which includes output features and hyperparameters for each layer in the backbone and branch networks.

Before being fed into the backbone network, the input image is first resized to $3 \times 256 \times 256$ in dimensions. The input features are down-sampled using a max-pooling layer after the first convolution layer, producing a feature map with dimensions of $64 \times 64 \times 64$. This feature map is then processed by four sets of convolutions, each consisting of four convolution layers separately. The backbone network ultimately results in the creation of a $512 \times 8 \times 8$ feature map as depicted in table I. The following equation 2 provides the size of the output feature (sf_o) map after each convolution

Layer	K	S	P	# of K	Output
Backbone Network					
conv1	(7, 7)	(2, 2)	(3, 3)	64	[64,128,128]
maxpooling	(3,3)	(2,2)	(1,1)	64	[64,64,64]
layer1-conv1	(3, 3)	(1, 1)	(1, 1)	64	[64,64,64]
layer1-conv2	(3, 3)	(1, 1)	(1, 1)	64	[64,64,64]
layer1-conv3	(3, 3)	(1, 1)	(1, 1)	64	[64,64,64]
layer1-conv4	(3, 3)	(1, 1)	(1, 1)	64	[64,64,64]
layer2-conv1	(3, 3)	(2, 2)	(1, 1)	128	[128,32,32]
layer2-conv2	(3, 3)	(1, 1)	(1, 1)	128	[128,32,32]
layer2-conv3	(3, 3)	(1, 1)	(1, 1)	128	[128,32,32]
layer2-conv4	(3, 3)	(1, 1)	(1, 1)	128	[128,32,32]
layer3-conv1	(3, 3)	(2, 2)	(1, 1)	256	[256,16,16]
layer3-conv2	(3, 3)	(1, 1)	(1, 1)	256	[256,16,16]
layer3-conv3	(3, 3)	(1, 1)	(1, 1)	256	[256,16,16]
layer3-conv4	(3, 3)	(1, 1)	(1, 1)	256	[256,16,16]
layer4-conv1	(3, 3)	(2, 2)	(1, 1)	512	[512,8,8]
layer4-conv2	(3, 3)	(1, 1)	(1, 1)	512	[512,8,8]
layer4-conv3	(3, 3)	(1, 1)	(1, 1)	512	[512,8,8]
layer4-conv4	(3, 3)	(1, 1)	(1, 1)	512	[512,8,8]
Landmarks (Input: 518 $\times$ 8 $\times$ 8)					
	(3, 3)	(1, 1)	(1, 1)	512	[512,8,8]
Conv+FC (68)	(1,1)	(1,1)	(0,0)	68	X: [68]
Conv+FC (68)	(1,1)	(1,1)	(0,0)	68	Y: [68]
Age Branch			**Gender Branch**		
Operation	K	Output	Operation	K	Output
Conv	(3,3)	[512,8,8]	Conv	(3,3)	[512,8,8]
FC1		128	FC1		128
FC2		128	FC2		128
FC3		128	FC3		128
FC4		56	FC4		56
FC5		56	FC5		56
FC6		56	FC6		56
FC7		10	FC7		2

TABLE I: Tabular description of the architecture for Backbone network and three target branches. With the details of layers: operation, filter size (K), stride (S), padding (P), number of filters (# of K), and output feature size for each layer

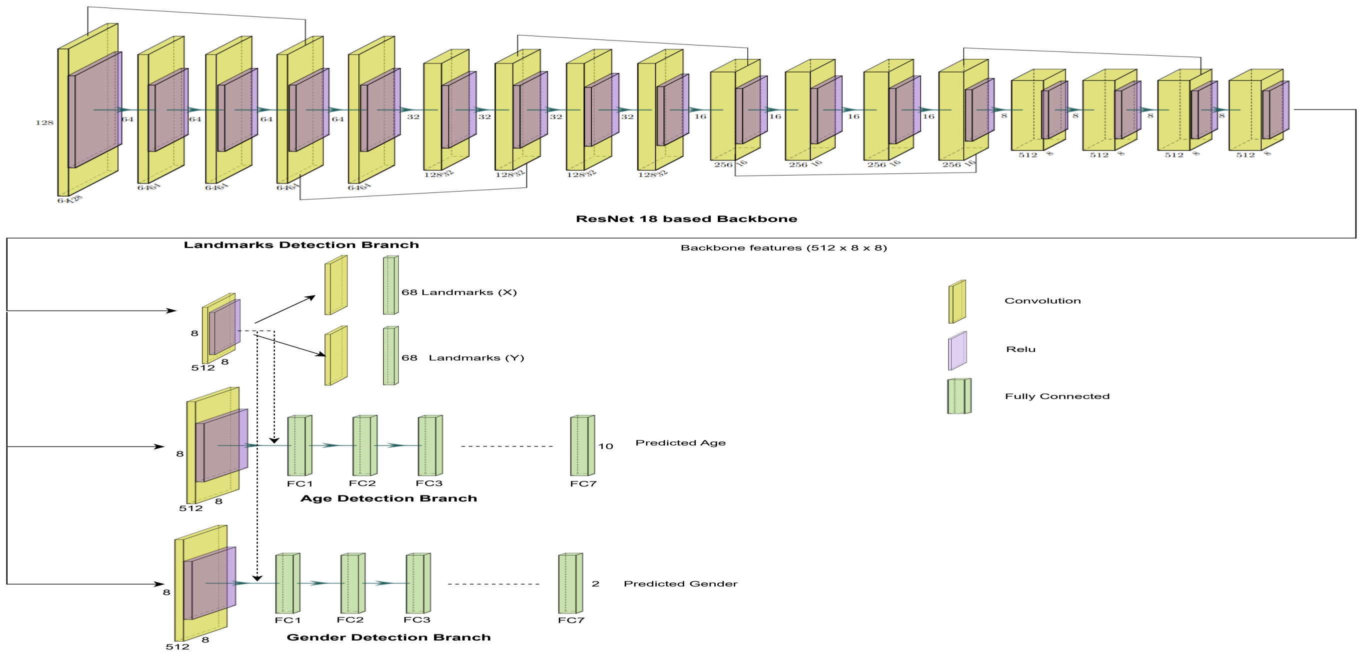


Fig. 3: A simplified illustration of the proposed network. Showcasing the Backbone network generating features with dimensions $518 \times 8 \times 8$ and sharing these shared features to three target branches.

operation on the input feature of size (sf_i). Where K, P and S shows the kernel size, padding and stride, respectively.

$$sf_o = \frac{sf_i - K + 2P}{S} + 1 \tag{2}$$

The feature map derived from the backbone network acts as a shared feature map across all target branches. As depicted in the table I, these backbone features of size $512 \times 8 \times 8$, are transferred to the landmarks branch, where they are processed independently for the x and y axes through two convolution layers to produce 68 landmarks separately (row 2 and 3 in landmarks section). Similarly, the backbone features along with landmarks features are used in the age and gender branches, moving through seven fully connected layers to produce outputs that include 10 classes for age and 2 classes for gender.

Figure 3 illustrates the structure of our proposed network. It is composed of various layers, including fully connected, activation, and convolution layers. Four sets of convolutional layers are grouped together in the backbone network. To provide the shared features, these layers process the input image by using RELU activation and convolution operation. Following that, these common features are processed through three distinct branches, each of which is designed to generate a target result. We have used ResNet in the backbone and skip connections are a fundamental part of ResNet. The connecting lines above and below the network are skip connections that are used to minimize the vanishing gradient issue in each residual block. With the help of these connections, information from one level can be added directly to the output of deeper levels, skipping multiple layers in the network. This technique makes sure that the gradient information from the input may pass through the network more readily during back-propagation, which keeps the gradient information from drastically decreasing as it passes through deep layers.

IV. Performance Evaluation

A. Experimental Setup

The PyTorch open-source software was used to develop the proposed network [26]. An Nvidia T4 instance with a total of 2560 GPU cores and 16GB of GPU memory was used for training. It took four and a half hours to train all the branches based on the features that were learned after fine-tuning the backbone network for face feature points.

We used a batch size of 4 and an initial learning rate of 0.001 for training. A stochastic gradient descent optimizer with the cross-entropy loss function was used to train the network. Deep networks benefit greatly from cross entropy as it measures the difference between true and predicted probabilities in a gradient-friendly manner. The cross-entropy loss function can be expressed by the equation 3, in which E represents entropy, T_i is the true probability, and P_i is the predicted probability.

$$E(T, P) = \sum_{i}^{Classes} T(i) log(P(i)) \tag{3}$$

B. Dataset

We have performed evaluation on two recent datasets for age and gender prediction Adience Benchmark [27] and UTKFace [29].

The Adience dataset is made up of photos that are automatically submitted to Flickr by phones. These photos were uploaded without going through manual filtration, which led to extremely unrestricted viewing conditions that mirrored the challenges in internet images in real life. As such, there are significant variations in head posture, lighting, and quality for the Adience dataset. The selected samples from the Adience dataset are displayed in Figure 4, offering insight into the complexity and diversity of this benchmark.

Fig. 4: Sample images from Adience Benchmark

There are about 26,000 photos total in the Adience benchmark, representing 2,284 subjects. Table II shows the statistics of individual genders in the Adience dataset based on 8 age groups with M for male, F for female and B is for both (image containing both genders).

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60>**
M	745	928	934	734	2308	1294	392	442
F	682	1234	1360	919	2589	1056	433	427
B	1427	2162	2294	1653	4897	2350	825	869

TABLE II: Statistics of Adience Dataset

UTKFace dataset is a recent good-sized dataset for diverse age groups ranging from 0 to 116. There is a wide age span from infants to elderly persons. The images in UTKFace cover good variation in terms of head pose, lightenning conditions and quality. For testing purposes, we split the data into 80% and 20% splits. Approximately 10k photos with evenly distributed age and gender are used in the experiment. Figure 5 shows some samples from UTKFace dataset.

Fig. 5: Sample images from UTKFace Benchmark

C. Results

Using the two benchmark datasets stated above, we performed a comparison of our proposed system against results that have already been published on these datasets. Our system was applied on the Adience dataset, and in highly restricted conditions, it achieved an accuracy of 61% for age detection. Even with the extremely difficult images, our system performed better than the state-of-the-art. Table III shows the results for Age classification for Adeince Dataset.

Method	Complexity	Accuracy
LBP [27]	Traditional Feature Learning	45.1
CNN [8]	Convolution Neural Network	50.7
RAN [30]	Dedicated Residual Network	57.3
MSFCL [32]	Three Stage Convolution Network	64.7
MultiHeadCNN (Our)	Shared Backbone Features	61.7

TABLE III: Age Detection Results on Adience Dataset

Similarly, we used the Adience benchmark to assess the performance of our gender detection method. Our results for gender detection on the Adience dataset are shown in the table IV. These results provide an accurate depiction of the performance our suggested approach achieves on a challenging dataset with wide variations in posture and illumination.
The primary benefit of the system is its ability to detect age and gender simultaneously, resulting in increased efficiency and a lighter network.

Method	Complexity	Accuracy
LBP [27]	Traditional Feature Learning	77.8
GRA_NET [35]	Gated Residual Attention Network	81.4
LW-MTCNN [33]	Light Weight Multi-Task CNN	85.16
MultiHeadCNN (Our)	Shared Backbone Features	90.3

TABLE IV: Gender Detection Accuracy on Adience Dataset

As mentioned earlier, we evaluated our proposed approach across two datasets. For testing on UTKFace, the complete dataset was first divided into 80-20 halves, with 80% employed for training and 20% for testing. Table V shows how the age and gender results compare with earlier research.
The results in table V clearly demonstrate that the proposed system yields promising results for a simple dataset with less variation in pose and it achieved accuracy scores of 94.01% for age and 99.86% for gender.

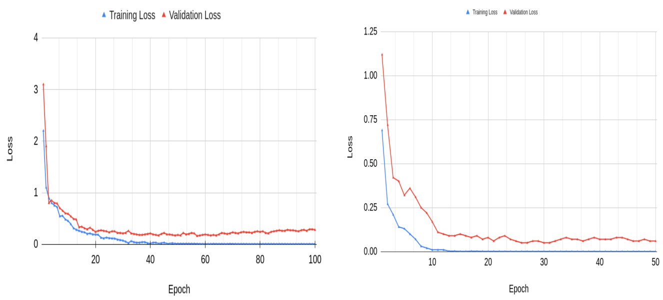
Method	Gender	Age
Mivolo [28]	98.5	72.64
Finetuned Facenet [34]	96.1	64
Multi-Task CNN [31]	98.23	70.1
Residual Attention Network [30]	97.5	85.4
Gated Residual Attention Network [35]	99.2	93.7
Improved CNN [36]	99.8	94.01
MultiHeadCNN (Our)	99.9	99.7

TABLE V: Evaluation Results on UTKFace Dataset

The training and validation losses for the age and gender networks are shown in the loss graphs shown in Figure 6. The lines on the graph indicate that the age detection training loss starts at 2.2 and 3.1, respectively, and gradually converges to 0.003 and 0.28. Similarly, gender detection training loss drops from 0.6 to 0.002, while validation loss falls from 1.5 to 0.09.

V. Conclusion

We acknowledge that there are few works that achieved better results in terms of accuracy for age prediction on the Adience dataset (Complex). This accuracy is attributed to



(a) Age Detection (b) Gender Detection

Fig. 6: Training and Validation Loss Graphs

their higher complexity in terms of parameters and utilization of dedicated network for age and gender separately without feature sharing which makes them resource hungry. In future, we can include custom loss function and try different backbone architectures including attention mechanism to experiment with the efficiency and accuracy.

References

[1] Sai, Phyo-Kyaw, Jian-Gang Wang, and Eam-Khwang Teoh. "Facial age range estimation with extreme learning machines." Neurocomputing 149 (2015): 364-372.

[2] Sawant, Manisha, Shalini Addepalli, and Kishor Bhurchandi. "Age estimation using local direction and moment pattern (ldmp) features." Multimedia Tools and Applications 78 (2019): 30419-30441.

[3] Eidinger, Eran, Roee Enbar, and Tal Hassner. "Age and gender estimation of unfiltered faces." IEEE Transactions on information forensics and security 9.12 (2014): 2170-2179.

[4] Kancharlapalli, Tara Pranay, and Pulkit Dwivedi. "A Novel Approach for Age and Gender Detection using Deep Convolution Neural Network." 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2023.

[5] Karahan, Mehmet, et al. "Age and gender classification from facial features and object detection with machine learning." Journal of fuzzy extension and applications 3.3 (2022): 219-230.

[6] Kumar, B. Anil, and Neeraj Kumar Misra. "Masked face age and gender identification using CAFFE-modified MobileNetV2 on photo and real-time video images by transfer learning and deep learning techniques." Expert Systems with Applications 246 (2024): 123179.

[7] Upadhyay, Shailja, Joshika Malai Magal BV, and A. Deepa. "Human: Age and Gender Identification Using SFTA and Deep Learning Algorithm." 2024 2nd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT). IEEE, 2024.

[8] Levi, Gil, and Tal Hassner. "Age and gender classification using convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015.

[9] Hassan, Khaled Rahman, and Israa Hadi Ali. "Age and gender classification using multiple convolutional neural network." IOP Conference Series: Materials Science and Engineering. Vol. 928. No. 3. IOP Publishing, 2020.

[10] Becerra-Riera, Fabiola, et al. "Age and gender classification using local appearance descriptors from facial components." 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017.

[11] Dibeklioğlu, Hamdi, et al. "Combining facial dynamics with appearance for age estimation." IEEE Transactions on Image Processing 24.6 (2015): 1928-1943.

[12] Bukar, Ali Maina, Hassan Ugail, and David Connah. "Automatic age and gender classification using supervised appearance model." Journal of Electronic Imaging 25.6 (2016): 061605-061605.

[13] Guo, Guodong, et al. "Human age estimation using bio-inspired features." 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009.

[14] Bekhouche, Salah Eddine, et al. "Facial age estimation using bsif and lbp." arXiv preprint arXiv:1601.01876 (2016).

[15] Jabid, T.; Kabir, H.; Chae, O. Gender Classification Using Local Directional Pattern (LDP). In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 2162–2165.

[16] Shobeirinejad, A.; Gao, Y. Gender Classification Using Interlaced Derivative Patterns. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 1509–1512.

[17] Tapia, J.E.; Perez, C.A. Gender Classification Based on Fusion of Different Spatial Scale Features Selected by Mutual Information From Histogram of LBP, Intensity, and Shape. IEEE Trans. Inf. Forensics Secur. 2013, 8, 488–499

[18] Azzopardi, George, et al. "Fusion of domain-specific and trainable features for gender recognition from face images." IEEE access 6 (2018): 24171-24183.

[19] Benkaddour, Mohammed Kamel, Sara Lahlali, and Maroua Trabelsi. "Human age and gender classification using convolutional neural network." 2020 2nd international workshop on human-centric smart environments for health and well-being (IHSH). IEEE, 2021.

[20] M. Duan, K. Li, C. Yang, and K. Li, "A hybrid deep learning cnn–elm for age and gender classification," Neurocomputing, vol. 275, pp. 448–461, 2018.

[21] G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are cnn-based features for age and gender classification?" in 2016 International Conference of the Biometrics Special Interest Group (BIOSIG).

[22] S. Lapuschkin, A. Binder, K.-R. Muller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017

[23] P. Rodr ıguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. Gonz'alez, "Age and gender recognition in the wild with deep attention," Pattern Recognition, vol. 72, pp. 563–571, 2017.

[24] Abdolrashidi, Amirali, et al. "Age and gender prediction from face images using attentional convolutional network." arXiv preprint arXiv:2010.03791 (2020).

[25] Alonso-Fernandez, Fernando, et al. "Facial masks and soft-biometrics: Leveraging face recognition CNNs for age and gender prediction on mobile ocular images." IET Biometrics 10.5 (2021): 562-580.

[26] Wu, Peng. "Pytorch 2.0: The journey to bringing compiler technologies to the core of pytorch (keynote)." Proceedings of the 21st ACM/IEEE International Symposium on Code Generation and Optimization. 2023.

[27] Eidinger, Eran, Roee Enbar, and Tal Hassner. "Age and gender estimation of unfiltered faces." IEEE Transactions on information forensics and security 9.12 (2014): 2170-2179.

[28] Kuprashevich, Maksim, and Irina Tolstykh. "Mivolo: Multi-input transformer for age and gender estimation." International Conference on Analysis of Images, Social Networks and Texts. Cham: Springer Nature Switzerland, 2023.

[29] Zhang, Zhifei, Yang Song, and Hairong Qi. "Age progression/regression by conditional adversarial autoencoder." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[30] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164

[31] Vidyarthi, Pankaj, Sunita Dhavale, and Suresh Kumar. "Gender and age estimation using transfer learning with multi-tasking approach." 2022 2nd Asian Conference on Innovation in Technology (ASIANCON). IEEE, 2022.

[32] M. Xia, X. Zhang, L. Weng, and Y. Xu, "Multi-stage feature constraints learning for age estimation," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 2417–2428, 2020

[33] Lee, Jia-Hong, et al. "Joint estimation of age and gender from unconstrained face images using lightweight multi-task cnn for mobile applications." 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, 2018.

[34] Setiawan, Haris, Mudrik Alaydrus, and Abdi Wahab. "Multibranch Convolutional Neural Network For Gender And Age Identification Using Multiclass Classification And FaceNet Model." 2022 Seventh International Conference on Informatics and Computing (ICIC). IEEE, 2022.

[35] Garain, Avishek, et al. "GRA_NET: A deep learning model for classification of age and gender from facial images." IEEE Access 9 (2021): 85672-85689.

[36] Sharma, Neha, Reecha Sharma, and Neeru Jindal. "Face-based age and gender estimation using improved convolutional neural network approach." Wireless Personal Communications 124.4 (2022): 3035-3054.