IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 7, 2021, accepted May 28, 2021, date of publication June 3, 2021, date of current version June 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3085971

GRA_Net: A Deep Learning Model for Classification of Age and Gender From Facial Images

AVISHEK GARAIN[1], (Member, IEEE), BISWARUP RAY[1], PAWAN KUMAR SINGH[2], (Member, IEEE), ALI AHMADIAN[3,4], (Member, IEEE), NORAZAK SENU[4], AND RAM SARKAR[1], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, India
[2]Department of Information Technology, Jadavpur University, Kolkata 700106, India
[3]Institute of Industry Revolution 4.0, The National University of Malaysia (UKM), Selangor 43600, Malaysia
[4]Institute for Mathematical Research, Universiti Putra Malaysia, Serdang 43400, Malaysia

Corresponding authors: Ali Ahmadian (ahmadian.hosseini@gmail.com) and NORAZAK SENU (norazak@upm.edu.my)

This work was supported in part by the Fundamental Research Grant Scheme (FRGS) grant provided by the Ministry of Education, Malaysia, under Project FRGS/1/2018/STG06/UPM/02/2, and in part by the Universiti Putra Malaysia.

ABSTRACT The problem of gender and age identification has been addressed by many researchers, however, the attention given to it compared to the other related problems of face recognition in particular is far less. The success achieved in this domain has not seen much improvement compared to the other face recognition problems. Any language in the world has a separate set of words and grammatical rules when addressing people of different ages. The decision associated with its usage, relies on our ability to demarcate these individual characteristics like gender and age from the facial appearances at one glance. With the rapid usage of Artificial Intelligence (AI) based systems in different fields, we expect that such decision making capability of these systems match as much as to the human capability. To this end, in this work, we have designed a deep learning based model, called GRA_Net (Gated Residual Attention Network), for the prediction of age and gender from the facial images. This is a modified and improved version of Residual Attention Network where we have included the concept of Gate in the architecture. Gender identification is a binary classification problem whereas prediction of age is a regression problem. We have decomposed this regression problem into a combination of classification and regression problems for achieving better accuracy. Experiments have been done on five publicly available standard datasets namely FG-Net, Wikipedia, AFAD, UTKFAce and AdienceDB. Obtained results have proven its effectiveness for both age and gender classification, thus making it a proper candidate for the same against any other state-of-the-art methods.

INDEX TERMS Age identification, gender classification, gated residual attention network, facial image, MAE, regression.

I. INTRODUCTION

Age identification and gender classification play a pivotal role in our social lives. Every language in the world reserves different salutations for men and women, and very often different vocabularies are used when addressing elders compared to young people. These customs are largely dependent on one's ability to estimate these individual traits of a person: age and gender, which are obtained from the facial appearances. A vast number of application developers, iespecially after the growth in social media and social networks, are indulging themselves with automatic age identification. Age and gender are the most fundamental facial qualities in social interaction. Human's face contains features that determine identity, age, gender, emotions, and the ethnicity of people. Among these features, age and gender identification can be especially helpful in several real-world applications including visual surveillance, medical diagnosis (premature facial aging), human-computer interaction system, access control or soft biometrics, demographic

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

information collection, law enforcement, marketing intelligence, etc. it is necessary to identify the age and the gender from the facial images of the human beings. However, several problems in age and gender identification are still considered as open challenges to the researchers. Despite the progress, the computer vision community keeps making continuous improvement by introducing the new techniques that advance the state-of-the-art, age and gender predictions from the unfiltered real-life facial images are yet to meet the need of the commercial and real-world applications. Thus a robust and accurate method for the age and the gender identification tasks becomes an absolute necessity.

In the past few years, in order to intensify the ability to identify these attributes from facial images, many methods have been put forward. Previous researchers approached the task of estimating age or classifying gender from face images using individually designed feature vectors with statistical models ([6]) and machine learning based models ([3], [19]). However, even though a lot of these models have been designed, the individually designed features behave inadequately on the benchmark datasets having unconstrained images. Hence, in the more recent years, researchers have been started exploring the domain of convolutional neural network (CNN) based deep learning architectures for the task of age and gender prediction ([1], [8], [17], [22], [36]–[38], [49]), which help in automatic feature extraction from the input images. To improve such feature extraction ability, use of the long short-term memory (LSTM) units inserted between Residual Networks (ResNets) has also been found ([50]). In this paper, we have proposed an architecture which harnesses the capabilities of both classification and regression to solve the tasks of age estimation and gender prediction from the facial images. It is having a backbone support of Residual Attention Network modified with the addition of a new parameter called 'Gate' similar to the concept of gates in Gated Residual Units (GRUs), the only difference being the use of the 'Gate' on the higher level of components rather than applying it on individual units of the architecture. Impressive results obtained on various standard datasets confirm the credibility and precision of the architecture.

The rest of the paper is organized as follows. Some of the previous works have been described in Section II. Section IV defines the data collection and preparation processes that are used here. The working principle of the proposed method has been described in detail in Section V. This is followed by detailed analysis of results in Section VI. The concluding remarks are given in Section VII.

II. LITERATURE SURVEY

In the past few years, there have been several attempts to estimate the individual traits: age and gender from the facial images. A variety of features consisting both of facial shape and textures have been used by the researchers for the estimation of age and gender from facial images. Some of those features are: mean pixel values of channels, energy and entropy of filtered images, Histogram of Oriented Gradients (HOG) etc. Simple features such as density of edges obtained from an image using an edge detector have also been applied.

Reference [19] has proposed a 5-stage method which calculates distances among the global features and some hybrid ratios. Classification is performed using thresholds for the two genders from the ratios. For the FGNET dataset, the method achieves an accuracy of 95% for the gender recognition task and an accuracy of 79.31% for class wise prediction of ages i.e., groups of 1 to 12 years, 13 to 40 years and 41 to 80 years. Reference [3] has proposed a supervised appearance model (sAM) that improves on active appearance model (AAM) by replacing PCA with partial least-squares regression. The sAM model is used as a feature extractor for age and gender estimation from the facial images. Reference [34] have presented a study investigating the effects of image preprocessing, model initialization and architecture choice on identification of age and gender from facial images. The study has also visualized model's prediction strategies in given preprocessing conditions using the Layer-wise Relevance Propagation (LRP) algorithm. An accuracy of 92.6% is achieved for the gender classification task and 62.8% for the age estimation task on the Adience dataset. Reference [43] has proposed an architecture that makes a gradual refining of the feature through three feature constraint stages. Every stage of the algorithm involves continuous update of the feature center of its corresponding age range, and minimization of the distance between each age feature and feature center of the corresponding age range by means of feature constraint.

However, the accuracies yield by these models are much lower. As a viable alternative, thus, researchers have started applying various deep learning models for the said tasks. In the recent years, deep learning based models have shown reassuring performance in the field of age and gender identification, especially on unfiltered face images. Recently, methods such as simple CNN, MobileNet CNN, LSTM of recurrent neural network (RNN) architectures have been used by various researchers to estimate age and gender from facial images. Reference [23] has proposed a simple CNN based architecture that could be used for limited amount of learning data. The network consists of only three convolutional layers and two fully-connected layers with a very less number of neurons. The proposed method achieves highest accuracy of 86.8% for gender estimation task and highest accuracy of 50.7% for exact age estimation task on Adience dataset.

Reference [49] has presented a novel CNN based method for age group and gender estimation leveraging Residual networks of Residual networks (RoR). The two-system RoR-34 + IMDBWIKI ([32]) achieves an age estimation accuracy of 67.34% and a gender recognition accuracy of 93.24%. In the same year, [38] has predicted age, gender and fine-grained ethnicity of an individual by providing baseline results using a CNN. Results of two CNN architectures are given, showing potential design considerations that promotes region based feature extraction thus optimizing the network. Age and gender accuracies of

38% and 88% respectively are achieved by the method on Wild East Asian Face Dataset (WEAFD). The WEAFD is a new and unique dataset consisting mainly of labeled facial images of individuals from East Asian countries.

In the method proposed by [37], Transfer learning is explored by making use of VGG19 and VGGFace models with pre-trained weights and doing ablation study to increase the efficiency. A hierarchy of deep CNNs is tested, which classifies subjects on the basis of gender. A gender recognition accuracy of 98.7% and an MAE of 4.1 years are achieved by the proposed method on MORPH-II dataset. A CNN based architecture for joint age-gender identification method is proposed by [17]. For the CNN network, Gabor filter responses are used as inputs. back-propagation for an end-to-end architecture have been used in order to learn the weighting of Gabor-filter responses. The proposed method achieves age and gender estimation accuracies of 61% and 88% respectively. Reference [22] in the same year have presented an efficient CNN called light weight multi-task CNN (LMTCNN) for simultaneous age and gender identification. The LMTCNN uses depth-wise separable convolution to reduce the model size and save the inference time. The method achieves age and gender recognition accuracies of 44% and 85% respectively on Adience dataset.

Reference [36] has proposed a two-stage approach, where at first, the CNN predicts age and gender and also extracts facial representations suitable for face identification by using a modified MobileNet. At the second stage, the extracted facial representations are grouped using hierarchical agglomerative clustering technique. The proposed method achieves 94.1% gender recognition accuracy, and 5.44 MAE on UTKFace dataset. In the same year, [8] have proposed a Multi-Task CNN (MTCNN) with joint dynamic loss weight adjustment towards classification of age and gender from the facial images. The mean classification accuracy of the gender classification task for UTKFace dataset is 98.23%, and for BEFA challenge dataset it is 93.72%. The accuracy of the age classification task for UTKFace dataset is 70.1% and for the BEFA challenge dataset is 71.83%.

Reference [50], have proposed a novel method based on attention long short-term memory (AL) network for fine-grained age estimation. The method combines LSTM units with RoR models or ResNet thus building AL-ResNets or AL-RoR networks. Local features of age-sensitive regions are extracted using the networks. For the age group classification task, 67.83% accuracy is achieved on Adience dataset and an MAE of 3.357 is achieved on the 15LAP dataset. Reference [1] have introduced a novel end-to-end CNN approach in 2020, to identify age and gender from unfiltered faces. A two-level CNN architecture is used. 96.2% gender recognition accuracy and 83.1% age prediction accuracy are achieved by the proposed method on OIU-Adience dataset.

III. DISCUSSION ON THE PAST RESEARCHES

Comparative study of some age and gender prediction methods proposed in last few years is shown in Table 1. We have

TABLE 1. Comparative study of some Gender and Age classification methods proposed in past few years.

References	Database	Advantages	Limitations
Chang et al. (2011)	FG-NET database, MORPH Album 2 database	The information of relative order between ages is more reliably employed than conventional ways of using by OHRank.	Achieves higher MAE (more errors) for age detection than other neural network based methods.
Karimi & Tashk (2012)	FGNET dataset	Achieved high accuracy for gender recognition. Achieved suitable performance even if the utilized images were subjected to intrusive noises.	The task for age recognition produced lower accuracy and the ages were divided into groups for classification
Levi & Hassncer (2015)	Adience dataset	Reduced number of parameters thus reducing chances of overfitting. Also has reasonable accuracy for gender recognition task.	Lower accuracy for the age recognition task due to its simple design.
Samek et al. (2017)	Adience dataset	Achieved reasonable accuracy for gender recognition.	Produced a lower accuracy for the age recognition task
Zhang et al. (2017a)	IMDB-WIKI, ImageNet dataset	Achieved high accuracy for gender classification task. Works well for high-resolution facial images.	Produced a lower age detection accuracy.
Srinivas et al. (2017)	Wild East Asian Face Dataset (WEAFD)	Produced reasonable accuracy for gender classification task. One of a kind dataset consisting primarily of labeled face images of individuals from East Asian countries were designed for classification task.	Achieved very low accuracy for the age detection task.
Das & Dantcheva (2018)	UTKFace dataset, BEFA challenge dataset	Produced high accuracy for gender classification task.	Achieved lower accuracy for age classification task. Used limited amount of facial attributes.
Smith & Chen (2018)	MORPH-II dataset	Produced high accuracy for gender classification task.	The task for age recognition produced higher MAE (more errors). Trivial changes (tilt of head, etc.) in the facial images brought a significant change for the prediction task.
Hosseini et al. (2018)	Adience dataset	Produced reasonable accuracy for gender classification task. The network focused only on useful features as appropriate features were designed to reflect the age and gender correctly.	Achieved lower accuracy for the age detection task.
Lee et al. (2018)	Adience dataset	Can be realized on mobile devices with limited computational resources. Achieved reasonable accuracy for gender classification task.	Achieved very low age detection accuracy. Larger size does not work with datasets of unconstrained face images with face attributes.
Savchenko (2019)	UTKFace dataset	Achieved high accuracy for gender recognition. Training images are not required to have all attributes available.	The task for age recognition produced higher MAE (more errors)
Zhang et al. (2019)	Adience dataset, 15LAP dataset, MORPH and FGNET	Reasonable MAE after addition of local features with global features	Low age group classification accuracy.
Agbo-Ajala & Viriri (2020)	OIU-Adience benchmark	Achieved reasonable accuracy for gender recognition. Also handled some of the variability observed in unfiltered real-world faces	Produced a lower accuracy for the age recognition task

tried to find out the strengths and weaknesses of past works done in this domain over the years, and harnessing the newer resources and algorithms available to us, proposed our architecture.

A. MOTIVATION AND CONTRIBUTIONS

From the above discussion and Table 1, it is clear that the previous methods have a common shortcoming of higher MAE and lower accuracy mainly for the task of age estimation. The accuracy of the gender classification of the methods is also not as high as expected to bridge the gap between the human level and the machine level errors. Minor changes in alignment of the face in the images degrade the performance for some of the methods like the work proposed by [37]. However, the method worked quite well in classifying the gender from the images.

Some methods work well on higher resolution images ([49]) while some other methods have more reliably used the information of relative order between ages for the purpose of correct age identification ([4]). Some of the works are computationally efficient enough to be employed on mobile devices with limited computational resources ([22]). This makes the method very useful in practical scenarios. However, the most important requirement of such method is the precision level which can match to the human's ability.

Keeping in mind the strengths and weaknesses of the previous works, we have made the following contributions in our present work:

1) An architecture harnessing the capabilities of classification and regression for Age identification purposes.
2) Same architecture capable of performing the separate task of gender classification thus ensuring the model's versatility.
3) Introduction of the new concept of Gates for Residual Attention Network used as a backbone of the architecture.
4) Handling the poor performance caused by minor changes in facial orientation by applying attention masks through various channels covering as many combinations as possible.
5) Evaluated on 5 datasets having images of people belonging to different ethnic groups and various background.
6) Achieved lower MAE and higher identification accuracy for age and impressive performance in gender classification.

The overall workflow of our architecture is shown in Figure 1.

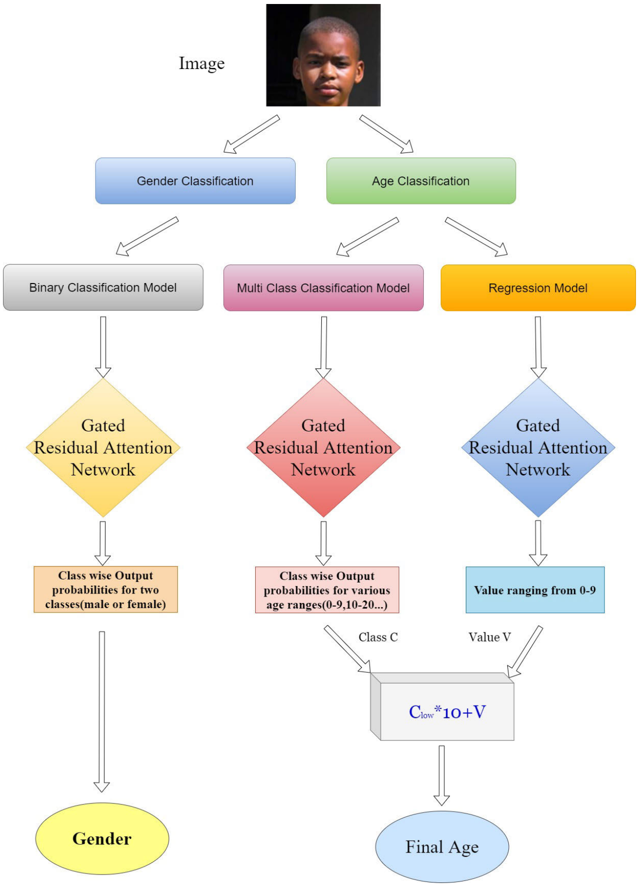


Image
Gender Classification
Age Classification
Binary Classification Model
Multi Class Classification Model
Regression Model
Gated Residual Attention Network
Gated Residual Attention Network
Gated Residual Attention Network
Class wise Output probabilities for two classes(male or female)
Class wise Output probabilities for various age ranges(0-9,10-20...)
Value ranging from 0-9
Class C
Value V
$C_{low}*10+V$
Gender
Final Age

FIGURE 1. Schematic diagram depicting the flowchart of the proposed model.

IV. DATASETS USED IN THE PRESENT WORK

A. FG-NET AGING DATASET

The FG-NET Aging dataset was developed as part of the project FG-NET (Face and Gesture Recognition Network) ([7]). The dataset consists of 1002 images of 82 different subjects with their ages varying from newborns to 69 years old. However, there is class imbalance resulting from over population of images belonging to ages between zero to 40 years in the database. The images with individuals at their more recent ages were the ones for which digital images were available. In most of the cases, the images were collected by scanning photographs of subjects found in personal collections.

Some challenging issues related to this dataset are mainly the quality of images depends on the photographic skills of the photographer and the quality of the imaging equipment that the photographer has used. Also, the quality of photographic paper and printing along with the condition of photographs also have a great impact on the dataset. Thus, the face images in the dataset display considerable variability in quality, illumination, resolution, viewpoint and expression. Another challenge to work upon the dataset is the presence of occlusion in the form of spectacles, facial hair and hats in a number of images. In particular, information about the age, gender, expression, pose, image quality and occlusions like moustaches, beards, hats or spectacles was recorded. Some sample images ([20]) are shown in Figure 2.

B. AFAD DATASET

The Asian Face Age Dataset (AFAD) ([29]) was developed for evaluating the performance of various age estimation

FIGURE 2. Sample facial images taken from FG-NET dataset ([7]).

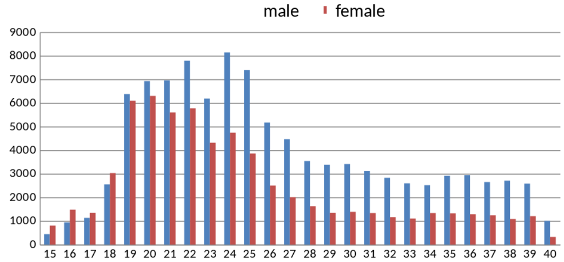


FIGURE 3. Distribution of images in AFAD dataset across age and gender labels ([28]).

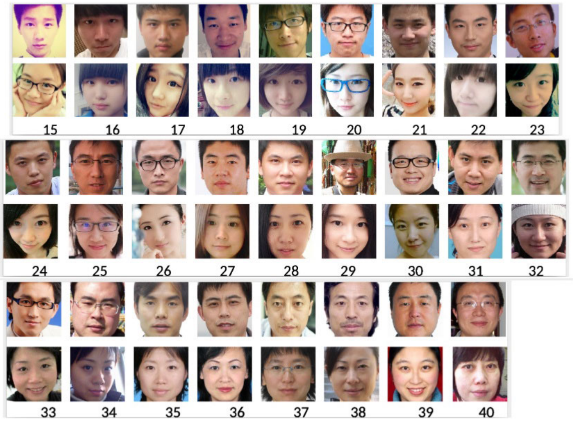


FIGURE 4. Sample facial images present in AFAD dataset ([29]).

FIGURE 5. Sample facial images in Wikipedia Age dataset ([33]).

models and architectures. It contains more than 160K facial images along with their corresponding age labels. This dataset has been designed for estimation of age on Asian faces, so all the facial images are of Asian people. It is to be noted that the AFAD is the largest dataset for age estimation till date. It is a perfect and well suited benchmark dataset to evaluate how deep learning methods can be adopted for age estimation. There are 164,432 labeled photos in the AFAD dataset with the ages varying from 15 to 40.

The AFAD dataset was built by collecting photos of users from a particular social network called RenRen Social Network (RSN). The RSN is a social media platform in China which is widely used by Asian students belonging to all levels of education be it middle school, high school, undergraduate, or graduate students. Even after leaving from school,most of the people still access the platform in order to connect and keep in touch with their old classmates. So, the age of the RSN users belongs to a wide range varying from 15-years to more than 40-years old. The distribution of images across various age groups and gender is shown in Figure 3 and sample images are shown in Figure 4.

C. WIKIPEDIA AGE DATASET

Public availability of the datasets comprising of face images are a challenging issue. If available, it is often of small to medium size and rarely exceeds tens of thousands of images. Moreover, collection of age information for them is quite a challenging task. So, a large dataset ([33]) of faces of various celebrities was collected for this purpose. The most popular 100,000 actors as listed on the IMDb website were taken and various metadata related to that person like date of birth, name and gender were (automatically) crawled from their profiles. This metadata was used to crawl all profile images from the pages of people from Wikipedia. The images which were not timestamped (the date when the photo was taken) were removed as they will not have any age information in them. It was assumed that the images with single faces were likely to show the actors and that the timestamp and date of birth are correct, thus enabling proper assignment of the biological (real) age to each such image. In total, 62,328 face images from 20,284 celebrities were obtained from Wikipedia.

D. UTKFace DATASET

The UTKFace dataset ([52]) is a large scale dataset which consists of face images with a very long age span ranging from 0 to 116 years old subjects. The dataset consists of more than 20K face images with metadata annotations of age, gender, and ethnicity. The images cover large variation in illumination, occlusion, pose, facial expression, resolution, etc. The dataset finds its use in a variety of tasks ranging from face detection, age estimation to age progression/regression and landmark localization, etc. Some sample images are shown in Figure 6.

E. AdienceDB DATASET

The images present in AdienceDB dataset ([9]) were crawled from Flickr.com albums, obtained by automatic upload from smartphones. After downloading the photos from the Flickr site, they were processed by first running the Viola and Jones face detector ([40]) on them. Thereafter, facial feature points were detected using a modified version of the code provided by the authors of ([53]). Many faces in the albums appeared at different roll angles and to avoid missing such

FIGURE 6. Sample facial images found in UTKFace Dataset ([52]).

faces, the process of detecting faces was applied to each and every image, rotated 360° degrees in steps of 5° increments. Finally, the images were manually labeled for gender, age and identity using both the images themselves and any available contextual meta information like image tags and associated text or additional photos in the same album etc.

The data distribution is shown in Table 2. Some sample images of the dataset are shown in Figure 7.

TABLE 2. Distribution of facial images considered from AdienceDB dataset ([9]).

Complete version (faces in the range of ±45° yaw from frontal)									
	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-	Total
Male	745	928	934	734	2308	1294	392	442	8192
Female	682	1234	1360	919	2589	1056	433	427	9411
T. Gnd.	1427	2162	2294	1653	4897	2350	825	869	19487
Total	2519	2163	2301	1655	4950	2350	830	875	
Front version (faces in the range of ±5° yaw from frontal)									
Male	557	691	738	501	1602	875	273	272	5824
Female	492	911	956	630	1692	732	295	309	6455
T. Gnd.	1049	1602	1694	1131	3294	1607	568	581	13649
Total	1843	1602	1700	1132	3335	1607	572	585	

FIGURE 7. Sample images in AdienceDB Dataset ([9]).

V. METHODOLOGY

A. GATED RESIDUAL ATTENTION NETWORK

The Residual Attention Network ([41]) used here has been constructed by combining and stacking multiple Attention blocks. Every block is further subdivided into two branches namely the Mask branch and the Trunk branch. The function of the Trunk branch is to perform feature processing, and it can easily be incorporated into any state-of-the-art network architectures.

In this architecture, we have used pre-activation Residual Unit ([16]) and ResNeXt ([44]) with gated activation as our Gated Residual Attention Network's basic unit to construct Attention block. For a given Trunk branch's output $\mathbb{P}(X)$ with input X, the Mask branch makes use of a bottom-up top-down approach ([2], [26], [30]) for learning the same size mask $\mathbb{K}(X)$ that softly adds weights to the output features $\mathbb{P}(X)$. The bottom-up top-down approach tries to mimic the fast feed-forward (top-down) and feedback attention (bottom-up) mechanisms. The output mask serves as the control gate for neurons of the Trunk branch similar to that of Highway Network ([39]). The output of Attention $\mathbb{O}$ is given as shown in Eq. 1.

$$\mathbb{O}_{i,c}(X) = \mathbb{K}_{i,c}(X) * \mathbb{P}_{i,c}(X) \tag{1}$$

where, i ranges over all spatial positions and $c \in \{1, \ldots, C\}$ is the index of the channel.

The attention mask that is present in the Attention blocks, not only serves in the feature selection procedure during forward inference, but also plays a vital role as a gradient update filter during back propagation. In the soft mask branch, the gradient of mask for the input feature is shown in Eq. 2.

$$\frac{\partial \mathbb{K}(X,\theta)}{\partial \phi}\mathbb{P}(X,\phi) = \mathbb{K}(X,\theta)\frac{\partial \mathbb{P}(X,\phi)}{\partial \phi} \tag{2}$$

where, θ are the mask branch parameters and ϕ are the trunk branch parameters.

This property makes Attention blocks robust to noisy labels. Mask branches have the property to prevent wrong gradients (from noisy labels) and to update Trunk parameters. Instead of stacking Attention blocks to model the architecture, a simpler approach is to make use of a single network branch in order to generate a soft weight mask which is similar to spatial transformer layer ([18]). In our model, features from different layers need to be modeled automatically by the different attention masks to capture the most effective features for age and estimation. Use of a single mask branch requires an exponential number of channels in order to capture all combinations of the different factors. A single Attention block only modifies the features once. If such scenarios arise where the modification fails on some parts of the image, the subsequent network modules are not able to get a second chance. The Gated Residual Attention Network alleviates the aforementioned problems. In Attention block, every Trunk branch has its own mask branch with the purpose to learn attention that is specialized for its features. Besides, for complex images, the incremental nature of stacked network configuration can gradually refine attention.

B. GATED RESIDUAL ATTENTION LEARNING

Just stacking up of Attention blocks in a naive manner which may lead to the performance drop. This mainly occurs due to two major reasons. Firstly, the dot product with mask range in between zero to one repeatedly results in degradation of the value of features in the deeper layers. Secondly, the soft mask can potentially break the good property of Trunk branch like the identical mapping of Residual Unit. We have used gated residual attention learning to ease the aforementioned problems. Similar to ideas in residual learning, if a soft mask unit can be designed so as to serve as identical mapping, the performances may be no worse than its counterpart without attention. Thus, we use a modified version of output $\mathbb{O}$ of Attention block as given in Eq. 3.

$$\mathbb{O}_{i,c}(X) = (\gamma + (1-\gamma) * \mathbb{K}_{i,c}(X)) * \mathbb{F}_{i,c}(X) \tag{3}$$

where, $\mathbb{K}(X)$ is in the range of [0,1], with $\mathbb{K}(X)$ approximating 0, $\mathbb{O}(X)$ will approximate original features $\mathbb{F}(X)$. The values γ and $(1-\gamma)$ signify the Gates of range [0,1] and are trainable parameters, controlling the effects of Mask branch and features generated by Deep Convolutional Networks on output $\mathbb{O}_{i,c}(X)$.

We call this method as Gated Residual Attention learning which is different from residual learning and the architecture therefore is termed as Gated Residual Attention Network (GRA_Net). Figure 8 shows the schematic working of a Gated Attention block.

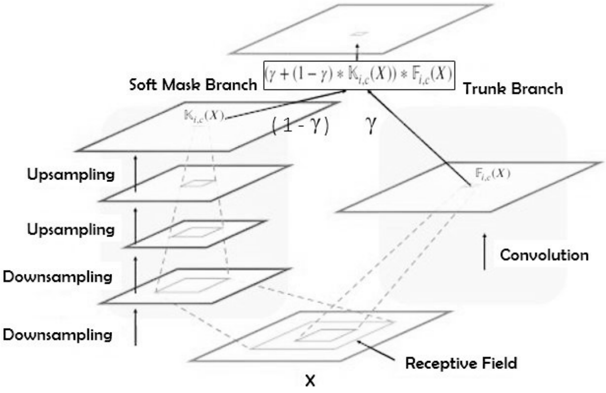


FIGURE 8. Schematic diagram representing Gated Attention block.

The effect of Gates on the output of Attention block can be inferred from the Table 3.

TABLE 3. Effect of Gates at extreme values on the output of Gated Attention Network.

γ	$1-\gamma$	$\mathbb{O}_{i,c}(X)$	**Remarks**
0	1	$\mathbb{K}_{i,c}(X) * \mathbb{F}_{i,c}(X)$	Full effect of Mask branch
1	0	$\mathbb{F}_{i,c}(X)$	No effect of Mask branch

In the original ResNet, residual learning is formulated as $\mathbb{O}_{i,c}(X) = X + \mathbb{F}_{i,c}(X)$, where $\mathbb{F}_{i,c}(X)$ approximates the residual function. In our formulation, $\mathbb{F}_{i,c}(X)$ represents the features generated by Deep Convolutional Networks. The mask branches, $\mathbb{K}(X)$, play the role of feature selector which primarily aim to keep the good features and suppress any kind of noises from the Trunk features.

Additionally, stacking up Attention blocks plays the role of backing up Gated Residual Attention learning by its incremental nature. This learning has the ability to keep good properties of original features, but also provides them the ability to bypass soft mask branch and forward to the top layers in order to weaken the mask branch's feature selection ability. Stacked Attention blocks can lead to gradual refinement of the feature maps.

As shown in Figure 9, the features become more and more useful with increasing depths. By using the Gated Residual Attention learning, increasing depth of the network can improve performance consistently.

C. SOFT MASK BRANCH

By making use of the previous attention mechanism concept as present in Deep Belief Network (DBN) like Restricted Boltzmann Machines(RBM) ([21]), the mask branch contains fast feed-forward sweep and top-down feedback steps. The feed-forward operation manages to quickly collect the global information available in the whole image, while the latter one tries to combine the same with the original feature maps.

In the CNN, the two steps unfold into bottom-up and top-down fully convolutional architectures. From the input, repetitive applications of max pooling is done in order to increase the receptive field with a rapid rate after a small number of Gated Residual Units. As soon as the lowest resolution is reached, the global information is then expanded by a symmetrical top-down architecture in order to guide the input features in each pixel position.

Linear interpolation tries to up sample the output after some Residual Units as seen in Figure 10. The number of bi-linear interpolation applied is the same in number as max pooling to make sure that the output size is the same as the input feature map. Then after two consecutive $1X1$ convolution layers, a sigmoid layer normalizes the output range to [0, 1].

Skip connections have also been added between the top-down and bottom-up components with the motive of capturing information from different scales without any loss of useful information from the knowledge of previous layers. The complete module is illustrated in Figure 10. The main aim of the mask branch is to improve the Trunk branch features in place of solving a complex problem directly.

D. SPATIAL AND CHANNEL ATTENTION

In this work, attention provided by the mask branch changes continuously by adapting with the features of the Trunk branch. However, attention in the mask branch can be restricted by making changes in the normalization step of the activation function before soft mask output. We have made use of 3 types of activation functions corresponding to Channel attention, Mixed attention and Spatial attention. The Mixed attention $\mho_1$ (refer Eq. 4) without any additional constraints makes use of a sigmoid function for each channel and each spatial position. Channel attention $\mho_2$ (refer Eq. 5) applies L2 normalization within all channels for each and

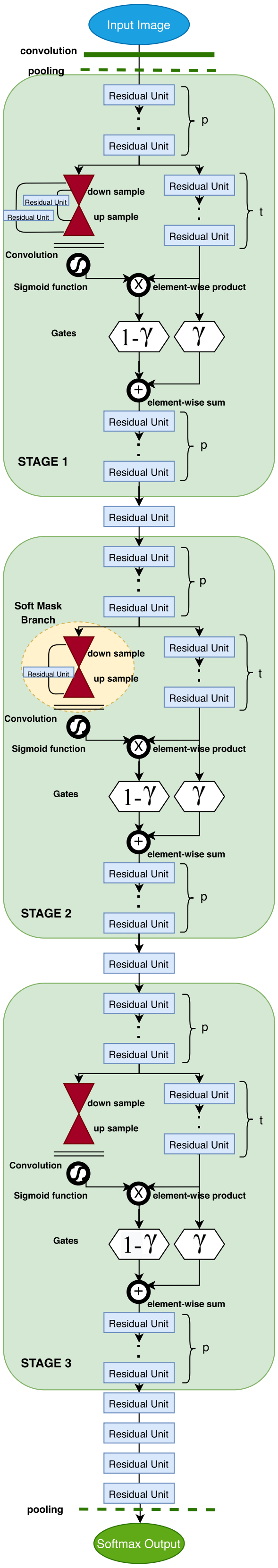


FIGURE 9. Overall architecture used in the present work for classification of age and gender from facial images.

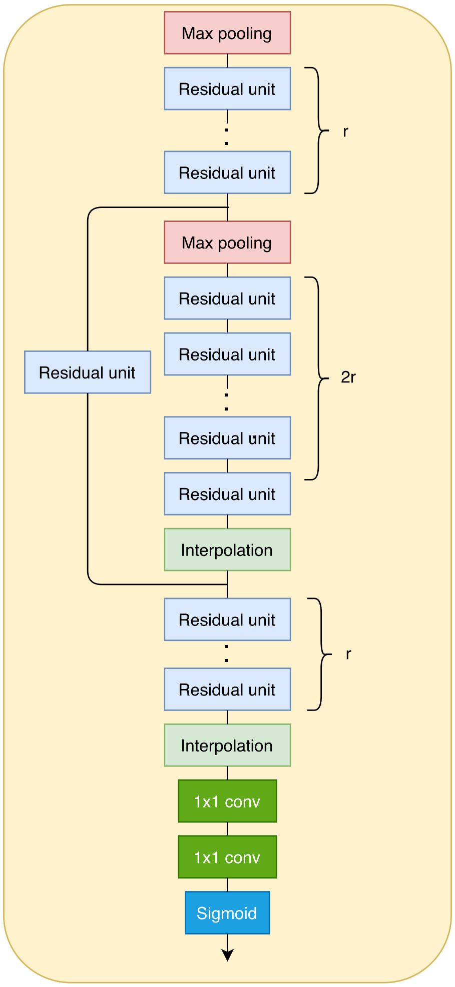


FIGURE 10. Illustration of Soft Mask Branch.

every spatial position in order to reduce spatial information. Spatial attention $\mho_3$ (refer Eq. 6) executes normalization within the feature map from each channel and then sigmoid activation in order to retrieve a soft mask related to spatial information only.

$$\mho_1(X_{i,c}) = \frac{1}{1+\exp(-X_{i,c})} \tag{4}$$

$$\mho_2(X_{i,c}) = \frac{X_{i,c}}{||X_i||} \tag{5}$$

$$\mho_3(X_{i,c}) = \frac{1}{1+\exp(-(X_{i,c}-mean_c)/std_c)} \tag{6}$$

where, i ranges over all spatial positions and c ranges over all channels. $mean_c$ and std_c denote the average and standard deviation of feature map from c^{th} channel respectively. X_i denotes the feature vector at the i^{th} spatial position.

E. REGRESSION

Only difference in this model is the output layer which is a Dense Layer consisting of a single node with linear activation and the labels being fed for training are the digits from 0 to 9.

F. LABEL DIVISION

The division of age label into various sub-labels is shown in Figure 11.

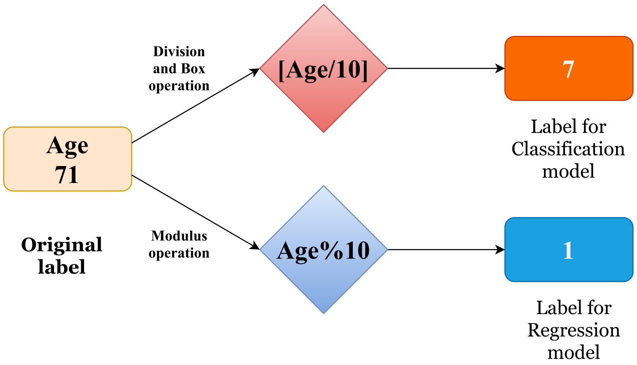

Division and Box operation
[Age/10]
7
Label for Classification model
Age 71
Original label
Modulus operation
Age%10
1
Label for Regression model

FIGURE 11. Decomposition of orignal age label into labels for different models.

VI. ANALYSIS OF RESULTS

A. PRE-PROCESSING AND PARAMETER SETTINGS

1) GENDER IDENTIFICATION

Facial images are first scaled and resized to 32 × 32 keeping the RGB channels intact. These images are then passed dynamically by an ImageDataGenerator object to the model for training purposes. The output layer consists of a softmax layer with two units depicting ''Male'' and ''Female'' classes. As this is a binary classification problem, so loss function used here is binary cross entropy. Optimization is done using Nadam (Adam with Nesterov momentum) optimizer with a learning rate of 0.001. The final model consisted of 33 million trainable parameters.

2) AGE IDENTIFICATION

Similar to the process of gender identification, the images are first scaled and resized to 32 × 32 keeping the RGB channels intact. Thereafter, these images are passed by an ImageDataGenerator object but to two different models, one for classifying the age category, and another one for classifying the regression value of exact age for the subject. As the classification model is used here a multi-class classification problem, so loss function used is categorical cross entropy and optimizer used is same as that used in gender identification. This model also consists of 33 million trainable parameters.

B. EVALUATION METRICS

For age estimation problem, in order to guarantee the accuracy of our algorithm and provide fair comparison with available state-of-the-art models, MAE ([35]) is taken as the evaluation metric, which minimizes the error between the estimated age and the ground truth label. MAE($\mathcal{J}(\mathcal{X})$ is given by Eq. 7.

$$\mathcal{J}(\mathcal{X}) = \frac{1}{M}\sum_{i=1}^{M}|\tilde{Y_i} - Y_i| \tag{7}$$

where, $\tilde{Y_i}$ = True Age and Y_i = Predicted Age for i^{th} data point.

Apart from MAE, the test accuracy for some datasets has also been measured for comparison purposes.

For gender classification, we have measured the test accuracy for the datasets which have labels for the same. As gender classification is basically a binary classification problem, so the evaluation metric ''accuracy'' is defined as shown in Eq. 8.

$$Accuracy = \frac{(tp + tn)}{(tp + tn + fp + fn)} \tag{8}$$

where, tp = True positive; fp = False positive; tn = True negative; fn = False negative.

C. EXPERIMENTAL RESULTS

1) AGE IDENTIFICATION

The MAEs attained by the proposed GRA_Net model for age identification over five benchmark datasets are shown in Table 4. The graphs representing the change in the MAE with the increase in the number of epochs for Wikipedia, AFAD, UTKFace, FG-Net and AdienceDB datasets are also illustrated in Figure 12 to Figure 16 respectively. In all the graphs, it can be clearly visualized that with the increase in number of epochs of the model, the MAE value decreases for each of the training and validation processes. This shows that the model has been properly trained and validated. However, there are no signs of saturation or steadiness in the MAE values for any of the training process graphs, which indicate that there is no case of overfitting. Also, the fact that the training MAE never gets lower than the validation MAE for every epoch further validates the absence of overfitting.

TABLE 4. MAEs achieved by the proposed GRA_Net model for age classification over five standard benchmark datasets.

Dataset	MAE
FG-NET	3.23
AFAD	3.10
Wikipedia Age	5.45
UTKFace	1.07
AdienceDB	10.57

From the graph present in Figure 12 for the Wikipedia dataset, there are a few fluctuations seen in the graph which indicate that there is an uniform distribution of facial images per age group with a few noises present in it. The graph for the AFAD dataset (shown in Figure 13) shows a uniform exponential graph, thereby indicating the fact that there is a uniform distribution of facial images per age group. Whereas the graph for the UTKFace dataset presented in Figure 14 shows numerous fluctuations, thus, indicating a presence of high noise and non-uniform distribution of images per age group. The graphs for FG-Net and AdienceDB datasets show a uniform distribution of facial images per age group with a less noisy graph.

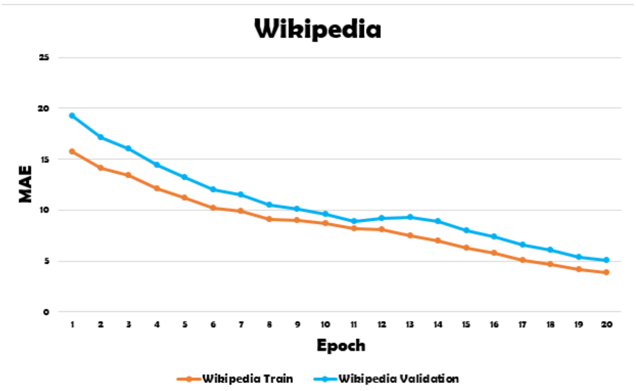


FIGURE 12. Graph showing the decrease of MAE per epoch for Wikipedia dataset.

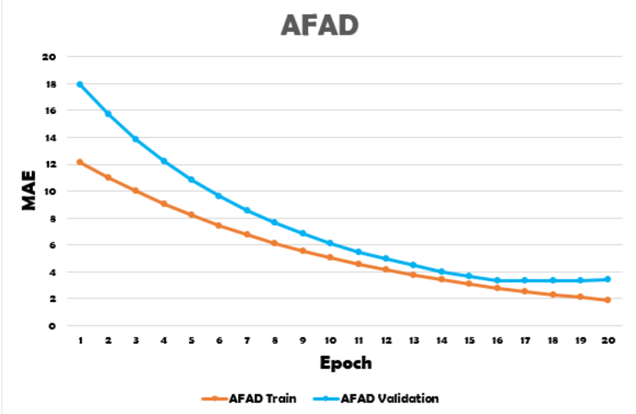


FIGURE 13. Graph showing the decrease of MAE per epoch for AFAD dataset.

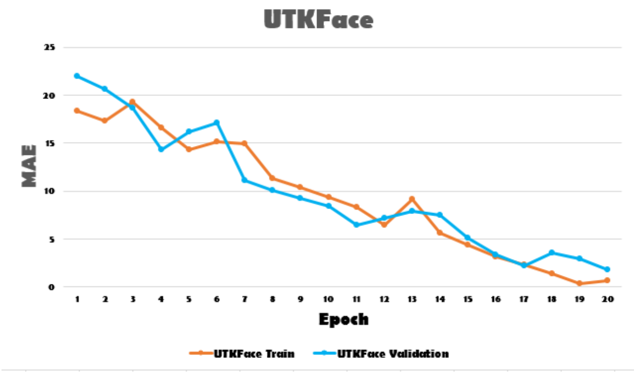


FIGURE 14. Graph depicting the decrease of MAE per epoch for UTKFace dataset.

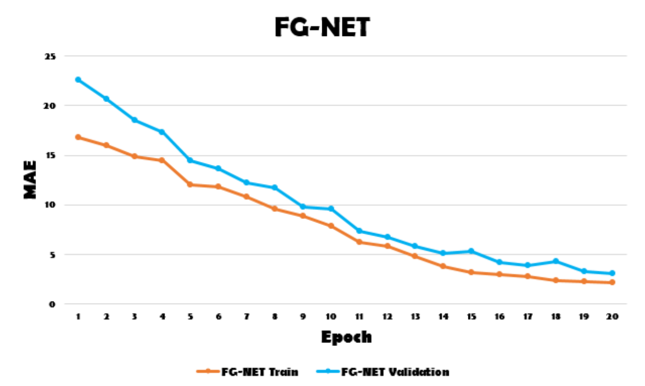


FIGURE 15. Graph illustrating the decrease of MAE per epoch for FG-NET dataset.

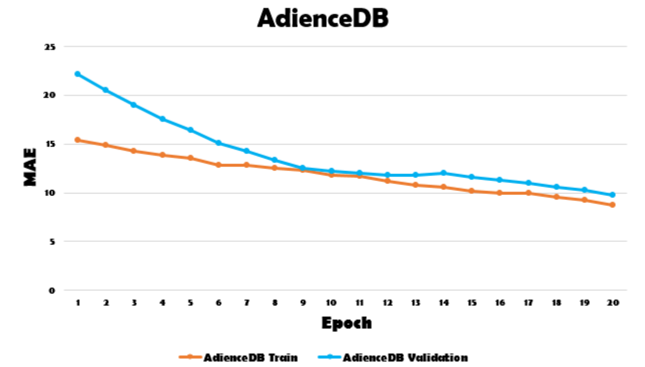


FIGURE 16. Graph illustrating the decrease of MAE per epoch for AdienceDB dataset.

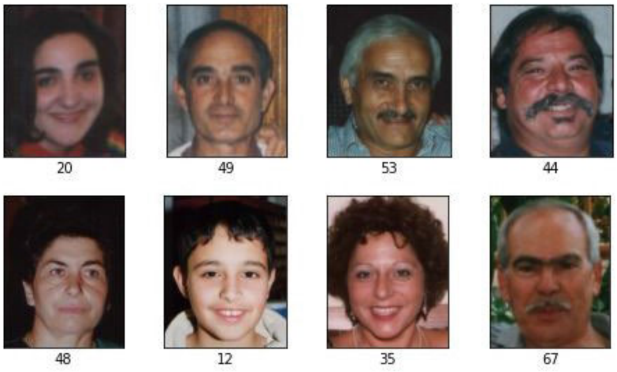


FIGURE 17. Some sample images(along with their predicted ages given below) taken from FG-NET dataset where our model identifies age correctly.

Some correct identification results with almost 100% precision are shown from Figs. 17-21.

All the facial images shown in Figure 17 are clear and do not show any signs of blurring in them. Except the 1st, 6th and 7th images, all of the facial images in Figure 17 have wrinkles present in them which could be a reason for the classifier to predict correctly. The presence of hair loss, facial hair could further validates the perfection of predicted ages for the 2nd, 3rd, 4th and 8th images. The greying of hair present in the 3rd and 8th images could ensure higher ages for these facial images. Whereas the absence of all these facial features in the other images could be a reason for the classifier to predict them as of lesser ages. The features representing those of a kid present in the 6th image validates the correct prediction.

In Figure 18, the last three facial images show presence of wrinkles and hair loss in them which is a prominent sign of

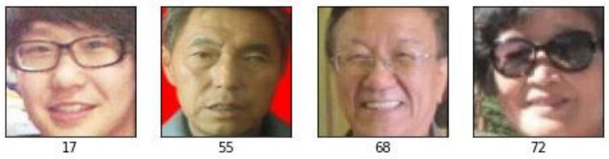


FIGURE 18. Some sample images (along with their predicted ages given below) taken from AFAD dataset where our model identifies age correctly.

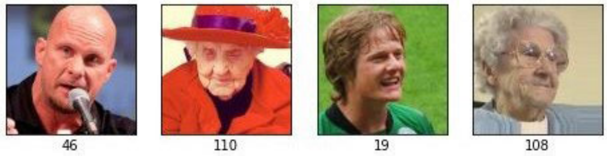


FIGURE 19. Some sample facial images (along with their predicted ages given below) taken from Wikipedia Age dataset where our model identifies age correctly.

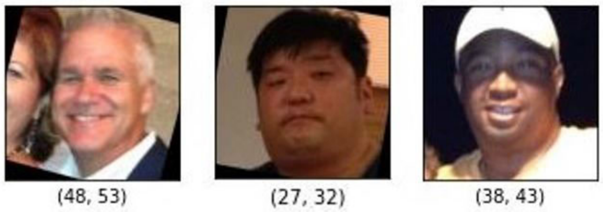


FIGURE 20. Some sample facial images (along with their predicted age groups given below) taken from AdienceDB dataset where our model identifies age correctly.

aging. The presence of a spectacle along with wrinkles could also help the classifier to predict a higher age for the last two images. Whereas the first image does not show any of the mentioned attributes thus facilitating the classifier to predict a lesser age for it.

The first image in Figure 19 shows signs of hair loss thus facilitating the classifier to predict a higher age. The presence of large amount of wrinkles and greying of hair could be a major factor for the classifier to predict a very high age for the 2nd and last images. Whereas the absence of all these features helps in correct prediction of a lower age for the 3rd image.

In Figure 20, the first facial image shows presence of hair loss and greying of hair which could help the classifier in predicting a higher age. Whereas no such attributes are present in the other two images, thus facilitating the classifier in predicting a lower age for them.

The presence of wrinkles in all of the images except the 2nd and 4th in Figure 21 could be a major factor for the correct prediction of a higher age in those images. Also the presence of greying of hair along with hair loss in the last three images could be the reasons for the correct prediction of an even higher age. The presence of the features a baby has in the 2nd image could be a reason for the correct prediction.

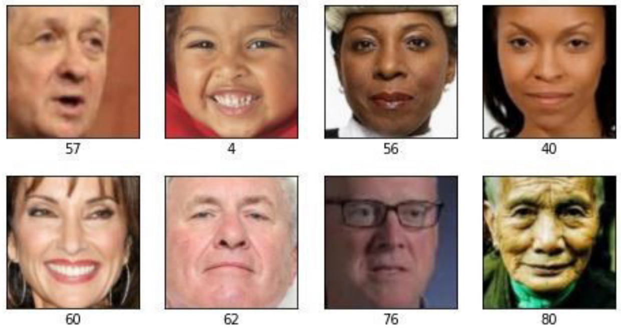


FIGURE 21. Some sample images (along with their predicted ages given below) taken from UTKFace dataset where our model identifies age correctly.

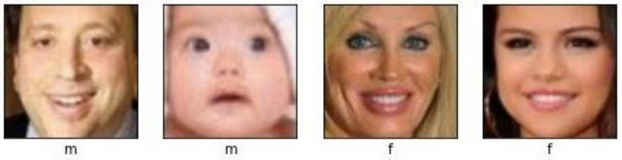


FIGURE 22. Some sample images (along with their predicted gender given below) taken from UTKFace dataset where our model identifies their gender correctly.

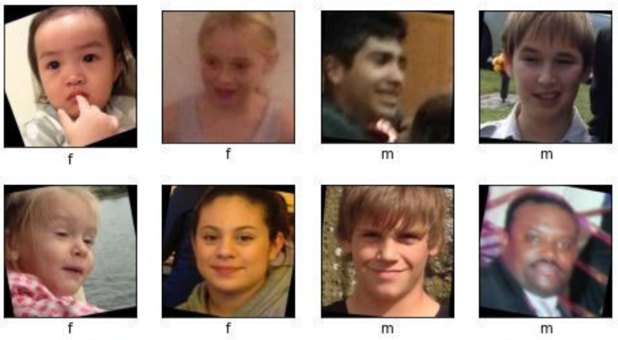


FIGURE 23. Some sample images (along with their predicted gender given below) taken from AdienceDB dataset where our model identifies their gender correctly.

2) GENDER CLASSIFICATION

The classification accuracies achieved by the proposed GRA_Net model for UTKFace and AdienceDB datasets are found to be 99.2% and 82% respectively. Some facial images showing correct gender identification for the UTKFace and Adience DB datasets are shown in Figure22 and Figure 23 respectively (m = Male, f = Female).

The presence of feminine features such as long hair, bigger eyelashes in the last two images could be the reason for correct prediction of female gender prediction for these images. Whereas the absence of these features in the other two images helps the model to predict them as male gender. The presence of a longer hair being a major feminine feature could be seen in the 1st, 2nd, 5th and 6th images, thereby helping the classifier to correctly predict their genders. Whereas the

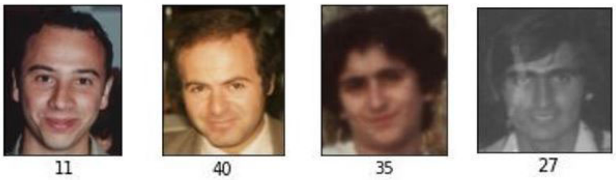


FIGURE 24. Some sample images(along with their predicted ages given below) taken from FG-NET dataset where our model identifies age incorrectly.

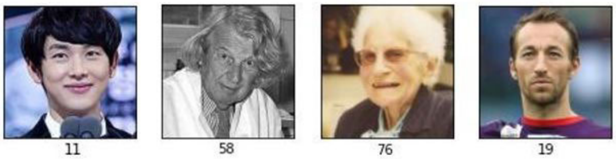


FIGURE 26. Some sample images (along with their predicted ages given below) considered from Wikipedia Age dataset where our model fails to identify the ages correctly.

presence of a shorter hair in the other images ensures the facial images to be of male gender. Also the presence of facial hair further validates male gender for the last image.

D. ERROR ANALYSIS

1) AGE IDENTIFICATION

Here, we have made a detailed analysis of the error cases done by our model on age identification for all the datasets.

In Figure 24, the correct ages for the people in the images are 21, 27, 15 and 16 respectively. The first image is wrinkle-free and devoid of some important features depicting higher ages like moustache and beards. This has resulted in misclassification of the age category, thus predicting $11(1 \times 10 + 1)$ instead of $21(2 \times 10 + 1)$. The second image has comparatively lesser hair and hair fall is in important aspect of aging. This might have led to the corresponding misclassification. The third and fourth pictures are a bit blurry. The mixing of hair with the background for the fourth image deferred the performance of the classifier too.

In Figure 25, the correct ages for the people in the images are 63, 19, 57 and 56 respectively. The first image has a side-view orientation which might have affected the performance of our model. The presence of wrinkles made it possible to predict an age which is very precise and close enough to the actual value (label value). The second one is a low quality image, thus resulting in blurriness and loss of features which might have been mistaken as presence of age marks, and leading to higher age value than the actual age. We can see that in the third and fourth figures, the spectacles are there but clarity of its presence is minimal. Also there are presence of wrinkles in the third image but that becomes absent due to softening filter applied to the image while capturing of the portrait. This has led to the estimation of higher age value for the third image compared to the fourth image. We can see that the hair color blends with the background which might have created a problem to the model, and accurate prediction is not possible.

In Figure 26, the correct ages for the people in the images are 26, 71, 95 and 33 respectively. People belonging to countries like Korea, China, Japan etc. look younger compared to people of other countries belonging to the same age group. It can be seen that our model tends to learn this feature too thus giving wrong prediction as seen in the first image. Presence of longer hairs and loss of features pertaining to the image being grayscale, aided to the mis-classification of age for the second image by making the model develop a perception of younger age. Though the third image is not grayscale but blurriness has led to covering up of some wrinkles and presence of lot of hair might have led to the identified wrong age. Athletes being more fit than any average individual because of their excessive physical activities tend to look younger and live longer. These characteristics led to smaller age prediction resulting in wrong prediction.

In Figure 27, the correct age groups for the people in the images are 38-43, 60-100 and 15-20 respectively. For this dataset, we have calculated the MAE by taking the average class age values for different age groups(continuous data groups). The side-view orientation, the presence of a hat as well as presence of lesser facial features resulting from smaller foreground to background ratio compared to the training images, might have led to wrong prediction for the first image sample. The age estimation like the one for the second image is the sole reason for producing highest MAE among all the datasets. It can be considered as a corner case, where our model fails as it is a random error which cannot be explained by any of the aforementioned reasons.

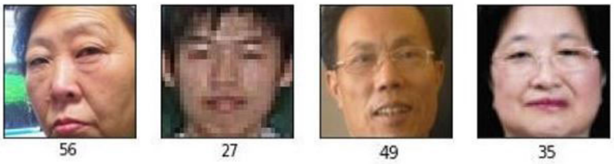


FIGURE 25. Some sample images (along with their predicted ages given below) of AFAD dataset where our model identifies age incorrectly.

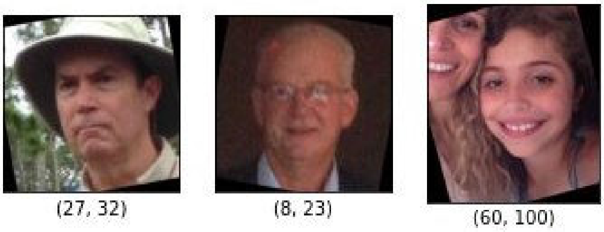


FIGURE 27. Some sample images (along with their predicted age groups given below) taken from AdienceDB dataset where our model identifies age incorrectly.

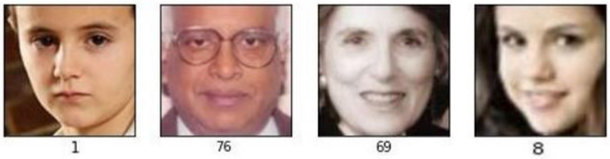


FIGURE 28. Some sample images (along with their predicted ages given below) of UTKFace dataset where our model identifies age incorrectly.

The mis-classification for the third image is explainable in the sense that it consists of two faces which it has not been trained upon, thus the prediction is quite a random one, though such random predictions have added up to the MAE, but the absence of such images might have given more exceptional results than the somewhat better results obtained.

In Figure 28, the correct ages for the people in the images are 11, 58, 53 and 18 respectively. In the first image the child looks very young. The characteristic features of a baby are present in the image, which might have resulted in wrong prediction of the classifier model though the regression model gives accurate result. The model might have learnt that the presence of spectacles and white hair together represents higher age values, the result of which can be seen in the wrong age prediction for the second image. The third image has wrinkles and almost zero visibility of hair, resulting in wrong prediction of age (higher value here). The fourth image consists of the same features as described for first image, and both the images are a bit side aligned lessening the prediction performance of our model.

2) GENDER CLASSIFICATION

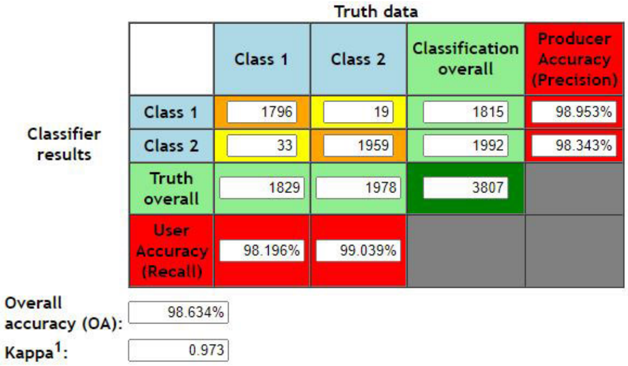
Here, we have made a detailed analysis of the error cases done by our model on gender prediction for the UTKFace and AdienceDB datasets. The accuracies attained by the proposed GRA_Net model for gender classification over two benchmark datasets are shown in Table 5.

Classifying gender of small children is a very challenging task even for humans, thus leading to higher Bayesian errors for this class of images. The reason is that the masculine and feminine features start to become distinguishable only after reaching the age of puberty. As seen in Figure 30, in both the images an important feature which is hair is absent, but this helps in distinguishing gender of a child. Thus the model faces difficulty in classifying the gender properly and hence, gives wrong results.

As seen in Figure 32, the image is that of a child with some drawings made by markers on her cheeks. As explained above, firstly the image is of a child thus increasing the difficulty of classification for the model. Secondly, the most accurate reasoning for mis-classification that can be explained with strong ground, is the idea of beard learnt by the model as an important feature for identifying males. In this image, our the model has mistaken the drawing as an actual feature of beards thus giving a wrong prediction.

TABLE 5. Accuracies achieved by the proposed GRA_Net model for gender classification over two standard benchmark datasets.

Dataset	Accuracy(%)
UTKFace	99.2
AdienceDB	81.4 ± 0.6



		Truth data			
		Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	1796	19	1815	98.953%
	Class 2	33	1959	1992	98.343%
	Truth overall	1829	1978	3807	
	User Accuracy (Recall)	98.196%	99.039%		

Overall accuracy (OA): 98.634%

Kappa[1]: 0.973

FIGURE 29. Confusion matrix produced by the proposed model for gender classification on UTKFace dataset.

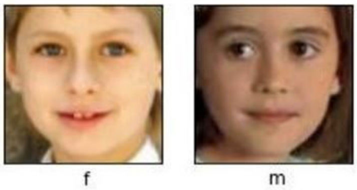
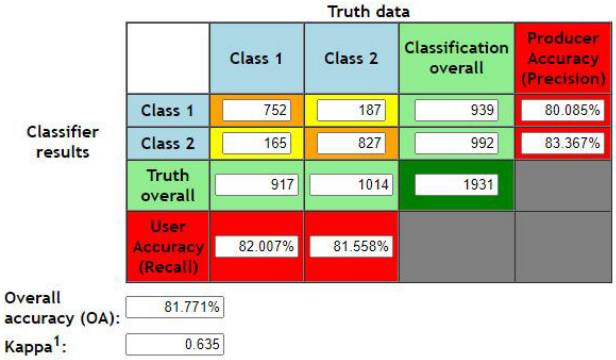


FIGURE 30. Some sample images (along with their predicted genders given below) of UTKFace dataset where our model identifies gender incorrectly.



		Truth data			
		Class 1	Class 2	Classification overall	Producer Accuracy (Precision)
Classifier results	Class 1	752	187	939	80.085%
	Class 2	165	827	992	83.367%
	Truth overall	917	1014	1931	
	User Accuracy (Recall)	82.007%	81.558%		

Overall accuracy (OA): 81.771%

Kappa[1]: 0.635

FIGURE 31. Confusion matrix generated by the proposed model for gender classification on AdienceDB dataset.

E. COMPARISON OF THE PROPOSED GRA_Net MODEL WITH SOME PREVIOUS METHODS

Tables 6- 9 show the performance comparison of our proposed GRA_Net model with some state-of-the art models for FG-Net, AFAD, UTKFace and AdienceDB datasets respectively.

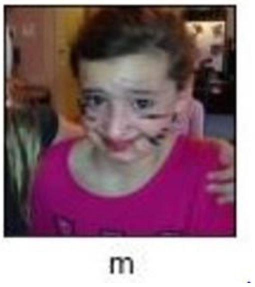


FIGURE 32. A sample facial image (along with the predicted gender given below) showing incorrect gender classification for AdienceDB dataset.

TABLE 6. Comparison of age identification results (measured in terms of MAE) with other methods for FG-NET dataset.

Method	MAE
Proposed model	**3.23**
MSFCL-LR (Xia et al. (2020))	2.79
MSFCL (Xia et al. (2020))	2.90
LSDML (Liu et al. (2017))	3.53
RAN (Wang et al. (2017))	4.05
CS-LBMFL (Lu et al. (2015))	4.36
CS-LBFL (Lu et al. (2015))	4.43
OHRank (Absolute cost) (Chang et al. (2011))	4.48
CA-SVR Lu et al. (2015)	4.67
PLO (Li et al. (2012))	4.82
MTWGP (Zhang & Yeung (2010))	4.83
PFA Guo et al. (2008b)	4.97
LARR (Guo et al. (2008a))	5.07
MHR (Qin et al. (2007)) with cost sensitivities	4.87
SSE (Yan et al. (2008))	5.21
RED-SVM (Chang et al. (2010))	5.24
RUN2 (Yan et al. (2007b))	5.33
GP (Zhang & Yeung (2010))	5.39
RankBoost (Yang et al. (2010))	5.67
LDL (Geng et al. (2013))	5.77
RUN1 (Yan et al. (2007a))	5.78
SVR (Niu et al. (2016))	5.91
AGES (Geng et al. (2007))	6.77
SVM	7.25
WAS (Geng et al. (2007))	8.06
KNN	8.24

It can be seen from Tables 6 - 9, that our model outperforms almost all previous models considered here for comparison in terms of MAE for age identification as well as accuracy in terms of gender classification. This proves the effectiveness of our model compared to the state-of-the-art models. Though works like MSFCL and MSFCL-LR ([43]) have better results owing to their higher complexity in terms of parameters as well as custom loss function. The inclusion of custom loss function in our architecture can be considered as part of our future work.

TABLE 7. Comparison of age estimation results (in terms of MAE) with some past methods for AFAD dataset.

Method	MAE
Proposed model	**3.10**
OR-MOCNN (Niu et al. (2016))	3.34
RAN (Wang et al. (2017))	3.42
BIFS+OHRank (Chang et al. (2011))	3.84
BIFS+LSVR (Guo et al. (2009))	4.13
BIFS+OR-SVM (Chang et al. (2010))	4.36
BIFS+CCA (Guo & Mu (2013))	4.40
CNN+LSVR (Wang et al. (2015))	5.56

TABLE 8. Comparison of both age and gender classification results with some past methods measured in terms of accuracy for UTKFace dataset.

Model	Gender(%)	Age(%)
Facenet	91.2	56.9
Finetuned Facanet (FFNet)	96.1	64
MTCNN	98.23	70.1
RAN (Wang et al. (2017))	97.5	85.4
Proposed model	**99.2**	**93.7**

TABLE 9. Comparison of both age and gender classification results with some past methods for AdienceDB dataset.

Method	Age(%)	Gender(%)
Proposed model	65.1 ± 2.1	81.4 ± 0.6
RAN (Wang et al. (2017))	57.3 ± 1.9	77.2 ± 0.4
LBP	41.4 ± 2.0	73.4 ± 0.7
FPLBP	39.8 ± 1.8	72.6 ± 0.9
LBP+FBLBP	44.5 ± 2.3	-
LBP+FBLBP+PCA 0.5	38.1 ± 1.4	76.1 ± 0.9
LBP+FBLBP+PCA 0.8	32.9 ± 1.6	-
LBP+FBLBP+Dropout 0.5	44.5 ± 2.6	-
LBP+FBLBP+Dropout 0.8	45.1 ± 2.6	-
MSFCL (Xia et al. (2020))	64.7	-
MSFCL-KL (Xia et al. (2020))	65.3	-

It is to be noted that no such previous work exists for the Wikipedia Age dataset which could be used for comparison purposes, so we have not included any comparison for it.

VII. CONCLUSION

Identifying the age and gender of the individuals we come across in our daily lives has an important role in our social lives too. For example, languages used to salute for men and women are very often different, and the words used to address the elders and the young are different too. We human beings are able to evaluate the individual's age and gender just from their facial appearances. Nowadays, many smart applications that include visual surveillance, medical diagnosis, and marketing intelligence, need to evaluate the same for individuals using their facial images. Here lies the importance of a robust and efficient methodology for gender and age estimation through the computing devices. To this end, in this paper, we have proposed a deep learning based model, named GRA_Net, for the purpose of age (regression problem) and gender (a binary classification problem) prediction from the facial images. We have considered the age prediction problem as a combination of classification and regression problems. Our proposed model has been evaluated on five publicly available standard datasets. Obtained results for both the tasks have been able to outperform many state-of-the-art methods.

There are still some rooms for improvement of the proposed model which can be done in the near future. For example, we have to do more work while identifying the gender of kids as in the tender age both the male and female individuals have many commonalities in the facial features. Also, our model needs to be more intelligent in estimating the age and gender when images are obstructive, partially viewed, bearing hat/glass/wig, and wearing some unusual make-up etc. Even facial images of different provinces of the world have different characteristics. Hence, we will have to make our model more adept towards this.

ANNEX

Keras implementation of Residual and Gated Attention Blocks.

```
from keras.layers import BatchNormalization
from keras.layers import Conv2D
from keras.layers import UpSampling2D
from keras.layers import Activation
from keras.layers import MaxPool2D
from keras.layers import Add
from keras.layers import Multiply
from keras.layers import Lambda

from keras import backend as K
from keras.engine.topology import Layer
import numpy as np

from keras.layers import add

class GatedLayer(Layer):

    def __init__(self, **kwargs):
        super(GatedLayer, self).__init__(**kwargs)

    def build(self, input_shape):
        # Create a trainable weight variable for this layer.
        self._gamma = self.add_weight(name=
            'gamma', shape=(1,),
            initializer='uniform',
            trainable=True)
        super(GatedLayer, self).build(input_shape)

    def call(self, x):
        return Add()([self._gamma, Multiply()([1-self._gamma, x])])

    def compute_output_shape(self, input_shape):
        return input_shape[0]


def residual_block(input, input_channels=None, output_channels=None, kernel_size=(3, 3), stride=1):
    if output_channels is None:
        output_channels = input.get_shape()[-1]
    if input_channels is None:
        input_channels = output_channels // 4

    strides = (stride, stride)

    x = BatchNormalization()(input)
    x = Activation('relu')(x)
    x = Conv2D(input_channels, (1, 1))(x)

    x = BatchNormalization()(x)
    x = Activation('relu')(x)
    x = Conv2D(input_channels, kernel_size, padding='same', strides=stride)(x)

    x = BatchNormalization()(x)
    x = Activation('relu')(x)
    x = Conv2D(output_channels, (1, 1), padding='same')(x)

    if input_channels != output_channels or stride != 1:
        input = Conv2D(output_channels, (1, 1), padding='same', strides=strides)(input)

    x = Add()([x, input])
    return x


def attention_block(input, input_channels=None, output_channels=None,
    ↪ encoder_depth=1):

    p = 1
    t = 2
    r = 1

    if input_channels is None:
        input_channels =
        ↪ input.get_shape()[-1]
    if output_channels is None:
        output_channels = input_channels

    # First Residual Block
    for i in range(p):
        input = residual_block(input)

    # Trunc Branch
    output_trunk = input
    for i in range(t):
        output_trunk =
        ↪ residual_block(output_trunk)


    # Soft Mask Branch

    ## encoder
    ### first down sampling
    output_soft_mask =
    ↪ MaxPool2D(padding='same')(input)
    ↪ # 32x32
    for i in range(r):
        output_soft_mask =
        ↪ residual_block(
        output_soft_mask)

    skip_connections = []
    for i in range(encoder_depth - 1):

        ## skip connections
        output_skip_connection =
        ↪ residual_block(
        output_soft_mask)
        skip_connections.append(
        output_skip_connection)
        # print ('skip shape:',
        output_skip_connection.
        get_shape())

        ## down sampling
        output_soft_mask =
        ↪ MaxPool2D(padding='same')(
        output_soft_mask)
        for _ in range(r):
            output_soft_mask =
            ↪ residual_block(
            output_soft_mask)

            ## decoder
    skip_connections = list(reversed(
    skip_connections))
    for i in range(encoder_depth - 1):
        ## upsampling
        for _ in range(r):
            output_soft_mask =
            ↪ residual_block(
            output_soft_mask)
        output_soft_mask =
        ↪ UpSampling2D()(
        output_soft_mask)
        ## skip connections
        output_soft_mask =
        ↪ Add()([output_soft_mask,
        ↪ skip_connections[i]])

    ### last upsampling
    for i in range(r):
        output_soft_mask =
        ↪ residual_block(
        output_soft_mask)

    output_soft_mask = UpSampling2D()(
    output_soft_mask)

    ## Output
    output_soft_mask =
    ↪ Conv2D(input_channels, (1,
    ↪ 1))(output_soft_mask)
    output_soft_mask =
    ↪ Conv2D(input_channels, (1,
    ↪ 1))(output_soft_mask)
    output_soft_mask =
    ↪ Activation('sigmoid')(
    output_soft_mask)

    # Attention: (γ + (1-γ)Xoutput_soft_mask) *
    ↪ output_trunk
    output =
    ↪ GatedLayer()(output_soft_mask)
    output = Multiply()([output,
    ↪ output_trunk])  #

    # Last Residual Block
    for i in range(p):
        output = residual_block(output)

    return output
```

ACKNOWLEDGMENT

The link to working code of our architecture can be found here [pypi][github].

REFERENCES

[1] O. Agbo-Ajala and S. Viriri, "Deeply learned classifiers for age and gender predictions of unfiltered faces," *Sci. World J.*, vol. 2020, pp. 1–12, Apr. 2020, doi: 10.1155/2020/1289408.

[2] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," 2015, *arXiv:1505.07293*. [Online]. Available: http://arxiv.org/abs/1505.07293

[3] A. M. Bukar, H. Ugail, and D. Connah, ''Automatic age and gender classification using supervised appearance model,'' *J. Electron. Imag.*, vol. 25, no. 6, Aug. 2016, Art. no. 061605, doi: 10.1117/1.JEI.25.6.061605.

[4] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, ''Ordinal hyperplanes ranker with cost sensitivities for age estimation,'' in *Proc. CVPR*, Jun. 2011, pp. 585–592.

[5] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, ''A ranking approach for human ages estimation based on face images,'' in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3396–3399.

[6] C. Chen, A. Dantcheva, and A. Ross, ''Impact of facial cosmetics on automatic gender and age estimation algorithms,'' in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 2, 2014, pp. 182–190.

[7] T. Cootes. (2014). *FG-Net Face and Gesture Recognition Network)*. [Online]. Available: http://www-prima.inrialpes.fr/FGnet/

[8] A. Das, A. Dantcheva, and F. Bremond, ''Mitigating bias in gender, age, and ethnicity classification: A multi-task convolution neural network approach,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–13.

[9] E. Eidinger, R. Enbar, and T. Hassner, ''Age and gender estimation of unfiltered faces,'' *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.

[10] X. Geng, C. Yin, and Z.-H. Zhou, ''Facial age estimation by learning from label distributions,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[11] X. Geng, Z.-H. Zhou, and K. Smith-Miles, ''Automatic age estimation based on facial aging patterns,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[12] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, ''Image-based human age estimation by manifold learning and locally adjusted robust regression,'' *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1178–1188, Jul. 2008.

[13] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, ''A probabilistic fusion approach to human age prediction,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2008, pp. 1–6.

[14] G. Guo and G. Mu, ''Joint estimation of age, gender and ethnicity: CCA vs. PLS,'' in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–6.

[15] G. Guo, G. Mu, Y. Fu, and T. S. Huang, ''Human age estimation using bio-inspired features,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 112–119.

[16] K. He, X. Zhang, S. Ren, and J. Sun, ''Identity mappings in deep residual networks,'' in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

[17] S. Hosseini, S. H. Lee, H. J. Kwon, H. I. Koo, and N. I. Cho, ''Age and gender classification using wide convolutional neural network and Gabor filter,'' in *Proc. Int. Workshop Adv. Image Technol.*, 2018, pp. 1–3, doi: 10.1109/IWAIT.2018.8369721.

[18] M. Jaderberg, K. Simonyan, and A. Zisserman, ''Spatial transformer networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[19] V. Karimi and A. Tashk, ''Age and gender estimation by using hybrid facial features,'' in *Proc. 20th Telecommun. Forum (TELFOR)*, Nov. 2012, pp. 1725–1728.

[20] A. Lanitis, ''Comparative evaluation of automatic age progression methodologies,'' *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, 2008, Art. no. 239480.

[21] H. Larochelle and G. E. Hinton, ''Learning to combine foveal glimpses with a third-order Boltzmann machine,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1243–1251.

[22] J. H. Lee, Y. M. Chan, T. Y. Chen, and C. S. Chen, ''Joint estimation of age and gender from unconstrained face images using lightweight multi-task CNN for mobile applications,'' in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, Art. no. 17877533.

[23] G. Levi and T. Hassner, ''Age and gender classification using convolutional neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 34–42, doi: 10.1109/CVPRW.2015.7301352.

[24] C. Li, Q. Liu, J. Liu, and H. Lu, ''Learning ordinal discriminative features for age estimation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2570–2577.

[25] H. Liu, J. Lu, J. Feng, and J. Zhou, ''Label-sensitive deep metric learning for facial age estimation,'' *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 292–305, Feb. 2018.

[26] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[27] J. Lu, V. E. Liong, and J. Zhou, ''Cost-sensitive local binary feature learning for facial age estimation,'' *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5356–5368, Dec. 2015.

[28] Z. Niu, M. Zhou, X. Gao, and G. Hua. *The Asian Face Age Dataset (AFAD)*. Accessed: Jun. 18, 2020. [Online]. Available: https://afad-dataset.github.io/

[29] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, ''Ordinal regression with multiple output CNN for age estimation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[30] H. Noh, S. Hong, and B. Han, ''Learning deconvolution network for semantic segmentation,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[31] T. Qin, X.-D. Zhang, D.-S. Wang, T.-Y. Liu, W. Lai, and H. Li, ''Ranking with multiple hyperplanes,'' in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2007, pp. 279–286.

[32] R. Rothe, R. Timofte, and L. V. Gool, ''DEX: Deep EXpectation of apparent age from a single image,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 10–15.

[33] R. Rothe, R. Timofte, and L. Van Gool, ''Deep expectation of real and apparent age from a single image without facial landmarks,'' *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 144–157, Apr. 2018.

[34] W. Samek, A. Binder, S. Lapuschkin, and K.-R. Müller, ''Understanding and comparing deep neural networks for age and gender classification,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1629–1638.

[35] C. Sammut and G. I. Webb, Eds., ''Mean absolute error,'' in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2010, p. 652, doi: 10.1007/978-0-387-30164-8_525.

[36] A. V. Savchenko, ''Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet,'' *PeerJ Comput. Sci.*, vol. 5, p. e197, Jun. 2019, doi: 10.7717/peerj-cs.197.

[37] P. Smith and C. Chen, ''Transfer learning with deep CNNs for gender recognition and age estimation,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2564–2571.

[38] N. Srinivas, H. Atwal, D. C. Rose, G. Mahalingam, K. Ricanek, and D. S. Bolme, ''Age, gender, and fine-grained ethnicity prediction using convolutional neural networks for the East Asian face dataset,'' in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 953–960, doi: 10.1109/FG.2017.118.

[39] R. K. Srivastava, K. Greff, and J. Schmidhuber, ''Training very deep networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[40] P. Viola and M. J. Jones, ''Robust real-time face detection,'' *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.

[41] 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.

[42] X. Wang, R. Guo, and C. Kambhamettu, ''Deeply-learned feature for age estimation,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 534–541.

[43] 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, doi: 10.1109/TIFS.2020.2969552.

[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, ''Aggregated residual transformations for deep neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[45] S. Yan, H. Wang, Y. Fu, J. Yan, X. Tang, and T. S. Huang, ''Synchronized submanifold embedding for person-independent pose estimation and beyond,'' *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 202–210, Jan. 2009.

[46] S. Yan, H. Wang, T. S. Huang, Q. Yang, and X. Tang, ''Ranking with uncertain labels,'' in *Proc. IEEE Multimedia Expo Int. Conf.*, Jul. 2007, pp. 96–99.

[47] S. Yan, H. Wang, X. Tang, and T. S. Huang, ''Learning auto-structured regressor from uncertain nonnegative labels,'' in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[48] P. Yang, L. Zhong, and D. Metaxas, ''Ranking model for facial age estimation,'' in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3404–3407.

[49] 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. 22492–22503, 2017, doi: 10.1109/ACCESS.2017.2761849.

[50] K. Zhang, N. Liu, X. Yuan, X. Guo, C. Gao, Z. Zhao, and Z. Ma, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3140–3152, Sep. 2020, doi: 10.1109/TCSVT.2019.2936410.

[51] Y. Zhang and D.-Y. Yeung, "Multi-task warped Gaussian process for personalized age estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2622–2629.

[52] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5810–5818.

[53] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

AVISHEK GARAIN (Member, IEEE) is currently pursuing the Bachelor of Engineering degree with the Department of Computer Science and Engineering, Jadavpur University. He was a Software Development Intern with the Samsung Research and Development Institute Bangalore and worked with the RNLU Team, Voice Research and Development Group. He is also a Product Development Intern with Renatus Meditech Pvt. Ltd. and WowExp Pvt. Ltd. He is an incoming Associate Engineer with Airbus India. He follows up several research journals to keep himself up-to-date with current research and also actively takes part in research at the undergraduate level. He has published 18 research articles in Scopus indexed journals and conferences. His research interest includes deep learning for natural language processing and computer vision.

BISWARUP RAY is currently pursuing the Bachelor of Engineering degree with the Department of Computer Science and Engineering, Jadavpur University. He was a Business Technology Analyst Intern with ZS Associates Pune. His research interest includes deep learning for natural language processing and computer vision.

PAWAN KUMAR SINGH (Member, IEEE) received the B.Tech. degree in information technology from the West Bengal University of Technology, in 2010, and the M.Tech. degree in computer science and engineering and the Ph.D. degree in engineering from Jadavpur University (JU), in 2013 and 2018, respectively. He also received the RUSA 2.0 fellowship for pursuing his postdoctoral research with JU, in 2019. He is currently working as an Assistant Professor with the Department of Information Technology, JU. He has published more than 75 research articles in peer-reviewed journals and international conferences. His current research interests include computer vision, pattern recognition, handwritten document analysis, image and video processing, feature optimization, machine learning, deep learning, and artificial intelligence. He is a member of The Institution of Engineers (India) and the Association for Computing Machinery (ACM) as well as a Life Member of the Indian Society for Technical Education (ISTE, New Delhi) and the Computer Society of India (CSI). He serves as an editorial board member, a reviewer, and a technical program committee member for a number of IEEE and Springer journals and conferences.

ALI AHMADIAN (Member, IEEE) received the Ph.D. degree from Universiti Putra Malaysia (UPM), in 2014, as the best postgraduate student. He is currently a Fellow Researcher with the Institute of Industry Revolution 4.0, UKM. As a young researcher, he is dedicated to research in applied mathematics. In general, his primary mathematical focus is the development of computational methods and models for problems arising in AI, biology, physics, and engineering under fuzzy and fractional calculus (FC); in this context, he have worked on projects related to drug delivery systems, acid hydrolysis in palm oil frond, and carbon nanotubes dynamics, Bloch equations and viscosity. He could successfully receive 15 national and international research grants and selected as the 1% top reviewer in the fields of mathematics and computer sciences recognized by Publons during 2017–2019. He is the author of more than 80 research articles published in the reputed journals, including IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Fuzzy Sets and Systems*, *Communications in Nonlinear Science and Numerical Simulation*, and *Computational Physics*. He also presented his research works in 38 international conferences held in Canada, Serbia, China, Turkey, Malaysia, and United Arab Emirates. He is a member of Editorial Board in *Progress in Fractional Differentiation and Applications* (Natural Sciences Publishing) and a Guest Editor in *Advances in Mechanical Engineering* (SAGE), *Symmetry* (MDPI), *Frontier in Physics* (Frontiers), and *International Journal of Hybrid Intelligence* (Inderscience Publishers). He was a member of programme committee in a number of international conferences in fuzzy field at Japan, China, Turkey, South Korea, and Malaysia. He is also serving as a referee in more than 80 reputed international journals.

NORAZAK SENU is currently an Associate Professor with the Institute for Mathematical Research, Universiti Putra Malaysia. He published more than 100 articles in the peer-reviewed international journals. His main interests include working on different types of differential equations and modeling real-world systems using such equations. He received several prizes for his research works from the Ministry of Education, Malaysia. He achieved a number of governmental grants to support his scientific works.

RAM SARKAR (Senior Member, IEEE) received the B.Tech. degree in computer science and engineering from the University of Calcutta, in 2003, and the M.E. degree in computer science and engineering and the Ph.D. degree in engineering from Jadavpur University, in 2005 and 2012, respectively. He joined the Department of Computer Science and Engineering, Jadavpur University, as an Assistant Professor, in 2008, where he is currently working as a Professor. He received the Fulbright-Nehru Fellowship (USIEF) for postdoctoral research in the University of Maryland, College Park, MD, USA, in 2014–2015. His current research interests include image processing, pattern recognition, machine learning, and bioinformatics.