International Journal of Engineering

Journal Homepage: www.ije.ir

Deep Multi-task Convolutional Neural Networks for Efficient Classification of Face Attributes

M. Rohani, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

PAPER INFO

Paper history:
Received 23 July 2023
Received in revised form 01 August 2023
Accepted 06 August 2023

Keywords:
Age Detection
Convolutional Neural Networks
Gender Classification
Multi-task Learning
Smile Detection

ABSTRACT

Facial feature recognition is an important subject in computer vision with numerous applications. The human face plays a significant role in social interaction and personology. Valuable information such as identity, age, gender, and emotions can be revealed via facial features. The purpose of this paper is to present a technique for detecting age, smile, and gender from facial images. A multi-task deep learning (MT-DL) framework was proposed that can simultaneously estimate three important features of the human face with remarkable accuracy. Additionally, the proposed approach aims to reduce the number of trainable network parameters while leveraging the combination of features from different layers to increase the overall accuracy. The conducted tests demonstrate that the proposed method outperforms recent advanced techniques in all three accuracy criteria. Moreover, it was demonstrated that multi-task learning (MTL) is capable of improving the accuracy by 1.55% in the smile task, 2.04% in the gender task, and 3.52% in the age task even with less available data, by utilizing tasks with more available data. Furthermore, the trainable parameters of the network in the MTL mode for estimating three tasks simultaneously increase only by about 40% compared to the single-task mode. The proposed method was evaluated on the IMDB-WIKI and GENKI-4K datasets and produced comparable accuracy to the state-of-the-art methods in terms of smile, age detection, and gender classification.

doi: *10.5829/ije.2023.36.11b.14*

NOMENCLATURE

M	Total nuber of traning sample	L_t	Cross-Entropy
N_t	Number of classes associated with each task	λ	Weight decay
C_i^t	Class score vector for the ith sample in the tth task	W	Output weights
S_i^t	Label corresponding to the ith sample in the tth task	μ_t	Significance factor in each loss function
X_i^t	One-hot Encoding of the ith sample	L_r	Learning rate
$\hat{X}_i^t$	Probability distribution in the tth task over the ith sample	$Decay\_step$	Steps the learning rate
β_i^t	The kind of sample	β_i^t	The kind of sample

1. INTRODUCTION

Deep learning (DL), a branch of artificial intelligence (AI), has made significant advancements in computer vision. Many AI developments and products have lifted living standards, increased productivity, and conserved social and human resources. In many instances, artificial intelligence has surpassed human skills. The study of human emotions and behavior; however, greatly benefited from the active research into automatic face detection [1, 2]. Hence, the present research addresses facial feature recognition, such as smile, emotion, and gender detection. Subsequently, each of these features is expressed as a task associated with the human face. In each task, facial images are used as inputs. In the smile task, the presence or absence of a smile is detected on each facial image. Then, the ages of the persons in the images are classified into six age groups. Finally, each person's gender is revealed. These tasks are typically investigated as separate problems, which makes model training challenging, especially when there is a lack of adequate training data. On the other hand, the

*Corresponding Author Email: hfarsi@birjand.ac.ir (H. Farsi)

Please cite this article as: M. Rohani, H. Farsi, S. Mohamadzadeh, Deep Multi-task Convolutional Neural Networks for Efficient Classification of Face Attributes, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 11, (2023), 2102-2111

information related to different facial analysis tasks frequently includes joint facial traits. Therefore, simultaneous learning from several facial datasets improves the detection accuracy of each task. The paper outlines the methodology for facial image analysis, which involves preprocessing images from different datasets using a convolutional neural network (CNN) to extract faces. A deep convolutional neural network (D-CNN) is then trained using a combination of common features for smile detection, age detection, and gender classification. To achieve this, the tasks are divided into three branches with their own loss functions, and a final loss function is defined by combining these. The approach allows each task to receive images from its corresponding dataset and share the features for training with other tasks. The second section of the paper offers a concise overview of CNNs and their application to the gender, smile, and age tasks. The section concludes with an examination of multi-task learning (MTL) methods. In section 3, a novel multi-task deep learning (MT-DL) framework is introduced, which enables the simultaneous estimation of three significant facial features with remarkable precision. A key aspect of the proposed approach is its focus on minimizing the number of trainable network parameters by proposing a new CNN, while effectively harnessing feature combinations from various layers to improve overall accuracy. Through comprehensive tests and evaluations, the proposed method consistently outperforms recent advanced techniques across all three accuracy criteria, underscoring its superiority and effectiveness in facial feature recognition tasks.

2. LITERATURE REVIEW

This section provides a comprehensive examination of the existing research on deep neural networks applied to smile detection, age detection, gender classification, and multi-task learning in the context of facial feature recognition. This section explores the methodologies, techniques, and approaches employed in previous studies, focusing on their utilization of deep neural networks for these specific tasks. Through critical analysis of the accomplishments, limitations, and advancements in these areas, this literature review establishes the foundation for the proposed method, highlighting the gaps that this present study aims to address and emphasizing the significance of the proposed approach in advancing the field of facial analysis.

2. 1. Deep Neural Networks Over the past ten years, there has been significant research in the field of computer vision, with a focus on deep learning techniques, which has led to the development of numerous methods. D-CNNs are common in the area of DL, such that they constitute a major DL method. In these networks, a large number of layers in these networks are trained with a powerful approach. This technique is extremely effective and is also frequently utilized in a variety of computer vision applications [3, 4]. The major CNN was AlexNet, introduced by Krizhevsk [5] fourteen years after LeNet in 2017. The network can be categorized as a shallow network as it has a total of eight layers, including three fully connected layers and five convolutional layers [5]. Subsequently, deeper neural networks were introduced. Recently, Google presented a network called Inception, which was based on deepening convolutional networks [6]. The VGG-16 network was introduced by Simonyan and Zisserman [7]. This network is highly popular due to its simple structure. Later, Microsoft presented a network named Residual Network, abbreviated as ResNet. In addition to the convolutional network, this network incorporates connections between layers to directly transfer inputs from one layer to the next, as well as errors during backpropagation, allowing for a faster and deeper training process without the need for intermediate layers [8]. After the introduction of deeper and more accurate networks, those such as ResNext and WideResNet were proposed. ResNext was a multi-branch version of ResNet, presented in 2017 by Xie et al. [9]. The aim of this network was to improve efficiency by increasing the network's width. In order to increase network width and decrease network depth, this network tried to improve efficiency by adding more filters. The performance of the presented neural networks was demonstrated in ImageNet, one of the most well-known computer vision competitions.

2. 2. Smile Detection Recent studies have concentrated on using neural networks to detect smiles in images. In research by Sang et al. [10], a network similar to the VGG network, named BK-Net, was designed. This is highly capable of smile detection. In 2018, Cui et al. [11] presented the extreme learning machine (ELM) with the aim of feature dimensionality reduction and focusing on the regions surrounding the mouth. Subsequently, in 2019, Vo et al. [12] extracted features from facial images using a CNN and carried out smile detection via an extreme gradient boosting (XGB) classifier. In the same year, Nguyen [13] employed the YOLO network for smile detection. In 2020, Wu et al. [14] proposed a method for fast smile detection in a working environment based on Multilayer Perceptron (MLP) network. Additionally, they used two different databases for training the network instead of one. In 2021, Hassen et al. [15] proposed a method based on MLP neural network and Cascade Classifier. The accuracy of these methods is challenged in low-light conditions and when the facial image is angled ao distored. In another study in 2022, Hassen et al. [16] used an ensemble classification approach and multiple classifiers to detect smiles. This method also has a high speed, but its accuracy decreases when dealing with low-resolution images [16]. In the same year in 2022, Liu et al. [17] utilized the MobileNetV2 network for smile detection.

2. 3. Age Detection

Recent years have witnessed the widespread use of CNNs for age detection [18]. To this end, they designed a network inspired by the VGG-16 architecture and implemented this model using the ResNet network [19]. In 2018, Rothe et al. [20] proposed a method for age estimation based on the ResNet. They then estimated the age of individuals using a unique network. They employed the Adience dataset to evaluate their methods. In 2019, Zhang et al. [21] combined the ResNet and LSTM networks into AL-ResNet networks in order to extract age-sensitive local regions. To accomplish this, they used a ResNet model that was pre-trained on the ImageNet dataset as a base model, and then identified age intervals using age-related datasets. Cao et al. [22] estimated age using two neural networks. First, they extracted features using a simple neural network, and then they estimated age using the Rank Consistent Ordinal Regression Network (RCORN) network. In another study, Xia et al. [23] extracted features using a neural network through multi-stage learning and estimated the age of individuals from facial images using a Feedforward Neural Network.

2. 4. Gender Classification

In recent years, the use of CNNs for age detection has considerably increased. Zhang et al. [24] utilized both ResNet and LSTM networks to extract age-related features from facial images, which were then combined to estimate gender. In 2018, Amit Dhomne et al. [25] proposed a gender detection approach based on the VGG network, which achieved acceptable accuracy. In a recent study, Nga et al. [26] employed the ImageNet network to detect gender and utilized a combination of features from multiple layers for improved performance. They believed that using pre-trained weights in the ImageNet network would improve their results. In 2022, Bekhet et al. [27] estimated individuals' gender using a proposed convolutional neural network with selfie images. However, a limitation of this approach is the low diversity of images in the dataset.

2. 4. Multi-Task Learning

MTL refers to multiple classifications with different objectives in different classes. In MTL, multiple tasks are simultaneously trained using CNNs. This type of learning exploits the similarities and differences between various tasks [28, 29]. In 2017, Ranjan et al. [30] presented an AlexNet-based MTL model named Hyper Face with face detection, landmark localization, facial gesture estimation, and gender estimation capabilities. These four tasks were trained simultaneously on a dataset named ALFW, and the network output was used to predict each of the tasks. In another study in multi task learning, Ran et al. [31] used multi-task learning to estimate human pose and shape and remove occlusions. In 2021, Savchenko [32] proposed a multi-task learning approach for facial expression and attribute recognition. The aim of the proposed method in this paper was to detect facial features and components. In 2022, Yu et al. [33] introduced a multi-task learning based approach for facial parameter detection. They used attention modules to focus on each part of the image in their method. The use of these modules improved the accuracy of detecting each parameter [33].

3. MATERIAL AND METHODS

The proposed framework utilizes multi-task learning to simultaneously learn and solve multiple tasks. Moreover, D-CNNs were simultaneously used for all the tasks and named the “shared convolutional neural network” (S-CNN). To this end, several independent datasets were used to train the network after being combined. The advantage of using several datasets along with the S-CNN is learning shared features from several datasets in different tasks. In addition, shared learning allows tasks with fewer data to use tasks with more data. Moreover, it seems that using features learned in shared learning leads to better results compared to the single-task mode. In the next step, the network was divided into three separate branches. Each of these branches was assigned to one task and sought to learn the features related to that task. In the end, an appropriate loss function was defined for each task. Figure 1 displays the combination of the datasets.

3. 1. Implementation of Method with MTL

The proposed framework has been presented and modified based on the BKNet network. In this method, the CNN was constructed by removing the last three fully connected layers of the BKNet and concatenating the output of the first layer with the last layer before the fully connected layers. Table 1 shows the BKNet architecture. In order to use the facial images from the datasets, one first needs to crop the face images from the original images. A multi-task convolutional neural network (MTCNN) was used for this purpose [34]. The comprehensive preprocessing begins by reading the images from a designated folder and utilizing MTCNN to identify the positions of faces within the images. Subsequently, a series of preprocessing steps, including face cropping, grayscale conversion, and resizing the images to a fixed dimension of 48x48 pixels. Furthermore, the quality and diversity of the dataset can be enhanced by incorporating additional preprocessing

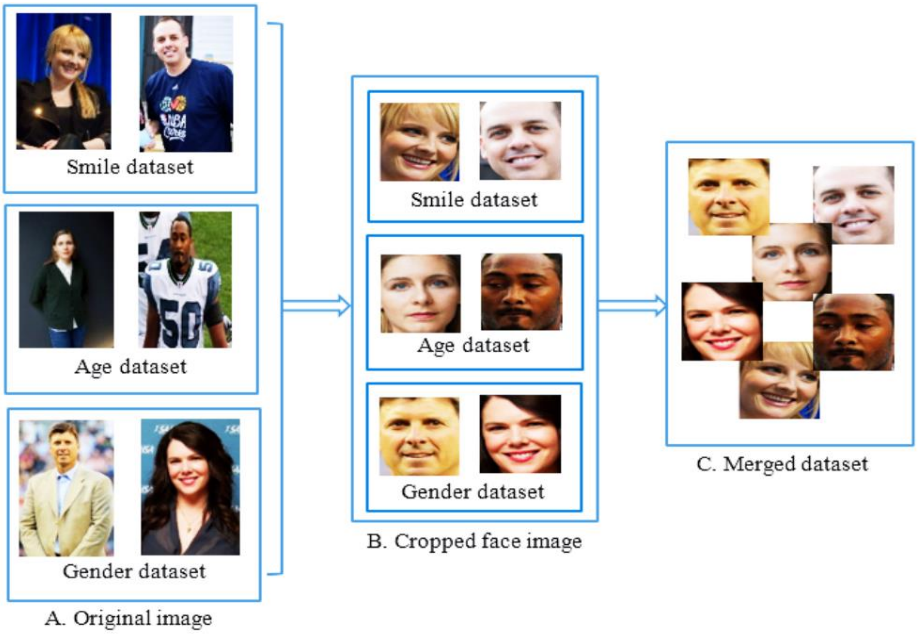


Figure 1. Dataset combination: original images (A); cropped facial images (B); and dataset combination (C)

TABLE 1. Architecture of the BKNet CNN [10]

Input image
Convolutional layer 64 filters 3×3, stride 1 and ReLU
Convolutional layer 64 filters 3×3, stride 1 and ReLU
Max pooling layer with a 2×2 filter and stride of 2
Convolutional layer 128 filters 3×3, stride 1 and ReLU
Convolutional layer 128 filters 3×3, stride 1 and ReLU
Convolutional layer 128 filters 3×3, stride 1 and ReLU
Max pooling layer with a 2×2 filter and stride of 2
Convolutional layer 256 filters 3×3, stride 1 and ReLU
Convolutional layer 256 filters 3×3, stride 1 and ReLU
Max pooling layer with a 2×2 filter and stride of 2
Convolutional layer 512 filters 3×3, stride 1 and ReLU
Convolutional layer 512 filters 3×3, stride 1 and ReLU
Max pooling layer with a 2×2 filter and stride of 2

techniques such as normalization and data augmentation. The inclusion of normalization ensures that the intensity values of the grayscale images are within a specific range, facilitating optimal model training. Figure 2 shows an illustration of a face image that the MTCNN has been cropped. Furthermore, the augmentation method was employed to increase the number of training data. This method reduces overfitting and improves learning. The augmentation was carried out using the Random Crop, Random Flip, and Random Rotate techniques.

3. 2. Shared Convolutional Neural Network

Four convolution blocks were used in this section. At first, the CNN was comprised of three blocks, each consisting of two convolutional layers with 32, 64, and 128 neurons, respectively, all with a 3 × 3 filter and a stride equal to 1. Each block had a max pooling layer with a 2 × 2 filter and stride equal 2. The fourth and final block had convolutional layers with 256 neurons and a 3 × 3 filter, stride equal 1, and the same max pooling layer. Also, the first block's output and the last block's output were concatenated. This allows the features of the first layer, which focuses on the overall facial features, to combine with those of the last layer, which focuses on the details,

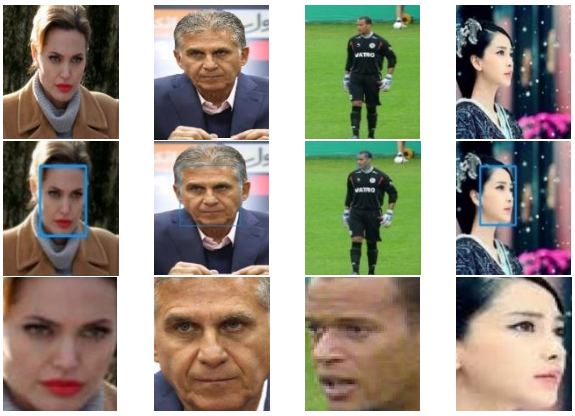


Figure 2. An example of a facial picture cropped by the MTCNN: original images (first row); face location (second row); and cropped images corresponding to the first row obtained from the WIKI dataset (third row)

to produce better results. Figure 3 presents the proposed method's block diagram.

3. 3. Separated Convolutional Neural Network

To facilitate smile detection, age detection, and gender detection, the network was divided into three sections following the S-CNN. The split CNN learns the characteristics of each task in the associated branch, whereas S-CNN learns the shared features of the three tasks from various datasets. Each individual layer consists of three fully connected layers, with the final layer containing N neurons . N is the number of classes for each task and the first two layers each containing 256 neurons. For smile detection, age detection, and gender detection, respectively, these parameters are equivalent to 2, 7, and 2. An activation function for the ReLU and a batch normalization function come after the first two fully connected layers. To prevent overfitting, the three fully connected layers utilized dropout. The first layer's output, which contains features that are closely related to the original images and have a significant impact on the output tasks, including the eyes, nose, and mouth, was concatenated with the last layer before the network was divided into three parts.

3. 4. Loss Functions

An effective CNN was presented in the proposed method for training from several datasets. In this section, all the data from different datasets were combined to form a larger shared training dataset. Each training datum may correspond to one or more tasks. Accordingly, the following points must be taken into account. M represents the total number of training samples collected from two datasets. N_t represents the number of classes associated with each task. C_i^t is the class score vector for the ith sample in the tth task. S_i^t is the true label corresponding to the ith sample in the tth task. X_i^t is the one-hot Encoding of the ith sample's true class label in the tth task ($X_i^t(S_i^t) = 1$). $\hat{X}_i^t$ is the probability distribution in the tth task over the ith sample. $\beta_i^t \in (0.1)$ indicates the kind of sample. ($\beta_i^t = 1$ if the ith sample matches the tth task; $\beta_i^t = 0$ otherwise) L_t represents the loss function that is specific to each task. L_t is named the Cross-Entropy, and according to Equation (1), the L_t corresponding to the tth task is defined.

$$L_t = -\frac{1}{M}\sum_i^M \left(\beta_i^t \sum_j^{N_t} X_i^t(j)\log(\hat{X}_i^t(j))\right) \quad (1)$$

In this equation, $\beta_i^t \in (0.1)$ states whether the label j is true for the ith sample or not. The value $\hat{X}_i^t(j)$ within the range of (0,1) represents the estimated probability of the jth label being true for the ith sample. To compute the overall loss function for the network, the weights generated by the loss functions of each task are added together. To mitigate overfitting in the network, the L2 regularization term is incorporated into the total loss function. The total loss function is expressed by Equation (2). In the equation, λ represents the weight decay, which is a hyperparameter that controls the strength of the L2 regularization term, and W refers to the output weights. Moreover, the tth task's significance factor in each loss function is denoted by the μ_t.

$$L_{total} = \sum_i^T \left(\sum_j^{N_t} \mu_t L_t + \lambda \|W\|^2\right) \quad (2)$$

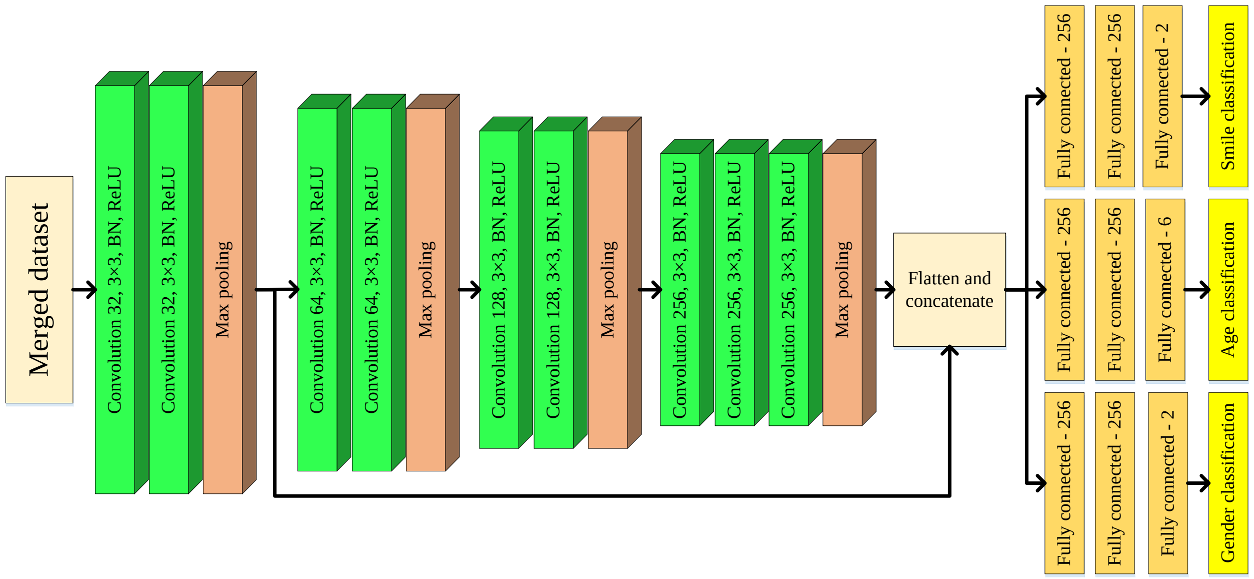


Figure 3. Block diagram of the proposed method

4. RESULTS AND DISCUSSION

4. 1. Dataset The GENKI-4K dataset is a widely used dataset for smile detection. It was published by the MPLab GENKI Database. This dataset consists of 4000 labeled images of the human face taken from individuals of various ages and races. 1838 of these images have the label "non-smiling," and 2162 have the label "smiling". This dataset was procured from the Internet, and the backgrounds of the images are different, which makes smile detection more challenging. In other words, unlike datasets that have identical backgrounds, this dataset was constructed from actual images. Another challenge involved in this dataset is that the smile is unclear in some of the images. Some previous studies have eliminated these images from the training and testing processes. These images may cause errors in the training process, complicate the network evaluation, and significantly reduce the overall accuracy. Despite this, all images in the GENKI-4K dataset were utilized for training and testing the proposed method. Figure 4 showcases some samples of images from the GENKI-4K dataset [33].

Also, the proposed method was trained and evaluated using the IMDB-WIKI dataset [34]. This dataset was introduced in 2016 to compensate for the shortage of images in related datasets. This dataset contains images of famous actors prepared from the IMDB website. In addition, it contains images from persons on Wikipedia. This dataset includes the date of birth, name, and gender of every person. Moreover, there are several images of each person in this dataset. In total, there are 460723 images of 20284 persons from IMDB and 62328 images of the same persons from Wikipedia, making up a total of 523051 images. Furthermore, there are six age classes (0-15, 16-25, 26-35, 36-45, 46-60, and 60-100 years) in this dataset. Also, the genders are labeled either male or female. Figure 5 presents samples of images from the IMDB-WIKI dataset.

4. 2. Evaluation Criteria The accuracy evaluation criterion measures the accuracy of the smile, gender, and age tasks. This criterion is calculated as the number of correct predictions to the total number of ground-truth labels. It assesses the proportion of face images correctly classified as smiling or not, the accurate age range, and correct gender identification. It is determined by Equation (3) [20, 35].

$$Accuracy = \frac{Number\ of\ accurate\ prediction}{Total\ number\ of\ prediction} \quad (3)$$

4. 3. Simulation Details The study utilized the GENKI-4K dataset to detect smiles and the IMDB-WIKI dataset for gender and age detection. Initially, the two datasets were merged to create a larger dataset, and 3000 samples from the GENKI-4K dataset were selected for training, while 1000 samples were reserved for testing purposes. Moreover, 20000 samples were utilized for testing and 150000 samples for training using the IMDB-WIKI dataset. The label corresponding to each task was assigned to every sample from each dataset. The vector composed of these labels is designated as β_i^t. Moreover, the batch size was considered to be 128. Additionally, all initial weights were taken into account using a Gaussian distribution with a variance of 0.01 and zero mean. It was assumed that L2 weight decay would be λ=0.01. Furthermore, the importance factors of all tasks were assumed to be $\mu_t = \mu_1 = \mu_2 = \mu_3 = 1$. The dropout rate was considered to be 0.5 for all the fully connected layers. An exponential function was used to reduce the learning rate. This rate is calculated in the pth step of Equation (4).

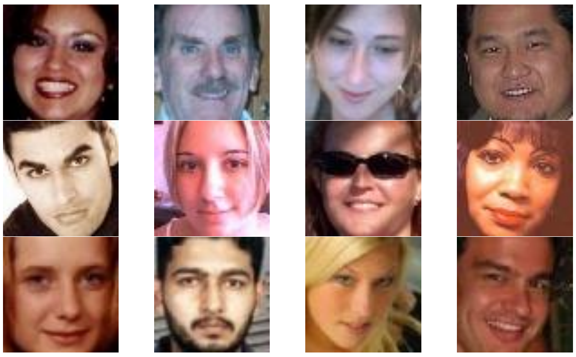


Figure 4. Samples of the GENKI-4K dataset's images [33]

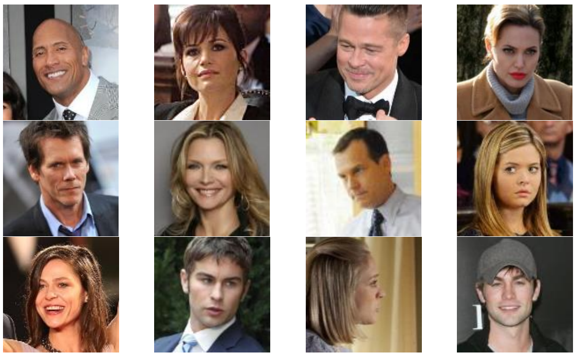


Figure 5. Samples of the images in the IMDB-WIKI dataset [34]

$$L_r = \text{Initial } L_r \times \text{Decay\_rate}^{\text{Step/Decay\_step}} \quad (4)$$

In this equation, L_r denotes the learning rate at the pth step. Moreover, $Decay\_step$ is the step in which the learning rate declines, and $Initial\, L_r$ is the initial learning rate. In the conducted experiments, $Initial\ L_r = 0.01$, $Decay_{step} = 2000$, and $Decay_{rate} = 0.8$. Also, the proposed model was executed for over 500 epochs.

Figure 6 shows the accuracy and loss function graphs for the three tasks for 100 epochs. In this paper, three groups of experiments were performed. The first network training method was single-task training. In this method, the network output was considered only for one task and was trained using the corresponding dataset.

4. 4. Results The evaluation of the network is presented in Table 2 under Config A (Single BK-Net). The second method, designated Config B, involves MTL without combining the features of the first and last convolutional layers (MTL-BK-Net). The third case, Config C, corresponds to the method proposed (MTL-MBK-Net) in Figure 3. The proposed method was initially trained on the two datasets. The results of the proposed method are compared to those of the state-of-the-art methods in Table 2. The proposed method MTL-MBKNet (Config C) achieved a smile detection accuracy of 96.63% on the IMDB-WIKI dataset according to the Table 2, which is superior to previous methods.

In addition, an accuracy of 95.08% was achieved in the single-task mode for the smile task, indicating that the smile task uses the other tasks to improve the results. In the gender task, the proposed method attained an accuracy of 93.20%, which outperformed the other methods. The RoR-34 method came in second place with an accuracy of 92.24% Additionally, the proposed method achieved a gender detection accuracy of 91.16% in the single-task mode. This difference in accuracy between the multi-task and single-task modes shows that the gender task has used the other tasks during training. In the age detection task, MTL-MBK-Net (Config C) produced the best accuracy at 68.92%, followed by AL-ResNet-34 with an accuracy of 67.83%. MTL performed better than single-task learning also in this task. MTL-MBK-Net (Config B) exhibited better results than Single BK-Net (Config A) in all the tasks. The proposed method suffered some errors during the evaluation in some tasks. These errors are present in smile, gender, and age detection. Figure 7 depicts the errors in the smile, gender, and age detection tasks. Specifically, the first, second, and third rows of the Figure represent the errors in smile, gender, and age detection, respectively. In smile detection, these errors can be attributed to the facial form and the presence or absence of the teeth during smiling or a special form of the smile in some persons. In gender detection, the errors may be due to the makeup used by some persons and their genetics. A major challenge in age detection is the makeup used by women, which modifies the lip, eye, and skin features.

The results of comparing the three proposed models based on the number of parameters of execution time and Fps are presented in Table 3. Single BK-Net (Config A) results in the lowest number of parameters, with only 2418146, and the fastest execution time of 6.23 seconds. It also has a relatively high processing speed of 20.54 Fps. MTL-BK-Net (Config B) provides 3731946 processing speed of 13.30 Fps. Finally, MTL-MBK-Net

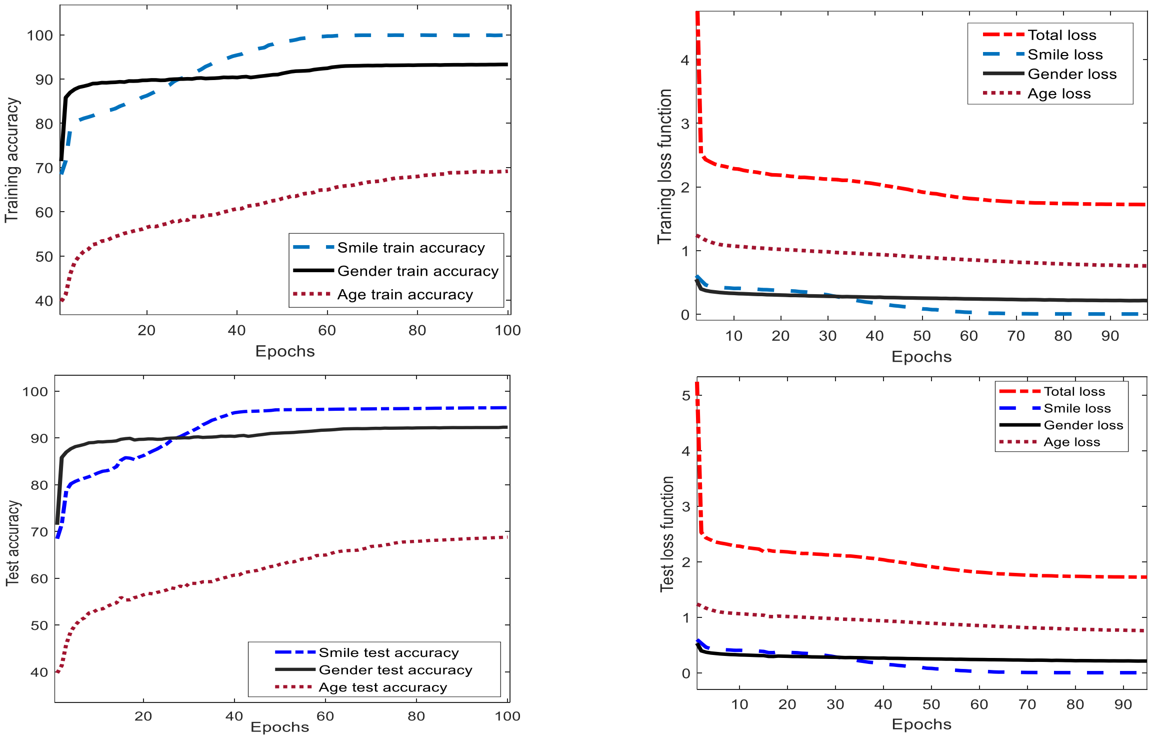
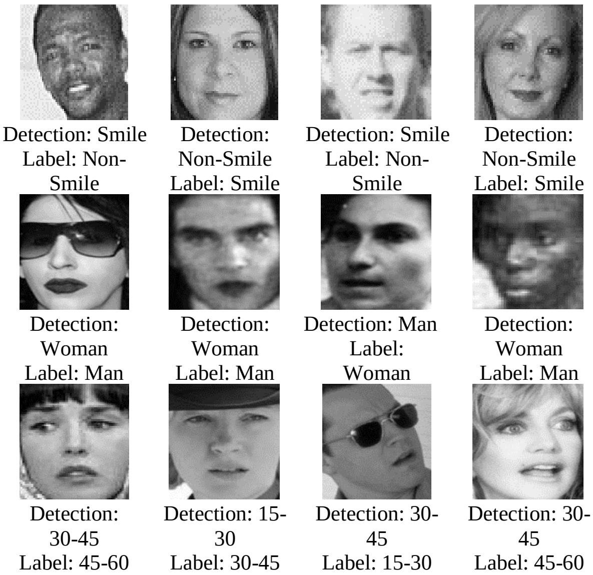


Figure 6. A view of the accuracy and loss functions for the three tasks during training and testing

TABLE 2. Numerical results of the proposed method and other state-of-the art methods

Task	Methods	Accuracy (%)
Smile accuracy comparison results	PDV-ELM [11]	93.42
	HF-SD [12]	93.60
	YOLO [13]	93.17
	MNV2 [17]	94.37
	Single BKNet (Config A)	95.08
	MTL-BKNet (Config B)	95.60
	MTL-MBKNet (Config C)	**96.63**
Gender accuracy comparison results	RoR-34 [24]	92.24
	TL-VGG19 [26]	91.09
	TL-DenseNet-169 [26]	70.48
	GSI-CNN [27]	89.00
	Single BKNet (Config A)	91.16
	MTL-BKNet (Config B)	91.28
	MTL-MBKNet (Config C)	**93.20**
Age accuracy comparison results	ResNet-34 [24]	64.63
	RoR-34 [24]	65.74
	AL-ResNet-34 [21]	67.83
	DEX-w [20]	64.00
	Single BKNet (Config A)	65.40
	MTL-BKNet (Config B)	67.81
	MTL-MBKNet (Config C)	**68.92**



Detection: Smile Label: Non-Smile

Detection: Non-Smile Label: Smile

Detection: Smile Label: Non-Smile

Detection: Non-Smile Label: Smile

Detection: Woman Label: Man

Detection: Woman Label: Man

Detection: Man Label: Woman

Detection: Woman Label: Man

Detection: 30-45 Label: 45-60

Detection: 15-30 Label: 30-45

Detection: 30-45 Label: 15-30

Detection: 30-45 Label: 45-60

Figure 7. Number of false detections by the proposed method

TABLE 3. Proposed models parameters and execution time comparison

Methods	Parameters	Execution time (s)	Fps
Single BK-Net (Config A)	2418146	6.23	20.54
MTL-BK-Net (Config B)	3731946	9.62	13.30
MTL-MBK-Net (Config C)	3953130	10.19	12.56

(Config C), has 3953130 parameters and takes 10.19 seconds to execute with a processing speed of 12.56 Fps. In order to evaluate three tasks, three single-task networks must be parallelized with each other. In this case, the parameters of the network are tripled and it has more parameters compared to the multi-task state. However, using multi-task learning causes that the parameter ratio of the network decreases and the computational burden reduces as well. Furthermore, according to Table 3, with the training of three single-task networks, the execution time of the network also increases compared to the multi-task state.

5. CONCLUSION

The present paper proposed a DL framework using MTL on facial images. In the proposed method, a D-CNN was shared to extract features related to smiling, gender, and age group from facial images from the well-known GENKI-4K and IMD-WIKI datasets. This improved the accuracy of smile detection, which had fewer available data than the other tasks. In addition, the proposed separated network was used along with a combination of layers to detect smile, age, and gender. In comparison to current methodologies used in both the multi-task mode and the single-task mode, the proposed framework achieved good results. In future studies, potential applications and challenges of the proposed architecture, including its scalability and robustness under different conditions and environments, will be investigated. Additionally, the use of more challenging datasets and auxiliary losses to improve the training procedure and increase the accuracy of neural networks in various tasks will be explored. Furthermore, the potential of the proposed framework for real-time face recognition in practical applications such as security, healthcare, and entertainment will be examined.

6. REFERENCES

1. Shahbakhsh, M.B. and Hassanpour, H., "Empowering face recognition methods using a gan-based single image super-resolution network", ***International Journal of Engineering,***

Transactions A: Basics, Vol. 35, No. 10, (2022), 1858-1866. doi: 10.5829/ije.2022.35.10a.05.

2. Firouzian, I., Firouzian, N., Hashemi, S. and Kozegar, E., "Pain facial expression recognition from video sequences using spatio-temporal local binary patterns and tracking fiducial points", ***International Journal of Engineering, Transactions B: Applications***, Vol. 33, No. 5, (2020), 1038-1047. doi: 10.5829/ije.2020.33.05b.38.

3. Charoqdouz, E. and Hassanpour, H., "Feature extraction from several angular faces using a deep learning based fusion technique for face recognition", ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 8, (2023), 1548-1555. doi: 10.5829/ije.2023.36.08b.14.

4. Thepade, S., Dindorkar, M., Chaudhari, P. and Bang, S., "Enhanced face presentation attack prevention employing feature fusion of pre-trained deep convolutional neural network model and thepade's sorted block truncation coding", ***International Journal of Engineering, Transactions A: Basics***, Vol. 36, No. 4, (2023), 807-816. doi: 10.5829/ije.2023.36.04a.17.

5. Krizhevsky, A., Sutskever, I. and Hinton, G.E., "Imagenet classification with deep convolutional neural networks", ***Advances in Neural Information Processing Systems***, Vol. 25, (2012). doi: 10.1145/3065386.

6. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2015), 1-9.

7. Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition", ***arXiv preprint arXiv:1409.1556***, (2014). doi: 10.48550/arXiv.1409.1556.

8. He, K., Zhang, X., Ren, S. and Sun, J., "Deep residual learning for image recognition", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), 770-778.

9. Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K., "Aggregated residual transformations for deep neural networks", in Proceedings of the IEEE conference on computer vision and pattern recognition. (2017), 1492-1500.

10. Sang, D.V., "Facial smile detection using convolutional neural networks", in 2017 9th International Conference on Knowledge and Systems Engineering (KSE), IEEE. (2017), 136-141.

11. Cui, D., Huang, G.-B. and Liu, T., "Elm based smile detection using distance vector", ***Pattern Recognition***, Vol. 79, (2018), 356-369. doi: 10.1016/j.patcog.2018.02.019.

12. Vo, T., Nguyen, T. and Le, C., "A hybrid framework for smile detection in class imbalance scenarios", ***Neural Computing and Applications***, Vol. 31, No. 12, (2019), 8583-8592. doi: 10.1007/s00521-019-04089-w.

13. Nguyen, C.C., Tran, G.S., Nghiem, T.P., Burie, J.-C. and Luong, C.M., "Real-time smile detection using deep learning", ***Journal of Computer Science and Cybernetics***, Vol. 35, No. 2, (2019), 135-145. doi: 10.15625/1813-9663/35/2/13315.

14. Wu, H., Liu, Y., Liu, Y. and Liu, S., "Fast facial smile detection using convolutional neural network in an intelligent working environment", ***Infrared Physics & Technology***, Vol. 104, (2020), 103061. doi: 10.1016/j.infrared.2019.103061.

15. Hassen, O.A., Abu, N.A., Abidin, Z.Z. and Darwish, S.M.,, "A new descriptor for smile classification based on cascade classifier in unconstrained scenarios", ***Symmetry***, Vol. 13, (2021), 805-816. doi: 10.3390/sym13050805.

16. Hassen, O.A., Abu, N.A., Abidin, Z.Z. and Darwish, S.M., "Realistic smile expression recognition approach using ensemble classifier with enhanced bagging", ***Computers, Materials & Continua***, Vol. 70, No. 2, (2022). doi: 10.32604/cmc.2022.019125.

17. Liu, Y., Liu, Z., Zhao, Y. and Xu, J., "A robust approach for smile recognition via deep convolutional neural networks", in 2022 7th International Conference on Image, Vision and Computing (ICIVC), IEEE. (2022), 60-64.

18. Mavaddati, S., "Voice-based age and gender recognition using training generative sparse model", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 31, No. 9, (2018), 1529-1535. doi: 10.5829/ije.2018.31.09c.08.

19. Zhang, K., Sun, M., Han, T.X., Yuan, X., Guo, L. and Liu, T., "Residual networks of residual networks: Multilevel residual networks", ***IEEE Transactions on Circuits and Systems for Video Technology***, Vol. 28, No. 6, (2017), 1303-1314. doi: 10.1109/TCSVT.2017.2654543.

20. Rothe, R., Timofte, R. and Van Gool, L., "Deep expectation of real and apparent age from a single image without facial landmarks", ***International Journal of Computer Vision***, Vol. 126, No. 2-4, (2018), 144-157. doi: 10.1007/s11263-016-0940-3.

21. Zhang, K., Liu, N., Yuan, X., Guo, X., Gao, C., Zhao, Z. and Ma, Z., "Fine-grained age estimation in the wild with attention lstm networks", ***IEEE Transactions on Circuits and Systems for Video Technology***, Vol. 30, No. 9, (2019), 3140-3152. doi: 10.1109/TCSVT.2019.2936410.

22. Cao, W., Mirjalili, V. and Raschka, S., "Rank consistent ordinal regression for neural networks with application to age estimation", ***Pattern Recognition Letters***, Vol. 140, (2020), 325-331. doi: 10.1016/j.patrec.2020.11.008.

23. Xia, M., Zhang, X., Weng, L. and Xu, Y., "Multi-stage feature constraints learning for age estimation", ***IEEE Transactions on Information Forensics and Security***, Vol. 15, (2020), 2417-2428. doi: 10.1109/TIFS.2020.2969552.

24. Zhang, K., Gao, C., Guo, L., Sun, M., Yuan, X., Han, T.X., Zhao, Z. and Li, B., "Age group and gender estimation in the wild with deep ror architecture", ***IEEE Access***, Vol. 5, (2017), 22492-22503. doi: 10.1109/ACCESS.2017.2761849.

25. Dhomne, A., Kumar, R. and Bhan, V., "Gender recognition through face using deep learning", ***Procedia computer science***, Vol. 132, (2018), 2-10. doi: 10.1016/j.procs.2018.05.053.

26. Nga, C.H., Nguyen, K.-T., Tran, N.C. and Wang, J.-C., "Transfer learning for gender and age prediction", in 2020 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan), IEEE. (2020), 1-2.

27. Bekhet, S., Alghamdi, A.M. and Taj-Eddin, I., "Gender recognition from unconstrained selfie images: A convolutional neural network approach", ***International Journal of Electrical & Computer Engineering***, Vol. 12, No. 2, (2022), 2066-2078. doi: 10.11591/ijece.v12i2.pp2066-2078.

28. Thung, K.-H. and Wee, C.-Y., "A brief review on multi-task learning", ***Multimedia Tools and Applications***, Vol. 77, (2018), 29705-29725. doi: 10.1007/s11042-018-6463-x.

29. Zhang, Y. and Yang, Q., "A survey on multi-task learning", ***IEEE Transactions on Knowledge and Data Engineering***, Vol. 34, No. 12, (2021), 5586-5609. doi: 10.1109/TKDE.2021.3070203.

30. Ranjan, R., Patel, V.M. and Chellappa, R., "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition", ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, Vol. 41, No. 1, (2017), 121-135. doi: 10.1109/TPAMI.2017.2781233.

31. Ran, H., Ning, X., Li, W., Hao, M. and Tiwari, P., "3d human pose and shape estimation via de-occlusion multi-task learning", ***Neurocomputing***, (2023), 126284. doi: 10.1016/j.neucom.2023.126284.

32. Savchenko, A.V., "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks", in

2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY), IEEE. (2021), 119-124.

33. Yu, W. and Xu, H., "Co-attentive multi-task convolutional neural network for facial expression recognition", ***Pattern Recognition***, Vol. 123, (2022), 108401. doi: 10.1016/j.patcog.2021.108401.

34. Wu, C. and Zhang, Y., "Mtcnn and facenet based access control system for face detection and recognition", ***Automatic Control and Computer Sciences***, Vol. 55, (2021), 102-112. doi: 10.3103/S0146411621010090.

35. Agbo-Ajala, O. and Viriri, S., "Deeply learned classifiers for age and gender predictions of unfiltered faces", ***The Scientific World Journal***, Vol. 2020, (2020). doi: 10.1155/2020/1289408.

Persian Abstract

چکیده

تشخیص ویژگی‌های انسان از روی چهره یکی از موضوعات مهم و دارای کاربردهای مختلف در حوزه بینایی کامپیوتر است. چهره انسان ویژگی‌های مهمی را در تعاملات اجتماعی و شخصیت شناسی افراد دارد. اطلاعات برجسته‌ای مانند هویت فرد، سن، جنسیت و احساسات را می‌توان با استفاده از چهره افراد تشخیص داد. در همین راستا در این مقاله روشی به منظور تشخیص سن، لبخند و جنسیت از روی تصویر چهره معرفی شده است. چهارچوب یادگیری عمیق چند وظیفه‌ای پیشنهاد شده است که می‌تواند به طور مشترک سه ویژگی مهم از چهره انسان را با دقت قابل توجهی تخمین بزند. آزمایش‌های انجام شده نشانگر این است که روش پیشنهادی نسبت به روش‌های پیشرفته اخیر در هر سه وظیفه در معیارهای رایج به دقت برتری دست یافته است. همچنین نشان داده شده است که یادگیری چند وظیفه‌ای قادر است با استفاده از داده‌های کمتر در یکی از ویژگی‌ها از سایر ویژگی‌های دارای داده‌های بیشتر بهره بگیرد تا دقت تشخیص را در آن ویژگی بالا ببرد. به منظور ارزیابی روش پیشنهادی از دو پایگاه داده IMDB-WIKI و GENKI-4K استفاده گردیده است.