International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Age-Group Classification from Facial Parts Using a Multi-Stream Convolutional Neural Network

Vijay Choudhary[1], Pankaj Pateriya[2]

Submitted: 05/11/2020 **Revised**: 19/12/2020 **Accepted**: 26/12/2020

Abstract: Age is a most significant part of human life. With time, there are many changes in the facial structure. The facial structure indicates the combination of the parts of the face (eyebrows, mouth, nose, eyes, and cheeks), shape, and texture of the skin. Several strategies have been proposed by researchers for age detection, although more work is still needed in this area for better results. In this study, we describe a method for estimating a person's age by using facial landmarks to crop facial features from an image of a person. The multi-stream CNN (convolutional neural network) model receives input from the facial regions that were clipped. The suggested approach uses age recognition methods on images that haven't been filtered. We are working on an Adience benchmark dataset that is primarily intended for identifying age and gender. Compared with processing complete facial images, the proposed approach processes cropped facial parts, reducing the amount of image data presented to each stream. The resulting performance is evaluated using the Adience benchmark dataset. The proposed architecture addresses overfitting through L2 regularization and dropout, while the data are evaluated using the reported cross-validation protocol.

Keywords: *Feature extraction, Age estimation, CNN, Classification Analysis, Dataset.*

1.0 Introduction

Artificial intelligence age detection has various applications in different fields, such as forensics (to find lost people), cosmetology, computer-human interaction, and access control. In the biological sense, molecules and cells of the human face are damaged over time. In an ageing progression, the human face is affected by several factors, such as physical conditions, health conditions, gender, and genetic factors. The face is a unique characteristic of a person. Over time, it illustrates prominent changes in its characteristics related to its pigment changes, mainly due to sun exposure, loss of muscle tone, skin thinning and the appearance of wrinkles, bones shrinking in size, and damage to cells due to external environmental effects. All these factors make it incredibly challenging to determine a person's age, and even more so for machines.

Estimating the age of a person using image datasets is the aim of researchers with the best performance. A lot of methods and techniques are used to estimate age. Automated age estimation consists of several processes, such as face recognition, face alignment, feature extraction from the face, and other age-related tasks. Accurate or most important feature extraction from images provides more accuracy for age detection. Various algorithms have been suggested by the researchers for feature extraction, such as initially introducing the anthropometric models, AAM (active appearance model), age manifold approach, age pattern subspace, and texture-based models. Yun Fu et al. [8], the researcher uses the age manifold technique to detect age from facial pictures with a regression model. Age-related prediction tasks can be categorized into age-group classification and age regression. In an age-group classification task, the target is represented as an age range (e.g., 2–5 or 6–10), and a multi-class classifier is used. In an age-regression task, the model predicts a numerical age value. We propose a deep learning approach for age-group classification from facial images. A CNN is a deep neural network architecture designed to learn spatial patterns from image data. CNNs learn hierarchical image representations through successive layers, from lower-level patterns to higher-level features. The CNN architecture works on a layering approach in which the input image is passed to the first layer (the "input layer") and a result is provided from the last layer (the "output layer"). The CNN model is also known as "ConvNets."

[1, 2] *Department of Computer Science & Engineering, IPS Academy, Institute of Engineering and Science, Indore 452012, INDIA*
vij.choudhary@gmail.com , ppateriya@gmail.com

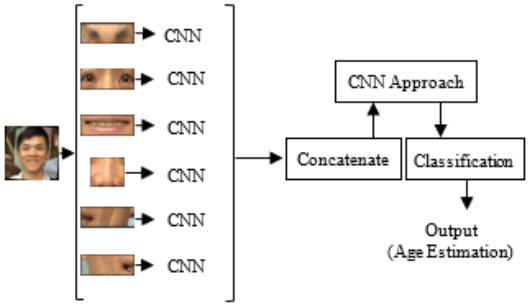

Figure 1. Flow of the age estimation process with facial parts

2.0 Literature Review

In recent years, estimating the human age has been an exciting topic in the research field. We have studied previous research papers and analyzed different age estimation techniques. Over time, several algorithms have been discovered with improved efficiency. Guodong Guo et al. [2] present the age-manifold learning technique for extracting the facial ageing patterns and introduce the LARR (locally adjusted robust regression) method. This method is designed for learning and predicting human age. O. Agbo-Ajala et al. [3] introduce a lightweight CNN model. It is designed to have fewer layers so that a person's actual and clear age can be determined from unrestricted facial images that can be deployed on smartphone devices. The result can be obtained from the MORPH-II, FG-NET, and APPA-REAL datasets. The age estimation process is done in three stages: feature extraction, selection, and classification. Noor F. Hasan et al. [4] proposed an LBP (local binary pattern) algorithm to extract features, and the FSM (feature selection method) selected the critical features to improve the accuracy. An SVM (support vector machine) is used to classify the faces and assign their corresponding ages.

The basic requirements of the problem of age estimation through the facial image are face identification and landmark detection. Vahid Kazemi et al. [6] introduce a one-millisecond face alignment for a single image. The researcher describes how a cluster of regression trees can be used to recover the location of the facial landmark from a sparse sub-cluster of pixel intensity values taken out of the input images. Siyu Han et al. [5] present a new Xception model for the age estimation technique. It provides higher accuracy as compared to VGG16, ResNet, DenseNet, InceptionResNetV2, and other deep learning CNN models. In this paper, the soft labelling technique is used for performance optimization. Hamdi Dibeklioglu et al. [7] focus on the smile expression to detect a person's age. In the videos, seven dynamic facial surface deformations (eyebrow, eye-side, eyelid, cheek, mouth, mouth-side, and chin) are tracked during a smile on the facial image. Hamdi proposed adaptive age grouping based on hierarchical age estimation. In this approach, they have divided the age estimation into two steps: the first is to estimate the age group. The second step is the estimation of a small adjustment to the age prediction model based on the exact age. Angeloni et al. [9], working on the CNN model with multi-stream technique by using an unconstrained image dataset, estimate the person's age from face parts (eye, nose, eyebrow, mouth, and cheek). Crop each facial part using landmarks and feed it into the individual CNN stream. Arwa S. et al. [11] categorized datasets into two kinds used in age estimation: constrained and unconstrained. The constrained data is defined as the correct captured image under the right conditions. While the unconstrained data are described as the images clicked in unfiltered conditions [12].

3.0 Methodology

In the proposed approach, multi-stream CNN architecture is done for determining the human age from individual images. CNN is a deep learning algorithm and type of feed-forward ANN (artificial neural network). Deep learning enables models to learn useful representations directly from training data. In deep learning techniques, feature extraction and classification both are simultaneously performed without human interference. Compared with approaches based primarily on handcrafted features, deep learning can learn feature representations directly from images, although computational cost depends on the architecture and training setup. One of the best reasons is to use the deep learning algorithm to handle large amounts of data. In CNN, artificial neurons, or nodes, receive input, process it, and send the processed data as output. Convolutional, pooling, and fully connected layers make up the conventional architecture of a CNN. With convolution, a series of convolutional filters are applied to the input images, each of which activates different features of the images. Pooling reduces the number of parameters the network needs to learn while doing nonlinear down-sampling to simplify the output. In contrast to a conventional neural network, CNN features shared weights and bias values that are similar for all neurons in the hidden layer of a particular layer.

There is a hidden layer between the input and output of the algorithm in which the function applies weights to the inputs and directs them through an activation function as outputs. ReLU (rectified linear unit) is one of the activation functions that maintains positive values while translating negative values to zero, enabling quicker and more efficient training. Since only the activated features are carried over to the following layer, this is frequently referred to as "activation". We are working on TensorFlow. It is a software library designed by the Google team. TensorFlow provides a collection of workflows to develop and train models using Python. The main expertise of the Keras library is to deal with models, layers, call-back API, optimizers, metrics, losses, data loading, and other functions. Whereas

TensorFlow is a framework layer for various programming languages, handling with tensors, gradients, and variables. The age-group classification process consists of face detection, face alignment, facial landmark localization, facial-part extraction, feature learning, and age-group classification. The age estimation process is illustrated in Fig. 2.

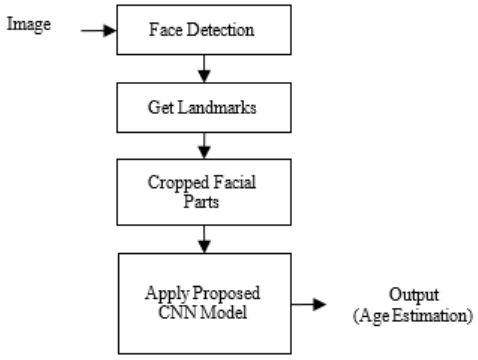

Figure 2. Age estimation process that indicates the flow of our method

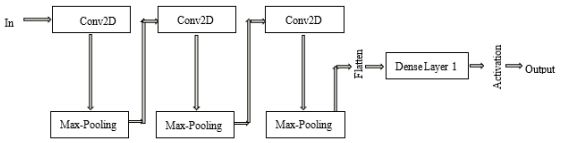

Figure 3. Block diagram of the CNN algorithm for each facial part

Face detection is the first step to estimate a person's age. In this process, the image is present in the format of RGB. The RGB image is nothing, but each pixel is a combination of the red, green, and blue channels. Initialization of facial landmark detector returned bounding box around the face.

In this work, we divide the image into facial parts and pre-process it. Manual annotation of facial landmarks or localization is a very complex process. Dlib facial landmark detector is an open-source toolkit that is trained on the iBUG-300 W dataset and contains machine learning algorithms. The toolkit returns 68 coordinates of 2D landmark location on the face. There are some points in the 68-landmark localization which are very useful for finding out the exact age. Using such points, align and crop facial parts (eyes, nose, eyebrows, mouth, and cheek) and resize them to format. To avoid overlap between facial parts, the coordinates of each part are different.

Eyebrows cover the area of coordinates between the 17 and 27 by a small portion horizontally and then before cropping, we aligned the facial part using the 21 and 22 coordinates. Similarly, the eyes cover the area of coordinates between 36 and 48 and are aligned with the eyes using 39 and 42 coordinates. The nose is a covered area between the 27 and 36 coordinates with alignment using 32 and 34. Points 48 to 68 coordinates are adopted for the mouth region and aligned the mouth region using 48 and 54. 1 to 17 coordinates cover the cheek area and alignment using 5 and 10 coordinates. Crop each facial part using the corresponding coordinates. This process is known as the pre-processing of an image. The above method pre-processes each facial part before applying the CNN model.

3.1 Age Estimation Architecture

In the proposed method, facial parts are processed by the individual CNN stream. Each stream follows the same approach shown in Fig. 3. Feature learning occurs before adding each facial part. Individual facial part feed into the separate CNN stream. In this architecture shown in Fig. 3, we are applying 3 convolution layers to extract lower-level features to higher-level features and generate feature maps. Most of the computation is performed on this layer. Each convolution layer performs the same calculation and is made up of a 3 × 3 kernel size with 32 filters. Stride is set to 1 for moving one pixel at a time. ReLU activation function is applying each convolution layer. Kernel regularizer, bias regularizer, and activity regularizer are set to L2 with 0.0001 value.

To reduce overfitting, regularization techniques such as L2 weight regularization and dropout are applied. Regularization helps control model complexity. L2 regularization adds a penalty proportional to the squared magnitude of the model weights to the training objective, discouraging excessively large weights. A variance-scaling initializer is used to initialize network weights while maintaining an appropriate scale of activations during training. After applying each convolution layer, max pooling is used for dimensionality reduction with 2 × 2 size. After applying 3rd max-pooling the height of cropped image will be 1 which cannot be further reduced because of this we cannot apply more than 3 convolution layers. The temporary output of the last pooling layer is flattened. Flatten output passed into the dense layer with 256 neurons as shown in Fig. 3. ReLU activation function is applied to the dense layer. The working activation function is to check whether a neuron's input is important or not to the network in the prediction's process. ReLU activation function is applied on the dense layer in each facial part stream. According to Angeloni et al. [9], classification is performed by the last layer / output layer of each stream individually as well as the last dense layer of 8-unit neurons.

Each CNN stream generates a temporary output from a dense layer shown in Fig. 3. We are taking each facial part output and concatenating using the multilayer perceptron, as shown in Fig. 1. Present concatenated output is passed into the first dense layer and applies dropout to reduce overfitting. The dropout rate is defined after the first dense layer and applied with a 0.5 value which means, that 50% of neurons are dropped from the FC (fully connected) layer. Again, output is fed into the second and third

dense layers of 256 neurons. The dense layer uses ReLU as an activation function. Before the final dense layer again apply the same dropout rate. To classify the input into the eight Adience age groups, a dense output layer with eight neurons uses Softmax activation to estimate the class probabilities.

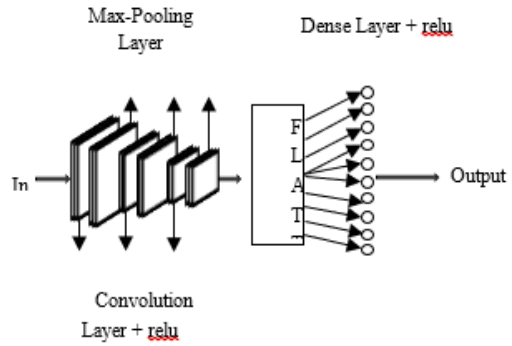

Figure 4. CNN architecture of our proposed method

Once the classification problem is tackled, sparse categorical cross entropy was selected as the loss function and the Adam algorithm was selected as an optimizer with learning rate is 0.001. The proposed method learns age-group representations directly from cropped facial-part images.

4.0 Dataset

We appraised our approach using the Adience benchmark dataset, which was developed to classify age and gender for face images clicked in real-world situations. The Adience benchmark dataset consists of unconstrained facial images which are uploaded without applying any filters. In the dataset, images are automatically added to Flickr.com from smart-phone devices or mobile phones. The dataset contains around 26580 images of 2284 people with 8 age labels and 2 gender labels. The age label is represented in format of (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60–above).

We have trained our model in images, captured in extreme variations such as head pose, lightning conditions quality, occlusion, and more. So that we can be able to generalize the model. Two variations are present in their dataset, first is frontal which is a roughly frontal face within approximately ±5° degree from a forward-facing image. Second is the complete set in which faces are within a yaw angle of ±45° degrees. The Adience dataset works on 5-fold cross-validation protocol [11].

Figure 5. Images of the Adience benchmark dataset

5.0 Implementation

We proposed multi-stream CNN approaches were implemented using TensorFlow library version 2.8.2, an end-to-end open-source platform for deep learning. The network training for 1-fold is required for about 40 minutes without using GPU. The training was conducted on Intel Core i3-5005U, CPU 2.00 GHz, and 2.00 GHz processor. The proposed model runs on Colab notebook in Python language. The execution time is noted from the input (facial part) process to the output generated by the CNN model.

5.1 Experimental Result

Table 1. Adience dataset images with each fold accuracies of our proposed approach

Train	Test	Validation	Dataset	Accuracy
2876	800	319	Fold 1	54%
2549	720	288	Fold 2	45%
2249	625	250	Fold 3	55%
2369	659	263	Fold 4	52%
2480	690	275	Fold 5	46%

The proposed approach explained in Section 3 generates specified results. The model was trained on the Adience benchmark dataset. We must provide better accuracy, images are split into three sections, 72% for the training set, 08% for validation data, and 20% for testing data. After we are detecting the landmark location, the image of the face fragment is pre-processed, which means it is resized according to its appropriate size and we enlarged the fragment region by a specific percentile. The eyebrow region enlarges by 3% from the horizontal direction and resizes it 228×33 pixels. Similarly, eyes resized in 202×38 pixels and enlarge is the same as the eyebrow region. The same thing was the nose, mouth, left cheek, and right cheek regions enlarged by 40%, 8%, 2%, and 2% respectively, and resized to 103×73, 114×66, 114×25, and 114×25 respectively in [10]. The resized images originally

pass into the CNN model. We reduced the size of the input because the image size takes a lot of time to process and even after adding all the facial parts, their sum is less than the image size. Image size in the Adience dataset is at least 200×200 pixels equals (40,000) parameters but the sum of facial parts equals (35,943) parameters which takes less time in processing.

For this reason, we introduce a new CNN architecture in which we train the data (facial part images) in limited labels. 'Why do we use limited labels?' If there is a problem of overfitting due to covering all the data points then we use limited labels. The number of parameters is present in Table 2. The parameter is nothing but a weight that is calculated using this formula:

Convolution Layer Parameter = ((Filter width × Filter height × Previous layer filters) + 1) × Number of Filters

Dense Layer Parameters = (number of input units × number of output units) + number of output units

Our proposed multi-stream CNN approach has a 3053448 parameter which is better than Angeloni et al. [9] approach that contains 4840744 parameters. We have taken only the important parameters because unattended parameters only increase the training time.

Sparse-Categorical accuracy class calculates the accuracy same as categorical accuracy used for sparse targets. Our proposed model has 8 target age-group labels (0–2, 4–6, 8–13, 15–20, 25–32, 38–43, 48–53, 60–above) according to Adience dataset, renamed as (0, 1, 2, 3, 4, 5, 6, 7) integer values. Categorical accuracy specifies target labels as a one-hot encoded vector e.g. 2 classes ((0, 1), (1, 0)) but our target labels as an integer format (1, 2 … 7) then we must use sparse-categorical accuracy. Table 1 shows us the exact accuracy of each fold, yielding a 5-fold average accuracy of 50.4% across the dataset.

The first three columns represent the images that help the model train, test, and validation data separated by a specific ratio. The cross-entropy loss function predicts the error that occurs during the training process.

Table 2. Number of parameters of our proposed approach

Facial Part	Conv2d #1	Conv2d #2	Conv2d #3	Dense Layer
Eyebrows	896	9248	9248	426240
Eyes	896	9248	9248	565504
Nose	896	9248	9248	631040
Mouth	896	9248	9248	590080
Left Cheek	896	9248	9248	98560
Right Cheek	896	9248	9248	98560
Concatenate Facial Parts				
Dense Layer	393472			
Dense Layer	65792			
Dense Layer	65792			
Output Layer	2056 (Softmax)			
Total Parameter	**3053448**			

At the compiled time our multi-labeled classification approach applying the sparse cross-entropy loss function to evaluate the gradients. Working of gradients are used to update the weight of the CNN. Using 0.001 learning rate, Adam optimizer updates the individual learning rate for each parameter and evaluates the first (mean) and second (uncentered variance) moments of the gradients. It means, at the time of training process, Adam optimizer automatically learn a new weight of training data. History of our model define the evaluation process of the CNN architecture with 20 epochs. Each epoch is defined as the total number of iterations that all the training data are utilized in one cycle to train our proposed model.

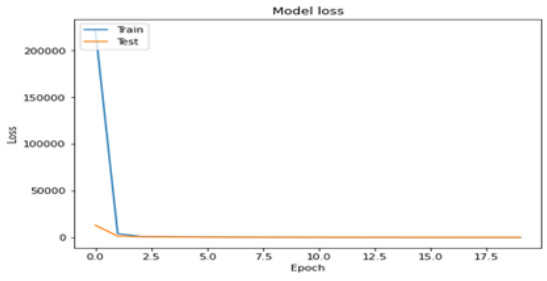

Figure 6. Graph between the model loss and epoch

Graph 1 (Fig. 6) (Model Loss) shows the process of testing and training between the epochs and loss. Graph 2 (Fig. 7) (Model Accuracy) identifies the

accuracy with respect to epochs. After then we evaluate the model with non-trainable data images to give better accuracy without GPU.

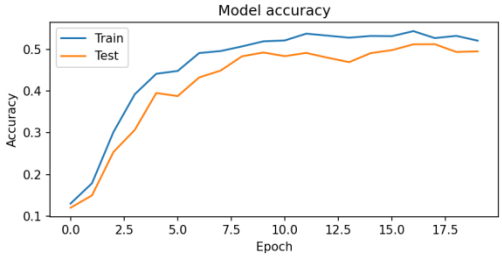

Figure 7. Graph between the model accuracy and epoch

In another deep analysis of our model, we must calculate the confusion metrics under the 8-age group. The predicted model compared with the testing data result is shown in Table 3.

Table 3. Confusion matrix of the proposed approach over the eight age groups (Fold 1)

Age Group	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60+
0-2	54	14	4	1	9	7	3	1
4-6	27	33	3	2	10	2	2	3
8-13	0	2	21	2	5	2	1	0
15-20	0	2	1	6	7	2	0	1
25-32	21	30	9	17	266	54	31	19
38-43	6	6	6	1	22	42	9	6
48-53	1	1	2	0	8	2	6	3
60+	0	1	0	0	2	1	0	1

This confusion matrix is made from Fold 1. We analyzed that the 25–32, 0–2, and 38–43 age groups give better results whereas the 8–13 age group is giving good results as compared to 15–20, 48–53, and 60+. They provide the worst result because of the unavailability of images in the Adience dataset.

To better understand the evaluation of our trained model we need to calculate the score of the term precision, recall, F1-score, and support. Using predicted and testing data along with the target data, the classification algorithm prepared the classification report presented in Table 4.

Table 4. Classification report of our proposed work

Age Group	Precision	Recall	F1-score	Support
0-2	0.58	0.50	0.53	109
4-6	0.40	0.37	0.39	89
8-13	0.64	0.46	0.53	46
15-20	0.32	0.21	0.25	29
25-32	0.60	0.81	0.69	329
38-43	0.43	0.38	0.40	112
48-53	0.26	0.12	0.16	52
60+	0.20	0.03	0.05	34
Accuracy			**0.54**	800
Macro avg.	0.43	0.36	0.37	800
Weighted avg.	0.50	0.54	0.51	800

Table 5. Comparison of the proposed approach with selected age-estimation methods

Authors	Datasets	Used Algorithm	MAE	Acc.	1-off Acc.
Guodong Guo et al. 2008 [2]	FG-NET	Age manifold + LARR	5.07	–	–
O. Agbo-Ajala et al. 2020 [3]	MORPH-II	Lighter CNN Model	2.31	–	–
Noor F. Hasan et al. 2020 [4]	FG-NET	LBP + FSM + SVM	4.51	–	87.69
Vahid Kazemi et al. 2014 [6]	HELEN + LFPW	Regression tree	4.9	–	–
H. Dibeklioglu et al. 2015 [7]	Uva-NEMO smile + disgust database	LBP + dynamic	4.38	–	89.58
Angeloni et al. 2019 [9]	Adience	Multi-Stream CNN	–	51.03	83.41
Our Proposed Model	**Adience**	**Multi-Stream CNN**	**–**	**50.4**	**–**

6.0 Conclusion

The proposed method describes age-group classification using a multi-stream CNN model with a set of facial parts (eyebrow, eye, nose, mouth, and cheek) as input. Initially, detect the face from the image and its landmarks. Afterward, align the five facial parts, resize them, and crop them. Angeloni et al. [9] have used four facial parts in their approach. We added one extra facial part (the cheek) to our method. The resizing portion of each part has the same pixel size. Individual streams are fed by pre-processed facial parts. Stream is nothing but a CNN model, which produces the output. Combined the outputs of each facial part and again fed them into the dense layer. The Softmax function computes the probability of each age-group class. This process is different from the rest of the methods and easy to learn. Models that have already been trained include ResNet, ZFNet, GoogLeNet, and AlexNet. As opposed to this strategy, where the minimum size of the facial part is 25×114 pixels, the pre-trained model's minimum picture size is up to 190×190 pixels. As a result, the pre-trained model cannot be used in this situation. The proposed CNN model achieved a reported average accuracy of 50.4% on the Adience benchmark dataset. The dataset contains approximately 26000 images uploaded by smartphone devices. This is the publicly available dataset.

The dataset contains eight age groups with an imbalanced class distribution. Our approach has fewer parameters even after adding an extra part. Future work on this problem will involve adjusting the size of face parts and the number of facial parts.

The reported accuracy reflects the difficulty of age-group classification on the unconstrained Adience dataset. Use of a GPU can reduce training and inference time; accuracy improvements would instead require changes such as improved training, data augmentation, architecture, or optimization. Apart from this, we can use this method in many contexts related to age-related tasks, access control, and the interaction between humans and computers.

References

[1] El Dib, M. Y., & Onsi, H. M. (2011). Human age estimation framework using different facial parts. Egyptian Informatics Journal, 12(1), 53–59.

[2] Guo, G., Fu, Y., Huang, T. S., & Dyer, C. R. (2008, January). Locally adjusted robust regression for human age estimation. In 2008 IEEE Workshop on Applications of Computer Vision, 1–6, IEEE.

[3] Agbo-Ajala, O., &Viriri, S. (2020). A lightweight convolutional neural network for real and apparent age estimation in unconstrained face images. IEEE Access, 8, 162800–162808.

[4] Hasan, N. F., & Mahdi, S. Q. (2020, April). Facial features extraction using LBP for human age estimation based on SVM classifier. In 2020 International Conference on Computer Science and Software Engineering (CSASE), 50–55, IEEE. DOI: 10.1109/CSASE48920.2020.9142061.

[5] Han, S. (2020, August). Age estimation from face images based on deep learning. In 2020 International Conference on Computing and Data Science (CDS), 288–292, IEEE. DOI: 10.1109/CDS49703.2020.00063.

[6] Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1867–1874. DOI: 10.1109/CVPR.2014.241.

[7] Dibeklioğlu, H., Alnajar, F., Salah, A. A., & Gevers, T. (2015). Combining facial dynamics with appearance for age estimation. IEEE Transactions on Image Processing, 24(6), 1928–1943.

[8] Fu, Y., Xu, Y., & Huang, T. S. (2007, July). Estimating human age by manifold analysis of face

pictures and regression on aging features. In 2007 IEEE International Conference on Multimedia and Expo, 1383–1386, IEEE. DOI: 10.1109/ICME.2007.4284917.

[9] Angeloni, M., de Freitas Pereira, R., & Pedrini, H. (2019). Age estimation from facial parts using compact multi-stream convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. DOI: 10.1109/ICCVW.2019.00366.

[10] Del Coco, M., Carcagni, P., Leo, M., Spagnolo, P., Mazzeo, P. L., &Distante, C. (2017). Multi-branch CNN for multi-scale age estimation. 19th International Conference on Image Analysis and Processing – ICIAP, Catania, Italy, September 11–15, 2017, Proceedings, Part II 19, 234–244, Springer International Publishing.

[11] Al-Shannaq, A. S., & Elrefaei, L. A. (2019). Comprehensive analysis of the literature for age estimation from facial images. IEEE Access, 7, 93229–93249. DOI: 10.1109/ACCESS.2019.2927825.

[12] Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security, 9(12), 2170–2179. DOI: 10.1109/tifs.2014.2359646.