ASEAN Engineering Journal

Full Paper

EVALUATION OF SIMULTANEOUS IDENTITY, AGE AND GENDER RECOGNITION FOR CROWD FACE MONITORING

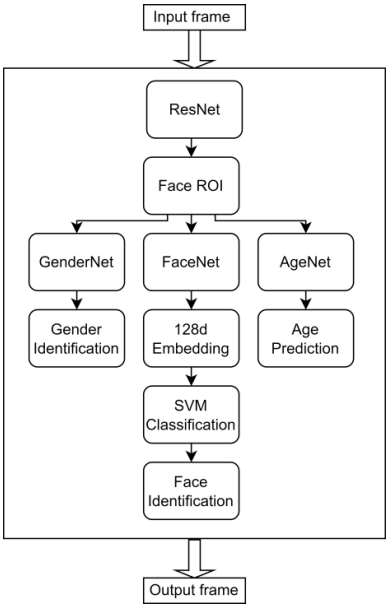
Intiaz Mohammad Abir*, Hasan Firdaus Mohd Zaki, Azhar Mohd Ibrahim

Department of Mechatronics Engineering, Kulliyyah of Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Article history
Received
10 September 2021
Received in revised form
28 April 2022
Accepted
18 August 2022
Published online
28 February 2023

*Corresponding author
intiaz.abir@live.iium.edu.my

Graphical abstract



Abstract

Nowadays, facial recognition combined with age estimation and gender prediction has been deeply involved with the factors associated with crowd monitoring. This is considered to be a major and complex job for humans. This paper proposes a unified facial recognition system based on already available deep learning and machine learning models (i.e., FaceNet, ResNet, Support Vector Machine, AgeNet and GenderNet) that automatically and simultaneously performs person identification, age estimation and gender prediction. Then the system is evaluated on a newly proposed multi-face, realistic and challenging test dataset. The current face recognition technology primarily focuses on static datasets of known identities and does not focus on novel identities. This approach is not suitable for continuous crowd monitoring. In our proposed system, whenever novel identities are found during inference, the system will save those novel identities with an appropriate label for each unique identity and the system will be updated periodically in order to correctly recognise those identities in the future inference iterations. However, extracting the facial features of the whole dataset whenever a new identity is detected is not an efficient solution. To address this issue, we propose an incremental feature extraction based training method which aims to reduce the computational load of feature extraction. When tested on the proposed test dataset, our proposed system correctly recognizes pre-trained identities, estimates age, and predicts gender with an average accuracy of 49%, 66.5% and 93.54% respectively. We conclude that the evaluated pre-trained models can be sensitive and not robust to uncontrolled environment (e.g., abrupt lighting conditions).

Keywords: Age estimation, crowd monitoring, deep learning, facial recognition, gender prediction

1.0 INTRODUCTION

One of the most trending topics nowadays in artificial intelligence is the facial recognition technology. It is an important biometric tool to detect and verify a person through learning algorithms. Biometrics can be used to recognize and authenticate an entity using a distinctive and person-specific collection of identifiable and verifiable information. There are also several reasons behind the current growing interest in facial recognition; namely growing public attention for safety, the use of digital identity authentication etc. However, face identification tasks are almost always focused on pre-trained known identities while ignoring novel identities. This can pose a major bottleneck when it comes to real-time and large-scale crowd monitoring where new identities always appear in the scene.

There are many existing face recognition, age estimation and gender prediction models with high accuracies. However, most of these models were not evaluated on real-life conditions (e.g., multiple faces in an image, imbalanced dataset). This paper proposes a deep learning-based system to simultaneously perform face identification, age estimation and gender prediction. The proposed method utilizes the already available deep learning and machine learning models instead of self-developed models. The proposed pipeline consists of ResNet-based detector [1] for face detection, FaceNet [2] for

extracting facial features, Support Vector Machine [3] for identity classification, AgeNet [4] for age estimation and GenderNet [4] for gender prediction. Additionally, we devise an incremental feature extraction method instead of extracting the features of all the faces in the dataset to lower the computational load. In the proposed system, whenever new identities are detected in the frame, the identities will be saved for pre-processing (i.e., face alignment, clustering, and labelling). After pre-processing is done, the features (i.e., 128-d embeddings using FaceNet) will be extracted from the new faces. This newly extracted embeddings will then be merged with the existing old embeddings without the need to re-extract all the embeddings from scratch. Then, a SVM model will be trained using the merged embeddings. This method allows the model to be updated periodically. Finally, we propose a multi-face, realistic and challenging test set that contains images taken in varying lighting conditions and varying poses to evaluate the proposed system.

The rest of the paper proceeds as follows: Section 2 explores the related works. After that, the methodology of the research is discussed in Section 3. In Section 4, the results are analysed and discussed. Finally, Section 5 concludes the paper.

2.0 RELATED WORKS

On automatic face recognition the works of [5], [6] are considered one of the pioneers. Finding the positions of a group of landmarks of face and measuring relative locations and distances between those by using specialized edge and contour detectors were proposed by them. These are referred to as geometry-based methods. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) which are known as statistical subspaces methods later gained popularity. These are also known as holistic methods. The idea of finding eigenvectors (also known as eigenfaces) by applying PCA to a group of training face images was first proposed in [7]. Eigenvectors result in the most variance in data distribution. Price et al. proposed a holistic method that is based on LDA [8]. For face recognition, Support vector machines (SVMs) [3] have also been used. Locality Preserving Projections (LPP), an approach related to PCA and LDA was proposed in [9]. Joint Bayesian method [10], where sum of two independent Gaussian variables represents a face image instead of image differences has been proposed in [11]. This method achieved a 92% accuracy, which is the highest reported by a holistic method, on the Labelled Faces in the Wild (LFW) dataset [12]. Feature-based approaches which consist of matching these local features across face images was proposed in [13]. This was referred to as modular eigenfaces method. Another popular method named elastic bunch graph matching (EBGM) method was proposed in [14], which represented a face using a graph of nodes which contain Gabor wavelet coefficients that extracted around a set of predefined facial landmarks. In [15], a method using histograms of oriented gradients (HOG) in place of Gabor wavelet features was proposed that outperforms the EBGM. Learning local features from training samples have been the focus for some feature-based methods [16]. Hybrid method was created by combining holistic and feature-based methods. Some approaches simply combine the two techniques without having any interaction [13] but the most popular approach extracts local features like LBP, Scale-invariant feature transform (SIFT) and projects these features onto a lower-dimensional and discriminative subspace (e.g., using PCA or LDA) [17], [18]. Recently, Deep Learning (DL) based methods became popular for face recognition due to the advancement of computer processing power and data storage capacity. The most common type of DL methods for FR are Convolutional Neural Networks (CNNs) which are end-to-end trainable models and have the advantage of being trained with a large amount of data which can learn the variations of face representations existent in the training data. One of the training approaches of CNN models treats the problem as a classification problem and each subject corresponds to a class which can recognize faces not present in the training dataset after being trained [19]. The process of bottleneck features learning by optimizing distance metric between triplets [2] of faces is another common approach. After being trained with data in large amounts, DL methods for face recognition became state-of-the-art. For example, Facebook's DeepFace [19], had an accuracy of 97.35% on the LFW benchmark. A massive dataset of 4.4 million faces from 4030 subjects was used to train the CNN with SoftMax loss. [20] achieved results the same as [19] by training 60 different CNN on patches where the dataset contained 202,599 face images of 10,177 celebrities. [21] carried out a thorough study of various CNN architectures and the results showed that the trade-off between accuracy, speed and model size obtained with a 100-layer ResNet. Google trained a CNN model named FaceNet [2] using a dataset of 200 million face identities and 800 million image face pairs. They used a "triplet based" loss, applied at multiple layers, where a pair of two same (x; y) and a third different face z are compared with the goal of making x closer to y than z. This method has one of the highest accuracies currently on LFW which is 99.63%. Currently, FaceNet is one of the most popular method for solving face recognition problems and is vastly used by researchers in this field [22]–[36].

For a long time, problems associated with automated age extracting attributes have been in attention. By calculating ration between features of face (e.g., eyes, nose, face) early classification of age was done. Ratio between these is calculated by calculating sizes and distances to predict age with the use of conventional methods after localizing [37]. From studies done in the '90s [38] to modern approaches, age estimation of a person by analysing the geometry of faces like the pipeline used in [39] is common. For example, [39] combined Biologically Inspired Features (BIF), Canonical Correlation Analysis (CCA) and Partial Least Square (PLS) based methods. Face images were already being represented in BIF [40] that paved the way to later works such as [41] which demonstrated the performance of humans was matched by that automatic approach. Two-stage pipeline-like feature extraction as LBP [42] and then classifying using a SVM or multilayer perceptron (MLP) was the base of approaches before CNNs. CNNs, on the other hand, implements the aforementioned process in one step where both the extraction and the classification of categories of age is learned by the network [4], [43] or by performing regression of age [44]. Unlike age analysis, gender recognition through neural networks like the approach of [45], was already being proposed in the early '90s. In [45], an autoencoder and a classifier whose input was the encoded output layer of the autoencoder were proposed as two neural networks. Its reliance on cropping manually, scaling and face rotation in the picture captured in a

supervised environment was its biggest drawback. In [46], pipelines based on a feature extractor and stacked classifier were proposed which was inspired from age estimation methodology. In [47], a pre-trained network was fine-tuned followed by an SVM being trained with the use of deep features computed by CNN was used to perform gender recognition. For determining gender, the same CNN-based methods used to determine age were also used in [4], [48], and this demonstrated that CNNs have the ability to execute tasks just by modifying the data that is used for learning and nothing else. In recent years, many studies related to age and gender recognition using CNN-based methods have been done by researchers and many of the works achieved state-of-the art results [4], [37], [49]–[62]

In the abovementioned literature, it is evident that face recognition is done separately in contrast with gender and age prediction. In this research, we propose a unified system by combining the works done by [2], [4] for simultaneous face detection, recognition, gender prediction and age estimation. Because identity along with age and gender can help to monitor the crowd properly and their behaviour can be categorized by their age and gender. We also propose an incremental embeddings extraction method to train a SVM model efficiently. The primary aim of this research is to evaluate the proposed system using our newly proposed, realistic, and challenging test dataset. The images in the test dataset contains varying number of persons (e.g., three persons in an image) with varying poses and lighting conditions.

3.0 METHODS

Deep learning has become an important medium for detection and classification of various things including face recognition. This paper focuses on person identification, age estimation and gender prediction. Age estimation is either a classification or a regression problem which can detect age either from an image or real time stream. Whereas gender prediction and face recognition are a classification problem. Here, we employ ResNet-based [46] face detector for face detection, FaceNet [2] combined with Support Vector Machine [3] for person identification, AgeNet [4] and GenderNet [4] for age estimation and gender prediction respectively. Some brief details of FaceNet, AgeNet and GenderNet are provided in the next subsections. After that the flow of operation is described. The flow of operation is divided into two parts (i.e., training and inference). The first part describes the training procedure. After that the second part describes the inference procedure.

3.1 FaceNet

FaceNet was developed and introduced around 2015 to solve the existing hurdles in terms of face detection as well as verification by researchers at Google. A deep convolutional network is used by FaceNet in order to optimize the embeddings directly. This one-shot learning approach is different than earlier deep learning approaches that have an intermediate bottleneck layer. Face image is transformed by the FaceNet algorithm into 128-dimensional vector in Euclidean space which is also known as “Embeddings”. Implementation of face recognition and verification can be done using the FaceNet embeddings as feature vectors. To summarize, the distance between random and non-similar images would be much further away than the similar images. A batch input layer and a deep convolutional neural network is existent in the FaceNet network architecture. The deep convolutional network is followed by L2 normalization by which the face embeddings are provided. Then the process is followed by the triplet loss. Tight crops of the face area are used as input images. The output of FaceNet is directly trained to become a compact 128-dimensional embedding by using the triplet-based loss function. Three images (i.e., an anchor, a positive and a negative) are required to calculate triplet loss. Among these three images, a positive and a negative image are present where the positive image and the anchor have the same identity, but the negative image and anchor has different identity. The distance between an anchor and the positive image is minimized and the distance between an anchor and the negative image is maximized by the triplet loss. Thus, to learn good 128-dimensional embedding for each individual face, triplet loss is one of the best methods. This method has one of the highest accuracies currently on LFW [12] which is 99.63%.

3.2 AgeNet and GenderNet

The AgeNet and GenderNet model was developed by Levi and Hassner. A simple AlexNet-like architecture is used by the AgeNet and GenderNet model. This architecture learns eight age brackets (i.e., 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+) for age and two brackets (i.e., Male, Female) for gender. These age brackets are ununiform. This happens due to the Adience dataset [42] that was used to train the model. The Adience dataset defines the age ranges exactly like these brackets. It is always challenging to accurately predict the age (regression problem) of a person as issues associated with genetics, physical appearances, cosmetics, and the effect of plastic surgery are existent. Due to this reason, age estimation is formulated by detecting age brackets by AgeNet as classification which simplifies the task to some extent. On the other hand, predicting gender is comparatively easier due to its binary classification nature. The AgeNet-GenderNet model architecture comprises of only three convolutional layers and two fully connected layers with a small number of neurons. Each of these three convolutional layers are followed by a rectified linear operation and pooling layer. The first two convolutional layers also follow normalization using local response normalization. Finally, two fully connected layers are added, each containing 512 neurons. This method has achieved one of the highest accuracies currently on the Adience dataset. For age estimation it achieved 84.7% one off accuracy and 86.8% exact accuracy for gender prediction which significantly outperformed the then state-of-the art methods.

3.3 Flow of Training Operation

For training a machine learning or deep learning model, dataset plays a vital role. The manual dataset method is not feasible in the proposed system as our recognition task is dependent on dynamic dataset, i.e., the proposed system needs to detect the novel identities in real time during inference and include them into the continuously growing dataset in order to detect those novel identities in a future inference run. To tackle this issue,

we propose Automated Dataset Creator System (ADCS). When novel identities are detected during inference run, ADCS will save the facial image of those identities from the frame and then it will align and cluster those faces and label each individual with a unique ID. Based on the updated dataset, the 128-d embeddings of the new faces will be extracted incrementally and the SVM model will be retrained periodically. Therefore, when those novel identities are again detected in the future time step, the system will be able to detect those identities and label them with the given unique ID. For age estimation and gender prediction we do not need to train the model as we are already using pretrained model. In summary, ADCS works in seven steps which is illustrated in Figure 1. These are listed below:

- During inference, detect unknown faces and save them for further processing.
- Align all the faces.
- Cluster all the faces and label them with unique IDs.
- Add them to the dataset.
- Extract the embeddings of the new faces.
- Merge the new embeddings with the old embeddings.
- Retrain the SVM model.

However, in order to run the inference operation, the system must be trained/initialized using an initial dataset. Our initial dataset contains 11 classes of known identities and 1 class of unknown identities. The unknown identity class contains frontal face images of random people which were collected randomly. On the other hand, the other 11 classes of known identities contain images of 11 specific people which were taken in different lighting conditions and varying poses. Figure 2 shows a sample of the 11 known classes. The next subsections provide details about the workflow of ADCS.

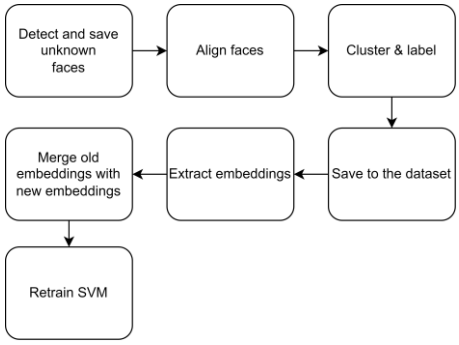


Figure 1 Workflow of our proposed Automated Dataset Creator System (ADCS)

3.3.1 Detecting and Saving Novel Identities

This step is to recognize novel identities in video streams. During inference, the video stream is feedforwarded to the ResNet face detection model to extract the face region of interest (ROI). Then, the face ROI is feedforwarded to the FaceNet model to output the 128-dimensional face embeddings. These facial embeddings are then fed as input to the SVM model for classification. A conditional criterion is then imposed where if the identity does not match with pre-trained "Known" classes and is more similar to the "Unknown" class, then it is labelled as "Unknown". Otherwise, it is given one of the labels from the existing pre-trained "Known" classes. The face ROI of the new identity is then saved as an image for alignment and clustering processes.

Figure 2 Sample data of eleven known classes

3.3.2 Face Alignment

After the novel identities are detected, they need to be aligned so that better clustering results can be achieved. Face alignment means digital image identification of the geometric facial structure and looking to secure canonical facial symmetry dependent on translation, size, and rotation. If there is a series of facial landmarks (input coordinates), the main aim is to transform the image into an output coordinate system. In this system, the facial area is in the centre of the image, the eye must be in a horizontal position and scaling is done to make sure faces remain visually similar for improved performance. Firstly, using a facial landmark model, the left and right eye regions are identified. Second step involves the computation of the centre of each eye which can be used as a parameter to re-align the rotation. After that, face rotation angle is calculated. In fourth step, the required or desired right eye is determined from left eye x coordinate. In step five, midpoint between two eyes or eye-centre is calculated. Final step is about aligning the face using the rotation matrix, which is generated using all the above-mentioned parameters. Figure 3 Demonstrates the result of face alignment.

Figure 3 Sample face alignment result. original face (left), aligned face (right)

3.3.3 Face Clustering

After face alignment is done the faces need to be clustered so that same faces can be grouped together and labelled using a unique id. While face recognition uses supervised learning for classification, face clustering involves unsupervised learning which consists of only faces with no classes. For the clustering task, we extract discriminative representation for the faces

which is an important criterion for the clustering algorithm. This involves extracting a 128-dimensional feature vector (referred to as encoding) for each image that will be used as a representation of the face. This process is done using a simplified version of ResNet DNN (Deep Neural Network). As for the clustering, Density Based Spatial Clustering of Applications with Noise (DBSCAN) is used. DBSCAN collects or amasses adjacent points packed from N-dimensional space. This creates a single cluster having adjacent points. Outliers are also handled well by DBSCAN. After extracting the face encodings DBSCAN is applied to cluster the encodings into unique clusters. Here, outliers are discarded. Each unique cluster represents unique class label where each face in a cluster is given the same label. After clustering is done, the faces are ready to be used for feature extraction. Figure 4 Visualizes the result of two sample clusters.

Figure 4 Sample face clustering result

3.3.4 Incremental Facial Features Extraction And SVM Training

For this step, usually the whole dataset is feedforwarded through FaceNet to extract the facial embeddings for each of the identities. Face detection and localization are done by feedforwarding the images through the ResNet model which provides the face ROI. After that the face ROI is transferred to the FaceNet model which creates a 128-dimensional facial embedding vector for each face. In the next step, these embeddings and respective labels are used to train a SVM model for classification. However, getting the embeddings of all the faces from scratch using FaceNet every time a new face is detected is not feasible in real scenario as the process is computationally heavy and time-consuming. To solve this problem, we propose an incremental feature extraction method. In this method, whenever new faces are detected, we will only extract the embeddings of the new faces and the old embedding data will be merged with the new embedding data and the combined data will be used to train an SVM. Thus, we can directly add the new embeddings to the old embeddings instead of getting the embeddings for all the identities from scratch. Using the proposed method, the model can be used in real scenarios as it reduces computational load and saves memory space. Figure 5 illustrates the process.

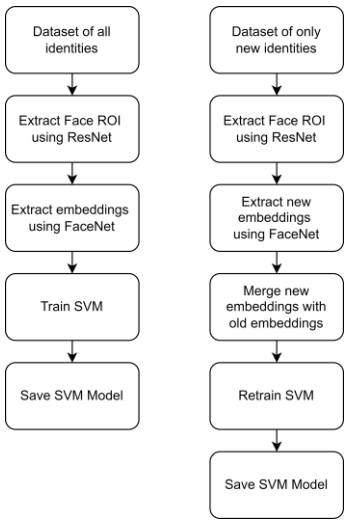


Figure 5 Workflow of conventional training (top), Workflow of our proposed incremental feature extraction based training (bottom)

3.4 Flow of Inference Operation

During inference, the first step is to detect the faces in an image/video using the ResNet based face detector which extracts the face region of interest (ROI). Then the extracted ROI(s) is passed to FaceNet, AgeNet and GenderNet simultaneously. FaceNet extracts the 128-dimensional facial embedding vectors for the faces. Then the embedding vectors are supplied to the SVM model for identity recognition. Concurrently, AgeNet estimates the age and GenderNet predicts the gender. Figure 6 illustrates the whole inference process. It is to be noted that face alignment algorithm is not applied during inference.

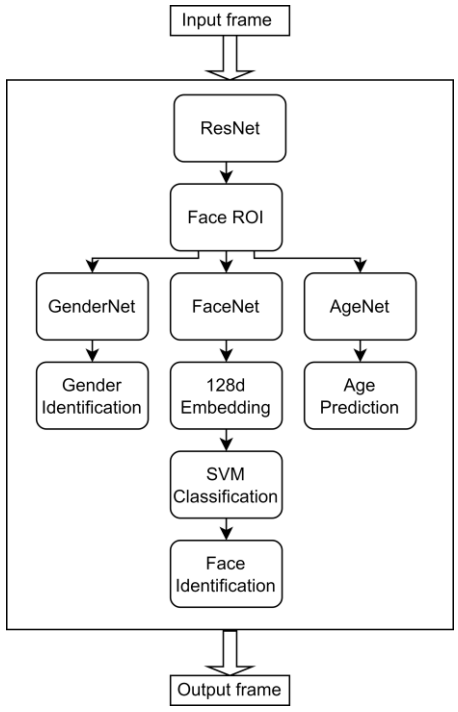


Figure 6 Flow of inference operation

4.0 RESULTS AND DISCUSSION

To evaluate the proposed system, we propose a realistic and challenging test dataset which contains images taken in bright light and dark light condition. In a real-life video stream, each frame can contain multiple faces. To reflect this, we have included images in our test dataset that contains varying number of faces ranging from one to four faces per image. We have evaluated the system for a total of 8 cases (e.g., two persons in a frame in dark light condition). Each case contains 25 images taken in varying poses. Our test dataset contains a total of 200 images (25 images for each case). It is to be noted that, compared to other publicly available dataset the proposed test set is very small. Thus, results can vary significantly even with adding or deleting even a few images. Figure 7 shows some samples of the proposed test dataset.

Figure 7 Sample data of the proposed test dataset

4.1 Identity Recognition Results

4.1.1 Accuracy

It can be observed from Figure 8A that the accuracy is 56% for a single person in dark light. The accuracy experiences a decline when multiple people reside in the frame. 68% accuracy can be achieved in bright light with a single person in the frame and rapid reduction of accuracy is observed with the increasing number of people in the frame which is illustrated in Figure 8B. Considering all scenarios, the average accuracy is 49%.

4.1.2 Precision

Precision is additionally referred to as reliability or repeatability. The precision for a single person in dark light and bright light is 57.14% and 63.44% respectively which decreases promptly as the number of people increases inside the frame. Figure 9 represents the simultaneous change in precision and condition.

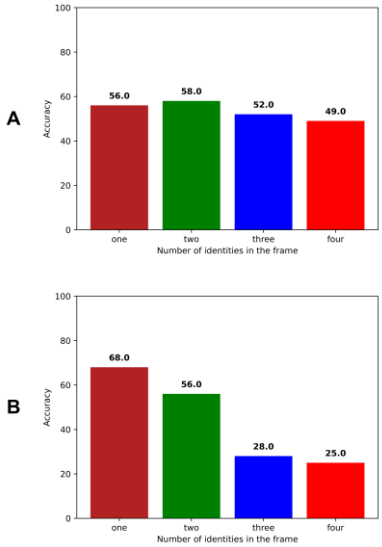


Figure 8. Accuracy in (A): dark light condition and (B): bright light condition

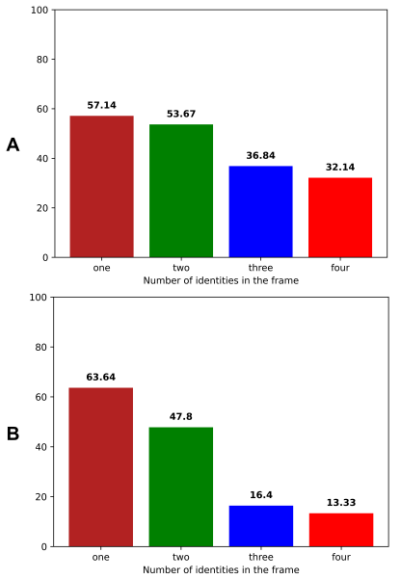


Figure 9 Precision in (A): dark light condition and (B): bright light condition

4.1.3 Recall

Recall is additionally referred to as true positive rate. As it can be observed from Figure 10 that the highest rate of recall is 65.20% in dark condition whereas the highest value in bright condition is 77.00% which is clearly higher than in dark condition. Similarly, the lowest value of dark condition and light condition is respectively 53.85% and 50.00%. Thus, it can be concluded that recall fluctuates in both conditions but is steadier in the dark light condition.

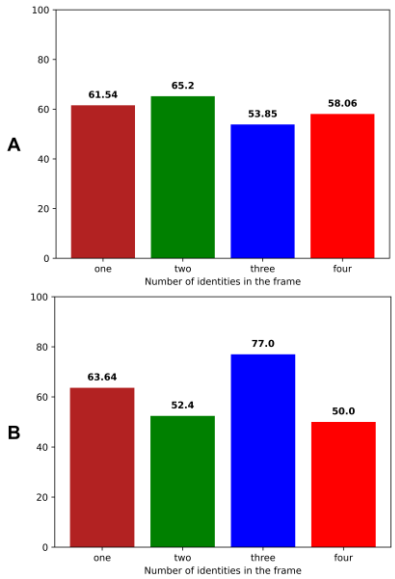


Figure 10 Recall in (A): dark light condition and (B): bright light condition

4.1.4 Specificity

True negative rate (TNR) is also called specificity. As demonstrated in Figure 11A, in dark light condition the specificity is 50.00% for a single person in the frame, as we increase the number of persons it remains almost linear and does not fluctuate much. In contrast, it can be seen in Figure 11B that in bright light condition the specificity is 71.43% for a single person in the frame, as we increase the number of persons it decreases rapidly and then becomes linear.

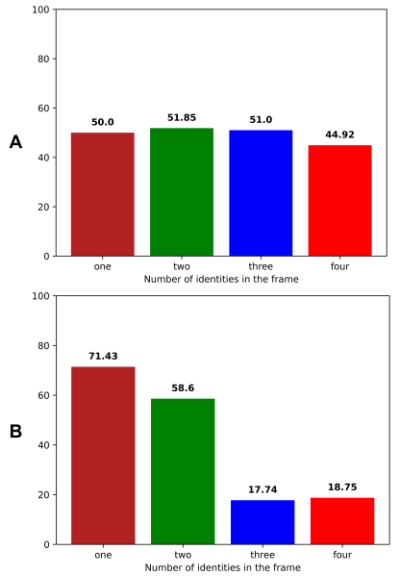


Figure 11 Specificity in (A): dark light condition and (B): bright light condition

4.2 Age Estimation and Gender Prediction Results

The one-off accuracy for age estimation does not follow any trend in both lighting conditions which is illustrated in Figure 12. As it can be observed that the highest value of the one-off accuracy is 78.66% in dark condition and the highest value in light condition is 89.33%. The lowest value in both condition is 54%.

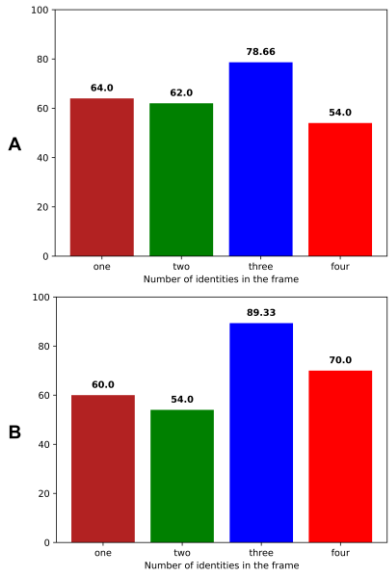


Figure 12 One-off accuracy in (A): dark light condition and (B): bright light condition

The exact accuracy for gender prediction, as illustrated in Figure 13., follows a linear trend in both light conditions. The highest values of both dark and bright light conditions are respectively 98% and 100%. The lowest value reads 81% for the dark light condition and 90% for the bright light condition.

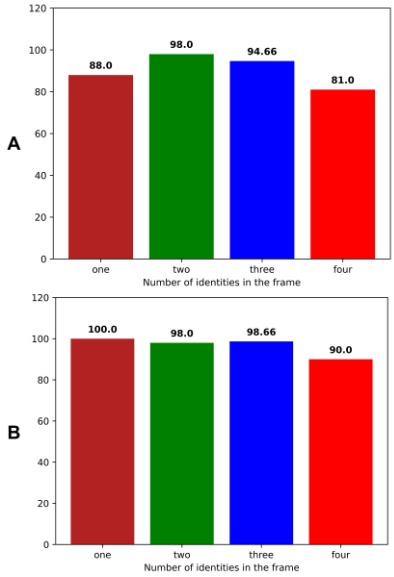


Figure 13 Exact accuracy in dark light condition (left) and bright light condition (right)

4.3 Discussion

Theoretically, the performance of identity recognition should be higher in bright light compared to dark light for face recognition systems. However, the identity recognition performance of the proposed system primarily relies on the lighting condition of the facial area rather than the overall lighting condition. As observed from the test images, the

overall lighting was better in bright light condition compared to dark light condition. However, the lighting was better in the facial area in the dark light condition compared to bright light condition. This situation caused the performance reduction in the bright light condition and the performance was better in dark light condition. Moreover, our test dataset is very challenging and heavily imbalanced as we did not collect the test dataset in a controlled manner. It is actually the aim of this research to use a challenging test set as most of the other studies usually use a test dataset which is collected in a controlled manner. Models usually perform very well in a controlled and clean test dataset but struggle in an uncontrolled and imbalanced dataset. Our test data for "3 people in a frame" and for "4 people in a frame" is heavily imbalanced and contains more unknown identities than known identities. For example, in some images out of four people three people are unknown. Due to these imbalances the system's performance is greatly reduced. For example, in "4 people in a frame and bright light condition" the number of total known identities were 20 and the number of total known identities were 80. The system generated 65 false positives in this case due to which the accuracy, precision and specificity decreased rapidly. On the other hand, as the number of known identities were significantly lower the model generated 10 false negatives. Because of this imbalance the recall is relatively higher than the accuracy, precision, and specificity. However, as our proposed test dataset is small, we believe by observing our proposed system's performance that if we increase the test dataset the fluctuation will be reduced significantly. The other possible factors for performance reduction can be frame size, distance from the camera, low resolution and pose variation. Moreover, the study by [25] also concluded that lower resolution can cause significant accuracy reduction on a FaceNet based system. This finding is also in line with our research.

From the outcomes of age and gender prediction, it can be concluded that both the one-off accuracy for age estimation and the exact accuracy for gender prediction are indifferent towards both the light condition and the number of persons in the frame. Rather it depends on the facial expression and pose angle. As both the models are pre-trained models, they do not depend on the new training dataset. In the original study [4], AgeNet achieved 84.7% one off accuracy and GenderNet achieved 86.8% exact accuracy on Adience dataset. After that, in a study conducted by [58], AgeNet achieved 45.7% exact accuracy and GenderNet achieved 87.32% exact accuracy on UTKFace dataset [63]. On the other hand, on our proposed test dataset, AgeNet achieved an average one-off accuracy of 66.5% for age estimation and GenderNet achieved an average exact accuracy of 93.54% under all conditions. By comparing the results, it can be concluded that our findings are aligned with previous studies. The accuracy can be further improved by aligning the faces before supplying them to the models for estimation.

5.0 CONCLUSION

In this paper, we have proposed a system by utilizing already available deep learning and machine learning models that is able to simultaneously identify and re-identify; estimate the age; and predict the gender of detected known and unknown people. Additionally, we have proposed a multi-face, realistic and challenging test dataset to evaluate the proposed system. On the proposed dataset, the system reported an average accuracy of 49% for identity recognition, 66.5% for age estimation and 93.54% for gender prediction. From the results, it can be concluded that the models (i.e, FaceNet and AgeNet) are not robust and can be sensitive to imbalanced and uncontrolled test dataset. The system fails to correctly identify the faces and estimate the age if the ideal conditions are not met. The FaceNet and the AgeNet are sensitive to facial lighting condition and pose variations. For example, the accuracy for identity recognition can be as low as 25% in some cases with abrupt lighting conditions and occluded pose. The proposed system certainly needs some modification so that it can be implemented is real life crowd monitoring. On the other hand, the GenderNet modelled performed impressively and achieved an average accuracy of 93.54% which is very high compared to the other two models. However, it is to be noted that, our proposed test dataset is very small so the results can vary significantly if new test images are added. We believe that the accuracy for identity recognition and age estimation can be considerably improved by implementing face alignment algorithm and image brightener algorithm. Moreover, the proposed system need be optimized and fine-tuned so that it can work on a CCTV feed in real time.

Acknowledgement

This research was partially supported by IIUM Research Acculturation Grant Scheme (IRAGS) Research Project IRAGS18-017-00. We would like to thank International Islamic University Malaysia for financial support under the KOE Postgraduate Tuition Fee Waiver Scheme 2020 (TFW2020).

References

[1] He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 770-778. DOI: https://doi.org/10.1109/CVPR.2016.90

[2] Schroff, F., Kalenichenko, D., and Philbin, J. 2015. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 815-823. DOI: https://doi.org/10.1109/CVPR.2015.7298682

[3] Jonsson, K., Kittler, K., Li, Y.P., and Matas, J. 2002. Support vector machines for face authentication. *Image and Vision Computing.* 20 (5-6): 369-375. DOI: https://doi.org/10.1016/S0262-8856(02)00009-4

[4] Levi, G., and Hassncer, T. 2015. Age and gender classification using convolutional neural networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 34-42. DOI: https://doi.org/10.1109/CVPRW.2015.7301352

[5] Kelly, M. D. 1973. Visual Identification of People by Computer. Department of Computer Science, Stanford University

[6] Takeo, K. 1973. Picture Processing by Computer Complex and Recognition of Human Faces. Kyoto University, Ph. D. thesis edition, 1973.

[7] Sirovich, L. and Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America A.* 4(3): 519. DOI: https://doi.org/10.1364/JOSAA.4.000519

[8] Price, J.R. and Gee, T.F. 2005. Face recognition using direct, weighted linear discriminant analysis and modular subspaces. *Pattern Recognition.* 38(2): 209-219. DOI: https://doi.org/10.1016/S0031-3203(04)00273-0

[9] He, X., Yan, S., Hu, Y., Niyogi, P. & Zhang, J.H. 2005. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 27(3): 328-340. DOI: https://doi.org/10.1109/TPAMI.2005.55

[10] Chen, d., Cao, X., Wang, L., Wen, F. & Sun, J. 2012. Bayesian face revisited: A joint formulation. *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2012. 7574: 566-579. DOI: https://doi.org/10.1007/978-3-642-33712-3_41

[11] Moghaddam, B., Wahid, W. and Pentland, A. 1998. Beyond eigenfaces: Probabilistic matching for face recognition. *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998.* 30-35.

[12] Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H. and Hua, G. 2016. Labeled faces in the wild: A survey. *Advances in face detection and facial image analysis, Springer.* 189-248. DOI: https://doi.org/10.1007/978-3-319-25958-1_8

[13] Pentland, A., Moghaddam, B. and Starner, T. 1994. View-based and modular eigenspaces for face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 84-91. DOI: https://doi.org/10.1109/CVPR.1994.323814

[14] Wiskott, L., Fellous, J.M., Krüger, N. and Von der Malsburg, C. 1997. Face recognition by elastic bunch graph matching. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 1296(7): 456-463. DOI: https://doi.org/10.1007/3-540-63460-6_150

[15] Albiol, A., Monzo, D., Martin, A., Sastre, J. and Albiol, A. 2008. Face recognition using HOG-EBGM. *Pattern Recognition Letters.* 29(10): 1537-1543. DOI: https://doi.org/10.1016/j.patrec.2008.03.017

[16] Lei, Z., Pietikainen, M. and Li, S. Z. 2014. Learning discriminant face descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 36(2): 289-302. DOI: https://doi.org/10.1109/TPAMI.2013.112

[17] Liu, C. 2004. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 26(5):* 572-581. DOI: https://doi.org/10.1109/TPAMI.2004.1273927

[18] Guillaumin, M., Verbeek, J. and Schmid, C. 2009. Is that you? Metric learning approaches for face identification. *Proceedings of the IEEE International Conference on Computer Vision.* 498-505. DOI: https://doi.org/10.1109/ICCV.2009.5459197

[19] Taigman, Y., Yang, M., Ranzato, M. and Wolf, L. 2014. DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 1701-1708. DOI: https://doi.org/10.1109/CVPR.2014.220

[20] Sun, Y., Wang, X. and Tang, X. 2014. Deep learning face representation from predicting 10,000 classes. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 1891-1898. DOI: https://doi.org/10.1109/CVPR.2014.244

[21] Deng, J., Guo, J., Xue, N. and Zafeiriou, S. 2019. ArcFace: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2019-June: 4685-4694. DOI: https://doi.org/10.1109/CVPR.2019.00482

[22] Wan, W. and Lee, H. J. 2017. FaceNet Based Face Sketch Recognition. *Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI.* 432-436

[23] Jose, E., Greeshma, M., Mithun Haridas, T. P. and Supriya, M. H. 2019, March. Face Recognition based Surveillance System Using FaceNet and MTCNN on Jetson TX2. *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS.* 608-613. DOI: https://doi.org/10.1109/ICACCS.2019.8728466

[24] Nyein, T. and Oo, A.N. 2019, November. University Classroom Attendance System Using FaceNet and Support Vector Machine. *2019 International Conference on Advanced Information Technologies, ICAIT.* 171-176. DOI: https://doi.org/10.1109/AITC.2019.8921316

[25] Golla, M.R. and Sharma, P. 2019. Performance evaluation of facenet on low resolution face images. *Communications in Computer and Information Science.* 839: 317-325. DOI: https://doi.org/10.1007/978-981-13-2372-0_28

[26] William, I., Ignatius, D. R., Moses Setiadi, Rachmawanto, E. H., Santoso, H. A., & Sari, C. A. 2019, October. Face Recognition using FaceNet (Survey, Performance Test, and Comparison). *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019.* DOI: https://doi.org/10.1109/ICIC47613.2019.8985786

[27] Frearson, M. and Nguyen, K. 2020, November. Adversarial Attack on Facial Recognition using Visible Light. DOI: https://doi.org/10.36227/techrxiv.13725634.v1

[28] Zeng, W. et al. 2020, January. Person Recognition Based on FaceNet under Simulated Prosthetic Vision. *Journal of Physics: Conference Series.* 1437(1): 012012. DOI: https://doi.org/10.1088/1742-6596/1437/1/012012

[29] Cahyono, F. Wirawan, W. and Fuad Rachmadi, R. 2020, September. Face recognition system using facenet algorithm for employee presence. *4th International Conference on Vocational Education and Training, ICOVET 2020.* 57-62. DOI: https://doi.org/10.1109/ICOVET50258.2020.9229888

[30] Moura, A. F. S., Pereira, S. S. L., Moreira, M. W. L. and Rodrigues, J. J. P. C. 2020, January. Video Monitoring System using Facial Recognition: A Facenet-based Approach. *IEEE Global Communications Conference, GLOBECOM 2020 – Proceedings.* 1-6. DOI: https://doi.org/10.1109/GLOBECOM42002.2020.9348216

[31] Nair, S. P., Abhinav Reddy, K., Alluri, P. K. and Lalitha, S. 2021. Face recognition and tracking for security surveillance. *Journal of Intelligent and Fuzzy Systems.* 41(5): 5337-5345. DOI: https://doi.org/10.3233/JIFS-189856

[32] Xu, J. 2021, February. A deep learning approach to building an intelligent video surveillance system. *Multimedia Tools and Applications.* 80(4): 5495-5515. DOI: https://doi.org/10.1007/s11042-020-09964-6

[33] Preetha, S., Sheela, S. V. 2021, December. Security Monitoring System Using FaceNet for Wireless Sensor Network

[34] Shirahatti, S. G. C, K. H. S, S and Bangari, S. R. 2021. Face Recognition System For Real Time Applications Using Svm Combined With Facenet And Mtcnn. *International Journal of Electrical Engineering and Technology (IJEET)*. 12(6): 328-335.

[35] Wu, C. and Zhang, Y. 2021, March. MTCNN and FACENET Based Access Control System for Face Detection and Recognition. *Automatic Control and Computer Sciences.* 55(1): 102-112. DOI: https://doi.org/10.3103/S0146411621010090

[36] Adhinata, F. D., Rakhmadani, D. P. and Wijayanto, D. 2021, April. Fatigue Detection on Face Image Using FaceNet Algorithm and K-Nearest Neighbor Classifier. *Journal of Information Systems Engineering and Business Intelligence* 7(1): 22-30. DOI: https://doi.org/10.20473/jisebi.7.1.22-30

[37] Özbulak, G., Aytar, Y. and Ekenel, H. K. 2016. How transferable are CNN-based features for age and gender classification?. *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI).* P-260: 1-6. DOI: https://doi.org/10.1109/BIOSIG.2016.7736925

[38] Kwon, Y. H. and Lobo, N. D. V. 1999. Age classification from facial images. *Computer Vision and Image Understanding.* 74(1): 1-21. https://doi.org/10.1006/cviu.1997.0549

[39] Guo, G. and Mu, G. 2013. Joint estimation of age, gender and ethnicity: CCA vs. PLS. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013.* 1-6. DOI: https://doi.org/10.1109/FG.2013.6553737

[40] Guo, G. and Mu, G., Fu, Y., Dyer, C. and Huang, T. 2009. A study on automatic age estimation using a large database. *Proceedings of the IEEE International Conference on Computer Vision.* 1986-1991.

[41] Han, H., Otto, C., Liu, X. and Jain, A. K. 2015. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 37(6): 1148-1161. DOI: https://doi.org/10.1109/TPAMI.2014.2362759

[42] Eidinger, E., Enbar, R. and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security.* 9(12): 2170-2179. DOI: https://doi.org/10.1109/TIFS.2014.2359646

[43] Chen, J. C., Kumar, A., Ranjan, R., Patel, V. M., Alavi, A. & Chellappa, R. 2016. A cascaded convolutional neural network for age estimation of unconstrained faces. *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems, BTAS 2016.* 1-8. DOI: https://doi.org/10.1109/BTAS.2016.7791154

[44] Huerta, I., Fernández, C., Segura, C., Hernando, J. & Prati, A. 2015. A deep analysis on age estimation. *Pattern Recognition Letters.* 68: 239-249. DOI: https://doi.org/10.1016/j.patrec.2015.06.006

[45] Golomb, B. A., Lawrence, D. T. and Sejnowski, T. J. 1991. Sexnet: A neural network identifies sex from human faces. *Advances in Neural Information Processing Systems 3.* 1(July): 572-577

[46] Shan, C. 2012. Learning local binary patterns for gender classification on real-world face images. *Pattern Recognition Letters.* 33(4): 431-437. DOI: https://doi.org/10.1016/j.patrec.2011.05.016

[47] Van De Wolfshaar, J., Karaaba, M. F. and Wiering, M. A. 2015. Deep convolutional neural networks and support vector machines for gender recognition. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015.* 188-195. DOI: https://doi.org/10.1109/SSCI.2015.37

[48] Yi, D., Lei, Z. and Li, S. Z. 2015. Age estimation by multi-scale convolutional network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).* 9005: 144-158. DOI: https://doi.org/10.1007/978-3-319-16811-1_10

[49] Rothe, R., Timofte, R. and Van Gool, L. 2018. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision.* 126: 144-157. DOI: https://doi.org/10.1007/s11263-016-0940-3

[50] Mansanet, J., Albiol, A. and Paredes, R. 2016. Local Deep Neural Networks for gender recognition. *Pattern Recognition Letters.* 70: 80-86. DOI: https://doi.org/10.1016/j.patrec.2015.11.015

[51] Agarwal, T., Andhale, M., Khule, A. and Borse, R. 2021. Age and Gender Classification Based on Deep Learning. *Techno-Societal 2020.* 425-437. DOI: https://doi.org/10.1007/978-3-030-69921-5_43

[52] Adhinata, F. D. and Junaidi, A. 2022, January. Gender Classification on Video Using FaceNet Algorithm and Supervised Machine Learning. *International Journal of Computing and Digital Systems.* 11(1): 199-208. DOI: https://doi.org/10.12785/ijcds/110116

[53] Agbo-Ajala, O. and Viriri, S. 2020. Deeply Learned Classifiers for Age and Gender Predictions of Unfiltered Faces. *Scientific World Journal.* DOI: https://doi.org/10.1155/2020/1289408

[54] Rouhsedaghat, M., Wang, Y., Ge, X., Hu, S., You, S. and Kuo, C. C. J. 2021, January. FaceHop: A Light-Weight Low-Resolution Face Gender Classification Method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics). 12668: 169-183. DOI: https://doi.org/10.1007/978-3-030-68793-9_12

[55] Garain, A., Ray, B., Singh, P. K., Ahmadian, A., Senu, N. and Sarkar, R. 2021. GRA_Net: A Deep Learning Model for Classification of Age and Gender from Facial Images. *IEEE Access.* 9: 85672-85689. DOI: https://doi.org/10.1109/ACCESS.2021.3085971

[56] Kharchevnikova A. S. and Savchenko, A. V. 2018, October. Neural Networks in Video-Based Age and Gender Recognition on Mobile Platforms. *Optical Memory and Neural Networks (Information Optics).* 27(4): 246-259. DOI: https://doi.org/10.3103/S1060992X18040021

[57] Duan, M., Li, K., Ouyang, A., Win, K. N., Li, K. & Tian, Q. 2020, May. EGroupNet: A Feature-enhanced Network for Age Estimation with Novel Age Group Schemes. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM).* 16(2). DOI: https://doi.org/10.1145/3379449

[58] Savchenko, A. V. 2019. Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output ConvNet. *PeerJ Computer Science.* 2019(6). DOI: https://doi.org/10.7717/peerj-cs.197

[59] Hosseini, S., Lee, S. H., Kwon,H. J., Il Koo, H. and Cho, N. I. 2018, May. Age and gender classification using wide convolutional neural network and Gabor filter. *2018 International Workshop on Advanced Image Technology, IWAIT 2018.* 1-3. DOI: https://doi.org/10.1109/IWAIT.2018.8369721

[60] Khan, K., Attique, M., Syed, I., Sarwar, G., Irfan, M. A. and Khan, R. U. 2019, June. A Unified Framework for Head Pose, Age and Gender Classification through End-to-End Face Segmentation. *Entropy 2019.* 21(7): 647. DOI: https://doi.org/10.3390/e21070647

[61] Duan, M., Li, K., Yang, C. and Li, K. 2018. A hybrid deep learning CNN-ELM for age and gender classification. *Neurocomputing.* 275: 448-461. DOI: https://doi.org/10.1016/j.neucom.2017.08.062

[62] Benkaddour, M. K. 2021, August. CNN Based Features Extraction for Age Estimation and Gender Classification. *Informatica.* 45(5): 697-703. DOI: https://doi.org/10.31449/inf.v45i5.3262

[63] Zhang, Z., Song, Y. and Qi, H. 2017, November. Age progression/regression by conditional adversarial autoencoder. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.* 4352-4360. DOI: https://doi.org/10.1109/CVPR.2017.463