IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 August 2023, accepted 1 September 2023, date of publication 6 September 2023, date of current version 12 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3312612

RESEARCH ARTICLE

A Study on the Effect of Ageing in Facial Authentication and the Utility of Data Augmentation to Reduce Performance Bias Across Age Groups

WANG YAO[1], (Graduate Student Member, IEEE), MUHAMMAD ALI FAROOQ[1], JOSEPH LEMLEY[1,2], (Member, IEEE), AND PETER CORCORAN[1], (Fellow, IEEE)

[1]College of Science and Engineering, University of Galway, Galway, H91 TK33 Ireland
[2]Xperi Corporation, Galway, H91 V0TX Ireland

Corresponding author: Wang Yao (w.yao2@universityofgalway.ie)

This work was supported in part by the Irish Research Council Enterprise Partnership Ph.D. Scheme under Grant EPSPG/2020/40, in part by the Data-Center Audio/Visual Intelligence On-Device (DAVID) Project (2020–2023) funded by the Disruptive Technologies Innovation Fund (DTIF), and in part by Xperi Corporation, Ireland.

ABSTRACT This work presents a study on the effects of aging on the performance and reliability of facial authentication methods. First, a brief review of the literature on the effect of age on face recognition algorithms is presented, followed by a detailed description of the face aging datasets. In contrast with some recent studies, we demonstrate significant variations in authentication robustness between age groups. The second part of this paper focuses on a comprehensive comparative assessment on the effects across age groups. Four different face recognition algorithms are studied of which three are state-of-the-art neural network based models and the fourth one is a conventional machine learning model. Two different age range threshold settings (±3 in Experiment Category A and ±5 in Experiment Category B) of the age groups are adopted in the experimental analysis to get a proper comparison. Moreover, a synthetic aging method has been incorporated to augment the age data. Experimental result shows that the older adults groups are easier to identify with higher levels of accuracy and robustness compared to other age groups, while younger adults are the most challenging and false authentications are more likely to occur.

INDEX TERMS Age effect, data augmentation, face recognition, FR evaluation.

I. INTRODUCTION

Automatic face recognition (FR) algorithms have been studied extensively in the literature and have become widely used in recent years [2], [3]. They are employed in many critical real-world applications including security systems and immigration controls. And in recent years most smartphones and some other consumer devices have adopted facial biometrics to provide access control in place of a conventional password. Facial authentication technology has become increasingly robust and is now at a point where it can be embedded into the hardware of even simpler consumer devices such as doorbell cameras to address privacy issues linked with personal biometrics. However, before such broad adoption can occur, this raises new issues with respect to the longer-term reliability and robustness of the technology and the potential for inbuilt bias due to the training of neural models on unbalanced datasets.

Recent research has shown that existing face recognition algorithms have performance biases due to various factors. These include age [4], gender [5], and ethnicity [6] which limit their robustness in applications that are more focused towards biometric security authentications. The potential factors leading to these biases are attributed to weaknesses in the training datasets and the resulting FR algorithms. In order to achieve the best possible performance, facial recognition

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

algorithms for a particular application or use case must be trained on a large-scale and balanced dataset. However, in the context of this research work, the existing large-scale publicly available datasets such as VGGFace2 [7], MS-celeb-1M [8], and CASIA [9] lack sufficient and balanced data regarding age variations of individual subjects. This can lead to a trained neural model which is insensitive to the age-linked characteristics of the subjects in the training dataset. In simple terms, younger age groups, especially children and teens, have less well-developed facial features compared with middle-aged or older adults.

Face recognition across the full span of human aging is a unique challenge, as age is a fundamental biological characteristic it is important to be able to adapt facial recognition and authentication algorithms to take into account different age demographics. Previous studies indicate that the aging effect is still an unsolved problem in this field [10], [11]. As a matter of fact, human facial characteristics change significantly due to the aging factor including hair, wrinkles, and other facial attributes.

Ideally, to design a biometric identification system that should remain effective for long-term usage, it is necessary to extract persistent features of the human face that can be further utilized for training advanced machine learning algorithms thus enabling to achieve of robust recognition. However, most of the large-scale facial datasets are based on celebrity persons who are mainly middle-aged. As a result, most of the publicly available datasets lack teens, younger, and elder age groups, especially children and older people. This implies that we need to adapt different models for teenagers, young adults, mature adults, and older adults; it also means that different models need to be trained for different demographics. But first, we have to understand plausible age effects. Therefore, it is essential to conduct an investigation of age group bias in state-of-the-art (SOTA) face authentication (FA)/face recognition (FR) methods.

Sawant and Bhurchandi [4] published a survey about the age effect as it is one of the essential techniques for age-invariant face recognition. Similarly, Abdurrahim et al. [12] presented age, gender, and race as essential factors affecting the effectiveness of automatic face recognition. Almost all these surveys concentrate on different factors that affect FR models and describe the age effect as one part of these factors. Thus they were not intended to do a comprehensive study on age bias. In addition, these studies are based on conventional machine learning based FR algorithms. Different from these earlier studies, our work mainly focuses on a study of the effect of face aging data on the performance of FR systems, including the effect of the age gap and the age group. Specifically, we provide a comprehensive study including the effect of face aging on FR algorithms and popular face aging datasets that are always used in age studies. In addition, we conduct a series of experiments about the effect of face-aging groups on the performance of various FR systems.

The main contributions of this research are as follows.

- This work conducts a comprehensive survey of the literature review that primarily focuses on the age effect on FR algorithms.
- We conduct experiments to determine the variations in performance/robustness/accuracy of SOTA FR models across different age groups and elaborate on reasons for inconsistent findings in previous studies whether because of the FR algorithms.
- Experiments associated with synthetic age data have shown that it is valid to use synthetic data as a way to augment the real-world age dataset and to investigate how aging affects FR models.
- Experiments results confirm that younger adults are more difficult to recognize than older adults even for neural facial FR algorithms.

The rest of the paper is organized as follows. Section II presents a review of the effect of face aging on the performance of face recognition models. Then the literature survey about datasets is presented in Section III. Subsequently, Section IV, Section V, and Section VI conduct experiments to evaluate the effect of state-of-the-art face recognition models on age groups. Specifically, the methodologies are introduced in Section IV. Section V presents the results of the experiment on original age data. Section VI extends the experiments on synthetic age data. Finally, we conclude the paper in Section VII and provide potential future research.

II. FACE RECOGNITION PERFORMANCE ACROSS AGE

There are two different approaches in the literature to investigate the effect of subject age on FR performances, that is, age gap (interval) and age group.

The age gap approach entails registering a subject at a particular age and subsequently evaluating the registered facial image against those recorded at different ages, thereby assessing the variability of the performance of the facial recognition model across individuals of varying ages. The age group method, on the other hand, involves quantifying the accuracy of facial recognition for distinct age groups, with the aim of determining the performance of the FR model within these specific age groups.

A. AGE GAP

A number of face aging studies in the literature have indicated that large age interval lengths cause significant degradation in face recognition accuracy. This large interval length is generally measured in years, rather than months or days, to produce substantial deterioration [13], [14]. The same studies have also shown that different feature extractors have various levels of sensitivity to aging intervals.

Ling et al. [15] were the first to propose a study on the effect of the age gap on face recognition performance, and their study focused on the task of passport renewal and passport verification. Their approach used traditional feature extraction methods to extract face features from real passport photos. Their experimental results conclude that the aging process increases the difficulty of the face recognition task and typically remains stable over a period of 4 to 10 years. Another study from them [1] shows that the additional difficulty of facial verification tests on age gaps becomes saturated after the age gap is larger than four years and this phenomenon remained valid for up to ten years.

Guo et al. [13] combined principal component analysis and elastic bunch graph matching to analyze age intervals in a very large Morph II database with 13,160 subjects of 54,675 face images and showed that recognition accuracy decreases with increasing age. They concluded that the facial recognition process is not linear with respect to aging and, in particular, they noted that the accuracy decreases dramatically when the age gap is greater than 15 years. Similarly, Deb et al. [16] studied two longitudinal mugshot datasets, PCSO and MSP from two different law enforcement sources, where the PCSO dataset contains 147,784 mugshots of 18,007 spanning 16 years and the MSP dataset contains 82,450 mugshots of 9,572 recidivists spanning for 13 years. The experiment results showed that age differences greater than 8.5 years resulted in a significant decrease in face recognition accuracy. Moschoglou et al. [46] tested face verification in AgeDB using the Centre Loss and Marginal Loss methods. The results show a significant decrease in accuracy over an age gap from 5 to 30 years.

Meng et al. [10] showed that robust features can be extracted using the Gabor wavelets feature, and features extracted by local binary pattern can achieve better recognition accuracy at large age intervals than features extracted by gradient orientation pyramid. Experiments by Bereta et al. [17] verified that Gabor wavelets and multi-scale block local binary pattern local descriptors provided the best recognition accuracy in the context of age. Similarly, Boussaad and Boucetta [18] examine the effects of age on face recognition performance across FGNET, using three approaches for comparison. A recent study by Negri et al. [54] provides a fine-grained analysis of the aging effect on different age intervals and shows that probabilistic linear discriminant analysis (PLDA) based approaches and a nonlinear version of pairwise support vector machine (NL-PSVM) could reduce the age sensitivity of facial features.

Furthermore, some interesting findings from other researchers related to this field are as follows. Ling et al. [1] experiment shows that individual age differences in the test and training sets can cause degradation in facial recognition performance. Otto et al. [19] discuss the effect of face aging on individual facial components including mouth, eyes, and nose, and they find that the nose is the most stable component during the face aging process. Their analysis shows that as individuals age, the upper facial region becomes more important for recognition performance than the lower facial region. Klare and Jain [20] explore whether a FR model that compensates for the aging effect compromises the performance of FR that has not undergone any aging. They have trained the models on datasets with age diversity and shown a decrease in performance in non-aging scenarios. Additionally, some other studies have investigated the effect of age intersecting with other factors on face recognition performance, such as expressions [21], gender [14], and race [22].

The FR algorithms used in most of these studies are based on pre-dates neural statistical machine learning. There is an assumption that neural networks are better than older methods when recognizing faces, but we can still learn about the potential weaknesses of neural networks from previous feature-based research efforts. Face recognition performance degrades when the difference between a pair of faces is more than a few years. The effects of other demographic factors overlaid with age should also be considered when analyzing performance.

B. AGE GROUP

Intuitively as people age their facial features evolve and become more defined; thus children can often look quite similar to humans, but as people grow into teens and mature adults it becomes easier to distinguish between them. However, some literature show that the outcomes of the impact of age group on face recognition performance are not consistent regarding which age group is more easily recognizable. In general, older methods of FR relied mainly on statistical machine learning and explicit feature extraction as opposed to DL where the model learns features in an ‘unconstrained’ way to achieve an end goal. These early studies have indicated that recognizing young individuals is a more arduous task.

Almost all the results for the face recognition vendor test (FRVT) 2002 illustrate that older people are easier to recognize than younger people [23]. Givens et al. [26] conducted three FR algorithms including principal component analysis (PCA), an interpersonal image difference classifier, and an elastic bunch graph matching algorithm on the FERET dataset and verified that older individuals are easier to recognize compared with younger ones. Beveridge et al. [27] studied 351 subjects from FRGC, indicating that males, older people, and subjects without glasses are easier to recognize under the aging effect. Similarly, another study by Akhtar et al. [22] getting consistent results that older subjects and males are easier to recognize. Boussaad and Boucetta [18] showed that the older group (>40 years) had the highest accuracy rate. Lui et al. [14] summarized existing results, indicating that while the magnitudes of the age effect are different for various algorithms, in 20 out of 22 results, older people are easier to recognize than younger people. Klare et al. [28] measure the performances of three commercial face recognition algorithms on three age groups and find that younger subjects (18 to 30 years old) are more difficult to recognize.

Recent studies that employ deep learning methods have yielded some different conclusions about which groups are more difficult to identify. In one study by Albiero et al. [29] it was shown that the FR results for older people (in the age bracket of 50 to 70 years old) using deep learning based face recognition models were less accurate than the results achieved on younger age groups (16 to 29 years old). Wu and Wang [30] present experimental results showing that middle-aged men are more difficult to identify than the youth and the elderly.

This inconsistency in recent studies on age groups using Neural Network (NN) based approaches is a key motivation to under this current study. NN based techniques have come to dominate computer vision research and have recently established themselves as gold standards [29], [37], [53] for FR algorithms. One of the rationales used for adopting NN models is that they can often find feature patterns that are too obscure to be identified by most humans and thus, arguably, can transcend human perception. Is it possible that NN models can provide improved performance for younger subjects where older FR algorithms have been challenged? Are the reasons for this inconsistency due to features extracted by earlier algorithms and deep learning methods? Or are they arising from the way age groups are divided?

In order to answer the above questions, this paper reclassifies different age groups, we have chosen more granular age groups than previous literature [18], [29] and divided them into seven age groups which include Group 1 (around 20 years old), Group 2 (around 30 years old), Group 3 (around 40 years old), Group 4 (around 50 years old), Group 5 (around 60 years old), Group 6 (around 70 years old), and Group 7 (around 80 years old). The 'around' here means different age range thresholds at ±3 years and ±5 years have been verified in our experiment. Multiple face recognition classifiers including ArcFace [31], CosFace [32], SphereFace2 [34] and LBP-based [36] are used to answer the question of whether different feature extractors have different recognition abilities for different age groups as detailed in Section V. In addition, one augmentation technique with synthetic aging to increase the amount of different age data is adopted in Section VI. This helps us to understand whether unbalanced age data in different age groups affect the performance of FR algorithms.

III. FACE AGING DATASETS

Quality databases are essential for the training and testing of face recognition systems. There are numerous face datasets accessible, but very few of them are created specifically to solve the aging issue. Building an aging database entails gathering multiple age-separated face photos of the same person, ideally in controlled (laboratory) conditions. This is a time-consuming and laborious task that should be executed over many years, ideally several decades. Although there are many existing large-scale face datasets, finding the appropriate datasets which can be used to solve the age problem is still a challenging task. This is because we have a limited amount of age-related data in these datasets.

There are several criteria that should be met when investigating datasets used in the process of facial aging. Firstly, there should be a sufficiently large number of subjects and face images across all age groups, from youngest (c.20 years) to oldest (c.80 years) age groups. Secondly, a balanced distribution of each age group should be available. Typically, there are fewer older data subjects and where datasets are drawn from public figures and celebrities there tend to be more subjects and samples in the 30 – 50 year age groups as successful public figures mostly emerge in these age groups. Third, facial images of a consistent minimum quality are available for each subject within a certain age range. For example, several images of a particular subject in his or her 20s-30s are needed to analyze the features of that subject. This section introduces and discusses face aging datasets which are often used in age-related research studies.

A. FG-NET

The FG-NET aging dataset [38] contains 1,002 images from 82 different subjects, with ages ranging from birth to 69 years. The dataset was collected mainly by scanning photographs of the subjects. The images in the database have considerable variation in resolution and image quality resulting from the resolution of the camera used to take the photographs and the resolution of the scanner. In addition, this dataset provides descriptions of the 68 facial landmark points for each image. FG-NET is counted as one of the most popular datasets in age-related studies. The major limitation of FG-NET is less number of subjects and research indicates that the accuracy of this dataset is nearly saturated [39], thus recent studies rarely use the FG-NET dataset as a benchmark dataset.

B. PCSO

The Pinellas County Sheriff's Office (PCSO) longitudinal dataset contains 147,784 images, collected from 18,007 criminals spanning from 1994 to 2010. This dataset provides metadata with a capture date of each image along with the date of birth for each subject. Each subject has at least five photographs of their face over a period of a minimum of five years. PCSO is significantly large in length and breadth, which made it a popular dataset for the research community on age-related studies such as age-invariant face recognition, however, this dataset is no longer publicly available.

C. MORPH

The Morph dataset [40] is one of the largest publicly available longitudinal face datasets collected by the Face Aging Group at the University of North Carolina. The dataset is divided into two parts: the commercial and the non-commercial version. The non-commercial release (Morph II) has a total of 55,134 images with 13,618 identities taken over 5 years. It contains people aged from 16 to 77 years old, with an average of four images per person. The Morph II dataset also records other information of each subject which includes gender, race, and whether the subject is wearing glasses. Literature [29], [41] shows that the Morph II dataset has some mislabelled data and given the filtering approach which could be used for our experimental works.

D. CACD

The Cross-Age Celebrity Dataset (CACD) contains 163,446 images of 2,000 subjects collected by Chen et al. [42]

It consists of celebrity images acquired in the time frame of 10 years from 2004 to 2013. The dataset is collected from Google Images using celebrity names and years as the keywords. Each image has metadata with the name, age, identity, year of birth, and whether it exists in the LFW dataset [43]. This is one of the publicly available large-scale facial aging datasets. However, there are many images that are mislabelled, have multiple faces in the image, or have no faces in the image, because all the images in CACD were collected directly from the Internet. This dataset can be beneficially used for training purposes. CACD-VS is a subset of the CACD dataset that has a more accurate label. CACD-VS contains 4,000 pairs of images with 10-fold which can be beneficial for face verification tasks.

E. UTKFace

The UTKFace dataset [44] is a large-scale face dataset with a long age span, which contains 20,000 face images in the wild. Each image has information about age, gender, ethnicity, and corresponding landmarks. This dataset includes large variations in pose, facial expression, illumination, occlusion, resolution, etc. This dataset can be used for variety of computer vision tasks, such as face detection, age estimation, age progression/regression, and landmark localization. However, this dataset does not provide information about the subjects, thus it is not applicable to explore age effects.

F. ADIENCE

The Adience dataset [39] contains 26,580 images of 2,284 subjects in the wild taken by mobiles. Each image has a subject label, gender label, and age label from eight age groups. This dataset captures all the variations such as appearance, noise, pose, and lighting. It is used as a benchmark dataset for face photos and for age and gender recognition studies.

G. IMDB-WIKI

The IMDB-WIKI dataset [45] contains 523,051 images from 20,284 subjects obtained from the IMDB and Wikipedia which is publicly available in 2015. This dataset provides labels about the date of birth, taken date, gender, face location, face score, second face score, celeb name, and celeb id. The approximate age can be obtained by calculating the taken date and the date of birth. Since the images in IMDB-WIKI are collected from the Internet, some information from labels is not accurate and this dataset should be cleaned first before use. The distribution of age groups in this dataset is unbalanced in the young and old age groups. And this dataset is a large dataset of celebrities which is always used as the training dataset.

H. AgeDB

AgeDB [46] was public in 2017 and is the first manually collected ''in-the-wild'' age dataset, which contains 16,488 images of various celebrities such as actors, writers, scientists, and politicians. Each image is annotated with identity, age, and gender attributes. There are 568 different subjects in total. The average number of images per subject is 29. And the average age range for each subject is 50.3 years. The minimum and maximum ages in the dataset are 1 and 101, respectively.

I. CAF

CAF dataset [55] is a large-scale cross-age face dataset with a large number of cross-age celebrities' faces. It is used for age-invariant face recognition. CAF dataset comprises 313k images in total. These face images are downloaded from several commercial image search engines including Google and Baidu. The authors have employed a public pre-trained age estimation model DEX [56] to obtain a rough age label for each image. CAF has minimized the noise data compared to IMDB-WIKI.

J. CAFR

The Cross-Age Face Recognition (CAFR) dataset [57] consists of 1,446,500 images annotated with age, identity, gender, race, and landmarks. CAFR is collected from real-world scenarios. The images in CAFR have various expressions, poses, occlusion, and resolutions. These images are annotated by using an off-the-shelf age estimator [56] and landmark localization method [58]. Then these annotations are manually rectified by professional data annotators. CAFR is one of the benchmark datasets for age-invariant face recognition.

K. LCAF

The large cross-age face (LCAF) dataset [49], [59] has 1.7M faces from cross-age celebrities. Huang et al. collected LCAF from MS-Celeb-1M dataset and used the public Azure Facial API to label the ages and genders of the images. They also build a subset of the cross-age face (SCAF) dataset, which contains 0.5M images with 12k individuals. These datasets could be used for training age-invariant face recognition and face age synthesis networks.

L. FFHQ-AGING

The FFHQ-Aging dataset [47] contains 70,000 original FFHQ images which is an extension of the NVIDIA FFHQ dataset [48]. It is designed for benchmarking age transformation algorithms. Each image is annotated with gender, age group, head pose, glasses, eye occlusion score, and full semantic map. These labels are acquired by different software platforms. As FFHQ has multiple resolutions, up to 1024 × 1024. This makes this dataset popular among age progression/regression tasks recently.

Table 1 shows the detailed numeric comparison of each of these face-aging datasets. In addition to the above datasets, there are some datasets dedicated to the study of children such as ITWCC [60], CLF [62], YFA [63], and ICD [61], which are mostly privately available. Among these aging databases, FG-NET, Morph II, and AgeDB are used in most common studies [17], [18], [29], [49] about aging experiments. We choose Morph II and AgeDB as the baseline

TABLE 1. Comparison of facial aging datasets.

Dataset	Subjects	Images	Images per Subject	Age range	Precision	Resolution	Year
FG-NET	82	1,002	6-18, Average-12	0-69	Accurate ages	349x384-522x577	2002
PCSO	18,007	147,784	Average-8	18-83	Accurate ages	N/A	N/A
Morph II	20,569	78,207	1-53, Average-4	15-77	Accurate ages	200x240,400x480	2006
Adience	2,284	26,580	Average-12	N/A	Age groups	816x816 (aligned)	2014
CACD	2,000	163,446	Average-82	16-62	Accurate ages	250x250	2015
IMDB-WIKI	20,284	523,051	Average-26	0-100	Accurate ages	N/A	2016
UTKFace	N/A	23,708	N/A	0-116	Accurate ages	200x200	2017
AgeDB	568	16,458	2-44, Average-29	1-101	Accurate ages	49x49-2050x2050	2017
CAF	4,668	313,986	Average-67	1-80	N/A	N/A	2018
FFHQ-Aging	N/A	70,000	N/A	N/A	Age groups	1024x1024	2020
CAFR	25,000	1,446,500	Average-58	1-99	Age groups	N/A	2022
SCAF	12,000	508,705	Average-42	N/A	Accurate ages	112x112	2023

datasets in this work as they have more subjects and images compared to the FGNET dataset. We have done two sets of experiments in this work. Firstly, we classify the dataset into different age groups and evaluate the performance of different age groups on various face recognition algorithms. Secondly, we introduce a synthetic age method to augment the real-world age datasets and evaluate the performance of synthetic age data on various face recognition algorithms.

IV. METHODOLOGY AND EVALUATION METRICS

There are some inconsistent results on the aging effect according to the age group in Section II. This discrepancy brings us to the following questions. How short-term age group affects non-deep learning and deep learning facial authentication models? Is it the reason that leads to differences in earlier research due to the different algorithms? Or is it due to the dataset? In this section, we introduce the adapted experiment methodology and evaluation metrics which could be used for answering the questions.

A. DATA AUGMENTATION WITH SYNTHETIC AGE SAMPLES

As discussed in Section III, it is difficult to find a large-scale real dataset that satisfied diversified aging samples including balance numbers of samples for individuals in different age periods due to the limitations of the existing aging dataset. Recently, researchers have mainly focused on generating photo-realistic face aging data [25], [49]. Inspired by these promising results, we plan to introduce the recent state-of-the-art synthetic age and aging method in our study to build an ideal synthetic aging dataset.

In this work, synthetic aging data is used to extend the real-world aging experiment in order to overcome the barrier of the imbalance distribution of data among different age groups in the existing datasets. As this is an initial work with synthetic aging effects, the SAM technique [25] is applied to synthesize photo-realistic aging data. The method achieves the state-of-the-art age transformation task from a single facial image by mapping the input image and the target age to the latent space of the StyleGAN generator.

B. FACE RECOGNITION ALGORITHMS

In this study, three state-of-the-art deep learning face recognition algorithms and one feature extraction-based face recognition algorithm are evaluated in our experimental work. This is done in order to answer whether the performance of FR algorithms on face images from different age groups is consistent with previous studies about the impact of age groups on FR performance.

1) ArcFace

ArcFace [31] was initially publicly available in 2018 as a 2D face recognition model. It has now been extended for tasks such as 3D face recognition, and the functionality is continuously maintained and updated, making it the most popular face recognition model within the research community. The public reference implementation of the ArcFace and pretrained model.[1] is used in this work, which provides a useful public baseline for performance comparisons [37].

2) CosFace

CosFace [32] is inspired by the initial SphereFace [33] and uses multi-class classification training and additive margin. It does not provide an official reference implementation. We used a public implementation framework[2] and trained on the VGGFace2 dataset [7] for 70,000 epochs with large margin cosine loss.

3) SphereFace

SphereFace2 [34] is recently published in 2022 and is considered as the latest generation of SphereFace [33]. It uses binary classification training and is robust to label noise. In this work, the pre-trained model 2 has been used and is trained on the VGGFace2 dataset [7]. Our experiments would be reproduced by the same model.

4) LBP

The Local Binary Pattern (LBP) based face recognition method was introduced by Ahonen et al. [36]. This method employs LBP histograms to extract face features, then recognition is performed using the nearest neighbor classifier in feature space with Chi-square as a dissimilarity measure.

[1] https://www.dropbox.com/s/ou8v3c307vyzawc/model-r50-arcface-ms1m-refine-v1.zip?dl=0

[2] https://github.com/ydwen/opensphere

TABLE 2. Evaluation metrics. Here TP means True Positive. TN means True Negative. FN means False Negative. FP means False Positive. P means the total number of Positive cases and N means the total number of Negative cases.

Measure	Description
Accuracy	$Accuracy = \frac{TP+TN}{P+N}$
Sensitivity	$Sensitivity = \frac{TP}{TP+FN} = \frac{TP}{P}$
Precision	$Precision = \frac{TP}{TP+FP}$
F1 score	$F1score = \frac{2TP}{2TP+FP+FN}$
Receiver Operating Characteristic (ROC)	The Receiver Operating Characteristic (ROC) curve [51] illustrates the performance of the classifier, which is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The higher the TPR and the smaller the FPR, the better the classifier effect. $FPR = \frac{FP}{FP+TN} = \frac{FP}{N}, \quad TPR = \frac{TP}{TP+FN} = \frac{TP}{P}$
Match Score Comparison	The genuine matching score is the match score from the same identity, and the imposter matching score is the match score from different identities. The smaller the overlap between the two distribution curves and the area enclosed by the axes, the better the differentiation performance of the features.

The features of LBP were commonly utilized in early-stage age studies. We have adopted LBP based FR method in this work as an earlier non-deep learning method, which is one of the commonly used methods for analyzing the aging effect [17], [20], [22]. In this work, we employ non-trainable LBP to extract face features and use Chi-square distance compare features.

C. EVALUATION METRICS

Several quantitative metrics shown in Table 2 have been employed to evaluate the age effect on face recognition methods. These metrics are widely used to evaluate the performance of neural networks [50].

V. EXPERIMENTS

In this section, our aim is to answer the question that how aging data impact the performance of face recognition algorithms and which age group has a larger effect on the deep learning based face authentication and conventional machine learning based face recognition method.

A. EXPERIMENT SETTING

This section presents the complete experimental work carried out in this study. We evaluate different aging groups from two different datasets which include AgeDB and Morph II using mentioned face recognition methods in section IV-B. We use two different age range thresholds to divide our age group in order to explore how small age range groups affect the FR results.

First, we have organized the images into 7 different age groups (20/30/40/50/60/70/80) with age range threshold setting of ±3 such that 17 ∼ 23, 27∼33, 37∼43, 47∼53, 57∼63, 67∼73, 77∼83. Second, in cases where the existing datasets do not have enough images for this small age range threshold, we introduce a larger age range threshold setting ±5 for each age group, which includes the age groups range from 15∼25, 25∼35, 35∼45, 45∼55, 55∼65, 65∼75, 75∼85. To evaluate the performance of face recognition models, an equal number of positive-identity pairs (PPs) and negative-identity pairs (NPs) are created. A positive-identity pair means that two images from one identity and a negative-identity pair means that two images from different identities. Table 3 shows the number of PPs / NPs and subjects for each age group from AgeDB and Morph II that is used in our experimental work. Figure 1 shows the example of two subjects in different age groups. We kindly mention that if there is only one image for a subject, this image will be deleted before PPs / NPs are created during the entire experiment.

B. ROC CURVE COMPARISON

The ROC curves for different age groups from AgeDB and Morph II datasets tested on ArcFace, CosFace, SphereFace2, and LBP are presented in Figure 2. Table 4 demonstrates the analysis of the ROC curves in Figure 2.

The analysis in Table 4 shows that the ROC curves performed consistently on the same age group with two age ranges on the same dataset using deep learning based face recognition models, although the accuracy of the different face recognition algorithms on the ROC curves is different. Besides, old people who are in Group 6 and 7 have a high performance on the deep learning based FR models, and young people who are in Group 1 and 2 have a lower performance on the deep learning based FR model. This result indicates that face recognition at different ages continues to have an impact on FR. The existing face recognition algorithms do not solve the age-invariant face recognition problem.

In addition, there are some minor different results for AgeDB and Morph II datasets from Table 4. The latest deep face recognition systems such as ArcFace, CosFace, and SphereFace2 are not effective in dealing with large age variations like the AgeDB dataset. For results test on Morph II, the high performance of these deep learning based ROC curves indicates that the age differences in the Morph II dataset have a low impact on the current stage of deep FR models. Also compared to the results from AgeDB and Morph II for the LBP algorithm, we could obtain that older people have high performance on the LBP algorithm, middle age people have the lowest performance in ROC curves on the LBP algorithm.

C. MATCH SCORE COMPARISON

The match score distribution results are depicted in Figure 3 which support the ROC curves in Figure 2. We can observe

TABLE 3. Division of age groups for AgeDB and Morph II dataset.

	DataSet	Age Group	1	2	3	4	5	6	7
Experiment Category A (EC-A)		Age range	[17,23]	[27,33]	[37,43]	[47,53]	[57,63]	[67,73]	[77,83]
	AgeDB	Number of pairs	1,123	3,020	2,882	2,231	3,112	1,105	465
		Subjects	168	335	357	343	279	203	86
	Morph II	Number of pairs	24,980	16,263	23,154	12,567	1,049	/	/
		Subjects	3,227	2,271	2,505	1,103	157	/	/
Experiment Category B (EC-B)		Age range	(15,25]	(25,35]	(35,45]	(45,55]	(55,65]	(65,75]	(75,85]
	AgeDB	Number of pairs	2,968	8,267	7,773	6,012	4,280	2,874	985
		Subjects	230	377	394	378	332	256	123
	Morph	Number of pairs	36,815	29,490	35,265	18,354	1,713	/	/
		Subjects	4,492	3,587	3,602	1,588	254	/	/

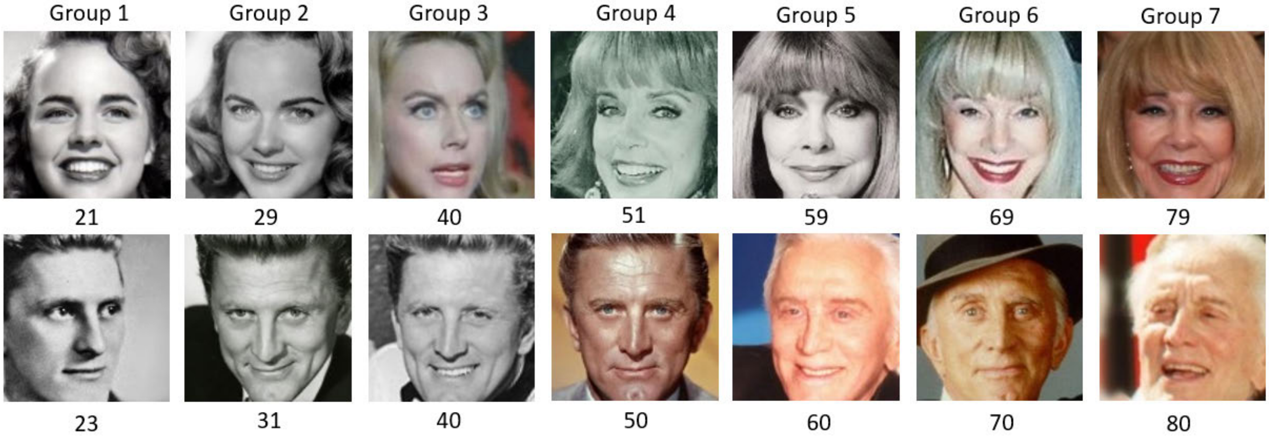


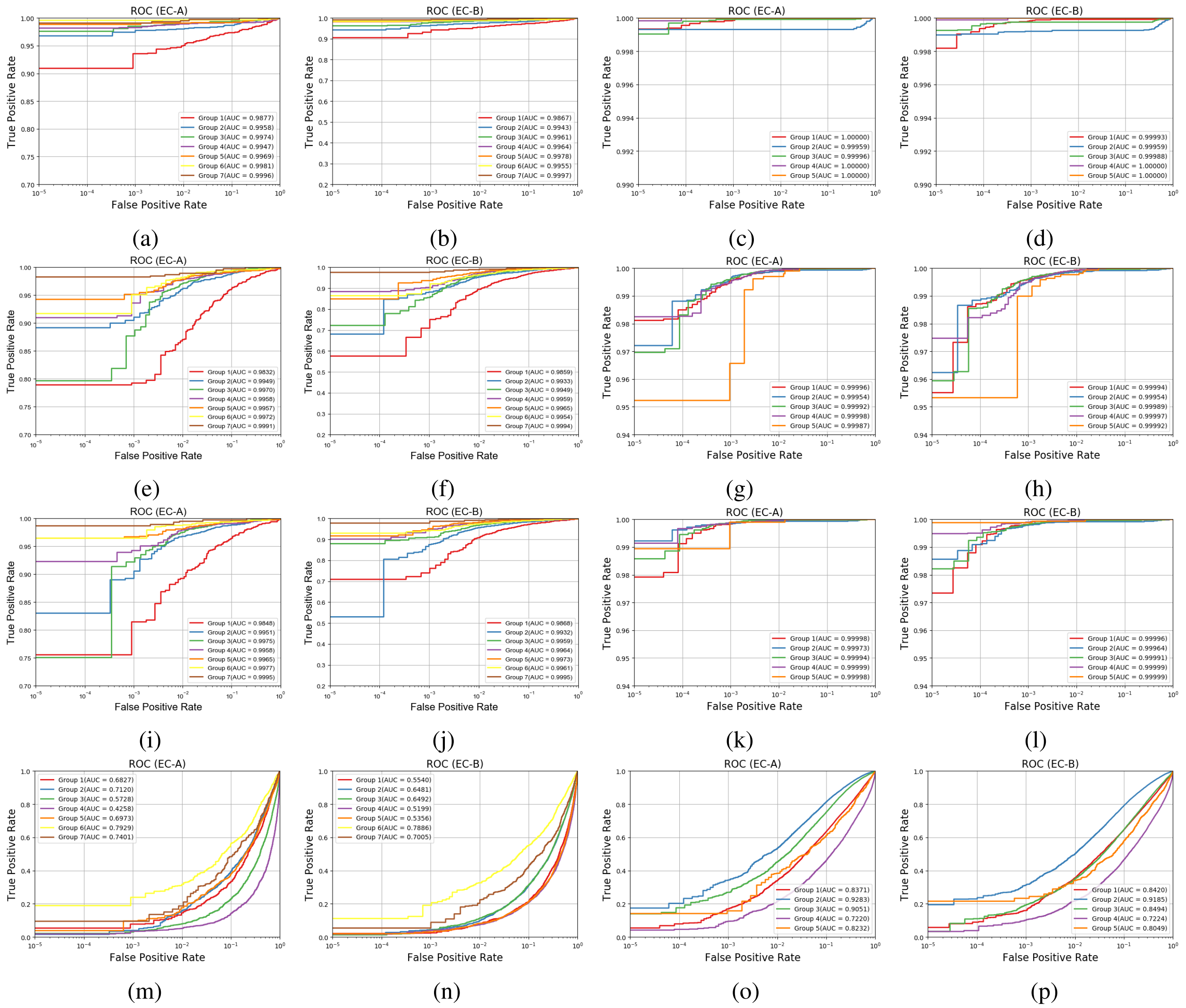
FIGURE 1. Example of image samples from different age groups. (Images from AgeDB [46].)

TABLE 4. The analysis of the ROC curves in Figure 2.

	AgeDB	Morph II
Deep learning based FR algorithms	Group 7 achieves the best performance by evaluating the ROC curves from (a)(b)(e)(f)(i)(j) in Figure 2. Group 1 has the worst performance by evaluating the ROC curves from (a)(b)(e)(f)(i)(j) in Figure 2. For the intermediate age groups, there was no clear indication of better performance at older ages.	All the AUC values for ArcFace, CosFace, and SphereFace2 are 99.9% for various age groups in the Morph dataset. The AUC value of Group 5 from Morph II is nearly 1, which means that older people have higher performance on the deep FR models.
LBP based FR algorithms	Group 6 and Group 7 have the best performance through the ROC curves in Figure 2 (m)(n) for the LBP algorithm, followed by Group 2. Group 4 has the lowest performance through the ROC curves in Figure 2 (m)(n).	Group 2 has the highest AUC values from the Morph II dataset in Figure 2 (o)(p). And the lowest AUC values from Morph II dataset for the LBP algorithm are found in Group 4.

that the central axis of the impostor distribution in different age groups is almost unchanged. This phenomenon illustrates that impostors have a weak impact on the performance of deep learning based face recognition models. Contrary to this, the genuine distribution shifted to higher similarity scores with an increase in the age group, especially in the AgeDB dataset.

The main factor affecting the variation in performance on face recognition tasks across age data has changed in the genuine distribution. From Figure 3, we can also observe that the Morph II dataset has less variability in score distributions for different age groups than the AgeDB dataset. Furthermore, the distance between the genuine distribution and the impostor distribution is larger in Morph II dataset than in the AgeDB dataset. This implies that FR could extract better features on Morph II dataset than that of the AgeDB dataset and the performance of the Morph II dataset on the FR models is better than the performance of AgeDB dataset on the FR models. Figure 4 shows the number of face images per subject and the latest image acquisition for each subject in years in the Morph II and AgeDB datasets. From Figure 4, we can find that 63.32% of the whole Morph II dataset has only 5 images or less, and 11,686 subjects (85.82%) only have a time span of 2 years or less. Thus, compared to the AgeDB dataset, the Morph II dataset cannot effectively reflect the aging effect of the same



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

FIGURE 2. The ROC curves for ArcFace (top row), CosFace (second row), SphereFace2 (third row), and LBPs(bottom row) of AgeDB(first & second column), Morph II(third & fourth column).

person on the face recognition models over a long time interval.

D. EXPERIMENTAL VALIDATION USING VARIOUS QUANTITATIVE METRICS

In this section, we evaluate the performance of FR models on various age groups of AgeDB and Morph II using four different quantitative metrics which include accuracy, as discussed in Section IV-C and shown in Table 2. The average accuracy value as depicted in Table 5 shows that the best recognition rate is mostly obtained in Group 5, Group 6, and Group 7 (around 60-80 years old), and the worst recognition rate is mostly obtained in Group 1 (around 20 years old).

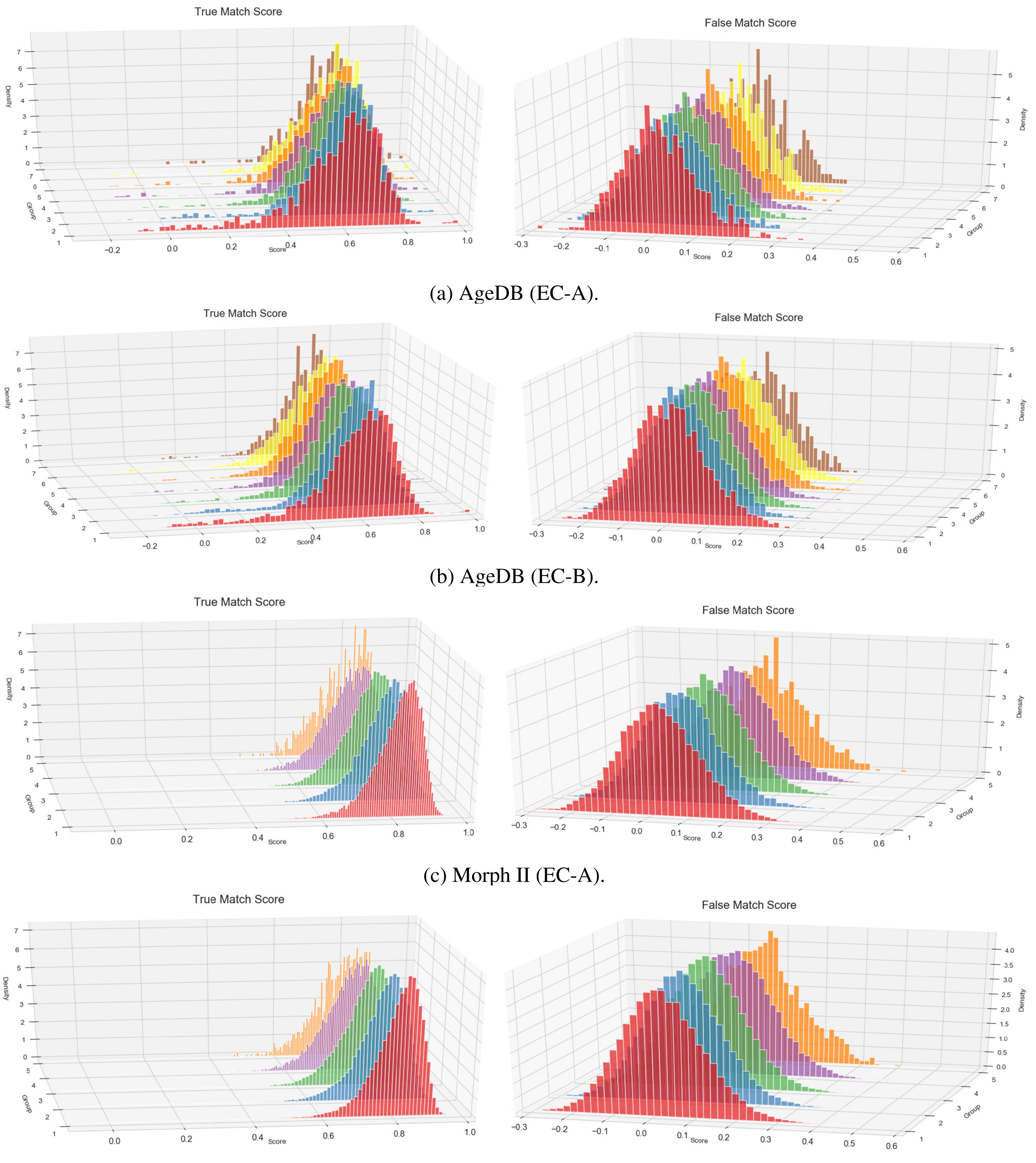
For the results of Group 2, Group 3, and Group 4, we cannot observe a significant trend of increasing accuracy. Also, there is no substantial difference in the accuracy of EC-A and EC-B, which indicates that the facial recognition system is not notably affected by age range within the upper and lower 2 years. (The upper and lower bounds of EC-A and EC-B differ by 2 years, respectively.)

Furthermore, Morph II achieves the highest accuracy rates using ArcFace, CosFace, and SphereFace2, with little variation between different groups. This suggests that the Morph II dataset is close to saturation for the recent deep learning based face recognition algorithms. i.e., the deep learning-based face recognition algorithms can very well identify the face images in the Morph II dataset.

Figure 5 shows the optimal F1 score, precision, and sensitivity of different age groups from AgeDB dataset by using the ArcFace model. From Figure 5 we can deduce that people from older age groups have the higher performance.

E. SUMMARY OF EXPERIMENTAL FINDINGS ON REAL-WORLD DATA

This section illustrates and quantifies the effect of data from various age groups on the face recognition model. From these experiments, the following findings can be obtained.



(a) AgeDB (EC-A).

(b) AgeDB (EC-B).

(c) Morph II (EC-A).

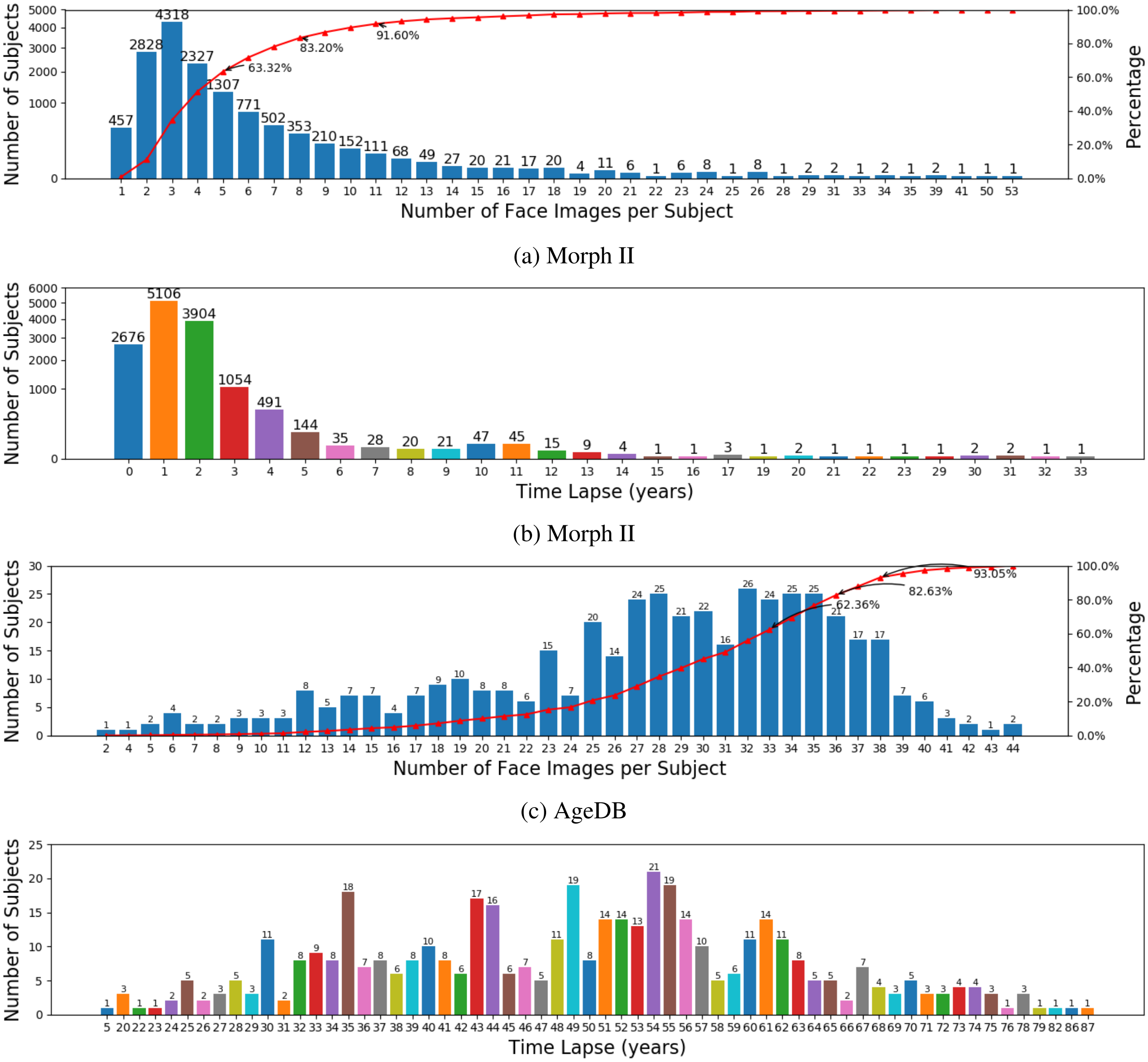
(d) Morph II (EC-B).

FIGURE 3. The match score distributions for ArcFace of AgeDB dataset and Morph II dataset.

(1) We concluded that the older adult group has a high performance on the face recognition models whereas the younger adult group has a low performance on the face recognition models. The performance of FR models is much better in people from older age groups (Group 6 and Group 7) as compared to people from the middle age group (Group 4 and Group 5). Subsequently, the performance of FR models on the middle age groups (Group 4 and Group 5) is comparatively better than the young adult groups (Group 1, Group 2, and Group 3).

(2) The genuine score distribution is the main factor that affects the variation of face recognition performance in various age groups.

(3) While deep learning based face recognition methods have achieved robust results, however, there is still room



(a) Morph II

(b) Morph II

(c) AgeDB

(d) AgeDB

FIGURE 4. **Analysis of AgeDB and Morph II Dataset. (a) and (c) shows the number of face images per subject. (b) and (d) shows the time span between the image of the youngest age and the image of the oldest age for each subject in years.**

for improvement when it comes to recognizing faces of all ages.

(4) Different age range threshold settings for age groups (EC-A and EC-B) can affect the accuracy of recognition. However, their overall trend is consistent, i.e., younger people are more difficult to recognize.

(5) Although the Morph II dataset has more amount of data, the AgeDB dataset is more suitable for the age and aging tasks because of the small number of images per identity and the short time-lapse in the Morph II dataset. The current aging dataset is still lacking in terms of the number of identities, the number of images per subject collected at different ages, and the low resolution of time-lapse images.

F. DISCUSSION OF THE DIFFERENCE WITH RELATED WORK

Recent study [29] on deep CNN face matches find that older people are harder to recognize while younger people are easier to recognize which is different from our conclusion. We analyzed the main differences between our findings and their study mainly due to the following reasons. (1) This paper employs different face recognition methods compared to [29] where authors have used FaceNet [35], VGGFace2(ResNet-50 trained on VGGFace2) [52], and ArcFace [31]. Popular deep learning based face authentications such as ArcFace, CosFace, and SphereFace2 and traditional LBPs-based FR method have been used in our work. (2) This paper uses different datasets as well as age group definitions. The age

TABLE 5. Recognition accuracy rates for AgeDB and Morph II evaluation (accuracy vs. age groups).

DataSet	FR model	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
AgeDB EC-A	ArcFace	97.19	98.85	99.25	99.21	99.48	99.82	99.57
	CosFace	94.79	97.86	98.14	98.48	98.52	98.69	99.14
	SphereFace2	95.41	97.96	98.39	98.52	98.75	99.19	99.35
	LBP	64.29	65.61	56.87	52.40	65.26	73.12	70.43
AgeDB EC-B	ArcFace	97.36	98.48	99.04	99.29	99.45	99.30	99.64
	CosFace	95.52	97.37	97.73	98.17	98.46	97.83	99.19
	SphereFace2	96.01	97.45	98.00	98.56	98.73	98.33	99.34
	LBP	56.25	61.58	61.60	55.50	55.76	72.67	67.06
Morph II EC-A	ArcFace	99.98	99.97	99.99	99.99	100.0	/	/
	CosFace	99.77	99.81	99.81	99.77	99.67	/	/
	SphereFace2	99.88	99.89	99.90	99.90	99.91	/	/
	LBP	77.37	85.76	83.07	68.95	76.12	/	/
Morph II EC-B	ArcFace	99.97	99.95	99.98	99.99	100.0	/	/
	CosFace	99.78	99.75	99.80	99.76	99.71	/	/
	SphereFace2	99.87	99.85	99.87	99.91	99.94	/	/
	LBP	77.78	84.73	78.06	69.08	74.90	/	/

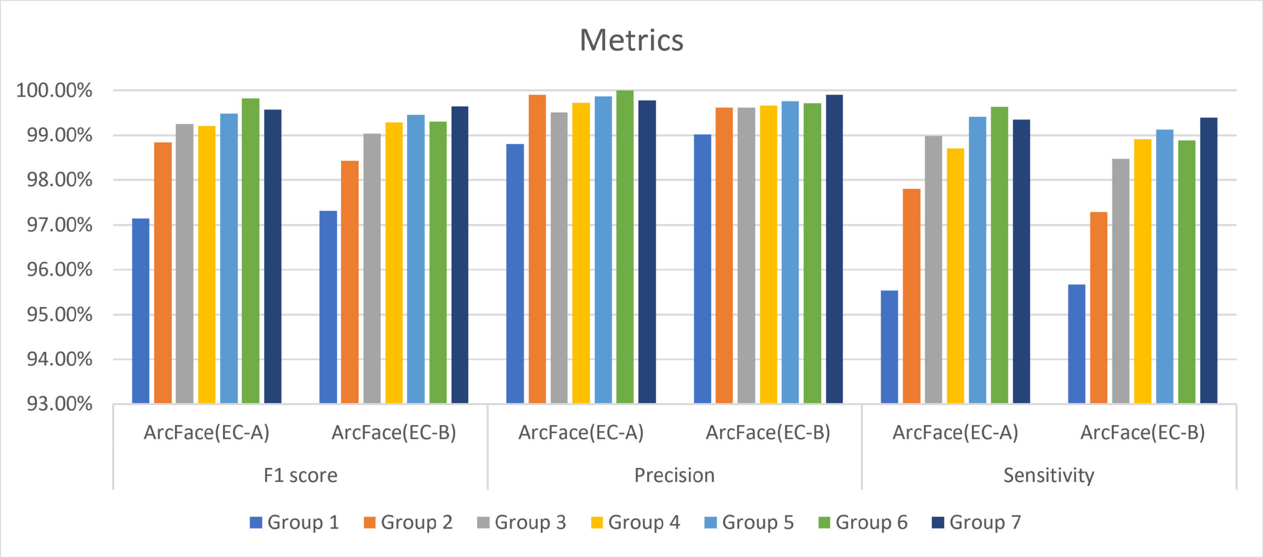


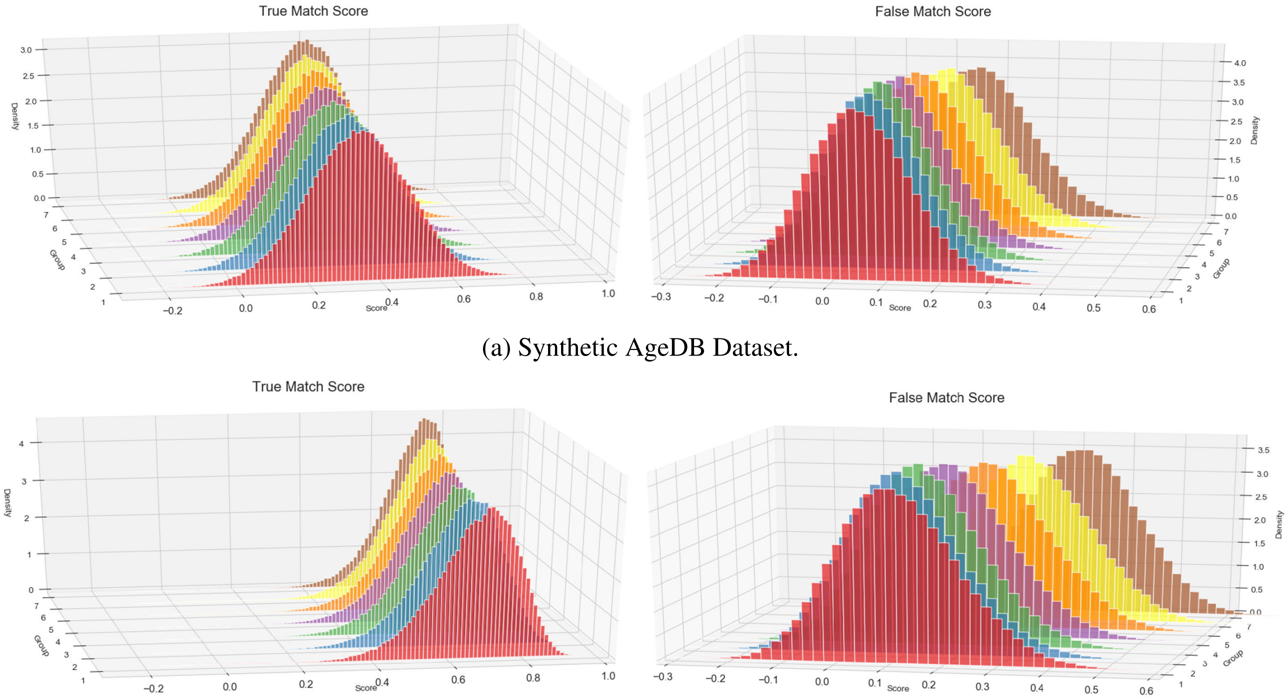
FIGURE 5. The F1 score, Precision, and Sensitivity for ArcFace of AgeDB.

gap between young (Group 1) and old (Group 7) as defined in our experimental design was larger than in previous studies. The age span between each age group in our experiment is the same, however, for most of the previous studies, it was different [18], [28], [29].

As compared to this research work authors in [29] have also used the Morph II dataset and ArcFace model. When only comparing the ROC curves on the whole Morph II dataset for the ArcFace model, it can be seen from their results that the best performance of the face recognition model is in the older and younger groups, while it decreases slightly in the middle-aged group. Our method has shown the best performance in Group 1 (around 20 years old), Group 4 (around 50 years old), and Group 5 (around 60 years old) in the Morph II dataset for the ArcFace model thus our outcomes are consistent with their method. This suggests that the reason for the inconsistency of our findings with them is mainly due to the use of different face recognition methods, as well as the design for age groups and the different datasets used in the experiments.

VI. ADDITIONAL EXPERIMENTS WITH SYNTHETIC-AGE DATA

The image pairs and identities are different within the various age groups in the experimental setting described in Section V because of the limitation of the real-world datasets that are available. This section investigates the use of synthetic aging as an augmentation technique to enlarge the original datasets. Synthetic data is now widely used in data-centric problems



(a) Synthetic AgeDB Dataset.

(b) Synthetic Morph II Dataset.

FIGURE 6. The match score distributions for ArcFace of synthetic AgeDB and synthetic Morph II.

and as there are some well-developed methods to age/de-age facial data samples this approach should help clarify the outcomes of the previous work on real data by providing more balanced age groups and enabling image pairs of the same data-subject to be used for our ROC studies.

We next evaluate the performance of balanced age groups on the face recognition models by introducing synthetic age data. These balanced age groups will have the same number of image pairs and the same identities. Firstly, we describe how to produce a balanced dataset by using synthetic age data. Next, we analyze and compare the match score distributions of the synthetic data with the real-world data. In addition, the impact of synthetic age data will be evaluated on the face recognition tasks.

A. EXPERIMENT SETTING

Firstly, all the face images from AgeDB and Morph II are aligned, cropped, and resized to 256 × 256 image resolution. Then various aging faces are generated by all these pre-processed face images via the SAM techniques [25]. To maintain consistency and balance in various age groups, synthetic faces were generated with different age groups which include 20, 30, 40, 50, 60, 70, and 80 years old. All 20-year-old faces will be in Group 1. Similarly, 30, 40, 50, 60, 70, and 80 years old faces are inserted in Group 2, Group 3, Group 4, Group 5, Group 6, and Group 7 respectively. It is important to mention that if any face could not be detected by using the MTCNN face detector [24], then all the correspondent faces are discarded. Then the AgeDB and Morph II are used to create PPs and NPs. The number of pairs from AgeDB is 176,008, and the number of pairs from Morph II is 135,537. The original pairs are generated from real-world AgeDB and real-world Morph II as the reference (Ref-Ori) in Figure 7. For different age groups, we replaced the original image pairs with the corresponding synthesized age pairs which are generated from the original real-world image pairs for evaluating the effect of the synthetic age images on face recognition algorithms.

B. MATCH SCORE ANALYSIS

Figure 6 shows the score distribution of synthetic aging generated by the SAM method [25]. We can observe that the impostor distribution and genuine distribution are significantly shifted relative to the original impostor distribution and genuine distribution. Secondly, observing the genuine distribution at synthetic age in Figure 6(a), it can be found that there is a tendency to shift to the right with increasing age, which is consistent with the moving trend of the genuine distribution for real age in Figure 3. Observing the impostor distribution of synthesis age data as shown in Figure 6(a), it can be noticed that the distribution shifts significantly to the right with increasing age. The synthetic Morph II dataset in Figure 6(b) has the same trend, the only difference is that the impostor distribution of the original Morph II dataset is far from the genuine distribution of the original Morph II dataset. This phenomenon suggests the synthetic Morph II dataset is better than the synthetic

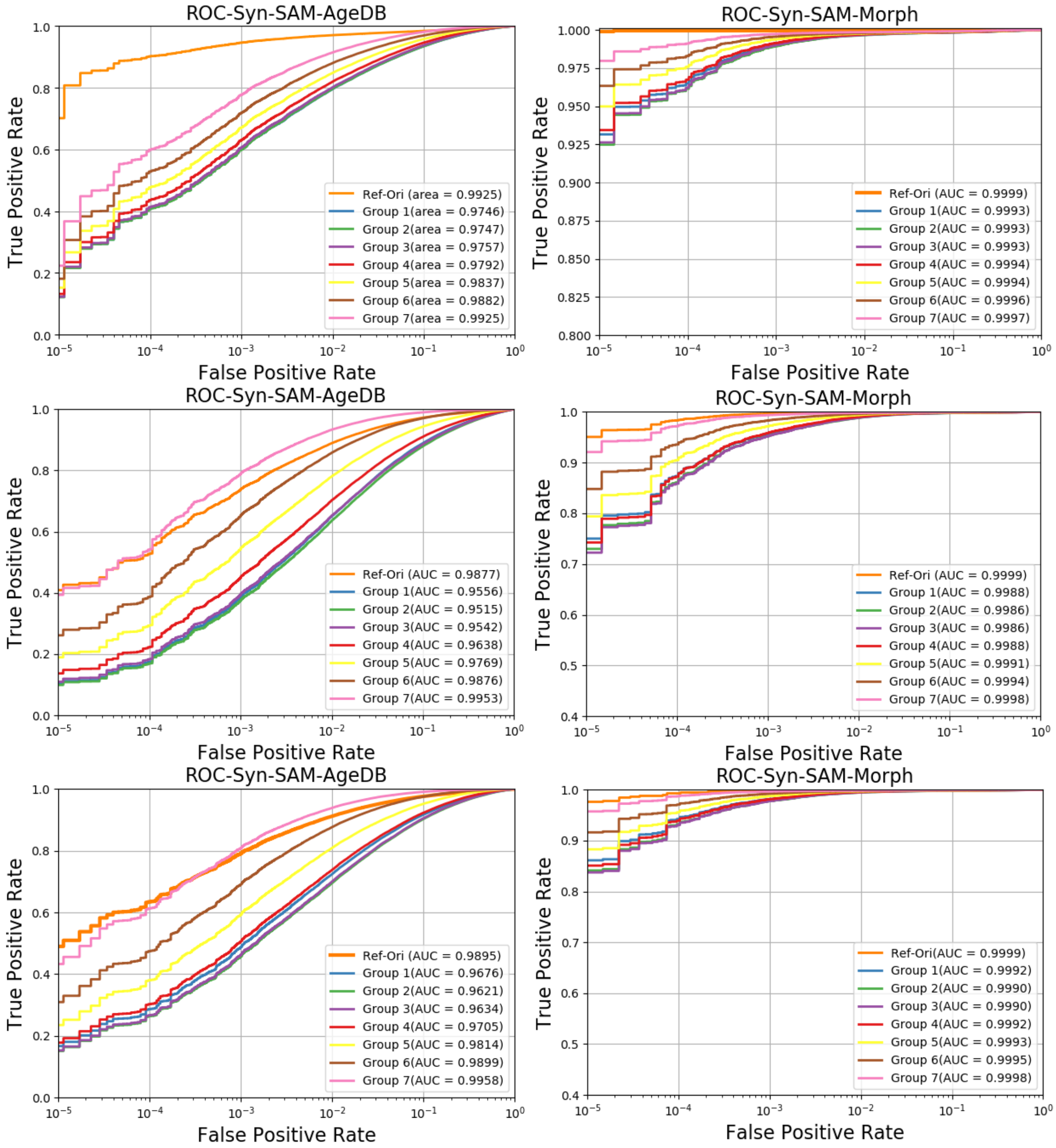


FIGURE 7. The revised ROC curves for ArcFace (first row), CosFace (second row), and SphereFace2 (third row) of synthetic AgeDB (first column) and synthetic Morph II (second column).

AgeDB in terms of the performance of the face recognition model.

From the above observation, first, we can conclude that as age increases, the impostor distribution shifts to the right to a much greater extent than the genuine distribution shifts to the right. That is, the distribution of impostor scores increases obviously with aging, and the distribution of genuine scores increases slightly with aging. This phenomenon implies a significant increase in the similarity scores of negative-identity pairs (i.e. different people), leading to a significant decrease in the recognition performance of the face recognizer for older people. Moreover, observing the data in different rows, it can be found that although the distribution is distinct, the trend is consistent, indicating that the phenomenon is related to the synthetic age algorithm and is independent of the different deep face recognizers. It is also implied that the main reason for the different performance of the synthetic age data generated by SAM and the real data on the face recognizer is that the synthetic age makes the faces more similar and the stronger the effect is applied the more similar they become. Furthermore, Figure 6 also supports that for the age data synthesized using SAM, the older the age the lower the performance of the face recognition.

C. ROC CURVE COMPARISON

As analyzed in Section VI-B, the main reason for the gap between the performance of the synthetic dataset and the real dataset is impostor distribution. For the real-age dataset, the impostor distribution remained consistent across different age groups. Thus, we used the NP of the original dataset to correct the NP of different age groups, which can roughly make it closer to the real data performance. The results are shown in Figure 7.

From Figure 7, the original reference ROC curves have the best performance. It indicates that there is still a gap between the synthetic data and the real data in terms of the performance of the face recognizer. Comparing the Ref-Ori ROC curves of different columns from Figure 7, it can be found that the first column has the best ROC curves, implying that ArcFace performs best in age-invariant face recognition among the three face recognition models - ArcFace, CosFace, and SphereFace2. Comparing the ROC curves for all age groups, it can be found that Group 7 (80-year-old) performed the best among all six results. Group 1 and Group 2 have poor ROC curves. This is consistent with Figure 2, indicating that the synthesized age data has similar properties to real age data.

From Figure 7, we can find that there is no significant trend of increasing the performance of the face recognizer during the change of the ROC curves from Group 1 (20-year-old) to Group 2 (30-year-old) and from Group 2 (30-year-old) to Group 3 (40-year-old). By observing the ROC curves corresponding to the age group from Group 3 to Group 4 and Group 4 to Group 5, there is a trend of increasing face recognition performance. In the process of ROC curve changes associated with the age groups of Group 5 to Group 6 and Group 6 to Group 7, a significant improvement in the performance of the face recognizer can be observed. This suggests that face recognizers are more susceptible to recognizing people who are older, but this is not a linear increasing process.

D. SUMMARY OF EXPERIMENTAL FINDINGS ON SYNTHETIC DATA

(1) The overall experiment analysis further illustrates that older people are prone to obtain better face recognition performance and this process is not a linear increase. The results from Section V are in line with this, showing that the number of images and the number of image pairs in a certain age group do not significantly affect the accuracy of the face recognition system.

(2) The synthetic age data generated by the SAM method poses an issue in terms of the identities of the synthetic data. The similarity score of the different identity samples increased with age. This implies that the distance between different identity classifications can be increased with increasing age to improve the authenticity of the synthetic age data distribution. Thus, increasing the synthetic data is useful for maintaining identity by decreasing the distribution of scores of different individuals.

(3) Although existing face recognizers do not solve the age-invariant face recognition problem, ArcFace has higher robustness for face recognition across ages compared to other methods.

VII. CONCLUSION AND FUTURE WORK

This paper investigated comprehensive studies of the aging effect and facial aging datasets. In addition to this, the impact of age group on the performance of face recognition algorithms is explored, along with an initial exploration of the use of synthetic age data to complement the real datasets. These experiments demonstrate that older people are more accurately identified, whereas younger individuals are more difficult to identify. To minimize the effect of data imbalance from the original dataset on these results, we employed synthetic age data to expand the number of samples for each age group. This further verifies that the face recognizer exhibits age-specific biases and it is more likely to distinguish between two older people and less likely to confuse them as being the same person. Among the datasets used in this study, this research illustrates that the AgeDB is a better option for investigating the impact of age, as it encompasses a considerable quantity of image samples for each individual and is more uniformly spread across various age groups, with a large interval between the minimum and maximum ages.

The SOTA face recognizers published in recent years commonly use hyper-spherical deep learning methods, and in this paper, the most widely used hyper-spherical deep learning based face recognizers which include ArcFace, CosFace, and SphereFace2 and a traditional LBP-based face recognizer are shortlisted to evaluate the performance of face recognizers for different age groups. The experimental outcomes show that the deep learning based face recognition algorithms perform consistently between different age groups, while the traditional FR method performs differently from the latest face recognition methods in different age groups. And this result may serve as an inspiration for future work in the related field.

In this study we have only explored one particular method of synthetic aging and the initial results presented in this study suggest that this method tends to reduce the FR performance as stronger aging effects are applied to the original data sample. We next plan to study a variety of synthetic aging methods including more robust methods that better account for identity preservation and explore their individual and ensemble effects on FR algorithms. It is also important to validate these algorithms through both quantitative and qualitative studies. Another potential direction includes the interactions between aging effects with other demographics such as race and gender. Ultimately our goal is to build a reference aging dataset derived from real data, but with augmented data that is consistent with the publicly available datasets. This will enable fine-tuning of FR models for specific age groups, with an ultimate goal to optimize performance across individual age groups, especially for younger adults.

REFERENCES

[1] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, ''Face verification across age progression using discriminative methods,'' *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 1, pp. 82–91, Mar. 2010.

[2] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, ''Past, present, and future of face recognition: A review,'' *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/8/1188

[3] J. I. Olszewska, ''Automated face recognition: Challenges and solutions,'' in *Pattern Recognition*, S. Ramakrishnan, Ed. Rijeka, Croatia: IntechOpen, 2016, ch. 4, doi: 10.5772/66013.

[4] M. M. Sawant and K. M. Bhurchandi, ''Age invariant face recognition: A survey on facial aging databases, techniques and effect of aging,'' *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 981–1008, Aug. 2019.

[5] C.-B. Ng, Y.-H. Tay, and B.-M. Goi, ''A review of facial gender recognition,'' *Pattern Anal. Appl.*, vol. 18, no. 4, pp. 739–755, Nov. 2015.

[6] K. S. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, ''Issues related to face recognition accuracy varying based on race and skin tone,'' *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 8–20, Mar. 2020.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, ''VGGFace2: A dataset for recognising faces across pose and age,'' in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, ''MS-Celeb-1M: A dataset and benchmark for large-scale face recognition,'' in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 87–102.

[9] D. Yi, Z. Lei, S. Liao, and S. Z. Li, ''Learning face representation from scratch,'' 2014, *arXiv:1411.7923*.

[10] C. Meng, J. Lu, and Y.-P. Tan, ''A comparative study of age-invariant face recognition with different feature representations,'' in *Proc. 11th Int. Conf. Control Autom. Robot. Vis.*, Dec. 2010, pp. 890–895.

[11] M. C. Agamez. (2016). *Aging Effects in Automated Face Recognition*. Open Access Theses. [Online]. Available: https://docs.lib.purdue.edu/open_access_theses/930

[12] S. H. Abdurrahim, S. A. Samad, and A. B. Huddin, ''Review on the effects of age, gender, and race demographics on automatic face recognition,'' *Vis. Comput.*, vol. 34, no. 11, pp. 1617–1630, Nov. 2018.

[13] G. Guo, G. Mu, and K. Ricanek, ''Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits,'' in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 3392–3395.

[14] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, ''A meta-analysis of face recognition covariates,'' in *Proc. IEEE 3rd Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2009, pp. 1–8.

[15] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, ''A study of face recognition as people age,'' in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[16] D. Deb, L. Best-Rowden, and A. K. Jain, ''Face recognition performance under aging,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 548–556.

[17] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat, ''Local descriptors in application to the aging problem in face recognition,'' *Pattern Recognit.*, vol. 46, no. 10, pp. 2634–2646, Oct. 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320313001398

[18] L. Boussaad and A. Boucetta, ''The aging effects on face recognition algorithms: The accuracy according to age groups and age gaps,'' in *Proc. Int. Conf. Artif. Intell. Cyber Secur. Syst. Privacy (AI-CSP)*, Nov. 2021, pp. 1–6.

[19] C. Otto, H. Han, and A. Jain, ''How does aging affect facial components?'' in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2012, pp. 189–198.

[20] B. Klare and A. K. Jain, ''Face recognition across time lapse: On learning feature subspaces,'' in *Proc. Int. Joint Conf. Biometrics (IJCB)*, Oct. 2011, pp. 1–8.

[21] M. S. Bartlett, *Face Image Analysis by Unsupervised Learning and Redundancy Reduction*. San Diego, CA, USA: Univ. California, 1998.

[22] Z. Akhtar, A. Rattani, A. Hadid, and M. Tistarelli, ''Face recognition under ageing effect: A comparative analysis,'' in *Proc. Int. Conf. Image Anal. Process.*, 2013, pp. 309–318.

[23] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, ''Face recognition vendor test 2002,'' in *Proc. IEEE Int. SOI Conf.*, Oct. 2003, p. 44.

[24] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, ''Joint face detection and alignment using multitask cascaded convolutional networks,'' *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[25] Y. Alaluf, O. Patashnik, and D. Cohen-Or, ''Only a matter of style: Age transformation using a style-based regression model,'' *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–12, Jul. 2021, doi: 10.1145/3450626.3459805.

[26] G. Givens, J. R. Beveridge, B. A. Draper, P. Grother, and P. J. Phillips, ''How features of the human face affect recognition: A statistical comparison of three face recognition algorithms,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun./Jul. 2004, p. 2.

[27] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper, ''Factors that influence algorithm performance in the face recognition grand challenge,'' *Comput. Vis. Image Understand.*, vol. 113, no. 6, pp. 750–762, Jun. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314209000022

[28] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, ''Face recognition performance: Role of demographic information,'' *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.

[29] V. Albiero, K. W. Bowyer, K. Vangara, and M. C. King, ''Does face recognition accuracy get better with age? Deep face matchers say no,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 250–258.

[30] S. Wu and D. Wang, ''Effect of subject's age and gender on face recognition results,'' *J. Vis. Commun. Image Represent.*, vol. 60, pp. 116–122, Apr. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1047320319300197

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ''ArcFace: Additive angular margin loss for deep face recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[32] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, ''CosFace: Large margin cosine loss for deep face recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[33] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, ''SphereFace: Deep hypersphere embedding for face recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.

[34] Y. Wen, W. Liu, A. Weller, B. Raj, and R. Singh, ''SphereFace2: Binary classification is all you need for deep face recognition,'' in *Proc. ICLR*, 2022, pp. 1–15.

[35] F. Schroff, D. Kalenichenko, and J. Philbin, ''FaceNet: A unified embedding for face recognition and clustering,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[36] T. Ahonen, A. Hadid, and M. Pietikäinen, ''Face recognition with local binary patterns,'' in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 469–481.

[37] W. Yao, V. Varkarakis, G. Costache, J. Lemley, and P. Corcoran, ''Toward robust facial authentication for low-power edge-AI consumer devices,'' *IEEE Access*, vol. 10, pp. 123661–123678, 2022.

[38] A. Lanitis, C. J. Taylor, and T. F. Cootes, ''Toward automatic simulation of aging effects on face images,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

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

[40] K. Ricanek and T. Tesafaye, ''MORPH: A longitudinal image database of normal adult age-progression,'' in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 341–345.

[41] B. Yip, G. Bingham, K. Kempfert, J. Fabish, T. Kling, C. Chen, and Y. Wang, ''Preliminary studies on a large face database,'' in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2572–2579.

[42] B.-C. Chen, C.-S. Chen, and W. H. Hsu, ''Cross-age reference coding for age-invariant face recognition and retrieval,'' in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 768–783.

[43] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, ''Labeled faces in the wild: A database for studying face recognition in unconstrained environments,'' in *Proc. Workshop Faces Real-Life Images Detection Alignment Recognit.*, 2008, pp. 1–15.

[44] Z. Zhang, Y. Song, and H. Qi, ''Age progression/regression by conditional adversarial autoencoder,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4352–4360.

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

[46] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, ''AgeDB: The first manually collected, in-the-wild age database,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1997–2005.

[47] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, ''Lifespan age transformation synthesis,'' in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2020, pp. 739–755.

[48] T. Karras, S. Laine, and T. Aila, ''A style-based generator architecture for generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[49] Z. Huang, J. Zhang, and H. Shan, ''When age-invariant face recognition meets face age synthesis: A multi-task learning framework,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7278–7287.

[50] S. Bazrafkan, S. Thavalengal, and P. Corcoran, ''An end to end deep neural network for iris segmentation in unconstrained scenarios,'' *Neural Netw.*, vol. 106, pp. 79–95, Oct. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360801830193X

[51] A. Mansfield, *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, document ISO/IEC 19795-1, 2006.

[52] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[53] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, ''A comprehensive study on face recognition biases beyond demographics,'' *IEEE Trans. Technol. Soc.*, vol. 3, no. 1, pp. 16–30, Mar. 2022.

[54] P. Negri, S. Cumani, and A. Bottino, ''Tackling age-invariant face recognition with non-linear PLDA and pairwise SVM,'' *IEEE Access*, vol. 9, pp. 40649–40664, 2021.

[55] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, ''Orthogonal deep features decomposition for age-invariant face recognition,'' in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 746–779.

[56] R. Rothe, R. Timofte, and L. Van Gool, ''DEX: Deep EXpectation of apparent age from a single image,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 252–257.

[57] J. Zhao, S. Yan, and J. Feng, ''Towards age-invariant face recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 474–487, Jan. 2022.

[58] J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim, ''Integrated face analytics networks through cross-dataset hybrid training,'' in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1531–1539.

[59] Z. Huang, J. Zhang, and H. Shan, ''When age-invariant face recognition meets face age synthesis: A multi-task learning framework and a new benchmark,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7917–7932, Jun. 2023.

[60] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, ''Face recognition algorithm bias: Performance differences on images of children and adults,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2269–2277.

[61] P. K. Chandaliya and N. Nain, ''ChildGAN: Face aging and rejuvenation to find missing children,'' *Pattern Recognit.*, vol. 129, Sep. 2022, Art. no. 108761.

[62] D. Deb, N. Nain, and A. K. Jain, ''Longitudinal study of child face recognition,'' in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 225–232.

[63] K. Bahmani and S. Schuckers, ''Face recognition in children: A longitudinal study,'' in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2022, pp. 1–6.

WANG YAO (Graduate Student Member, IEEE) received the B.Sc. degree in computer science and technology from Southwest University, China, in 2016, and the M.Sc. degree in control engineering from the University of Chinese Academy of Sciences (UCAS), in 2019. She is currently pursuing the Ph.D. degree with the University of Galway. She is also an Intern at FotoNation/Xperi. Her research interest focuses on computer vision.

MUHAMMAD ALI FAROOQ received the B.E. degree in electronic engineering from IQRA University, in 2012, the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST), in 2017, and the Ph.D. degree from the National University of Ireland Galway (NUIG), in 2022. He is currently a Postdoctoral Researcher with the University of Galway and a Machine Learning Research Intern with Xperi Corporation, where his work is focused on building large-scale synthetic datasets for real-world applications. His research interests include machine vision, computer vision, video analytics, machine learning, thermal imaging, and sensor fusion. He has won the prestigious H2020 European Union (EU) Scholarship as part of his Ph.D. research project and worked as a Consortium Partner in the EU-funded HELIAUS Project.

JOSEPH LEMLEY (Member, IEEE) received the B.S. degree in computer science and the M.S. degree in computational science from Central Washington University, Ellensburg, WA, USA, in 2006 and 2016, respectively, and the Ph.D. degree from the National University of Ireland, Galway. He is currently leading Xperi's Sensing Group, which develops novel algorithms and artificial neural networks for upcoming sensor technologies. His research interests include artificial intelligence, deep learning, and computer vision. He received the 2017 Best Paper Joint Award of the *IEEE Consumer Electronics Magazine* and the Best Paper Second Place Award at ICCE 2018 and other awards during previous years.

PETER CORCORAN (Fellow, IEEE) is currently the Personal Chair of Electronic Engineering with the College of Science and Engineering, University of Galway. He was the Co-Founder of several start-up companies, notably FotoNation (currently the Imaging Division, Xperi Corporation). He has more than 600 technical publications and patents, more than 100 peer-reviewed journal articles, 120 international conference papers, and a co-inventor on more than 300 granted U.S. patents. He is an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He has been a member of the IEEE Consumer Electronics Society for more than 25 years. He is the Editor-in-Chief and the founding Editor of *IEEE Consumer Electronics Magazine*.