Face Recognition Algorithm Bias: Performance Differences on Images of Children and Adults

Nisha Srinivas, Karl Ricanek
University of North Carolina Wilmington
Wilmington, North Carolina, USA
srinivasn, ricanekk@uncw.edu

Dana Michalski
Defence Science and Technology Group
Edinburgh, South Australia, Australia
dana.michalski@dst.defence.gov.au

David S. Bolme
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
bolmeds@ornl.gov*

Michael King
Florida Institute of Technology
Melbourne, Florida, USA
michaelking@fit.edu

Abstract

In this work, we examine if current state-of-the-art deep learning face recognition systems exhibit a negative bias (i.e., poorer performance) for children when compared to the performance obtained on adults. The systems selected for this work are five top performing[1] commercial-off-the-shelf face recognition systems, two government-off-the-shelf face recognition systems, and one open-source face recognition solution. The datasets used to evaluate the performance of the systems are both unconstrained in age, pose, illumination, and expression and are publicly available. These datasets are indicative of photo journalistic face datasets published and evaluated on, over the last few years. Our findings show a negative bias for each algorithm on children. Genuine and imposter distributions highlight the performance bias between the datasets further supporting the need for a deeper investigation into algorithm bias as a function of age cohorts. To combat the performance decline on the child demographic, several score-level fusion strategies were evaluated. This work identifies the best score-level fusion technique for the child demographic.

*This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

[1]According to NIST Face Recognition Vendor Tests, https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

1. Introduction

Typically, face recognition solutions have been developed to solve recognition problems for adults. These systems have presupposed that the general solution for recognition can be modeled fully through the use of adult faces. However, we know that children look different from adults and they are not simply scaled down versions of adults. In fact, the craniocomplex of a child is much different from an adult. While there has been a marked improvement in face recognition for adults over the last few years, predominantly driven by the adoption of deep learning techniques, recognition on children has not kept pace.

There still remains a paucity of research on the topic of child face recognition. This is possibly due to the lack of understanding regarding the mechanisms of facial changes that occur in children. From the scientific literature, we understand that there are fundamentally different mechanisms for facial aging between children and adults [10, 13]. When face recognition systems are not challenged by the construct of aging, i.e., the enrolled and probe images are sufficiently close in chronological age, there does not appear to be a sufficient differentiation in performance except for the youngest of ages. Does this difference in performance rise to the level of genuine algorithm bias? This work reviewed and expanded on previous child face recognition research to delve deeper into this question [3][15][2][1][4][11].

Most of the early research evaluating child face recognition has focused on constraining the face image for pose, illumination, and expression, except for the work of Ricanek et al. [15]. The work of Ricanek et al. and this current work looks at the practical problem of face recognition which is an "in-the-wild" problem. There are limited use cases for controlled face recognition in today's environ-

ment, e.g. immigration/border control. The typical use case is "in-the-wild" matching, e.g., social media tagging, digital photo books, etc. Some of the more pressing use cases are child exploitation, child abduction, and other law enforcement uses. This work through its comparison to a congruent adult corpus amplifies the problem of unconstrained child face recognition through the largest evaluation of the latest deep learning commercial-off-the-shelf (COTS) systems, government-off-the-shelf (GOTS) systems, and a single open-source solution. The five COTS systems used have all been top performers in the National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) over the years. Hence, this extends the general body of knowledge on algorithm bias as a function of age cohorts, children versus adults. This work will be extremely valuable to world governments and other organizations attempting to use face recognition to stymie the demonstrable trafficking of children and other types of child exploitation.

This work contributes to the body of scientific knowledge in the following ways:

1. It provides an evaluation of **eight face recognition systems**, comprising of five COTS[2] systems, two GOTS systems, and one open-sourced solution. Prior research focused on the performance of a single COTS algorithm or one COTS with one open-sourced algorithm.

2. It **extends the publicly available child face dataset**, known as In The Wild Child Celebrity (ITWCC) dataset, which is developed on child television and movie celebrities in the USA. This is a gender balanced dataset conducive for evaluation in the area of algorithm bias.

3. It provides a preliminary study on **fusion for child face recognition**.

4. It defines a **new type of age bias**, children versus adults.

5. It establishes **performance between similar face corpora for children and adults**.

2. Related Work

In this section, we examine and summarize the research published to date on child face recognition to illustrate the paucity of research and to establish the gaps. The earliest work in child face recognition is from Bharadwaj et al. [3] on a small non-longitudinal dataset containing 34 newborns with multiple image captures between 2 hours of birth and 8 to 15 hours post birth. The objective was to study the applicability and performance of face recognition to avoid baby switching or for identifying abducted infants. This work extracts and compares features at different levels of a Gaussian Pyramid (GP). SURF features are extracted at level 0 and LBP features are extracted at level 1 and level 2 of the GP. Finally, a similarity score is computed as a weighted sum of similarity scores obtained by comparing features in corresponding level of the GP. The earliest known work examining the problem of child face recognition using a longitudinal dataset was by Ricanek et al. [15]. This work focused on matching for the development stages of child, preadolescence, and adolescence. Match performance was examined on a longitudinal child corpus using a COTS algorithm and the standard set of hand-engineered algorithms circa 2014. The standard set of academic algorithms were PCA, LDA, LRPCA and cohort-LDA along with the state-of-the-art open-source face recognition algorithm, OpenBR.

There has also been some focus on evaluating the performance of face recognition algorithms for identifying newborns and toddlers. Best-Rowden et al. [2] evaluated the performance of a COTS algorithm against the Newborns, Infants, and Toddlers (NITL) dataset and studied the effects of age variation on the performance of the algorithm. The NITL dataset consists of face images of 314 children aged 0 to 4 years. The images captured exhibit a wide variation in lighting, pose, and expression. The study indicated that the performance of the COTS algorithm decreased as the age variation increased, with high verification accuracy observed when comparing images that were captured in the same session and a degradation in verification accuracy observed when comparing images captured in different sessions.

Basak et al. [1] investigated the use of multimodal biometrics for toddlers and pre-school children. The work indicates that fingerprint and iris outperform face recognition on this demographic. The authors developed the first multimodal dataset for children (aged 2 to 4 years) with face, iris, and fingerprint, known as the Children Multimodal Biometric Database (CMBD). The database was collected in two sessions. The time elapsed between the two sessions was 3 months. Images were captured over a span of two months in Session 1 and over a span of three months in Session 2. Face images were collected of 141 subjects in Session 1 and 118 subjects in Session 2. Each face image was captured with a DSLR camera with a resolution of 12.3 Megapixels. This work used a COTS algorithm (Verilook face recognition SDK) for two face only experiments. The first experiment matched a single enrolled image from Session 1 to a probe of five randomly selected images from Session 2. A time lapse of a few months to several months existed between the enrolled and probe images with a Genuine Accept Rate (GAR) of 18.96% at 0.1%. The second experi-

[2]These systems are all the latest version available as the date of writing.

ment examined the effects of enrolling multiple images of the subject. In this experiment, the performance increased and yielded a GAR of 26.46% at 0.1%.

Deb et al. [4] investigated the performance of a state-of-the-art COTS face algorithm and an open-source algorithm, FaceNet [17], on their Children Longitudinal Face (CLF) dataset, which has a total of $3,682$ images of 919 subjects aged 2 to 18 years old. The work examined and compared the performance of the two algorithms independently and by fusing the scores using the sum rule under both verification and identification scenarios, for age variations of 1 year, 3 years, and 7 years. Further, the work demonstrates the loss of performance as a function of age variation for both verification and identification scenarios. The simple score level sum fusion, which produced a marked difference in performance over any single algorithm, also indicated a strong performance loss as the age variation between images increased. A notable experiment conducted in this work was fine-tuning of FaceNet. FaceNet was constructed from the MS-Celeb [7] database composed of 10 million images of 100k adult celebrities. Tuning was based on a set of $3,294$ face images of $1,119$ children aged 3 to 18 years old. A performance increase over the untuned variant was demonstrated.

Michalski et al. [11] studied the effects of age (0 to 17 years) and age variation (0 to 10 years) on the performance of a COTS algorithm (NEC). The dataset used for the study has a total of $4,562,868$ face images of children captured under a controlled environment similar to standard passport or visa images. This work focuses on an operational use of face recognition, i.e., the use of matching child visitors to their e-passports, and the potential of assigning different thresholds for matching children at different ages. As stated in this work a single threshold may be selected for an operational system and potentially designed for adults. Further, this work highlights the challenge of face matching for children across age and age variation. This work concludes, using a granular heat map of "youngest age of child" (0 to 17 years) versus age variation (0 to 10 years), that it is harder to match infants, toddlers, and young children, regardless of age variations. However, at around age 8, age variation drives the match performance, i.e., after age 8, matching performance picks up for age variations of 0 to 2 years, but after two years it falls off sharply. When the child hits sexual maturity, the face shape becomes stable, the performance degradation due to age variation gracefully decreases. This work examines the largest child study to date on a single COTS algorithm, however, the dataset is not publicly available.

3. Objective

In this paper, we examine a set of eight face recognition systems to determine if a negative bias (i.e., poorer performance) exists for child faces. We also explore different score level fusion techniques to determine if we can improve performance. We evaluate five COTS face recognition systems, two GOTS, systems and an open source algorithm against the two datasets described in 4. These are all deep learning based algorithms that were trained on unconstrained adult faces. All the COTS solutions were top performers in the Ongoing FRVT conducted by NIST. The open-source algorithm uses a face detector based on Faster R-CNN [9] [14] and feature extractor based on the VGG-Face [12].

Three of the five COTS solutions were developed by RankOne, CyberExtruder, and Neurotech. We do not have permission to report the developers of the other two COTS systems. However, the developers of the anonymous algorithms have been a part of the face recognition space for a decade or more. They are consistently ranked high in the the NIST FRVT. We also do not have permission to report the developers of the two GOTS systems. The open sourced algorithm, Face Recognition from Oak Ridge (FARO), was developed at Oak Ridge National Laboratory. The systems are labeled from FR-A through to FR-H to maintain anonymity.

Further, we explore multiple fusion schemes to determine if we can improve performance. The different fusion schemes are defined as follows:

1. Fusion-A: The scores are normalized using z-score normalization and are fused using the sum rule. In the sum rule, the mean of the scores across all systems is chosen as the best score for a given comparison.
2. Fusion-B: The scores are normalized using min-max normalization and are fused using the sum rule.
3. Fusion-C: The scores are normalized using z-score normalization and are fused using the max rule. In the max rule, the maximum score across all systems is chosen as the best score for a given comparison.
4. Fusion-D: The scores are normalized using min-max normalization and are fused using the max rule.
5. Fusion-E: The scores are normalized using z-score normalization and are fused using the min rule. In the min rule, the minimum score across all systems is chosen as the best score for a given comparison.
6. Fusion-F: The scores are normalized using min-max normalization and are fused using the min rule.

4. Datasets

The number of child face databases available to the academic community is limited. The different datasets comprising of child faces used in different studies are listed in Table 1. Of the few publicly available datasets, the largest of these, the ITWCC dataset, was developed by the I3S Face Aging Group at the University of North Carolina Wilmington [15]. It was developed to research the problem of un-

Dataset	# of Subjects	# of Images	Age Range	Availability	Type
NITL [2]	314	3,144	0yrs-4yrs	Private	Wild
CMBD [1]	106	1,060	2yrs-4yrs	Private	Wild
CLF [4]	919	3,682	2yrs-18yrs	Private	Constrained
ITWCC [15]	304	1,705	0yrs-32yrs	Public	Wild
ITWCC−D1	**745**	**7,990**	**0yrs-32yrs**	**Public**	**Wild**
MORPH-II [16]	13,000	55,134	16yrs-77yrs	Public	Constrained
FG-NET [6]	82	1,002	0yrs-69yrs	Public	Wild
Adience [5]	2,284	26,580	0yrs-60+yrs	Public	Wild

Table 1: Datasets consisting of child face images that are used by researchers to explore the problem of face recognition.

constrained face recognition as it relates to children. The data corpus was designed to emulate photo-journalistic type captures while maintaining some basic requirements that make the dataset useful for face recognition evaluations. As the dataset's name, "in the wild", suggests, age, pose, illumination, expression, occlusions, etc. are not controlled. The initial dataset comprised of $34,323$ images corresponding to 745 subjects; 349 females and 396 males[3]. The dataset has a minimum of three images per subject and a maximum of 823 images per subject. There is an imbalance in the number of images for each subject and across age. This work addresses this imbalance by deriving a subset of images for this study known as ITWCC−D1. The criteria used to create the ITWCC−D1 dataset are:

1. The subject has at least a single image at three different ages.
2. Each subject has at most three images per age[4].

An adult dataset that has characteristics similar to the ITWCC−D1 dataset is the well known Labeled Faces in the Wild (LFW) dataset [8]. The LFW dataset has $5,749$ subjects and a total of $13,233$ images. But there are only $1,680$ subjects that have two or more images and the remaining $4,069$ subjects have only one image. Therefore, we use a subset of the LFW dataset (LFW−D1) for this study consisting of subjects that have more than one image. This subset has a total of $9,164$ face images from $1,680$ subjects. We compare the performance of the eight face recognition systems against the ITWCC−D1 dataset and the LFW−D1 dataset.

The composition of the ITWCC−D1 and the LFW−D1 dataset are shown in Table 2. Example images from the ITWCC−D1 dataset are shown in Figure 1 along with the subject's ages.

[3]Age and gender labels are available for the dataset.

[4]If there are more than three images per age, the images are ranked by image quality and the top three selected.

Dataset	#Images	#Subjects	μ and σ of age
ITWCC−D1 (Child Faces)	#M = $3,345$ #F = $4,645$ Total = $7,990$	#M = 396 #F = 349 Total = 745	μ = 12.753 σ = 3.984
LFW−D1 (Adult Faces)	$9,164$	$1,680$	Unknown

Table 2: Composition of the ITWCC−D1 and LFW−D1 datasets.

Figure 1: Images of subjects from the ITWCC−D1 dataset.

5. Evaluation Protocol

The protocol used to determine if face recognition systems exhibit a negative bias towards child faces is as follows:

1. The datasets used were not pre-processed for any of the evaluated solutions. Any type of image quality verification, image normalization, etc. is a function of the face recognition system being evaluated. The internal details of pre-processing of images prior to face detec-

tion and template generation are unknown to us.

2. Default parameters were used for all face recognition solutions. This work did not attempt to maximize performance of any system. Each algorithm performs its own image pre-processing, if any, face detection/segmentation, template generation and scoring.

3. Generate performance measures for verification and identification experiments as outlined below.

Figure 2 shows a schematic representation of the evaluation protocol.

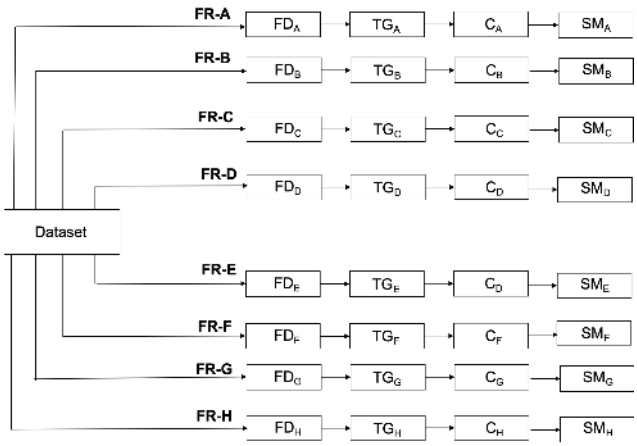


Figure 2: Evaluation protocol for each algorithm, where FD_i is face detector, TG_i is template generation, C_i is comparison function (score), and SM_i is score matrix.

6. Results

The goal of the study is to determine if face recognition systems exhibit a negative bias on child faces as compared to adult faces. Results are presented for both verification (1:1 matching) and identification (1:N matching). Receiver Operating Characteristics (ROC) curves are reported to compare performance in verification scenarios and Cumulative Match Characteristics (CMC) curves are used to compare performance in identification scenarios. For the ITWCC−D1 dataset, the target set consists of a single youngest image of the subjects and the query set consists of the remaining images of the subjects. This may include images of the subject at their youngest age, if multiple images at that age exist. There are 745 subjects in the target set and $7,239$ images in the query set. For the LFW−D1 dataset, since there are no age labels available, we enroll the first image of the subjects based on the naming convention in the target set and the remaining images of the subjects in the query set. There are $1,680$ subjects in the target set and $7,484$ images in the query set.

6.1. ITWCC−D1 Corpus

We first present the verification and identification performance results of the eight different systems against the ITWCC−D1 dataset as illustrated in Figures 3 and 4. The results indicate that the performance of face recognition systems are progressing for the child demographic compared to the prior work on COTS systems [4][1][15]. From Figure 3, we observe that FR-E performs best on the verification scenario followed closely by FR-F. Further, Figure 3 indicates that there is a large gap in verification performance across the suite of systems. The top two performers for the identification scenario, as illustrated in Figure 4, are FR-G and FR-E. FR-G with its retrieval rate of 82% is approximately 28% greater than the worst performer.

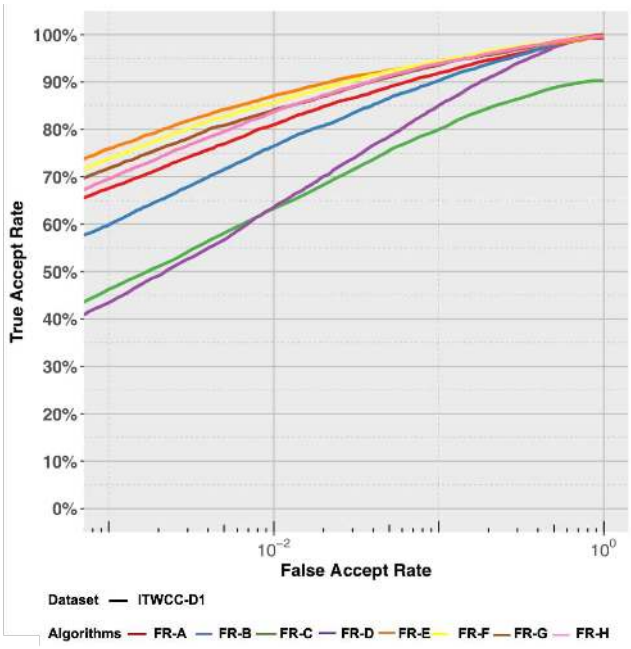


Figure 3: Verification performance of the eight systems when evaluated on the ITWCC−D1 dataset.

In Table 3, we compare the verification and identification rates of the individual systems with the different fusion schemes. Fusing scores from all the systems improves the verification rate by 2% using scheme Fusion-B at a False Accept Rate (FAR) of 0.1% and the identification accuracy increases by 4 to 5% at Rank-1 using scheme Fusion-B.

6.2. Algorithm Bias

Figure 5 shows the genuine and imposter distributions for each face recognition algorithm for the two datasets described in section 4, and clearly depicts the challenges of the child corpus. The overlap between the genuine and imposter distributions determine the performance of a biometric algorithm, with large overlaps in the distributions indi-

Face Recognition Algorithm	True Accept Rate		Rank-N Accuracy	
	FAR 0.1%	FAR 1%	Rank-1	Rank-5
FR-A	0.676	0.81	0.764	0.862
FR-B	0.598	0.765	0.68	0.811
FR-C	0.463	0.633	0.56	0.701
FR-D	0.434	0.636	0.563	0.729
FR-E	0.759	0.871	0.808	0.894
FR-F	0.738	0.856	0.787	0.887
FR-G	0.718	0.841	0.82	0.92
FR-H	0.695	0.837	0.774	0.912
Fusion-A	0.764	0.869	0.846	0.932
Fusion-B	**0.782**	0.878	**0.852**	0.932
Fusion-C	0.77	0.883	0.819	0.929
Fusion-D	0.711	0.832	0.783	0.902
Fusion-E	0.605	0.73	0.651	0.787
Fusion-F	0.462	0.632	0.53	0.729

Table 3: The verification rate and the identification accuracy for the eight systems and the different fusion schemes.

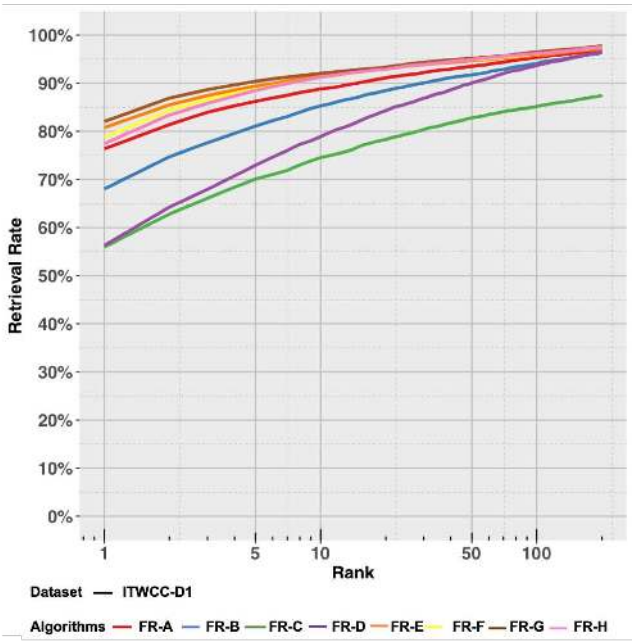


Figure 4: Identification performance of the eight systems when evaluated on the ITWCC−D1 dataset.

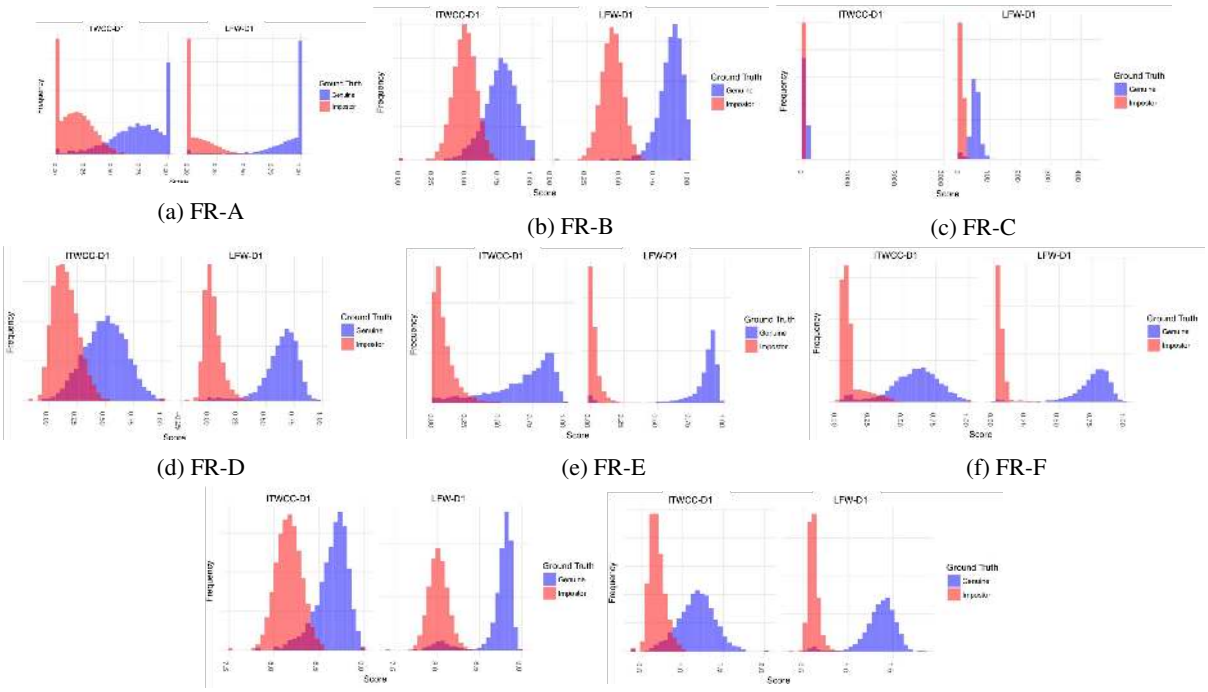
cating poorer performance. Comparing the distributions for each algorithm between ITWCC−D1 and LFW−D1, it is clear that there is a negative bias towards the ITWCC−D1 set. To illuminate the bias further, one only has to examine Figures 6 and 7, which distinctly indicates that these systems perform poorer on children than on adults. The level of negative bias is dependent on the face recognition algorithm. FR-E demonstrates the smallest bias for the verification experiment and FR-G exhibits the smallest bias in the identification experiments.

7. Conclusion

A clear bias exists when comparing performance measures between child and adult face datasets. The existence of bias as indicated by this work can be considerable and warrants a deeper understanding of the implications of face recognition for children. The number of scientific articles on this topic is demonstrably low, however, with the contribution from this work to extend and balance one of the only public child datasets, ITWCC, we believe more research is forthcoming.

We indicate that fusion results outperform individual systems. However, this work did not investigate optimizing the schemes for score-level fusion. The optimal scheme may well be a fusion between two of the reported systems instead of an all fusion approach taken here. The authors believe that a deeper investigation into fusion is warranted based on the results presented. Further, taking the lead from [4] a deep dive into retraining an open-sourced algorithm will yield a better performing algorithm; especially when fused with a COTS solution as indicated in [4].

The authors acknowledge that using the LFW dataset for the unconstrained adult corpus is challenging because we cannot guarantee that this prolific dataset was not used for training of the evaluated systems. However, NIST has documented the performance enhancements of face recognition over the last decade and the performance of the systems on LFW is consistent with reported results.

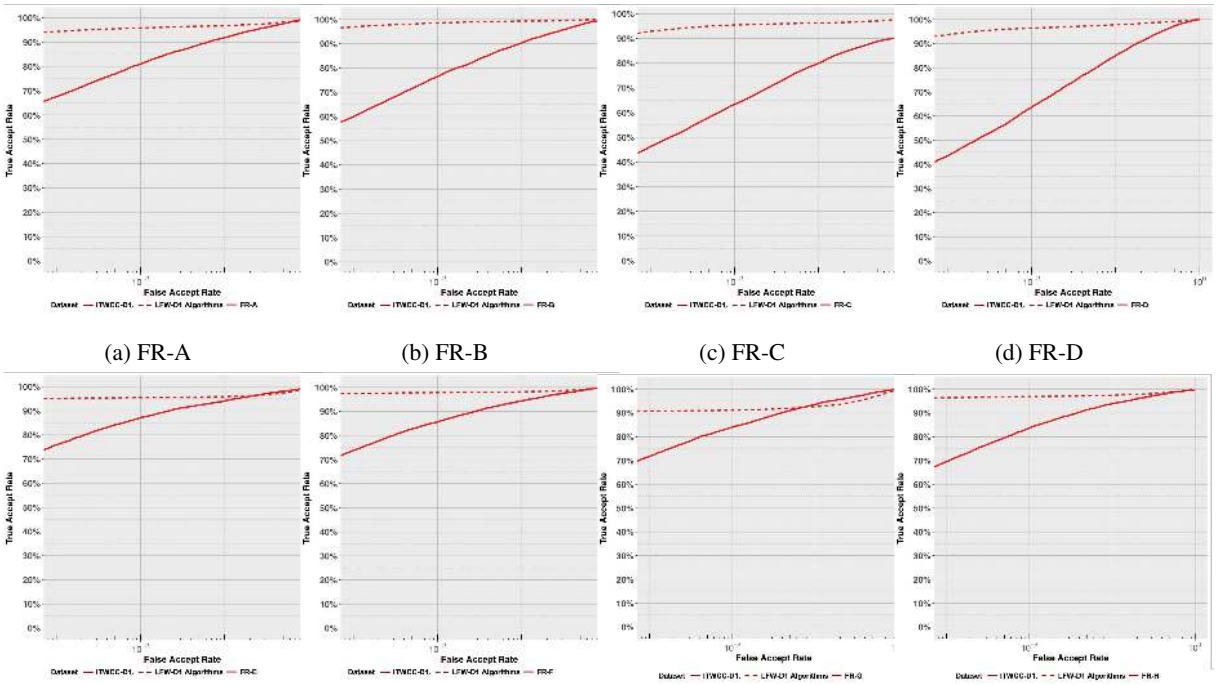


(a) FR-A (b) FR-B (c) FR-C

(d) FR-D (e) FR-E (f) FR-F

(g) FR-G (h) FR-H

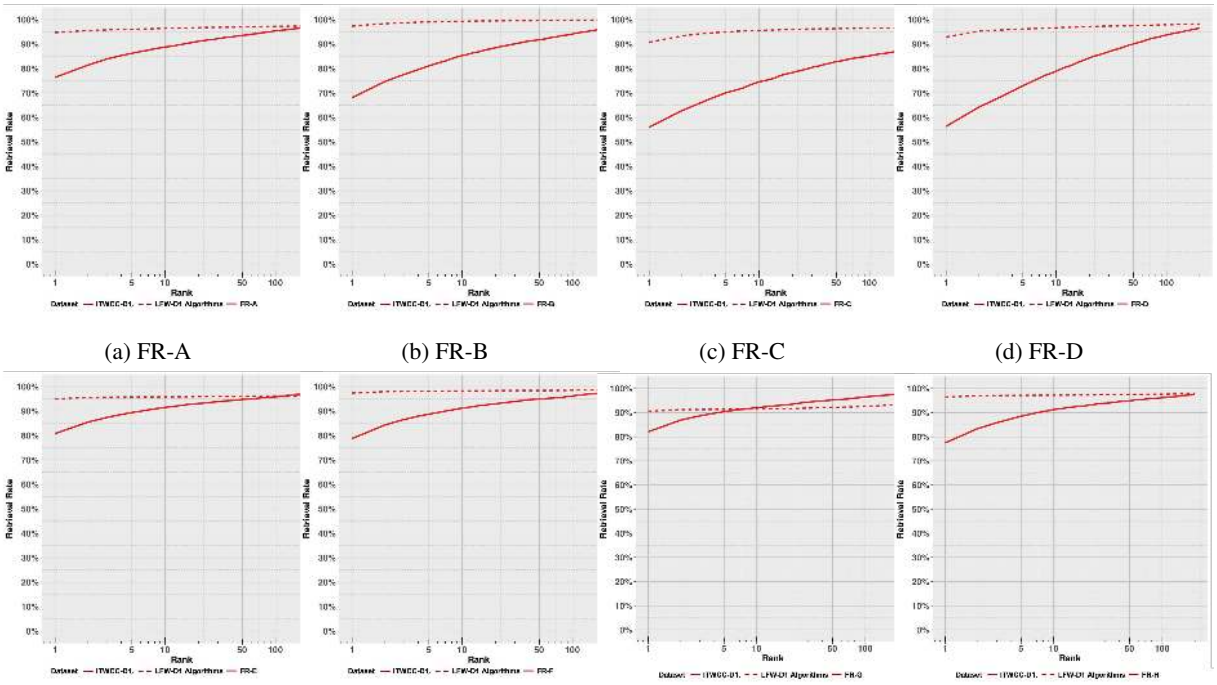
Figure 5: Genuine and imposter score distributions of the eight face recognition systems when evaluated against the ITWCC−D1 and the LFW−D1 datasets.



(a) FR-A (b) FR-B (c) FR-C (d) FR-D

(e) FR-E (f) FR-F (g) FR-G (h) FR-H

Figure 6: The verification performance of each algorithm on the ITWCC−D1 and LFW−D1 datasets.



(a) FR-A (b) FR-B (c) FR-C (d) FR-D

(e) FR-E (f) FR-F (g) FR-G (h) FR-H

Figure 7: The identification performance of each algorithm on the ITWCC−D1 and LFW−D1 datasets.

References

[1] Pratichi Basak, Saurabh De, Mallika Agarwal, Aakarsh Malhotra, Mayank Vatsa, and Richa Singh. Multimodal biometric recognition for toddlers and pre-school children. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 627–633, Oct 2017.

[2] Lacey Best-Rowden, Yovahn Hoole, and Anil K. Jain. Automatic Face Recognition of Newborns, Infants, and Toddlers: A Longitudinal Evaluation. In Arslan Brmme, Christoph Busch, Christian Rathgeb, and Andreas Uhl, editors, *Biosig 2016*, pages 87–98, Bonn, 2016. Gesellschaft fr Informatik e.V.

[3] Samarth Bharadwaj, Himanshu S. Bhatt, Richa Singh, Mayank Vatsa, and Sanjay K. Singh. Face Recognition for Newborns: A Preliminary Study. In *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, Sept 2010.

[4] Debayan Deb, Neeta Nain, and Anil K. Jain. Longitudinal Study of Child Face Recognition. In *2018 International Conference on Biometrics (ICB)*, pages 225–232, Feb 2018.

[5] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, Dec 2014.

[6] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Yuan Yao, and Shaogang Gong. Interestingness prediction by robust learning to rank. In *ECCV*, 2014.

[7] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In *European Conference on Computer Vision*. Springer, 2016.

[8] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. (07-49), October 2007.

[9] Huaizu Jiang and Erik Learned-Miller. Face detection with the faster r-cnn. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 650–657, May 2017.

[10] Frederick K. Kozak, Juan C. Ospina, and Marcela F. Cardenas. *Characteristics of normal and abnormal postnatal craniofacial growth and development*.

[11] Dana Michalski, Sau Yee Yiu, and Chris Malec. The Impact of Age and Threshold Variation on Facial Recognition Algorithm Performance Using Images of Children. In *2018 International Conference on Biometrics (ICB)*, pages 217–224, Feb 2018.

[12] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[13] Alastair Partington. *An industrial perspective on biometric age factors*. 2013.

[14] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[15] Karl Ricanek, Shivani Bhardwaj, and Michael Sodomsky. A Review of Face Recognition Against Longitudinal Child Faces. In Arslan Brmme, Christoph Busch, Christian

Rathgeb, and Andreas Uhl, editors, *BIOSIG 2015*, pages 15–26, Bonn, 2015. Gesellschaft fr Informatik e.V.

[16] Karl Ricanek and Tamirat Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345, April 2006.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015.