The Watchlist Imbalance Effect in Biometric Face Identification: Comparing Theoretical Estimates and Empiric Measurements

Pawel Drozdowski, Christian Rathgeb, Christoph Busch
da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
name.lastname@h-da.de

Abstract

Recently, different research groups have found that the gallery composition of a face database can induce performance differentials to facial identification systems in which a probe image is compared against up to all stored reference images to reach a biometric decision. This negative effect has been referred to as "watchlist imbalance effect" by the researchers and exhibits high relevance in real applications of biometrics, most prominently in identification searches against criminal databases and blacklists.

In this work, we conduct a detailed analysis of said effect. In particular, we compare empiric observations with theoretical estimates, based on the verification performance across demographic groups and the composition of the used gallery. The experimental evaluations are conducted by systematically varying the size and demographic composition of a cleaned subset of the academic MORPH database and utilising the state-of-the-art open-source ArcFace face recognition system.

1. Introduction

Automated systems (including biometrics) are increasingly used in decision making processes within various domains. In recent years, substantial media coverage of systemic biases inherent to such systems have been reported and hotly debated. In this context, a biased algorithm produces statistically different outcomes (decisions) for different groups of individuals, *e.g.* based on sex, age, and race [9]. Regarding biometric recognition, this means that false-positive identification and/or false-negative identification error rates can differ across demographic groups. The aftermath of demographic differentials can be unsettling – for example, a large study has reported disproportionally high arrest and search rates of African Americans based on decisions made by automatic facial recognition software [13]. In fact, the vast majority of works on measuring or achieving fairness in biometric systems is focused on facial recognition [8], although works for other biometric characteristics do exist (see *e.g.* [26, 6, 12, 24]). Attention has been brought to face since in that biometric modality, performance differentials mostly fall across points of sensitivity (*e.g.* race, sex), see figure 1 for examples of facial images of different demographic groups. Due to the enormous scope and scale of deployments of biometric facial (and other) recognition technologies (see *e.g.* [34, 11, 33, 27]), ensuring equitable treatment for all individuals and demographic groups is considered to be one of the most critical challenges in the area of biometrics [30, 28, 22].

Figure 1: Examples of different demographics which might exhibit performance differentials in a facial recognition system (images taken from a publicly available research database [23])

Demographic performance differentials can lead to differential outcomes which can be measured and quantified using standardised metrics [17, 21]. Additionally, metrics to measure the fairness (or equity) of a biometric system have been proposed by different research groups, *e.g.* in [31, 4, 14]. Fairness is related to the consequences of differential outcomes, *i.e.* mapping the response/behaviour of a (biometric) algorithm onto an application. In one of the largest evaluations of performance differentials in face recognition, large (orders of magnitude) differential outcomes have been observed for numerous commercial algorithms submitted to the NIST Face Recognition Vendor Test (FRVT) benchmark [15]. Many researchers have reported similar effects on various datasets, which motivated the proposal of numerous bias mitigation approaches, see [8] for a recent survey. However, it is important to note that the vast majority of works that focused on measuring demographic performance and outcome differentials in face recognition solely considered biometric verification (one-to-one comparison), with a few notable exceptions, *e.g.* [15, 25].

If a biometric system is operated in identification mode (one-to-many search), differentials can result from an uneven composition of demographics in the gallery which may be referred to as *watchlist imbalance effect* [32]. Specifically, false matches can occur with significantly higher probabilities within the same demographic group [17]. Consequently, false matches become more likely in biometric identification trials for demographic groups that are overly represented in a gallery. This effect has also been mentioned in [25].

In this work, we conduct a detailed analysis of the watchlist imbalance effect. We perform an empiric evaluation on the academic MORPH face database using a well-known face recognition system, which we subsequently compare and discuss w.r.t. theoretical estimates based on a recently proposed model. To this end, identification performance in terms of false-positive errors is measured for subsets of the database of systematically varying in size and demographic composition regarding the subjects' sex and skin colour. The results obtained from theoretical and empirical measurements are compared and discussed.

The rest of this paper is organised as follows: section 2 describes the watchlist imbalance effect in biometric systems. Section 3 summarises the experimental setup of this work. Results are reported and discussed in section 4. Conclusions are drawn in section 5.

2. The Watchlist Imbalance Effect

Biometric systems typically operate in one of two ways, namely verification and identification as defined in [21]): *biometric verification*, referring to the "process of confirming a biometric claim through biometric comparison"; *biometric identification*, referring to the "process of searching against a biometric enrolment database to find and return the biometric reference identifier(s) attributable to a single individual". In the latter case, two main scenarios can be distinguished: closed-set identification, for which all individuals known to have a reference in the enrolment database are enrolled in the system, and open-set identification, for which the biometric capture subject may or may not have a reference in the enrolment database.

In many cases, an exhaustive search (*i.e.* comparing a probe against all the enrolled subjects) is required in order to reach a decision. This is because biometric data has no inherent logical ordering, the samples acquired from the same subject (even within short time intervals) are almost never exactly identical (*i.e.* they are fuzzy), and the biometric feature vectors are typically high-dimensional. However, this naïve approach quickly runs into two non-trivial problems: as the number of enrolled subjects increases, the system response times become gradually slower, thus requiring optimisations and/or investment into larger hardware architectures; the probability of running into false positives increases. The latter issue makes biometric identification generally more challenging compared to biometric verification [7]. In the formulas below, a typical (in numerous practical deployments of biometric identification) exhaustive search-based retrieval is considered, *i.e.* without workload reduction strategies such as binning based on demographic attributes[1].

The probability of at least one false positive identification event occurring in an identification scenario, *i.e.* the false-positive identification error rate (FPIR), can according to Daugman [3], be estimated as:

$$P_N = 1 - (1 - P_1)^N, \tag{1}$$

where N is the number of enrolled subjects and P_1 the false positive probability of a one-to-one template comparison (verification). Equation 1 assumes that the false positive probability of a one-to-one comparison, *i.e.* false match rate (FMR), is equal across all enrolled subjects. This is an unrealistic assumption in general. In order to be able to account for such cases in the following descriptions and equations, let $P_1(x, y)$ be the false positive probability of a one-to-one comparison between subjects from demographic groups x and y, and $\mathbb{D}$ be the set of all the considered demographic groups.

Recently, several large empiric studies, [15, 17, 32] have shown that for face recognition false matches generally occur with higher probability within the same demographic groups, *i.e.* $\forall x, y \in \mathbb{D} : P_1(x, x) > P_1(x, y) \land x \neq y$. Furthermore, the probability of false matches within respective groups may also differ, *i.e.* $\exists x, y \in \mathbb{D} : P_1(x, x) \neq P_1(y, y) \land x \neq y$. In this context, Howard *et al.* [17] introduced the conceptual framework defining the terms "differential performance" and "differential outcomes" w.r.t. both false-positive and false-negative errors; this framework is

[1] Note, that the described effects due to gallery imbalance are also expected to appear when such strategies, *e.g.* binning, are applied.

quickly gaining acceptance in the field and is expected to form the basis of the ISO/IEC IS 19795-10 [20] which is currently under development.

Sirotin *et al.* [32] were the first to highlight the theoretical consequences of FMR (P_1) differentials within and between demographic groups to FPIR (P_N) for different demographic groups. They further demonstrated fact that these FPIR differentials will be observed as a consequence of gallery imbalance, not FMR imbalance. It is important to note that demographically unbalanced databases are likely common in numerous real-world applications [8], for example due to historical and societal biases being propagated [13].

The formula for probability of at least one false positive occurring in an identification for a probe of demographic group x against a gallery with *exactly two* demographic groups can be derived as:

$$P_N(x) = 1 - (1 - P_1(x,x))^{N_x} \cdot (1 - P_1(x,y))^{N_y}, \quad (2)$$

where N_x and N_y are the numbers of enrolled subjects belonging to demographic group x and y, respectively. This equation can be further extended to estimating the probability of at least one false positive occurring in an identification for a probe of demographic group x against a gallery with *more than two* demographic groups as follows:

$$P_N(x) = 1 - (1 - P_1(x,x))^{N_x} \cdot \prod_{y \in \mathbb{D}} (1 - P_1(x,y))^{N_y}, \quad (3)$$

where the gallery consists of subjects belonging to the same demographic group (x) and a set of $\mathbb{D}$ other demographic groups.

In summary, works in this area have found following factors to be relevant w.r.t. the demographic differentials in a biometric identification scenario:

- The probability of errors in one-to-one comparisons within the demographic groups.
- The probability of errors in one-to-one comparisons across the demographic groups.
- The demographic composition of the gallery.

Pereira *et al.* [4] and Grother *et al.* [14] suggested that demographic differentials in different scenarios may be expressed and reasoned about using compound metrics (*e.g.* a ratios, sums, and products) taking into account the observed error rates. While initially proposed for biometric verification, such metrics can also be adapted for biometric identification; for instance, a ratio $\frac{\max_{\forall d \in \mathbb{D}} P_N(d)}{\min_{\forall d \in \mathbb{D}} P_N(d)}$ might be computed [4, 14]. The ratio is equal to 1 if no differentials were observed, *i.e.* the system being equitable across the demographic groups; higher values show the system to be inequitable across the demographic groups and express the effect magnitude of the observed differentials. This thus provides a simple and intuitive way of reasoning about inequity across the demographic groups.

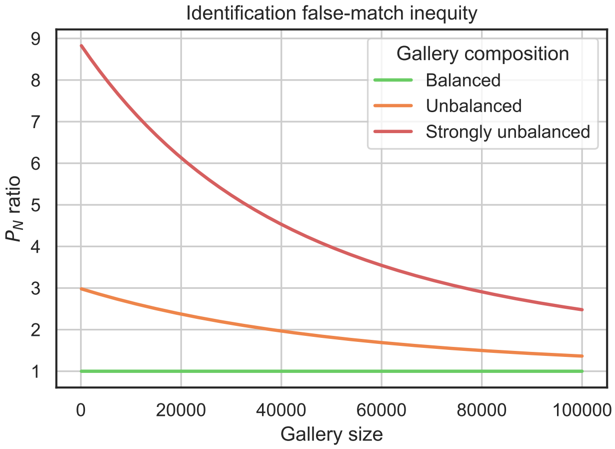
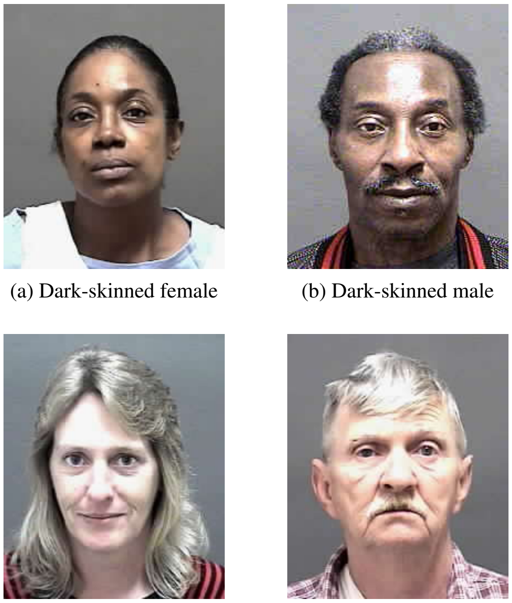


Figure 2: Illustration of the watchlist imbalance effect

For the purpose of illustrating the aforementioned effect in biometric identification, consider a system with two demographic groups (x and y) and assume a low $P_1(x,x) = 0.5 \cdot 10^{-3}$, $P_1(y,y) = 0.5 \cdot 10^{-3}$, and $P_1(x,y) = 10^{-5}$. The fact that P_1 across the groups is much lower is to be expected as described previously in this section. Furthermore, since identical error rates can be expected for both demographic groups, the system could, in aggregate, be considered equitable in the biometric verification scenario. For the identification scenario, consider for the purposes of this example three hypothetical gallery compositions: 1) *balanced*, with a 50% – 50% split between x and y, 2) *unbalanced*, with a 75% – 25% split between x and y, 3) *strongly unbalanced*, with a 90% – 10% split between x and y. Figure 2 plots the aforementioned inequity measure as a function of the size of the gallery (*i.e.* the number of template comparisons in biometric identification). In case of a balanced gallery, the P_N is identical for both demographic groups, *i.e.* the system is equitable. However, as the gallery becomes unbalanced, the P_N strongly increases for the demographic group which is predominantly represented in the gallery.

3. Experimental Setup

The academic MORPH dataset [29] was selected for the experimental evaluation. Although initially aimed at evaluating ageing in face recognition systems, this dataset has recently been used in numerous studies relating to demographic differentials [2, 25] due to its large size, relatively constrained image acquisition conditions, and the presence of groundtruth labels (from public records) for sex, race, and age of the subjects. Example images from the dataset are shown in figure 3.



(a) Dark-skinned female

(b) Dark-skinned male

(c) Light-skinned female

(d) Light-skinned male

Figure 3: Example images from the used dataset

For the experiments, in order to minimise the confounding factors such as head pose and image quality, a subset of the images was selected based on approximate conformance with ICAO requirements for passport images [19]; furthermore, twins, duplicates, and images with erroneous labels were removed as in [1]. The resulting subset contains around 13,000 subjects with the demographic distribution as shown in figure 4. The subjects whose reference data was enrolled into the gallery were selected randomly, keeping in mind the required numbers for the used gallery compositions (the balanced, unbalanced, and strongly unbalanced splits w.r.t. sex and skin colour, see section 2). The galleries of different sizes were created incrementally in such a way, that larger galleries contain all the data present in the corresponding smaller galleries (*e.g.* a gallery for a given demographic split with the size of 200 subjects contains all the subjects from the corresponding 100-subject gallery, plus additional 100 subjects *etc.*). This procedure and corresponding recognition experiments were repeated using 10-fold cross-validation; 95% confidence intervals are shown in the plots.

Face recognition was carried out using a popular face recognition system (ArcFace [5]), which achieves excellent biometric performance in popular large-scale face recognition benchmarks. The open-source code and pre-trained model "LResNet100E-IR,ArcFace@ms1m-refine-v2" provided by the authors were used[2]. The model is based on a ResNet-100 backbone and the additive angular margin loss function; its training was conducted using a cleaned subset of the MS-Celeb-1M dataset [16]. The evaluations conducted in this paper follow protocols and metrics standardised by ISO/IEC [21].

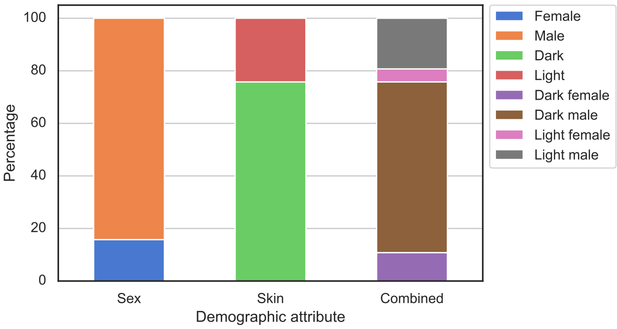


Figure 4: Overview of the demographics in the used dataset

4. Results and Discussion

A baseline verification experiment was conducted to establish a decision threshold corresponding to a fixed FMR value of 0.1% which is recommended as a security level by FRONTEX in several operationally relevant scenarios [10]. Table 1 shows the corresponding observed error rates for comparisons within and across demographic groups[3].

Table 1: Verification performance (in %)

(a) Sex

FMR(F, F)	FMR(M, M)	FMR(F, M)
0.414	0.115	0.031

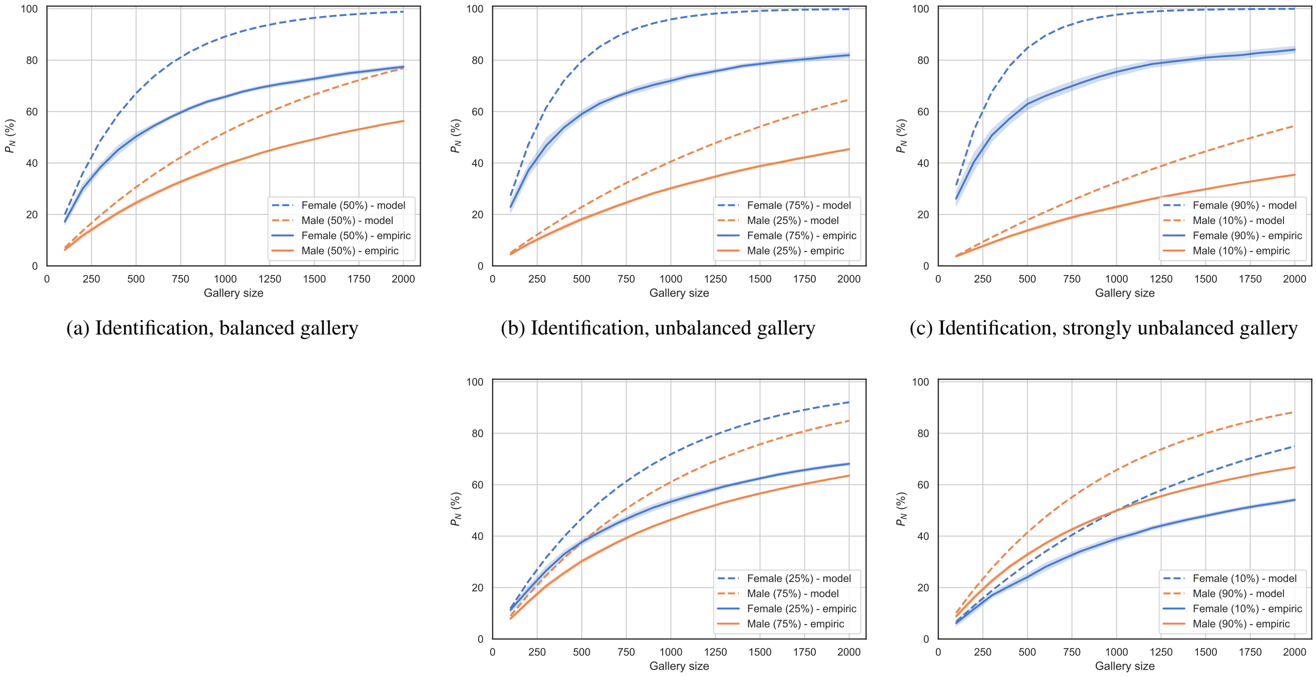
(b) Skin

FMR(D, D)	FMR(L, L)	FMR(D, L)
0.166	0.021	0.004

In the results of the biometric verification experiment, it can be observed that the false-match probabilities were almost 4 times higher within the female group than within the male group, and around 8 times higher within the dark-skinned group than within the light-skinned group. The probability of false matches across the demographic groups (both sex and skin) was an order or two orders of magnitude lower than within the respective groups. Those results mirror many previous results in this area [8].

[2] https://github.com/deepinsight/insightface

[3] Note that the number of template comparisons varies across the depicted six combinations of demographic attributes.



(a) Identification, balanced gallery

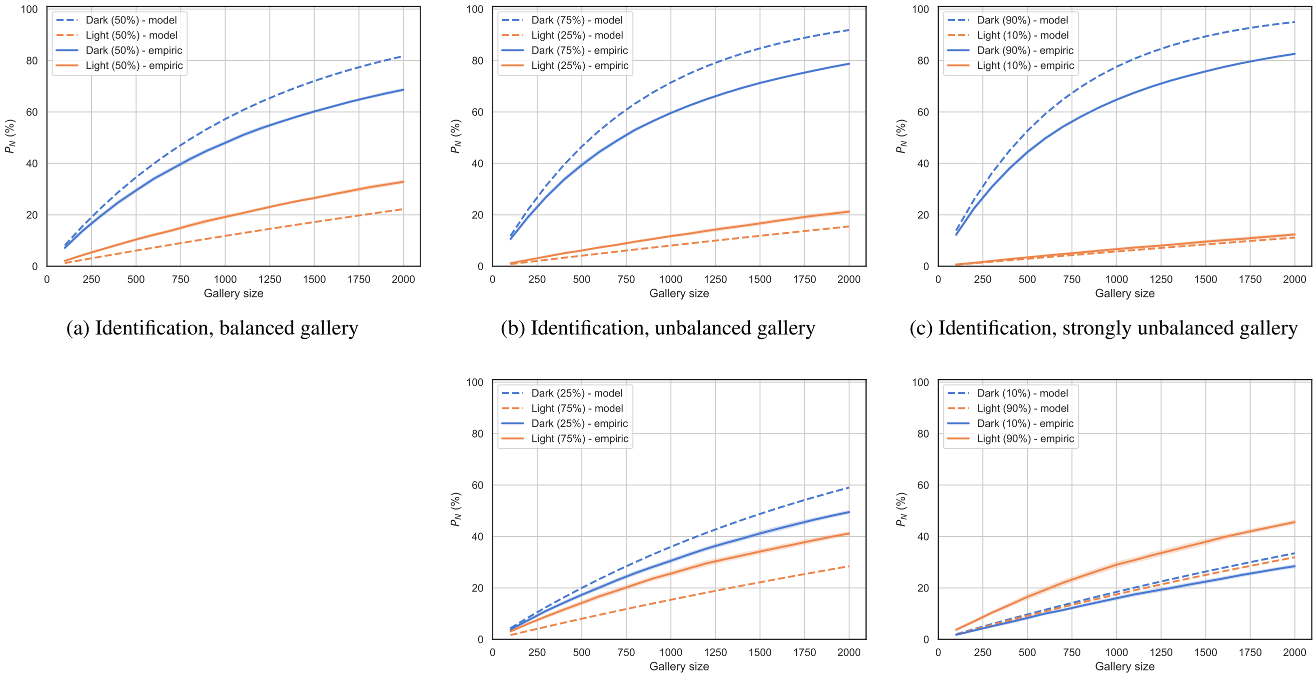
(b) Identification, unbalanced gallery

(c) Identification, strongly unbalanced gallery

(d) Identification, unbalanced gallery

(e) Identification, strongly unbalanced gallery

Figure 5: Results for varying gallery size and composition based on subjects' sex



(a) Identification, balanced gallery

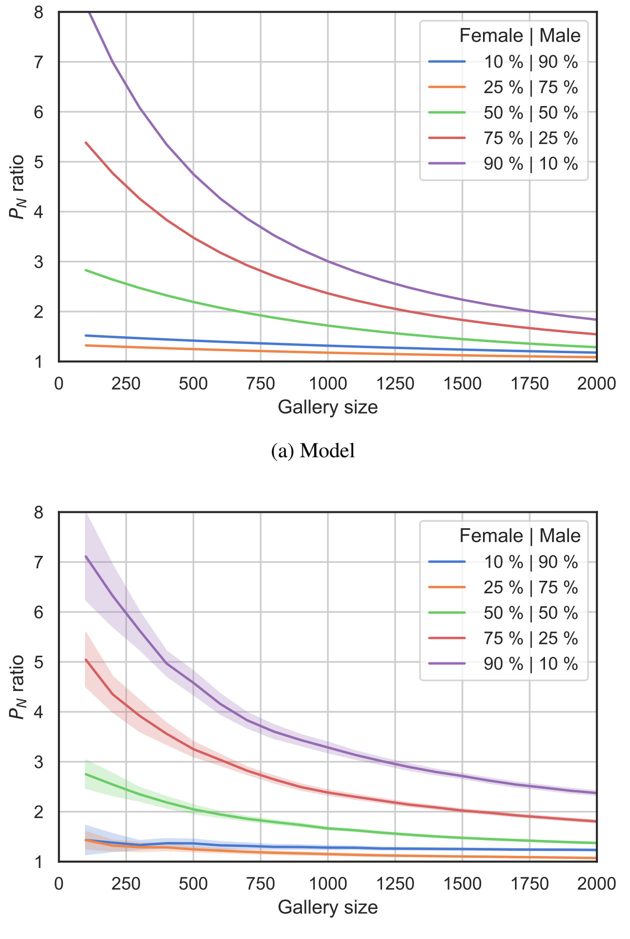
(b) Identification, unbalanced gallery

(c) Identification, strongly unbalanced gallery

(d) Identification, unbalanced gallery

(e) Identification, strongly unbalanced gallery

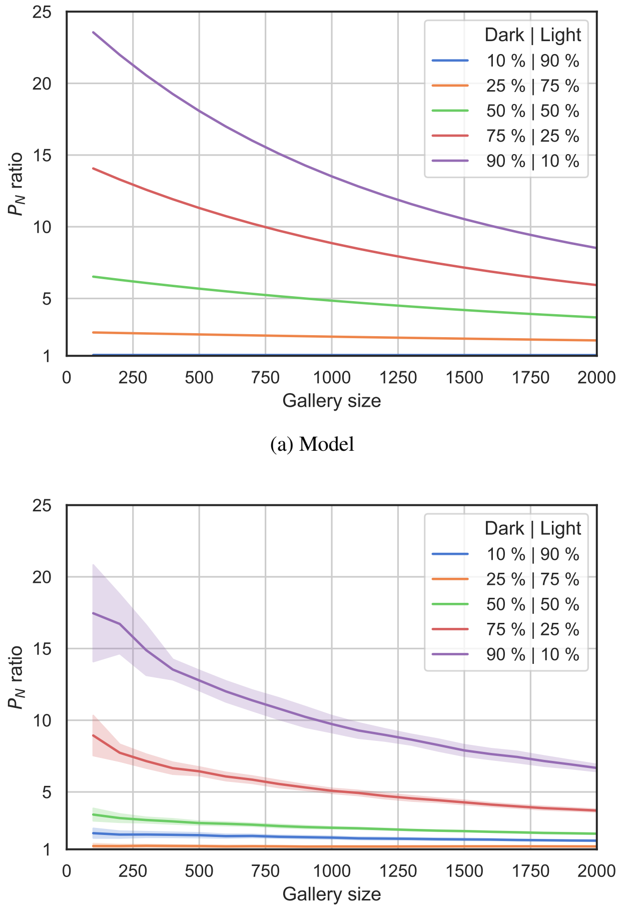
Figure 6: Results for varying gallery size and composition based on subjects' skin tone



(a) Model

(b) Empiric

Figure 7: False-match inequity for varying gallery size and composition based on subjects' sex



(a) Model

(b) Empiric

Figure 8: False-match inequity for varying gallery size and composition based on subjects' skin

Figures 5 and 6 show[4] the results of the experiments. Specifically, a balanced, as well as an unbalanced and strongly unbalanced gallery composition with 75% and 90% subjects belonging to one of the demographic groups, respectively, were used. Interesting effects can be observed, especially for the unbalanced galleries:

Gallery balanced as expected, in this case, the inequity present in the verification is perpetuated – P_N for females remains higher than for males. Same effect is observed w.r.t. dark-skinned and light-skinned individuals.

Gallery unbalanced towards female or dark-skinned this case exacerbates the inequity present in the verification. In figures 5b and 5c, P_N for females increases very quickly with the growing size of the gallery. Conversely, P_N for males increases quite slowly. Similar effect is observed in figures 6b and 6c, where P_N quickly increases for dark-skinned individuals, while its increase is very slow for light-skinned individuals. The gap between the two groups (*i.e.* the inequity) is very large.

Gallery unbalanced towards males or light-skinned although P_1 within the male group is much lower than within the female group, P_N for males is nearly identical to that of females in figure 5d, whereas for a strongly unbalanced gallery shown in figure 5e, it even exceeds that of the females. An analogous effect can be observed in figures 6d and 6e for P_N of light-skinned individuals.

Tables 2 and 3 show the differences between the theoretical estimates following the model described in section 2 with empiric results obtained in section 4. The theoretical

[4] This type of visualisation was used in [32] to illustrate the theoretical estimates described in section 2.

Table 2: Differences (in percentage points) between theoretical estimates and average empiric results based on subjects' sex

Gallery size	Gallery Composition (Female / Male)				
	10% / 90%	25% / 75%	50% / 50%	75% / 25%	90% / 10%
100	0.5 / 1.3	0.6 / 1.1	2.7 / 0.8	4.3 / 0.5	5.1 / 0.2
500	5.2 / 8.6	9.3 / 7.4	16.8 / 6.1	20.6 / 4.7	21.8 / 4.1
1000	11.0 / 15.8	18.5 / 14.7	23.4 / 12.4	23.9 / 10.4	22.2 / 9.5
1500	16.7 / 20.0	22.6 / 19.1	23.7 / 17.4	20.6 / 15.3	18.7 / 14.7
2000	20.9 / 21.5	24.0 / 21.3	21.4 / 20.5	17.9 / 19.3	15.9 / 19.0

Table 3: Differences (in percentage points) between theoretical estimates and average empiric results based on subjects' skin

Gallery size	Gallery Composition (Dark / Light)				
	10% / 90%	25% / 75%	50% / 50%	75% / 25%	90% / 10%
100	0.3 / -1.8	0.6 / -1.4	1.0 / -0.9	1.2 / -0.4	1.6 / -0.1
500	1.4 / -7.4	2.7 / -6.1	5.1 / -4.3	7.2 / -2.0	8.3 / -0.6
1000	2.5 / -11.6	5.4 / -10.2	9.3 / -7.4	11.8 / -3.7	12.8 / -0.9
1500	3.9 / -12.9	7.6 / -12.0	11.9 / -9.4	13.5 / -4.8	13.6 / -1.1
2000	5.0 / -13.7	9.5 / -12.7	13.1 / -10.6	13.1 / -5.8	12.4 / -1.2

model generally tends to overestimate the P_N value somewhat. The empiric P_N results for the light-skinned individuals are the only ones higher than their corresponding theoretical estimates. Possible reasons for it may include the low initial FMR for this group or the presence of Doppelgängers (*i.e.* random zero-effort impostors with high comparison scores) in the dataset. This effect should be investigated closer in future studies of this subject. These uncertainties notwithstanding, the inequities predicted by the model for the identification scenario (figures 7a and 8a) very closely resemble those obtained in the actual experiments (figures 7b and 8b). In other words, while the theoretical model may exhibit some inaccuracies in the precise prediction of the P_N values, it very accurately predicts the relative differences of P_N (*i.e.* the inequity) across the considered demographic groups.

Depending on the gallery composition, the false-match inequity for a given group in the evaluated identification scenario can be higher than in the verification case. In the evaluated case, females were disadvantaged in the verification, which was further exacerbated (as evidenced by the purple and red lines in figure 7b) in scenarios where gallery contained more females. An analogous effect was observed for dark-skinned individuals and galleries unbalanced towards them (see the purple and red lines in figure 8b). On the other hand, a gallery containing mostly males has disadvantaged this group in identification despite their comparatively low FMR in verification (as evidenced by figure 5e and the blue line in figure 7b). Unbalancing the gallery towards light-skinned individuals has likewise led to this group being disadvantaged in identification despite their comparatively low FMR in verification (see figure 6e and the blue line in figure 8b).

5. Conclusion and Future Work

Due to the rapid technological progress, scale, and scope of the deployments of automatic decision systems (including biometrics), there exists a massive potential of both beneficial and harmful applications. The latter includes (unintentional) discrimination based on individuals' demographic properties. The notion of "fairness" is a very complicated and nuanced topic, spanning across ethical, legal, social, technological, and other disciplines [18]. In the context of biometric recognition systems, a key technological component in this broader debate is the act of quantifying demographic differentials and trade-offs, as well as identifying sources thereof. The results and insights obtained by researchers in this area will help informed discussions and policy decisions by the relevant stakeholders, thus potentially contributing to more equitable quality of service and usability for all the system users. Beyond contributions of such technical evaluations, potential policy decisions regarding this topic will inevitably require deep considerations from a very broad and interdisciplinary perspective, including but not limited to ethics, law, and sociology.

This work conducted a systematic empiric evaluation of false-positive demographic differentials in biometric identification under varying size and demographic composition of the enrolment database. The obtained results were analysed and discussed in detail w.r.t. theoretical estimates stemming from a recently proposed model. The evaluation concentrated on the so-called "watchlist imbalance effect" [32], *i.e.* the fact of the gallery being unbalanced towards a certain demographic group (*e.g.* containing many more males than females). This effect has been shown to exhibit a profound influence on the equity of biometric identification systems. Furthermore, it demonstrates that demographically equitable biometric verification systems do not necessarily guarantee demographically equitable identification systems. Future works may extend the scope of this preliminary study w.r.t. generalisability, *e.g.* by considering more datasets and face recognition systems, as well as decision thresholds corresponding to different system security settings. Furthermore, should a larger (in terms of the number of available subjects) dataset of high-quality images and annotations become available, it would enable assessing the theoretical predictions and empiric measurements for intersectional demographic groups (*i.e.* combining the sex, skin colour, and potentially other demographic attributes).

The described effects can have a profound impact on applications of biometric systems, as demographic distributions in real databases, *e.g.* watchlists, may be imbalanced (*e.g.* due to certain policies, laws, or historical conditions [13]). Thus, future works may consider ways of mitigating the "watchlist imbalance effect", *e.g.* by application of dynamic decision thresholds depending on the demographic properties of the compared templates or artificially balancing the gallery, *e.g.* with synthetic data. In cases where equitable biometrics are striven for, a new trade-off becomes apparent – optimising for overall accuracy or equity across demographic groups?

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] V. Albiero and K. W. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *British Machine Vision Conference (BMVC)*, pages 1–13. BMVA, September 2020.

[2] V. Albiero, K. S. Krishnapriya, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 81–89. IEEE, March 2020.

[3] J. Daugman. Biometric decision landscapes. Technical Report UCAM-CL-TR-482, University of Cambridge - Computer Laboratory, January 2000.

[4] T. de Freitas Pereira and S. Marcel. Fairness in biometrics: a figure of merit to assess biometric verification systems. *arXiv preprint arXiv:2011.02395v2*, March 2021.

[5] J. Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *Computing Research Repository (CoRR)*, pages 1–11, January 2018.

[6] P. Drozdowski, B. Prommegger, G. Wimmer, R. Schraml, C. Rathgeb, A. Uhl, and C. Busch. Demographic bias: A challenge for fingervein recognition systems? In *European Signal Processing Conference (EUSIPCO)*, pages 825–829. EURASIP, April 2020.

[7] P. Drozdowski, C. Rathgeb, and C. Busch. Computational workload in biometric identification systems: An overview. *IET Biometrics*, 8(6):351–368, November 2019.

[8] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *Transactions on Technology and Society (TTS)*, 1(2):89–103, June 2020.

[9] Editorial. Algorithm and blues. *Nature*, pages 1–1, September 2016.

[10] eu-LISA. Best practice technical guidelines for automated border control ABC systems. Technical Report TT-02-16-152-EN-N, European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, September 2015.

[11] European Commission. Smart borders. `https://ec.europa.eu/home-affairs/what-we-do/policies/borders-and-visas/smart-borders_en`, 2018. Last accessed: August 5, 2021.

[12] G. Fenu, H. Lafhouli, and M. Marras. Exploring algorithmic fairness in deep speaker verification. In *Computational Science and Its Applications (ICCSA)*, pages 77–93. Springer, September 2020.

[13] C. Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, October 2016.

[14] P. Grother. Demographic differentials in face recognition algorithms. In *Virtual Events Series – Demographic fairness in biometric systems*. EAB, March 2021. `https://eab.org/files/videos/2021-03-15_EAB-virtual-events-series/03-Grother-NIST-210315.pdf`.

[15] P. Grother, M. Ngan, and K. Hanaoka. Ongoing face recognition vendor test (FRVT) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology, December 2019.

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision (ECCV)*, pages 87–102. Springer, October 2016.

[17] J. J. Howard, Y. B. Sirotin, and A. R. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *International Conference on Biometric, Theory, Applications and Systems (BTAS)*. IEEE, September 2019.

[18] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Conference on Fairness, Accountability, and Transparency (FAT)*, pages 49–58. ACM, January 2019.

[19] International Civil Aviation Organization. Machine readable passports – part 9 – deployment of biometric identification and electronic storage of data in eMRTDs. Technical Report 9303, ICAO, 2015.

[20] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC WD 19795-10 information technology – biometric performance testing and reporting – part 10: Quantifying biometric system performance variation across demographic groups. unpublished draft.

[21] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 19795-1:2021. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework*. International Organization for Standardization and International Electrotechnical Committee, April 2021.

[22] A. K. Jain, D. Deb, and J. J. Engelsma. Biometrics: Trust, but verify. *arXiv preprint arXiv:2105.06625v2*, May 2021.

[23] K. Kärkkäinen and J. Joo. FairFace: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[24] A. Krishnan, A. Almadan, and A. Rattani. Probing fairness of mobile ocular biometrics methods across gender on VISOB 2.0 dataset. In *International Conference on Pattern Recognition Workshops and Challenges (ICPR)*, volume 12668 of *Lecture Notes in Computer Science*, pages 229–243. Springer, January 2021.

[25] 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. *Transactions on Technology and Society (TTS)*, 1(1):8–20, March 2020.

[26] E. Marasco. Biases in fingerprint recognition systems: Where are we at? In *International Conference on Biometrics: Theory Applications and Systems (BTAS)*, pages 1–5. IEEE, September 2019.

[27] L. Pascu. Global biometrics market to surpass $45B by 2024, reports Frost & Sullivan. https://www.biometricupdate.com/202003/global-biometrics-market-to-surpass-45b-by-2024-reports-frost-sullivan, March 2020.

[28] C. Rathgeb, P. Drozdowski, N. Damer, D. C. Frings, and C. Busch. Demographic fairness in biometric systems: What do the experts say? *arXiv preprint arXiv:2105.14844v1*, May 2021.

[29] K. Ricanek and T. Tesafaye. MORPH: a longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition (FGR)*, pages 341–345. IEEE, April 2006.

[30] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yaday. Some research problems in biometrics: The future beckons. In *International Conference on Biometrics (ICB)*, pages 1–8. IEEE, June 2019.

[31] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, and I. Rahwan. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. *arXiv preprint arXiv:1912.01842*, December 2019.

[32] Y. B. Sirotin and A. R. Vemury. Demographic variation in the performance of biometric systems: Insights gained from large-scale scenario testing. In *Virtual Events Series – Demographic fairness in biometric systems*. EAB, March 2021. https://mdtf.org/publications/EAB2021-Demographics.pdf.

[33] Thales. DHS's automated biometric identification system IDENT - the heart of biometric visitor identification in the USA. https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/customer-cases/ident-automated-biometric-identification-system, January 2021. Last accessed: August 5, 2021.

[34] Unique Identification Authority of India. Aadhaar dashboard. https://www.uidai.gov.in/aadhaar_dashboard/. Last accessed: August 5, 2021.