Demographic Bias: A Challenge for Fingervein Recognition Systems?

P. Drozdowski†, B. Prommegger†, G. Wimmer†, R. Schraml†, C. Rathgeb†, A. Uhl†, and C. Busch†

† da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
† Multimodal Signal Processing and Security Lab, University of Salzburg, Austria

Abstract—Recently, concerns regarding potential biases in the underlying algorithms of many automated systems (including biometrics) have been raised. In this context, a biased algorithm produces statistically different outcomes for different groups of individuals based on certain (often protected by anti-discrimination legislation) attributes such as sex and age. While several preliminary studies investigating this matter for facial recognition algorithms do exist, said topic has not yet been addressed for vascular biometric characteristics. Accordingly, in this paper, several popular types of recognition algorithms are benchmarked to ascertain the matter for fingervein recognition. The experimental evaluation suggests lack of bias for the tested algorithms, although future works with larger datasets are needed to validate and confirm those preliminary results.

Index Terms—Biometrics, Fingervein Recognition, Bias

I. INTRODUCTION

Automated systems (including biometrics) are increasingly used in decision making processes within various domains, some of which have traditionally enjoyed strong antidiscrimination legislation protection (see e.g. [1]).

Fig. 1: Overview of the conducted experiments

In recent years, substantial media coverage of systemic biases inherent to several 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 ethnicity [2]. For biometric recognition specifically, it means that the score distributions and therefore the chances of false positives and/or false negatives may vary across the groups. This, in consequence, may impact the outcomes on the system/application level – for example, one recent study claimed disproportionally high arrest and search rates for certain groups based on decisions made by automatic facial recognition software [3].

Although some studies have approached bias measurements and ensuring fairness in various machine learning contexts (see e.g. [4] and [5]), for computer vision and biometrics in particular, this remains a nascent field of research. In [6], it was reported that facial recognition algorithms tend to exhibit higher biometric performance with individuals from ethnic groups corresponding to the area of development of the algorithm; presumably due to training data availability. In [7] and [8], some facial biometrics algorithms were shown to exhibit lower recognition and classification performances for certain groups of individuals (in particular women and non-white people), whereas [2] conducted a large-scale benchmark of commercial and academic algorithms controlling for various demographic (and other) attributes. Proof-of-concept studies into bias mitigation for facial soft biometric classification using neural networks were presented in e.g. [9] and [10].

While facial recognition is certainly the most widely covered and discussed biometric characteristic recently, also with some existing preliminary studies in the context of bias and fairness, this topic remains even less explored for other biometric characteristics. In general, algorithmic bias is considered (by some influential researchers) to be one of the, as of yet unresolved, challenges in biometric systems [12]. For fingervein specifically, a small study evaluating the biometric performance w.r.t. sex and age of the subjects (as a part of a paper presenting a new dataset) has been conducted in [13].

Intuitively, the demographic covariates could be proxies for certain anatomical features which might influence the performance of fingervein recognition systems (e.g. the thickness of the finger). Otherwise, currently no studies benchmarking the potential biases in fingervein algorithms have been reported in the scientific literature. In this paper, a benchmark is conducted to address the following two questions, see figure [1]:

- Do score distributions computed by fingervein recognition algorithms on disjoint groups of data instances (i.e. based on metadata attributes, such as subject sex, age, and others) exhibit statistically significant differences?
- Do these results persist across fundamentally different types of fingervein recognition algorithms?
The remainder of this paper is organised as follows: the experimental setup is described in section II. The results of the evaluation are presented in section III, while concluding remarks and a summary are given in section IV.

II. EXPERIMENTAL SETUP

In this section, the used datasets (subsection II-A) and data processing pipelines (subsection II-C) are presented.

A. Datasets

Four publicly available fingerprint databases with metadata labels were used. They are listed in table I, while example images are shown in figure 2. The datasets were chosen based on the presence of the metadata information (sex and age), as well as the presence of samples from both hands and different fingers.

![Example images from the chosen datasets](image)

**TABLE I: Summary of the chosen datasets**

| Database  | Subjects | Instances | Samples |
|-----------|----------|-----------|---------|
| MMCBNU [14] | 100      | 600       | 6000    |
| PLUS [13], [15] | 78       | 468       | 2340    |
| UTFVP [16]   | 60       | 360       | 1440    |
| VERA [17]    | 110      | 220       | 440     |

B. Feature Types

All experiments are executed using four fundamentally different types of vein recognition schemes:

1) **Vein pattern**: The first two of the used techniques, Maximum Curvature (MC [18]) and Principal Curvature (PC [19]), aim to separate the vein pattern from the background resulting in a binary image.

2) **Keypoints**: In contrast to the vein pattern techniques, key-point based techniques try to use information from the most discriminative points as well as considering the neighborhood and context information of these points by extracting key-points and assigning a descriptor to each key-point. We used a Scale-Invariant Feature Transform (SIFT [20]) based technique with additional key-point filtering. All details of this method are described in [21].

3) **Texture**: The used approach is an adapted version of [22]. It combines Log Gabor convolution magnitude and local binary patterns (LBP [23]).

4) **Deep learning**: LightCNN with triplet loss (LCNN [24]) is a small neural network which is trained using the triplet loss, a special loss function that enables the identification of subjects that were not included in the training set. By using a more advanced selection of the triplets (input images) for training (hard triplet online selection) and by omitting the supervised discrete hashing of the CNN outputs, the results could be improved w.r.t. [24]. It should be noted that the LCNN was trained according to its actual purpose, the recognition of individuals based on the captured vein images. If the net would be trained according to the chosen groups, such as sex or age, the results might look different.

**TABLE II: Biometric performance of the used methods**

| Dataset  | LCNN | LBP | MC | PC | SIFT |
|----------|------|-----|----|----|------|
| MMCBNU   | 2.2  | 1.3 | 1.8| 1.7| 2.0  |
| PLUS     | 4.5  | 3.6 | 0.5| 0.2| 0.8  |
| UTFVP    | 7.0  | 1.5 | 0.2| 0.4| 1.5  |
| VERA     | —    | 3.2 | 1.8| 2.3| 2.6  |

C. Data Processing Pipelines

The finger vein recognition tool-chain consists of the following components: (1) For finger region detection, finger alignment and ROI extraction an implementation that is based on [25] is used. (2) To improve the visibility of the vein pattern, High Frequency Emphasis Filtering (HFE) [26], Circular Gabor Filter (CGF) [27], and simple CLAHE (local histogram equalisation) [28] are used during pre-processing. (3a) For the simple vein pattern based feature methods, MC and PC, the binary feature images are compared using a correlation measure, calculated between the input images and in x- and y-direction shifted and rotated versions of the reference image as described in [29]. (3b) The SIFT based method applies feature extraction and comparison as proposed in [21], and (3c) the LBP based approach as described in [22], where the LDP features are replaced with LBP features from [23], respectively. (3d) For LCNN, the ROIs extracted in (1) have been resized to the required input size of 256×256. For separating training from input data, a 2-fold cross validation has been applied. Due to a limited number of samples, a training on the VERA data set was not possible. The biometric performance of the used methods is summarised in table II.

III. EVALUATION

The conceptual overview of the conducted experiments is shown in figure 1. The descriptive statistics (mean – µ and standard deviation – σ) of the score distributions w.r.t. the metadata-based groupings are computed in subsections III-A to III-C. Additionally, for comparison, table III shows supplementary files (e.g. comparison scores) are available for download at [http://wavelab.at/sources/Drozdowski20a/](http://wavelab.at/sources/Drozdowski20a/).
the same statistics for all template comparisons without any metadata-based grouping. The algorithms have different ranges of scores. In the context of the conducted bias benchmark, only intra-algorithm comparisons are meaningful – therefore, score normalisation is not required. The results are only listed where available (e.g. the VERA dataset only contains two samples per finger and therefore it is not possible to train a CNN using the triplet loss function).

TABLE III: Score distribution statistics without metadata-based grouping

| Dataset   | Algorithm | Score μ Genuine | Score σ Genuine | Score μ Imposter | Score σ Imposter |
|-----------|-----------|-----------------|-----------------|------------------|------------------|
| MMCBNU    | LCNN      | 0.72477         | 0.26189         | 0.12129          | 0.07989          |
| LBP       | 0.84812   | 0.78781         | 0.02009         | 0.00805          |
| MC        | 0.27205   | 0.12055         | 0.04891         | 0.01730          |
| PC        | 0.41059   | 0.30588         | 0.02066         | 0.01654          |
| SIFT      | 0.39081   | 0.01510         | 0.16538         | 0.02434          |
| PLUS      | LCNN      | 0.70367         | 0.25433         | 0.13798          | 0.08999          |
| LBP       | 0.80335   | 0.12704         | 0.16154         | 0.05897          |
| MC        | 0.25004   | 0.12464         | 0.03946         | 0.00874          |
| PC        | 0.41261   | 0.30209         | 0.03030         | 0.01361          |
| SIFT      | 0.35114   | 0.01119         | 0.14678         | 0.01196          |
| UTFVP     | LCNN      | 0.69913         | 0.32198         | 0.11472          | 0.08033          |
| LBP       | 0.84664   | 0.81038         | 0.01344         | 0.00471          |
| MC        | 0.23834   | 0.11789         | 0.03854         | 0.00733          |
| PC        | 0.41269   | 0.30209         | 0.03030         | 0.01361          |
| SIFT      | 0.31269   | 0.00937         | 0.14795         | 0.01062          |
| VERA      | LCNN      | 0.81926         | 0.78861         | 0.01101          | 0.00417          |
| LBP       | 0.24056   | 0.11351         | 0.04569         | 0.00947          |
| MC        | 0.38741   | 0.29499         | 0.03025         | 0.01156          |
| PC        | 0.23220   | 0.00562         | 0.14523         | 0.00822          |

A. Sex

In this experiment, the data instances are grouped by the subject sex (male and female) with the genuine and impostor score distributions computed within the groups. Table IV shows the descriptive statistics of those score distributions.

B. Age

In this experiment, the data instances are grouped by the subject age (into buckets) with the genuine and impostor score distributions computed within the groups. Table V shows the descriptive statistics of those score distributions.

C. Finger/Hand

In this experiment, the data instances are grouped by the finger (index, middle, ring) or hand (left, right) with the genuine and impostor score distributions computed within the groups. Tables VII and VIII show the descriptive statistics of those score distributions.

D. Summary

In order to check whether the small differences reported in previous tables are statistically significant, the standard score (Z-score) is computed as shown in equation 1 (absolute value of the means is used, since only the magnitude of the difference is interesting in this case). Using Z-score is possible, since all-against-all comparisons were conducted and as such, the whole population of the comparison scores for each database/algorithm is known.

\[
Z = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}
\]

This computation is done for all the relevant pairs of score distributions for all the experiments. In other words, for each database and recognition algorithm, all permutations of the genuine and impostor distribution pairs are considered within the respective metadata attribute. Table VI shows the medians and maximums of the computed Z-scores. Overall, the Z-scores (medians) are very low and those of the outliers (maximums) are relatively low. In other words, statistically significant differences are not present for any of the score distribution pairs within their respective experiments (database, algorithm, and metadata attribute). The biometric performance evaluations (in verification and closed-set identification modes) have also not revealed any statistically significant differences.

IV. Conclusion

As shown in the evaluation in section IIII, statistically significant biases in score distributions w.r.t. the sex and age of
the data subjects, as well as the chosen finger/hand have not been detected for the five fingerprint recognition algorithms tested on the four datasets. Accordingly, no impact on the biometric performance in neither the biometric verification nor biometric identification mode has been discovered. The results thus indicate that various fundamentally different classes of fingerprint recognition algorithms might be suitable for application irrespective of the tested meta-parameters. This also points to a potential advantage in certain application scenarios of the vascular characteristics in comparison to others (e.g., face, see section 1), for which potential biases have been reported in the literature. An obvious limitation of this work is the size of the used fingerprint datasets – unfortunately, no larger ones (with metadata present) are currently publicly available. Hence, an important avenue of future research in this area would be the acquisition of a more sizeable dataset and a validation of the scalability of those preliminary results.

Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE, the LOEWE-3 BioBiDa Project (594/18-17), the FFG KIRAS project AUTFingerATM under grant No. 864785, and the FWF project “Advanced Methods and Applications for Fingerprint Recognition” under grant No. P 32201-NBL.
| Dataset | Algorithm | Hand | Score $\mu$ Genuine | Impostor | Score $\sigma$ Genuine | Impostor |
|---------|-----------|------|---------------------|---------|----------------------|---------|
| MMICBNU | LCNN      | Left | 0.73241             | 0.27535 | 0.11214              | 0.08600 |
|         |           | Right| 0.72972             | 0.28659 | 0.10971              | 0.08007 |
|         | LBP       | Left | 0.84654             | 0.78811 | 0.02014              | 0.01735 |
|         |           | Right| 0.84969             | 0.78784 | 0.01993              | 0.01819 |
|         | MC        | Left | 0.26829             | 0.12023 | 0.04914              | 0.01738 |
|         |           | Right| 0.27581             | 0.12124 | 0.04838              | 0.01735 |
|         | PC        | Left | 0.40823             | 0.30551 | 0.02880              | 0.01697 |
|         |           | Right| 0.41295             | 0.30658 | 0.02710              | 0.01618 |
| SIFT    | Left      | 0.37725             | 0.01505 | 0.16594              | 0.02398 |
|         |           | Right| 0.40438             | 0.01532 | 0.16370              | 0.02490 |
| PLUS    | LCNN      | Left | 0.71348             | 0.25962 | 0.13772              | 0.09873 |
|         |           | Right| 0.70663             | 0.29911 | 0.13051              | 0.09379 |
|         | LBP       | Left | 0.79772             | 0.38143 | 0.16295              | 0.06064 |
|         |           | Right| 0.80837             | 0.37036 | 0.15992              | 0.06330 |
|         | MC        | Left | 0.28116             | 0.12307 | 0.03837              | 0.00744 |
|         |           | Right| 0.28375             | 0.13202 | 0.03763              | 0.00768 |
|         | PC        | Left | 0.43033             | 0.31581 | 0.02397              | 0.01204 |
|         |           | Right| 0.43320             | 0.31342 | 0.02495              | 0.01422 |
| SIFT    | Left      | 0.45395             | 0.01591 | 0.14634              | 0.01645 |
|         |           | Right| 0.47223             | 0.01610 | 0.14504              | 0.01828 |
| UTFVP   | LCNN      | Left | 0.69651             | 0.33332 | 0.11244              | 0.17393 |
|         |           | Right| 0.69369             | 0.33540 | 0.11329              | 0.10294 |
|         | LBP       | Left | 0.84549             | 0.81043 | 0.01359              | 0.00483 |
|         |           | Right| 0.84780             | 0.81076 | 0.01319              | 0.00476 |
|         | MC        | Left | 0.23586             | 0.11704 | 0.03855              | 0.00711 |
|         |           | Right| 0.24082             | 0.11883 | 0.03838              | 0.00752 |
|         | PC        | Left | 0.39946             | 0.28287 | 0.02852              | 0.01059 |
|         |           | Right| 0.40367             | 0.28388 | 0.02764              | 0.01052 |
| SIFT    | Left      | 0.30254             | 0.00917 | 0.15592              | 0.01036 |
|         |           | Right| 0.32287             | 0.00996 | 0.15148              | 0.01107 |
| VERA    | LBP       | Left | 0.81147             | 0.78964 | 0.01143              | 0.00402 |
|         |           | Right| 0.81446             | 0.78940 | 0.01036              | 0.00432 |
|         | MC        | Left | 0.23471             | 0.11385 | 0.04772              | 0.00959 |
|         |           | Right| 0.24641             | 0.11477 | 0.04277              | 0.00939 |
|         | PC        | Left | 0.38461             | 0.29064 | 0.03262              | 0.01128 |
|         |           | Right| 0.39022             | 0.29487 | 0.02740              | 0.01167 |
| SIFT    | Left      | 0.21957             | 0.00583 | 0.14676              | 0.00826 |
|         |           | Right| 0.24482             | 0.00612 | 0.14257              | 0.00868 |