A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

Can point-cloud based neural networks learn fingerprint variability?

Dominik Söllinger[1], Robert Jöchl[1], Simon Kirchgasser[1], Andreas Uhl[1]

Abstract: Subject- and environmental-specific variations affect the fingerprint recognition process. Quality metrics are capable of detecting and rating severe degradations. However, measuring natural variability, occurring during different fingerprint acquisitions, is not in the scope of these metrics. This work proposes the use of genuine comparison scores as a measure of variability. It is shown that the publicly available PLUS-MSL-FP dataset exhibits large natural variations which can be used to distinguish between different acquisition sessions. Furthermore, it is showcased that point-cloud (set) based neural networks are promising candidates for processing fingerprint imagery as they provide precise control over the input parameters. Experiments show that point-cloud based neural networks are capable of distinguishing between the different sessions in the PLUS-MSL-FP dataset solely based on FP minutiae locations.

Keywords: fingerprint similarity, fingerprint variability, fingerprint ageing, deep learning, point-cloud.

1 Introduction

The ISO/IEC 19795-1:2006 standard [IS06] states that “Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as template ageing”. However, the main reasons for template ageing are still undefined, as the sources have not been investigated in detail. So far, only ”high-level” explanations for these effects have been given in various studies. For example, it was said that changes in the individual’s behaviour and characteristic of each subject can cause template ageing. However, this definition is general and doesn’t precisely explain the changes across various acquisition. For instance, these variations can be caused by changes in the placement of the finger on the sensor plate, the pressure applied to the sensor, the finger conditions (e.g., injuries, diseases [Dr12]) and biological finger ageing (i.e., [Mo07, GHB18]). On the other hand, there also exist environmental-specific conditions, e.g. changes in the ambient light, the temperature, the acquisition protocol (sensor cleaning,...). Often these conditions remain constant during an acquisition session but change across different acquisition sessions.

In several studies ([YJ15, Ki18, KHB21, KKU21]), the presence of template ageing is explained or questioned, inter alia, using state-of-the-art quality metrics (e.g., NFIQ2.0).

[1] Department of Artificial Intelligence and Human Interfaces, University of Salzburg, AT, {dsoellinger, rjoechl, skirch, uhl}@cs.sbg.ac.at

However, it has not yet been explicitly discussed whether and to what extent the above-mentioned factors ultimately influence the measured quality scores. For example, it seems plausible that a loss of collagen in the course of ageing ([Mo07]) leads to slight changes in the quality measure. Depending on how sensitively the used quality metric reacts to changes, the detection of template ageing effects might be possible or not. Hence, it is important to deduce which properties of fingerprints are influenced by such factors.

Fingerprint (FP) quality metrics are designed to describe the suitability of a FP image for FP processing. However, their purpose is not to quantify changes in FP templates, e.g., positional changes influencing the comparison scores. The first part of this work is aimed at proposing a simple metric (Genuine Comparison Scores — GCS) for measuring the variability between sets of FP of fingerprints samples. Based on a publicly available FP dataset, it is showcased that the GCS is suited for measuring the variability between different acquisition sessions.
Furthermore, the current study is the first one to evaluate if specialised neural network based approaches (point-set or graph-based architectures) are capable of learning fingerprint-specific variances ”encoded” in the FP minutiae. In contrast to conventional CNN applied on 2D imagery, those architectures provide precise control over the input features used for learning, such as minutiae locations or angles. As a proof of concept, it is shown that the applied architecture based on point sets can distinguish between different acquisition session and is not misguided by non-subject-specific features (e.g., sensor noise).

The remainder of this paper is organised as follows: The next Section 2 introduces the GCS similarity metric and show its functionality based on the PLUS-MSL-FP dataset [KKU21]. Section 3 explains the network architecture and experimental setup used in this work. Finally, the experimental results are discussed in Section 4 followed by the conclusion in Section 5.

2 Genuine comparison score similarity metric

As already mentioned in Section 1 no dedicated metric has yet been proposed to measure variability among FPs. Quality metrics are capable of measuring how qualified an imprint is for FP comparison. Yet, it is necessary to quantify the extent of changes detectable in FP samples collected over time.
In this work, the use of Genuine Comparison Scores (GCS) is proposed to measure how a subject’s FP evolves over time by taking into account intra-subject variations. Using the GCS as a measure for FP similarity, this analysis can not only be done on intra-session basis, but also on inter-session basis. It also can be studied how comparison scores change over time and if there is some overall trend (caused by physical ageing, behavioural changes, etc.). To compute the GCS metric for different collections of FP samples, firstly, comparison scores need to be computed for all genuine FP pairs using FP comparison

tools such as Innovatrics ANSI[5], Neurotechnology VeriFinger SDK[6] or NBIS[7]. Next, the obtained scores for each collection of FP samples need to be aggregated using an aggregation function such as mean, median, minium or maximum dependent on the properties that should be studied. In this work, averaging is used to aggregate the genuine comparison scores.

In the current study the GCS is utilised to evaluate the similarity between different acquisition sessions contained in the PLUS-MSL-FP dataset. A detailed description of this publicly available dataset, including 10 sensors in total (named EikonT, EikonS, IBCol, IBCurve, NB, RealScan, LumiM, LumiV, URU, Zvetco), can be obtained from [KKU21]. All comparison scores used in this work were computed with the VeriFinger SDK 12.1. Note that the GCS are only calculated using fingers (subjects) which occur in all of the four PLUS-MSL-FP dataset sessions.

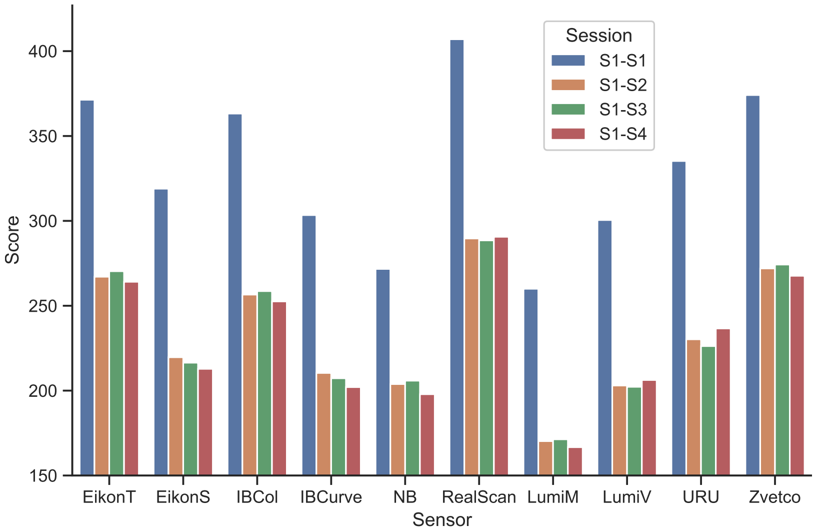


Figure 1: Average genuine comparison score for different sessions. Each impression of Session 1 is matched against every mated sample in another session.

Figure 1 shows the GCS obtained by the comparison of all FP impressions in Session 1 with all mated impressions in Session k where k denotes a session number from 1-4. Figure 2 shows the same evaluation but for all FP impressions in Session 2 compared with impressions from Session $k = 2, 3, 4$. It can be observed that the average intra-session GCS is always considerably higher than the corresponding inter-session GCS. However, this behaviour can easily be justified by visually comparing the FPs of different sessions. Apparently subject's FPs within a session have a similar appearance with respect to factors such as alignment, contact pressure, etc.. A visual comparison of FPs captured in different sessions can be found in the Appendix in Figure 6. Note that the high intra-session GCS

[5] https://www.innovatrics.com; accessed on 2022-08-16)

[6] https://www.neurotechnology.com/verifinger.html; accessed on 2022-08-16)

[7] https://www.nist.gov/services-resources/software/nist-biometric-image-software-nbis; accessed on 2022-08-16)

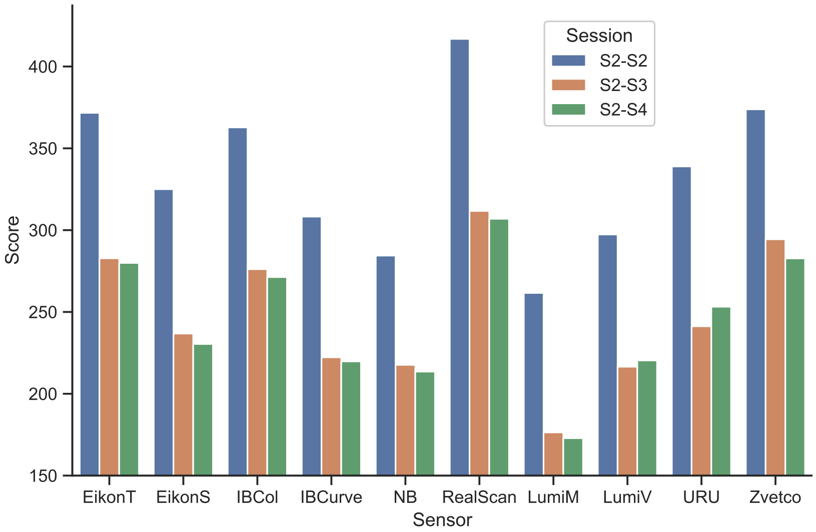


Figure 2: Average genuine comparison score for different sessions. Each impression of Session 2 is matched against every mated impression in another session.

do not result from FP impressions being compared to itself which obviously will lead to high GCS.

Interestingly, comparing the intra-session scores from Session 1 to Session 2 with regard to the average GCS, a similar behaviour of the FPs can be found, although it is known that subjects were allowed to hold the sensor in their hands from the second session on, while they were prohibited in the first session.

Looking at the average inter-session GCS in detail, a possible trend can be recognized depicted by an average score decrease if the time-interval between two FPs being captured increases. In Figure 1, in case of 7 out of 10 sensors the FP impression similarity becomes worst when samples from Session 1 are compared with samples from Session 4 (longest time-interval between two sessions). In case of RealScan, the similarity score remains relatively stable across all sessions with Session 1-4 exhibiting the highest similarity. Only in case of LumiV and URU, the similarity of Session 1-4 comparison clear outperforms the other session comparisons. A similar behaviour can be seen when looking at comparisons with samples from Session 2. In Figure 2, the tendency for FP impression similarity to decrease over time can be found in 8 out 10 cases with GCS of samples from Session 2 and 4 being compared to be the worst. Again, only LumiV and URU do not agree with the overall trend. Apart from subject- and environmental-specific variations in general, this overall trend might also be explainable by template ageing as sensor specific influences have been ruled out [KU16, JU21]. Interestingly, the annual GCS drop is in line with the study [GHB18] on FP ageing which reported an annual decrease of the comparison scores by around 2-3%.

3 Network architecture and experimental setup

In the previous section, a high difference in GCS for intra- and inter-session comparisons was observed, while quality is fairly stable across sessions as reported in [KKU21] for the same dataset. As already discussed, subject- and/or environmental variations can lead to changes in the FP templates which subsequently affect the recognition performance (template ageing), but are not detectable as quality degradation. Yet, it is unclear if a neural network can learn such FP-specific variations. Traditional convolutional neural networks (CNNs) operating directly on 2D FP imagery might be sensitive to all kinds of variations, including non-subject specific effects such as sensor noise. Hence, it is desirable to have control over the features considered during learning. Point-cloud or graph-based approaches might be better suited for this task as input features can be precisely controlled.
As a proof of concept, a neural network architecture used for point cloud classification (PointNet [Qi17]) is adapted to directly process a set of minutiae locations. The proposed architecture is therefore called Minutiae Position Net (MPNet) in this work. It is trained to separate FPs acquired during different acquisition session solely based on minutiae point locations. Thus, the result of the session prediction can only be affected by minutiae point-specific variations.

3.1 Minutiae Position Net (MPNet)

MPNet is a network architecture that can process unordered 2D point sets and is essentially a simplified version of PointNet [Qi17] where the two registration submodules have been omitted. Figure 3 shows the building blocks of MPNet. To bring the minutiae points obtained from the Verifinger SDK in a suitable format for MPNet, at first 40 points have been randomly sampled [8] from the minutiae point set. Then the set of minutiae locations $\{(x_i, x_i)\}$ where $i \in [0, 40)$ is transformed as follows: (i) Normalize the range between $0-1$ by dividing through the image size. (ii) Center the point set around the origin. In other words, we update each minutiae point (x_i, y_i) with $x_i = x_i - \big[\min\big(\{x_0, ..., x_N\}\big) + \max\big(\{x_0, ..., x_N\}\big)\big]/2$ and $y_i = y_i - \big[\min\big(\{y_0, ..., y_N\}\big) + \max\big(\{y_0, ..., y_N\}\big)\big]/2$.

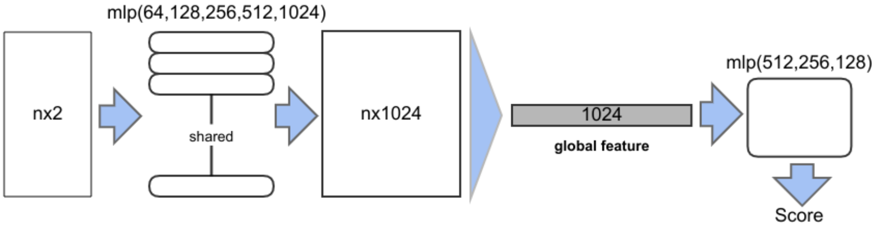


Figure 3: MPNet takes n points as input and aggregates point features by max pooling. The output is classification scores for two classes (sessions). ”mlp” stands for multi-layer perceptron, numbers in bracket are layer sizes. Batchnorm is used for all layers with ReLU.

The network is trained using a Binary Cross Entropy loss and an ADAM [KB17] optimizer ($lr_{max} = 10^{-3}$) over a period of 15000 epochs with a batch size of 1024. Training

[8] Number chosen empirically by analysing the number of minutiae in the FPs. Random sampling guarantees that can get ”seen” if the number of minutiae in a FP exceeds 40

samples are augmented by randomly rotating (±180), horizontal / vertical shifting (±0.1) and jittering (±0.02) of the point set.

3.2 Sampling Strategies

In the course of the experimental evaluation two different sampling strategies for training and test data are applied. They are described in the following:

Random-Stratified (RS): As the naming implies FP imprints are sampled completely randomly but in a stratified way to follow the class distribution. In this scenario, at least one imprint from each finger and session contained in the test set is most likely also included in the training set (e.g., the first imprint of a subject's thumb in Session 1 is contained in test set, while the second imprint of the thumb is contained in the training set). Hence, the model is allowed to learn finger-specific properties. 80% of the imprints are used for training and 20% of the imprints are used for testing.

User-Strict (US): In this scenario, subject-wise sampling is performed to ensure that no imprints from a subject in the test set are present in the training set. If the model now learns finger-specific features, it will result in overfitting. Hence, when evaluating the model's performance only generalizable features, like environmental-specific ones, can be used to correctly predict the session. 80% of the users are used for training and 20% of the users are used for testing.

Experiments for each sampling strategy are run three times using different training and test sets in each run. The reported performance is the average performance (average F1-score) of all three runs.

3.3 Evaluation metric

The separation between two acquisition sessions can be interpreted as a two-class (binary) classification problem. Thus, several metrics can be considered for evaluation, e.g., precision, recall, accuracy and F1-score. As it is not possible to exclude class imbalances and false positives and false negatives are important, the F1-score (harmonic mean of the recall and precision) is used in this work as defined in Equation 1, where TP denotes the true positives, FP the false positives and FN the false negatives.

$$\mathrm{F1} = \frac{2TP}{2TP+FP+FN} \tag{1}$$

Note that an F1-score greater than 0.5 does not imply a performance better than random. However, it is possible to derive a lower bound for a random predictor as described below in Equation 2, where $|Y|$ denotes the number of classes (in the binary case $|Y|$ = 2). Furthermore, α represents the relative number of samples from one class and $(1-\alpha)$ the relative number of samples belonging to the other class, while N depicts the total amount of samples.

$$\frac{2\frac{\alpha N}{|Y|}}{2\frac{\alpha N}{|Y|}+\frac{(1-\alpha)N}{|Y|}+(\alpha N-\frac{\alpha N}{|Y|})} = \frac{2\alpha}{2\alpha+(1-\alpha)+(\alpha|Y|-\alpha)} = \frac{2\alpha}{1+\alpha|Y|} = \frac{2\alpha}{1+2\alpha} \tag{2}$$

Subsequently, for each of the given two classes a lower bound for a random prediction can be computed. These values are used in Section 4 as so called ”prediction by chance” in Figures 4 and 5, while the other values represent the F1 scores for each sensor contained in the utilised PLUS-MSL-FP dataset.

4 Results

The Figures 4 (RS sampling) and 5 (US sampling) show the average F1-scores over three runs for the session classification problem. The vertical dashed line separates the different session-specific experiments, i.e., discriminating FPs between sessions S1 and S2, S1 and S3, S1 and S4, S2 and S3, S2 and S4, as well as S3 and S4. The average F1-scores for a given class are illustrated by the different points (lines). For example, in the leftmost plot of Figure 4 S1 and S2 should be separated. Depending on which session is considered as the positive class, two different F1-scores can be obtained. The points in the S1 column represent the F1-scores, where S1 is the positive class. In case of the S2 column, session S2 is the positive class. The different colours correspond to the sensors used for acquisition. Finally, the solid black lines represent the random prediction boundary (see Equation 2).

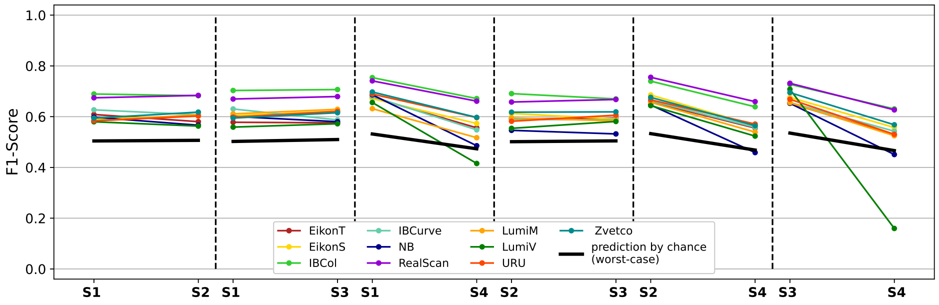


Figure 4: Average F1-scores for the RS sampling scenario computed from three runs.

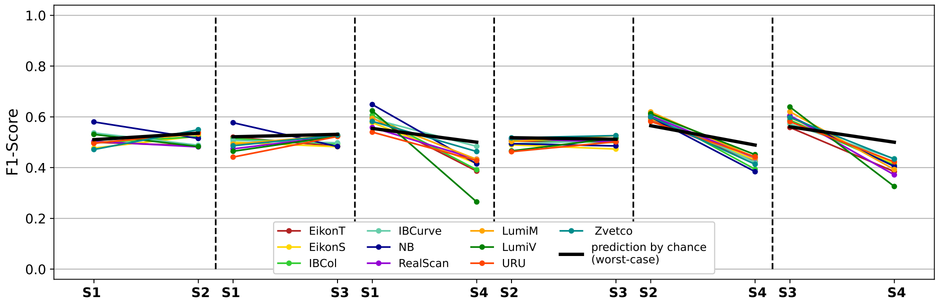


Figure 5: Average F1-scores for the US sampling scenario computed from three runs.

It can be seen in the RS sampling scenario (Figure 4) that the architecture is capable of separating the different sessions fairly consistently (except for LumiV and NB in case of the rightmost plot) as the random predictor is outperformed. In case of US sampling (see Figure 5) a separation is impossible. Note, that the US strategy is more complex as only session-specific environmental differences can be learned by the network (as no subject-specific properties between training and test dataset are shared). Based on the observation that a session separation is not possible in case of US, session-specific environmental variances contained in the dataset can be excluded. However, as a separation is possible in case of RS, this allows to conclude that subject-specific properties contain enough discriminative information to distinguish between sessions. In other words, there exists a subject-specific property (but not a generalisable one across different subjects), which is characteristic for a single session.

Based on the GCS analysis performed in Section 2, the initial hypothesis was that FP minutiae allow to separate between sessions to a certain extent. Hence, the obtained results are in line with this hypothesis. Although, it is not clear if these positional variations are introduced by reasons resulting in template ageing or by other natural variances (e.g., sensor contact pressure). Furthermore, it can be concluded that environmental-specific variations that might affect minutiae positions are of less importance. Otherwise, a separation in the US sampling scenario would also be possible (at least for some sensors).

5 Conclusion

This work proposes the use of a genuine comparison scores (GCS) as a measure for FP variability, while quality measures might not be suited for this purpose. By design, the GCS metric only considers features relevant for FP comparison, in particular minutiae points. Subsequently, the amount of variability exhibited in the PLUS-MSL-FP dataset is measured by applying GCS. Later on, the usefulness of point-cloud based neural networks is demonstrated by separating the different acquisition session contained in the dataset, solely using minutiae locations. It turned out, that MPNet is capable of separating between sessions in case random stratified sampling is employed. Hence, it can be concluded that only finger-specific variances are important for this separation.

In future work, additional features, such as minutiae angle, are planned to be included as input parameter. Furthermore, session separation can be considered as a multi-class classification problem. Additionally, it seems plausible that given a longitudinal FP dataset captured over a long period of time, point-cloud based neural networks can be used to study physical FP ageing by trying to learn an ”age” feature space in which impressions captured close in time are also close to each other in the feature space. Point-cloud based neural networks could also be considered for FP synthesis to improve the diversity of generated FPs or to preserve the identity of a subject. For example, it would be conceivable to use a point-cloud based neural network as an encoder for the inversion of StyleGAN2 [Ka20]. The encoder can be trained to directly transfer minutiae points into the latent space of StyleGAN2 and obtain the corresponding subject in the latent space.

References

[Dr12] Drahansky, Martin; Dolezel, Michal; Urbanek, Jaroslav; Brezinova, Eva; Kim, Tai-hoon: Influence of skin diseases on fingerprint recognition. Journal of Biomedicine and Biotechnology, 2012, 2012.

[GHB18] Galbally, Javier; Haraksim, Rudolf; Beslay, Laurent: A study of age and ageing in fingerprint biometrics. IEEE Transactions on Information Forensics and Security, 14(5):1351–1365, 2018.

[IS06] ISO, ISO: IEC 19795-1: Information technology–biometric performance testing and reporting-part 1: Principles and framework. ISO/IEC, Editor, 1(3):5, 2006.

[JU21] Joechl, Robert; Uhl, Andreas: Apart from In-Field Sensor Defects, are there Additional Age Traces Hidden in a Digital Image? In: Proceedings of the IEEE Workshop on Information Forensics and Security (WIFS2021). Montpellier, France, pp. 1–6, 2021. accepted.

[Ka20] Karras, Tero; Aittala, Miika; Hellsten, Janne; Laine, Samuli; Lehtinen, Jaakko; Aila, Timo: Training Generative Adversarial Networks with Limited Data. In: Proc. NeurIPS. 2020.

[KB17] Kingma, Diederik P.; Ba, Jimmy: , Adam: A Method for Stochastic Optimization, 2017.

[KHB21] Kessler, Roman; Henniger, Olaf; Busch, Christoph: Fingerprints, forever young? In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 8647–8654, 2021.

[Ki18] Kirchgasser, Simon; Uhl, Andreas; Castillo-Rosado, Katy; Estévez-Bresó, David; Rodríguez-Hernández, Emilio; Hernández-Palancar, José: Fingerprint Template Ageing Revisited-It's the Quality, Stupid! In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE, pp. 1–9, 2018.

[KKU21] Kirchgasser, Simon; Kauba, Christof; Uhl, Andreas: The PLUS Multi-Sensor and Longitudinal Fingerprint Dataset: An Initial Quality and Performance Evaluation. IEEE Transactions on Biometrics, Behavior, and Identity Science, pp. 1–13, 2021.

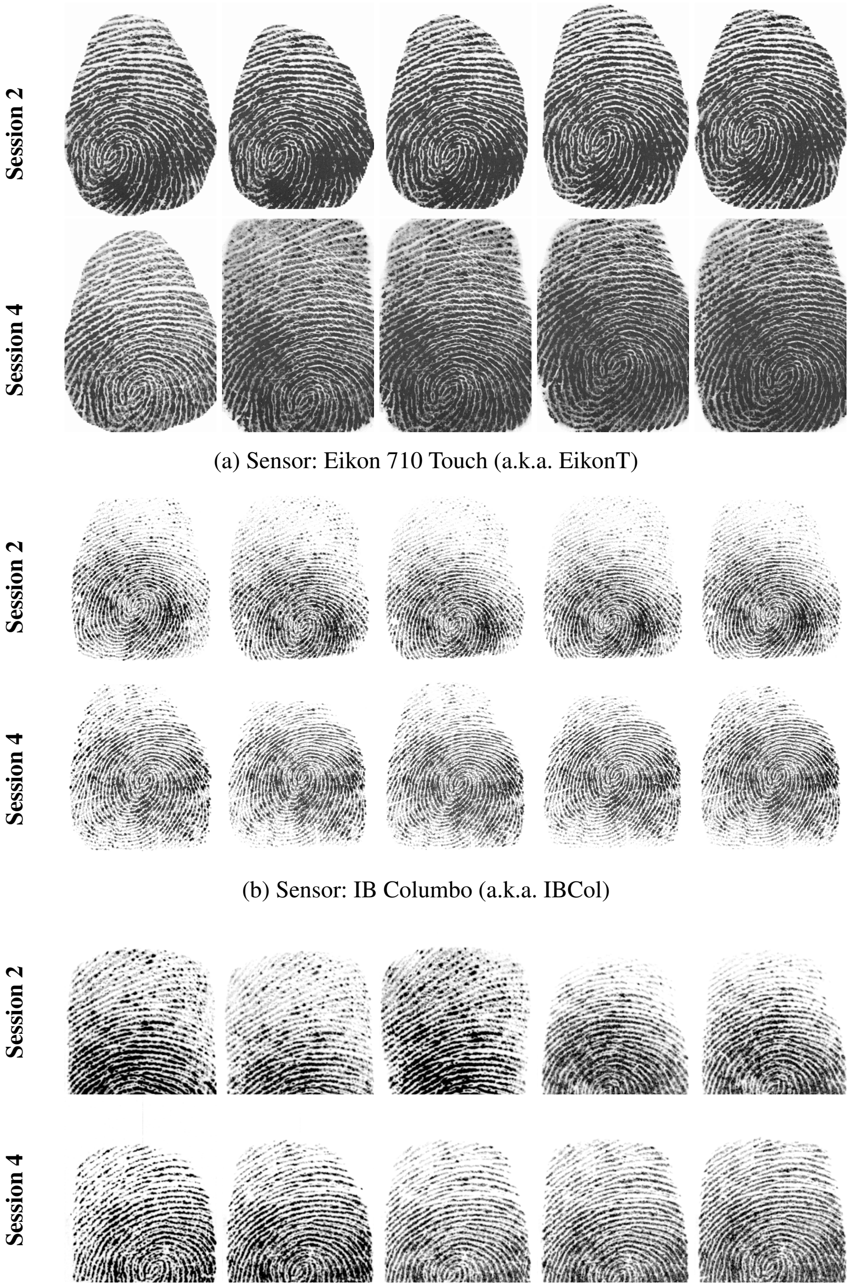
[KU16] Kauba, Christof; Uhl, Andreas: Fingerprint Recognition under the Influence of Sensor Ageing. IET Biometrics, 4(6):245–255, 2016.

[Mo07] Modi, Shimon K; Elliott, Stephen J; Whetsone, Jeff; Kim, Hakil: Impact of age groups on fingerprint recognition performance. In: 2007 IEEE Workshop on Automatic Identification Advanced Technologies. IEEE, pp. 19–23, 2007.

[Qi17] Qi, Charles R; Su, Hao; Mo, Kaichun; Guibas, Leonidas J: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660, 2017.

[YJ15] Yoon, Soweon; Jain, Anil K: Longitudinal study of fingerprint recognition. Proceedings of the National Academy of Sciences, 112(28):8555–8560, 2015.

Appendix

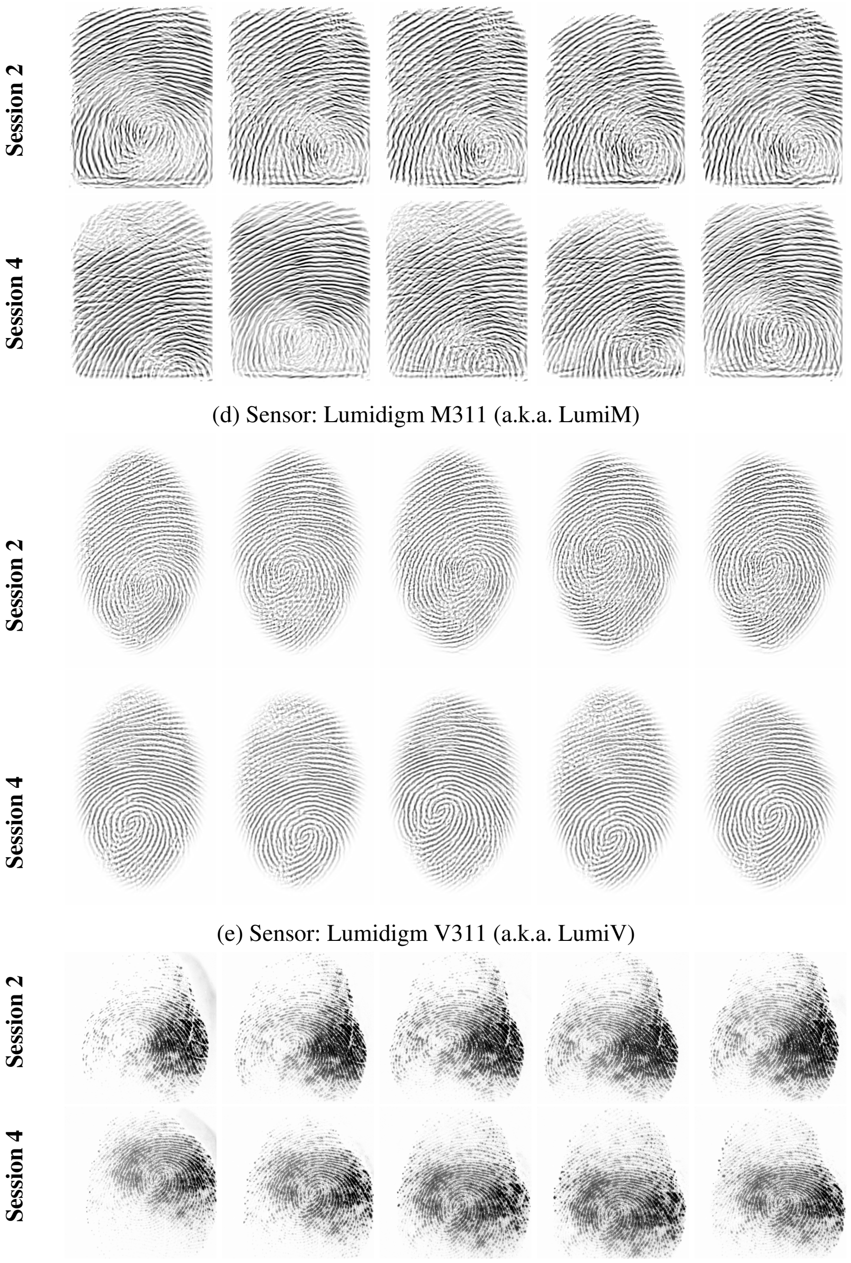


(a) Sensor: Eikon 710 Touch (a.k.a. EikonT)

(b) Sensor: IB Columbo (a.k.a. IBCol)

(c) Sensor: IB Curve (a.k.a. IBCurve)

Figure 6: Illustrative comparison of FPs captured in Session 2 and 4 of the PLUS-MSL-FP dataset with different sensors. The shown FPs originate from the left thumb of User 8.

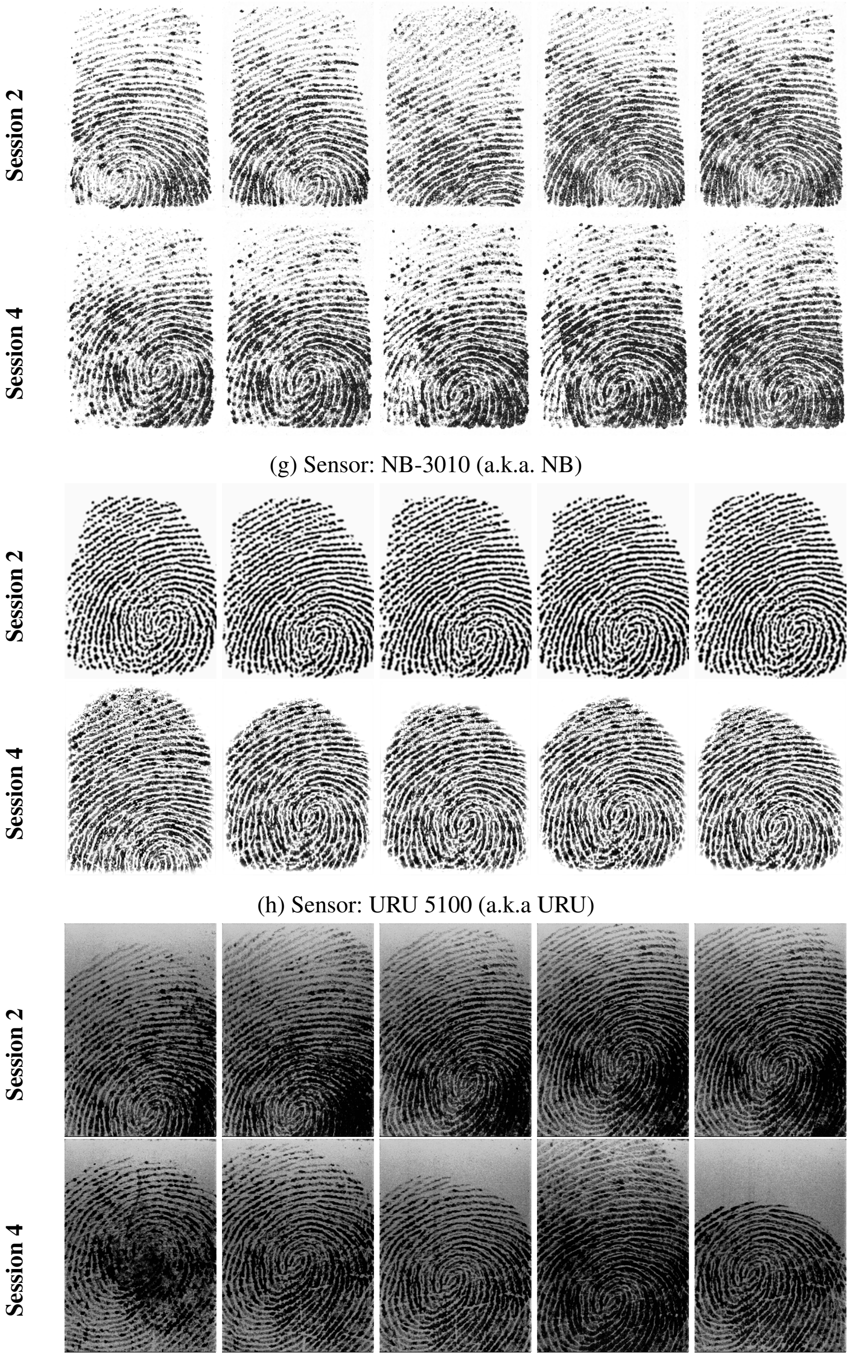


(d) Sensor: Lumidigm M311 (a.k.a. LumiM)

(e) Sensor: Lumidigm V311 (a.k.a. LumiV)

(f) Sensor: RealScan G1 (a.k.a. RealScan)

Figure 6: Illustrative comparison of FPs captured in Session 2 and 4 of the PLUS-MSL-FP dataset with different sensors. The shown FPs originate from the left thumb of User 8.



(g) Sensor: NB-3010 (a.k.a. NB)

(h) Sensor: URU 5100 (a.k.a URU)

(i) Sensor: Zvetco P5000 (a.k.a. Zvetco)

Figure 6: Illustrative comparison of FPs captured in Session 2 and 4 of the PLUS-MSL-FP dataset with different sensors. The shown FPs originate from the left thumb of User 8.