A. Brömme, C. Busch, A. Dantcheva, K. Raja, C. Rathgeb and A. Uhl (Eds.): BIOSIG 2020,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2020 1

Privacy Evaluation Protocols for the Evaluation of Soft-Biometric Privacy-Enhancing Technologies

Philipp Terhörst[12], Marco Huber[1], Naser Damer[12], Peter Rot[3], Florian Kirchbuchner[1], Vitomir Struc[3], Arjan Kuijper[12]

Abstract: Biometric data includes privacy-sensitive information, such as soft-biometrics. Soft-biometric privacy enhancing technologies aim at limiting the possibility of deducing such information. Previous works proposed several solutions to this problem using several different evaluation processes, metrics, and attack scenarios. The absence of a standardized evaluation protocol makes a meaningful comparison of these solutions difficult. In this work, we propose privacy evaluation protocols (PEPs) for privacy-enhancing technologies (PETs) dealing with soft-biometric privacy. Our framework evaluates PETs in the most critical scenario of an attacker that knows and adapts to the systems privacy-mechanism. Moreover, our PEPs differentiate between PET of learning-based or training-free nature. To ensure that our protocol meets the highest standards in both cases, it is based on Kerckhoffs's principle of cryptography.

Keywords: Face, soft-biometric privacy, privacy-enhancing technologies, evaluation protocols.

1 Introduction

Recent works on soft-biometrics showed that privacy-sensitive information, such as gender, age, ethnicity, or even health can be deducted from biometric data of an individual [DER16, Te19c]. However, for many applications, biometric data is expected to be used for recognition purposes only, and extracting such information without the user's agreement raises major privacy issues [Ki13]. Consequently, this kind of data is given special protection, e.g. by the European Union with its General Data Protection Regulation [CotEU16]. Soft-biometric privacy aims at suppressing this privacy-sensitive information in biometric data, to prevent a potential misuse *(function creep)* of this information. Previous works proposed several solutions to this problem. However, since these works consider different evaluation metrics and attack scenarios, a meaningful comparison is difficult. In this work, we propose a standardized framework for evaluating the performance of PETs on soft-biometric privacy. We introduce propose privacy evaluation protocols (PEPs) for learning-based and training-free scenarios. Following the Kerkhoff principle, our PEPs build on the critical scenario of a function creep attacker that knows and adapts to the system's privacy-mechanism. Our PEPs include a detailed description of the data handling, the choice and the training of the attack estimators, as well as, robust and meaningful evaluation metrics for both aspects of soft-biometric privacy, suppressing privacy-risk information and maintaining recognition ability.

[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
[3] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia

2 Related Work

Previous works on soft-biometric privacy either proposed solutions at the image-level [Su11, OR14, MR17, Mi18, MRR18, MRR19] or at template-level [MFV19, Te19a, Te19b, Te20b, Te20a]. At the image-level, Suo et al. [Su11] proposed a gender conversion approach that exchanges facial components of a given face with similar components of the opposite gender. Othman and Ross [OR14] proposed an image-based approach by applying face morphing. To disturb the original gender of an individual they morph the individuals‘ image with an image of the opposite gender. In [Ch18], imperceptible noise was used to suppress k attributes at the same time. However, this noise is trained to suppress attributes from only one specific neural network classifier and consequently, may not generalize to other classifiers. Mirjalili and Ross [MR17] iteratively perturb face images until the soft-biometric attribute assigned by arbitrary estimator flips. More recently, Mirjalili et al. [Mi18] used semi-adversarial networks (SAN) to suppress the gender information in images. SANs are auto-encoders with adversarial training that aim to maximize the performance of a face matcher and to minimize the performance of an estimator. In [MRR18] and [MRR19], the authors extended the idea of SANs to (a) an ensemble of SANs and (b) combining a diverse set of SAN models to compensate for each other's weaknesses.

Recently, template-based solutions received a lot of attention. In 2019, Terhörst et al. [Te19b] proposed similarity-sensitive noise transformations to suppress privacy-sensitive attributes in face representations in an unsupervised manner. Concurrently, Morales et al. [MFV19] introduced SensitiveNets, a network that suppresses target information in face templates based on triplet loss learning. In [Te19a], Terhörst et al. proposed Incremental Variable Elimination (IVE). IVE iteratively determines the most privacy-risk variables and deletes them from the face template. Bortolato et al. [Bo20] proposed PFRNet, a face template learning framework that disentangles identity from soft-biometrics to enhance privacy. In [Te20b], Terhörst et al. proposed Negative Face Recognition (NFR). This unsupervised approach stores only complementary identity information to enhance the user's privacy. Exploting the structural differences between face recognition (use-case) and facial attribute estimation (attack scenario), same authors proposed a privacy-preserving face recognition approach based on minimal information units (PE-MIU) [Te20a].

The following list summarizes the limitations of previous works and demonstrates the need for a standardized evaluation protocol:

- **Violation of Kerkhoffs‘ principle**: Most previous works [Su11, OR14, MR17, Mi18, MRR18, MRR19, Ch18] assume an attacker with only restricted resources and knowledge about the systems privacy mechanism.
- **Gender focus**: Most previous works focus mostly on the evaluation of the binary characteristic gender. The effectiveness of categorical or continuous attributes, such as race and age, is not well investigated [Su11, OR14, MR17, Mi18, MRR18, MRR19].
- **Non-robust evaluation metrics:**: Evaluation metrics (accuracy) used in most previous works [MFV19, OR14, Ch18] are sensitive to the underlying data distribution and thus, vulnerable to unbalanced data.
- **Non-standardized evaluation process**: Due to no established evaluation protocols, a meaningful comparison of PETs is difficult.

3 Attack Scenario

For the privacy evaluation protocol, we assume the following attack scenario: the attacker gained unauthorized access to the face templates or images (stored or transmitted) used to recognize individuals. The attacker may have extensive knowledge of how these were created and what method was used to enhance the privacy of the users. Moreover, the attacker may have access to computational power and an annotated face dataset. Accordingly, we follow Kerckhoffs's principle known from cryptography, which Shannon formulated as *„the enemy knows the system being used"* [Sh49]. The attacker's objective is the function creep of the privacy-sensitive information of the individuals for an unknown purpose.

4 Framework / Protocol

In this section, we propose three soft-biometric *privacy evaluation protocols (PEPs)*. We distinguish between the evaluation of training-free (PEP-TF) and learning-based (PEP-LB) PETs. The learning-based PETs need additional data about the suppressed attributes for the training. PEP-TF requires no additional training and can be directly applied to the data. For the learning-based scenario, we suggest an additional (third) loose protocol (PEP-LBL) if the amount of data is not sufficient to perform the strict evaluation protocol (PEP-LBS).

4.1 Preliminary

The first step of the protocol is to split the data set in approximately equally sized folds k with $k \geq 3$. This split should preserve the statistical distribution of the data set and enforce subject-exclusiveness. This means that images of an individual are not distributed over multiple folds but only included in one fold exclusively. This is done to ensure that virtual attackers learn abstract soft-biometric information and do not rely on learned identity information when predicting soft-biometric attributes. The folds are used to perform k-fold cross-validation. The number of folds used for training, development (parameter tuning), and testing are specified in an extended notation: PEP-LBS-N_{train}-N_{dev}-N_{test}. The N values indicate the number of folds for the specific step. For instance, PEP-LBS-2-1-2 would indicate that the learning-based and loose protocol was performed with two folds as training set, one fold for hyperparameter-tuning, and two folds for testing. After splitting the data in the different folds, the feature vectors are scaled to unit-length and further normalized. Feature normalization, such as z-score or min-max scaling, is applied in the same way as the protocol presented below. These two steps ensure a meaningful start for the attack estimators.

4.2 PEP-LBS: Learning-based and Strict Evaluation Protocol

The *learning-based and strict privacy-enhancing protocol (PEP-**LBS**)* assures that the same data is not used multiple times during the evaluation process. The protocol assumes that the PET includes a training process. Therefore, the original data set is divided into three parts D_{train}, D_{dev}, and D_{test} (which all may consist of multiple folds). The D_{train} set is used to train the PET and the D_{dev} to fine-tune possible hyper-parameters of the method. The D_{test} is transformed using the trained and fine-tuned privacy-enhancing method and further divided into the three subsets: T_{train}, T_{dev}, and T_{test}. It is important to note that T_{train},

T_{dev}, and T_{test} are subsets of the transformed D_{test} and not the transformed D_{train} and D_{dev}. T_{train} is then used to train the different FCEs. T_{dev} is used to fine-tune the hyper-parameters of these FCEs. The T_{test} set is used to evaluate the performance of the PETs in regard of its recognition performance and the suppression performance on the FCEs. A schematic view of the PEP-LBS protocol can be seen in Figure 1a. When using the PEP-LBS we recommend to choose the number of folds in the test subset, $N_{test} \geq 3$.

(a) PEP-LBS (b) PEP-LBL

Fig. 1: Schematic of the data handling of both learning-based protocols PEP-LB.

4.3 PEP-LBL: Learning-based and Loose Evaluation Protocol

In PEP-LBS, dividing the D_{test} into T_{train}, T_{dev}, and T_{test}, requires an appropriate large test set D_{test} and thus, a large amount of data. Since this is often not available, we introduce the *learning-based and loose protocol (PEP-**LBL**)*. In this protocol, the data separation is loosened. This comes at the cost of a partial overfit of the PET on T. The D_{train} subset is used to train the privacy-enhancing method. The D_{dev} subset is used to adjust the hyper-parameters of the PET. Afterwards, all three subsets D_{train}, D_{dev}, and D_{test} are transformed using the PET into T_{train}, T_{dev}, and T_{test}. T_{train} is used to train the estimators of the attacker and T_{dev} to fine-tune the parameters of the estimators. T_{test} is only used to evaluate the PET. The loose protocol provides a trade-off if splitting the test set D_{test} to evaluate the FCEs would lead to too small subsets that meaningful results cannot be obtained. To prevent this, the train set D_{train} and the development set D_{dev} are used twice, once in their unaltered templates/images to train and fine-tune the PET and once in their transformed ones T_{train} and T_{dev} to train and fine-tune the attack estimators. A schematic view of the PEP-LBL protocol is shown in Figure 1b.

4.4 PEP-TF: Training-free Evaluation Protocol

The proposed *training-free evaluation protocol (PEP-**TF**)* assumes that the PET does not require a training phase. Therefore, the three parts of the original data set, D_{train}, D_{dev}, and D_{test} are directly transformed by the PET to the modified templates/images T_{train}, T_{dev}, and T_{test}. T_{train} is used to train the different FCEs, T_{dev} is used to adjust the hyper-parameter of those estimator models and T_{test} is then used to evaluate the performance of the privacy-enhancing method. An illustration of the PEP-TF protocol is shown in Figure 2.

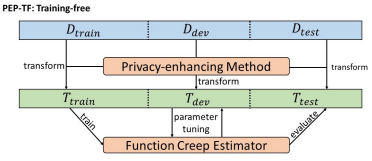

Fig. 2: Illustration of the data handling for the training-free protocol PEP-TF.

4.5 Function Creep Estimators

In the proposed attack scenario, the function creep attacker deploy *function creep estimators (FCEs)* to determine privacy-sensitive attributes that were previously obscured by the transformation through the PET used. These FCEs are trained and fine-tuned as described in the used protocol. The hyperparameter tuning can be done, for example, via Grid Search, Random Search or Bayesian Optimization. In Section 4.1, the pre-processing of the data was already described.

The template-based approaches are evaluated using the extracted feature representations of the face images. For the possible FCEs, we recommend well-known estimators that should be used as a baseline to assess the quality of the PETs. These include *random forest*, *support-vector machines*, *k-nearest neighbors* and *logistic regression*. This choice is based on (a) their membership to different kinds of machine learning models and (b) the fact that these perform evidently well on face templates [Te20b, Te19a]. Each FCE is independently trained twice: first, on the unmodified data and second, on the transformed data that was modified by the PET. This allows us to compare the performance of the estimators without having noise due to different test samples. The training of several different estimators is intended to ensure the robustness of the PET for different kind of attacks. Please note that another attack scenario might come from regenerating a face image from a template and manually investigating this. However, patterns of privacy-sensitive information in templates are generally easily detectable due to the feature entanglement during the learning process.

In contrast to PET based on template-level, image-based approaches have to deal with large-scale and more restricted feature spaces. Image-based approaches have the advantage that, for many attributes, the modified representations can be evaluated by humans as well. However, the choice of function creep estimators should additionally include machine-based solutions since these solutions might catch suspicious artifacts that humans are not aware of. Due to the large-scale nature of images, (a) CNN approaches [KSH12] should be used as potential FCEs or (b) a combination of lower-dimensional handcrafted features, such as LBPH [AHP06], with the proposed template-based estimators.

5 Evaluation

So far, the protocol descriptions focus on the data handling and the training of PETs and FCEs. Based on this, this section describes how the PETs can be robustly evaluated in regard to the FCEs. The challenge of soft-biometric privacy describes a trade-off between maintaining the recognition performance of face representations and suppressing the predictability of privacy-sensitive attributes within these. To evaluate both aspects of the trade-off, the attribute estimation and the recognition results of the modified and unmodified face representations are compared. For the evaluation of the attribute suppression performance, the predictions of the FCEs on the un/modified representations of T_{test} are used. The evaluation of the recognition performance is based on the un/modified representations of T_{text}.

The recognition performance is the most important factor of recognition systems, since it measures its major purpose. We recommend to evaluate these in terms of *receiver operating characteristic* (ROC) curves with *false non-match rates* (FNMR) at a different *false*

match rates (FMR) as defined in the ISO standard [ISO]. ROC curves provide a broad performance overview independent of the application and allow to compare the recognition performance of the unmodified baseline with the PETs. For more specific comparisons, FNMR at a fixed FMR of 10^{-3} or smaller can be used as recommended by the European Border Guard Agency Frontex [Fr17].

To evaluate the suppression performance of PETs, we recommend the use of the *balanced accuracy*. This balanced accuracy is equivalent to the standard accuracy definition with class-balanced sample weights. This allows an unbiased performance measure on testing data with unbalanced attribute information. The suppression performance of PETs can be compared by providing the attribute estimation performance of the unmodified baseline and compare these with the estimation performances of the FCEs on the PET-modified representations. For a single value comparison on the suppression performance the *suppression rate* [Te20b] can be used. The suppression rate $\gamma = \frac{acc_{org} - acc_{mod}}{acc_{org}}$ is defined by the difference in prediction accuracy between unmodified (acc_{org}) and PET-modified (acc_{mod}) representations.

To measure the benefit of applying a PET, the *privacy gain identity loss coefficient (PIC)* [Te19b] is a suitable tool. The PIC is defined as $PIC = \frac{AE'-AE}{AE} - \frac{RE'-RE}{RE}$ where AE and AE' denote the attribute prediction errors of an FCE. RE and RE' define the recognition errors with and without the privacy-enhancement of the face representations. In Figure 3, equipotential lines for different PIC-values are shown and visualize the trade-off. The PIC values the relative error of the FCE prediction with the relative error of the recognition performance. Consequently, it directly measures the benefit of using the PET such that a higher coefficient states a higher benefit.

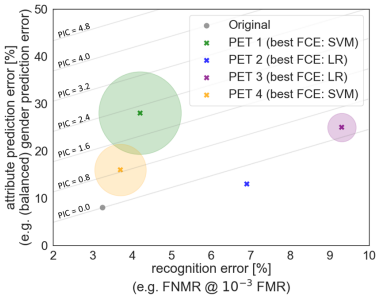

Fig. 3: Example of a recognition-attribute plot [Bo20]. The attribute prediction error is shown over the recognition error for the unmodified baseline and the different PETs. The attribute error refers to the most successful FCE. The size of the shaded areas refer to the PIC coefficient for a PET. Additionally, equipotential lines for different PIC-values are shown in grey.

To visualize the worst-case privacy-enhancing performance, we recommend the use of recognition-attribute plots [Bo20], as shown in Figure 3. This plot shows the recognition error (e.g. the FNMR at 10^{-3}FMR) over the balanced prediction error of an attribute (e.g. gender). The attribute prediction error refers to the most successful FCE, to simulate the most critical attack scenario. In the plot, the unmodified baseline is shown, as well as the PETs under the specification of the most successful FCE. This allows a complete evaluation of the trade-off between suppressing an attribute and maintaining the recognition performance. To further visualize the benefit of applying a PET, the size of the shaded areas around a PET represents its PIC coefficient.

6 Conclusion

Extracting privacy-sensitive information, such as demographics or health information, about an individual from biometric data without consent is considered a major privacy issue. Recent works proposed PETs under different evaluation processes, metrics, and considered attack scenarios. This makes a meaningful comparison of these methods challenging. To enhance the comparability of PETs, and thus enhance the development of this field, we propose PEPs in the most critical attack scenario of a function creep attacker that knows and adapts to the systems privacy-mechanism. We propose three PEPs to ensure sufficient use of the data concerning the nature of the evaluated PET. This includes efficient and independent data handling, training of PETs and FCEs, and robust evaluation metrics for both aspects of soft-biometric privacy.

Acknowledgement 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.

References

[AHP06] Ahonen, Timo; Hadid, Abdenour; Pietikainen, Matti: Face Description with Local Binary Patterns: Application to Face Recognition. IEEE Trans. Pattern Anal. Mach. Intell., 28(12):2037–2041, December 2006.

[Bo20] Bortolato, B.; Ivanovska, M.; Rot, P.; Krizaj, J.; Terhörst, P.; Damer, N.; Peer, P.; Struc, V.: Learning Privacy-Enhancing Face Representations through Feature Disentanglement. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG). IEEE Computer Society, Los Alamitos, CA, USA, pp. 45–52, may 2020.

[Ch18] Chhabra, Saheb; Singh, Richa; Vatsa, Mayank; Gupta, Gaurav: Anonymizing k Facial Attributes via Adversarial Perturbations. In (Lang, Jérôme, ed.): Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. ijcai.org, pp. 656–662, 2018.

[CotEU16] Council of the European Union, European Parliament: , Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[DER16] Dantcheva, Antitza; Elia, Petros; Ross, Arun: What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Trans. Information Forensics and Security, 11(3):441–467, 2016.

[Fr17] Frontex: Best Practice Technical Guidelines for Automated Border Control (ABC) Systems. 2017.

[ISO] : Information technology - Biometric performance testing and reporting - Part 1: Principles and framework. Standard, International Organization for Standardization.

[Ki13] Kindt, Els J.: Biometric Data, Data Protection and the Right to Privacy. In: Privacy and Data Protection Issues of Biometric Applications: A Comparative Legal Analysis. Springer Netherlands, Dordrecht, 2013.

[KSH12] Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E: ImageNet Classification with Deep Convolutional Neural Networks. In (Pereira, F.; Burges, C. J. C.; Bottou, L.; Weinberger, K. Q., eds): Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc., 2012.

[MFV19] Morales, Aythami; Fiérrez, Julian; Vera-Rodríguez, Rubén: SensitiveNets: Learning Agnostic Representations with Application to Face Recognition. CoRR, abs/1902.00334, 2019.

[Mi18] Mirjalili, Vahid; Raschka, Sebastian; Namboodiri, Anoop M.; Ross, Arun: Semi-adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In: 2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018. IEEE, pp. 82–89, 2018.

[MR17] Mirjalili, Vahid; Ross, Arun: Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In: 2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017. IEEE, pp. 564–573, 2017.

[MRR18] Mirjalili, Vahid; Raschka, Sebastian; Ross, Arun: Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers. In: 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018. IEEE, pp. 1–10, 2018.

[MRR19] Mirjalili, Vahid; Raschka, Sebastian; Ross, Arun: FlowSAN: Privacy-Enhancing Semi-Adversarial Networks to Confound Arbitrary Face-Based Gender Classifiers. IEEE Access, 7:99735–99745, 2019.

[OR14] Othman, Asem A.; Ross, Arun: Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity. In (Agapito, Lourdes; Bronstein, Michael M.; Rother, Carsten, eds): Computer Vision - ECCV 2014 Workshops - Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II. volume 8926 of Lecture Notes in Computer Science. Springer, pp. 682–696, 2014.

[Sh49] Shannon, Claude E: Communication theory of secrecy systems. Bell system technical journal, 28(4):656–715, 1949.

[Su11] Suo, Jin-Li; Lin, Liang; Shan, Shiguang; Chen, Xilin; Gao, Wen: High-Resolution Face Fusion for Gender Conversion. IEEE Trans. Systems, Man, and Cybernetics, Part A, 41(2):226–237, 2011.

[Te19a] Terhörst, Philipp; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Suppressing Gender and Age in Face Templates Using Incremental Variable Elimination. In: 2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019. IEEE, pp. 1–8, 2019.

[Te19b] Terhörst, Philipp; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. Appl. Intell., 49(8):3043–3060, 2019.

[Te19c] Terhörst, Philipp; Huber, Marco; Kolf, Jan Niklas; Zelch, Ines; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Reliable Age and Gender Estimation from Face Images: Stating the Confidence of Model Predictions. In: 10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, Florida, USA, September 23-26, 2019. IEEE, 2019.

[Te20a] Terhörst, Philipp; Riehl, Kevin; Damer, Naser; Rot, Peter; Bortolato, Blaz; Kirchbuchner, Florian; Struc, Vitomir; Kuijper, Arjan: PE-MIU: A Training-Free Privacy-Enhancing Face Recognition Approach Based on Minimum Information Units. IEEE Access, 8:93635–93647, 2020.

[Te20b] Terhörst, Philipp; Huber, Marco; Damer, Naser; Kirchbuchner, Florian; Kuijper, Arjan: Unsupervised Enhancement of Soft-biometric Privacy with Negative Face Recognition. CoRR, abs/2002.09181, 2020.