A False Sense of Privacy: Towards a Reliable Evaluation Methodology for the Anonymization of Biometric Data

Simon Hanisch
Center for Tactile Internet (CeTI),
Technical University Dresden
Dresden, Germany
simon.hanisch@tu-dresden.de

Julian Todt
KASTEL,
Karlsruhe Institute of Technology
Karlsruhe, Germany
julian.todt@kit.edu

Jose Patino
Cerence[*]
Burlington, United States
jose.patino@cerence.com

Nicholas Evans
Digital Security Department,
EURECOM
Biot - Sophia Antipolis, France
evans@eurecom.fr

Thorsten Strufe
KASTEL,
Karlsruhe Institute of Technology
Karlsruhe, Germany
thorsten.strufe@kit.edu

ABSTRACT

Biometric data contains distinctive human traits such as facial features or gait patterns. The use of biometric data permits an individuation so exact that the data is utilized effectively in identification and authentication systems. But for this same reason, privacy protections become indispensably necessary.

Privacy protection is extensively afforded by the technique of anonymization. Anonymization techniques protect sensitive personal data from biometrics by obfuscating or removing information that allows linking records to the generating individuals, to achieve high levels of anonymity. However, our understanding and possibility to develop effective anonymization relies, in equal parts, on the effectiveness of the methods employed to evaluate anonymization performance.

In this paper, we assess the state-of-the-art methods used to evaluate the performance of anonymization techniques for facial images and for gait patterns. We demonstrate that the state-of-the-art evaluation methods have serious and frequent shortcomings. In particular, we find that the underlying assumptions of the state-of-the-art are quite unwarranted. State-of-the-art methods generally assume a difficult recognition scenario and thus a weak adversary. However, that assumption causes state-of-the-art evaluations to grossly overestimate the performance of the anonymization. Therefore, we propose a strong adversary which is aware of the anonymization in place. This adversary model implements an appropriate measure of anonymization performance. We improve the selection process for the evaluation dataset, and we reduce the numbers of identities contained in the dataset while ensuring that these identities remain easily distinguishable from one another. Our novel evaluation methodology surpasses the state-of-the-art because we measure worst-case performance and so deliver a highly reliable evaluation of biometric anonymization techniques.

KEYWORDS

privacy, biometric, methodology, anonymization, evaluation

1 INTRODUCTION

Biometric data is rich in sensitive personal information that can be used to identify individuals and infer private attributes. Usage examples of biometric data are face recognition [11], gait recognition [87], inference of medical conditions [34], and inference of character traits based on eye motion [37]. Thus, the utility provided by biometric data is undeniable. Users upload images to social media, share videos with friends, and use online services for health tracking. However, users also care about personal privacy, and in this connection, the use of biometric data poses a real and serious threat. Biometric data allows to draw conclusions about the sensitive personal information of users without their explicit consent.

In order to protect the privacy of the individuals whose biometrics are captured, techniques have been developed that perturb biometric data and so obfuscate or remove their sensitive personal information, or information that may allow for linking them to their generating individuals. Besides protection, these techniques are also designed to maintain the practical utility of the biometric use case. In short, privacy-protecting techniques make the trade-off between utility and protection. The representative technique in identity protection is anonymization.

Identity protection through anonymization ultimately depends on reliable evaluation of anonymization performance. A reliable evaluation methodology for anonymization will accurately assess the level of protection afforded against identification by the anonymized data. Reliable evaluation begins at the assumptions made about an attacker. These assumptions must be robust, because otherwise the evaluation methodology will likely deliver grossly inaccurate estimates of anonymization performance. The result will be a false sense of privacy, and the consequence will be the erosion of user trust. Moreover, any inaccuracy or even error in an evaluation methodology will detrimentally affect advances in the research. Flaws in the methodology may feed into future research and thus hinder or even arrest the development of advanced anonymization techniques. The upshot is this: Only when a biometric anonymization technique has been convincingly evaluated can researchers

[*] Work done while at EURECOM.

Proceedings on Privacy Enhancing Technologies 2024(1), 116–132

https://doi.org/10.56553/popets-2024-0008

improve on existing techniques or provide secure applications to users.

In this paper, we assess the state-of-the-art evaluation methodology for the anonymization of biometric data. In particular, we assess the evaluative methods for face anonymization and gait anonymization. Our choice for face anonymization is based on the fact that there are many widely-employed techniques in application. On the other hand, we have chosen gait anonymization precisely because the techniques here are fewer and relatively seldom in application. Moreover, the 3D time-series data in gait anonymization allow us to push the envelope of the simpler two-dimensional space of the face image. We acknowledge that the comprehensive evaluation of any anonymization technique is only possible when utility is also taken into consideration. However, for this paper, we have narrowed our scope to the improvement of the methods evaluating privacy protection of anonymization only.

Our assessment of the state-of-the-art in evaluation for the anonymization of biometric data shows that these methods often fail at convincingly evaluating the performance of the privacy protection.

The state-of-the-art methods have been uncritically adopted from the evaluation methodology for biometric recognition. In biometric recognition, the problems employ many identities and difficult biometric samples (e.g., profile photos or nearly indistinguishable identities). In anonymization, on the other hand, a difficult problem has a small number of identities which are very diverse, thus making the identities easier to differentiate but more difficult to anonymize.

Furthermore, the state-of-the-art methods rely on weak adversary models. These methods assume that the attacker is unaware of the anonymization mechanisms in place. For example, a method will use pre-trained recognition models which perform well on clear data. However, such models prove incapable of adapting to data modifications performed by an anonymization technique. Consider this straightforward scenario: An anonymization technique for a face image performs consistently the same block permutation. This anonymization can easily be removed with the inverse permutation. However, the permutation will go unnoticed by a recognition model pre-trained on the clear data. Moreover, if only a single recognition is used, then that will jeopardize the reliability of the evaluation. Although a given anonymization technique may successfully degrade the performance of one recognition system, *other* systems classifying *other* feature vectors may be more robust or even largely unaffected. However, the use of just a single recognition system is the norm among state-of-the-art evaluation methods.

We conclude that improvements to the state-of-the-art evaluation for the anonymization of biometric data will also improve the privacy protection offered by anonymization techniques. To this end, we offer specific recommendations for the improvement of state-of-the-art evaluation methods. We draw special attention to just three of our recommendations here. First, we recommend training and (where in use) pre-training recognition systems on anonymized data. Second, we recommend considering multiple and different recognition systems. Third, we recommend choosing smaller datasets in an informed manner.

The contributions of our paper are as follows:

- We assess the current state-of-the-art evaluation methodology for biometric data anonymization and point to fatal flaws in the evaluation methodology.
- We update the state-of-the-art evaluation methodology. Our methodological improvements involve (1) retraining the recognition system on anonymized data, (2) using multiple recognition systems to evaluate the anonymization, and (3) generating evaluation datasets that are challenging to anonymize and consequently reliable for the evaluation of the anonymization performance.
- We test our methodological improvements on the biometric traits face and gait with extensive experimentation. Our evidence supports the conclusion that our improved methodology delivers reliable evaluations of biometric data anonymization.

Here we outline the organization of our paper. In Section 2, we introduce the related work and continue in Section 3 by giving the background. In Section 4, we present our improved methodology for the evaluation of anonymization techniques. In Section 5, we setup our experiments on both face and gait recognition, and in Section 6 we analyze our results to show how our methodology improves upon the state-of-the-art. In Section 7, we discuss our findings and explore the future work in this important area of privacy research. And in Section 8, we draw our conclusions.

2 RELATED WORK

Biometric recognition spans dozens of biometric traits and hundreds of techniques, but the methodology for evaluating the performance of these techniques has been assessed by only a very few works.

Goga et al. [15] assess the methodology for evaluating matching techniques of profiles from different social media platforms. They find that evaluation commonly overestimates the performance of the approaches by using an unrealistic methodology. Granger and Gorodnischy [16] describe the methodology that should be applied to evaluate the performance of biometric recognition for video surveillance applications. For the evaluation of stylometric authorship attribution, Stolerman et al. [76] make the case that an open-set model should be applied since in a realistic scenario the actual author might not be on the suspect list. Brennan et al. [9] propose adding attacks to the methodology of stylometry evaluation because most methods cannot defend against attacks. These investigations of the evaluation methodology in different fields have shown that wrong assumptions lead to an overestimation of performance. In the case of anonymization, overestimation of performance may give users false assurances of privacy because, in fact, their identities are actually left unprotected. In this paper we similarly look at a current evaluation methodology, highlight issues and propose solutions.

Template protection is a very specific kind of privacy protection because it removes all possible attribute inferences while still allowing identity verification. One specialized evaluation methodology [68] for template protection specifies the properties (irreversibility, unlinkability, and confidentiality) which any template protection scheme must achieve to be deemed secure. However, this evaluation methodology is not directly applicable to our work because anonymization seeks, on the contrary, *to remove* the connection between identity and data.

Le et al. [39] discuss how to evaluate privacy-utility trade-offs for face anonymization, but their focus is exclusively on measuring the utility and not privacy.

Recent works [23, 77, 81] propose attacks on biometric data anonymization that use machine learning to reverse the obfuscation of images. These results show that the method is highly effective even when a human observer cannot recognize anything at all in the image. The reversal of anonymization is indeed comparable to the training of a recognition system on anonymized data. However, we consider training recognition systems on anonymized data the more straightforward way to test whether identifying information remains in the anonymized data. Further, we also consider the reduction of the dataset.

In the context of the VoicePrivacy challenge [82], other recent works have investigated the evaluation methodology of speaker anonymization. Noé et al. [59] also propose a framework to evaluate and compare speech pseudonymization approaches using ZEBRA [52] and voice similarity matrices [58]. ZEBRA aims at creating a worst-case metric to evaluate speaker anonymization and voice similarity matrices allow to compare how well specific identities are anonymized. Bonastre et al. [6] propose a benchmarking methodology to test speaker recognition against spoofing and anonymization. We investigate whether some of the methodological improvements to the evaluation of speaker anonymizations, like training recognition systems with anonymized data, can be applied to a wider range of biometrics like face and gait data.

In sum, many improvements to the evaluation methodologies of different research fields have been proposed. However, for the anonymization of biometric data, we find that multiple improvements can still be made to evaluation methodology, such as anonymized data in the training dataset and a more challenging anonymization scenario.

3 BACKGROUND

In this Section, we define basic terminology required for our work.

3.1 Biometrics, Inference, and Recognition

Biometric traits (also called biometric characteristics [1]) are properties of a human that either capture the physiology of a human (e.g. face, iris, fingerprint) or its behavior (e.g. voice, gait, heartbeat). *Soft biometric traits* (e.g. age, sex, weight) are insufficiently entropic to positively identify an individual. However, the combination of soft biometrics can suffice to identify an individual.

Due to the unique nature of biometric traits for each human being they can be used for privacy-invasive inferences. We distinguish between two privacy threats. By the term *identity inference* we mean that the identity of an individual is inferred. By the term *attribute inference* we mean that only a specific private attribute (e.g. age, sex, medical condition) is inferred.

In biometric recognition, identity inference and attribute inference are made operative in a system that learns an inference on representative samples for each class. For each biometric sample to be classified, a biometric recognition system returns a list of possible classes, where each class has been assigned its own separate likelihood. In closed-set recognition, the sample must belong to one of the classes in the dataset, while in open-set recognition the sample may belong to an unknown class.

3.2 Anonymization

The aim of *anonymization* is to protect an individual's identity. During the process of anonymization, information is removed or perturbed that is specific to an individual. Hence, anonymization prevents an adversary from using the data to infer the class corresponding to an individual (i.e. identification). In contrast to anonymization, *pseudonymization* is aimed at retaining some connection between identity and data in order to link the data to an alternative identifier. Besides preventing identification, an anonymization or an pseudonymization also seeks to retain *utility* (i.e. usefulness) of the data. Most often a trade-off between privacy and utility must be made.

An example scenario where biometric anonymization is used is the publication of images in newspapers where the identity of the person in the image should be protected. The basic ways to achieve anonymization here are to remove the identifying information (e.g., cropping the face out of the image), to coarsen the identifying information (e.g., pixelating the face), or to perturb the identifying information (e.g., adding noise to it). In most cases, it is not necessary to completely delete the identifying information, but rather to delete enough so that the person cannot be uniquely identified.

4 IMPROVING THE EVALUATION METHODOLOGY

In this Section, we aim to achieve a reliable evaluation methodology for the anonymization of biometric data. Our premise is that an evaluation methodology for anonymization techniques should be pessimistic and assume a strong adversary based on the worst-case performance of the anonymization technique. To improve the evaluation methodology for the anonymization of biometric data, we proceed in two steps.

First, we present our adversary model and then analyze the shortcomings of the state-of-the-art for evaluating biometric data anonymization. Overall, we find that the evaluation of anonymization performance has been uncritically adopted from the evaluation of recognition systems.

Second, we make three suggestions for improvement. We suggest (1) that recognition systems be trained with anonymized data, (2) that anonymization performance be tested against different recognition systems, and (3) that evaluation datasets consist of recognition problems more challenging to anonymization performance.

4.1 Adversary Model

We investigate the efficacy of anonymizing biometric data, in other words, of preventing biometric recognition. Hence, we consider a scenario in which a user provides his biometric data to a service provider to receive some utility. Examples of this kind of service are step counters based on gait data (e.g. for exercise/activity monitoring), a medical service that analyses the heart rate of users, a social media platform that is being used to publish images of the user, or a website that tracks the mouse movements of a user. Privacy-minded users will try to protect themselves or others from privacy inferences and therefore anonymize their data before sending it to the service.

We consider an adversary who gets full access to the data set submitted to the service, either because the attacker actually is the service provider or because the service provider leaks the data set. The adversary's goal is to perform privacy inferences on the data set under attack. For this, the adversary has access to a training data set that contains labeled biometric data and can be used to train a biometric recognition system for this task. For the training data set, we consider that it can consist of both clear data or anonymized data. We believe that this is a realistic assumption since the adversary will likely learn about the applied anonymization (similar to Kerckhoffs's principle) and can then apply the anonymization to the training set. Alternatively, the adversary might use scraped anonymized data, for example, from social media.

4.2 State-of-the-Art Evaluation Methods for the Anonymization of Biometric Data

We began by gaining an overview of the problems of the state-of-the-art evaluation methods. To this end, we assessed the papers covered in two recent surveys [21, 70] on the topic of biometric data anonymizations. Next, to gain a closer perspective on the field of face anonymization, we analyzed works published from 2018 [12, 27, 30, 46, 67, 73, 88], and one work from 2005 [57]. We included as many works as we could find which appeared at *USENIX Security*, *Privacy Enhancing Technologies Symposium* (PETs), and *Data and Applications Security and Privacy.* As a recent work [38] of 2023 testifies to the persistence of the said methodological flaws to this day. An expanded survey of the current methodology can be found in the Appendix A.

Our survey shows that the methods for evaluating techniques of biometric recognition or anonymization use the same recognition systems, the same datasets, and the same evaluation scenarios. This unquestioned reuse of the same attacker model, dataset, and scenario is highly problematic and will undermine the reliability of any evaluation of anonymization performance. Our reasoning is as follows. In biometric recognition, an evaluation method presents challenging scenarios to the recognition system. Identities are hard to distinguish from one another, the number of identities to be distinguished is high, the biometric samples are poor in quality, an open-set scenario is used, and imposters are introduced to mislead recognition systems. However, in biometric anonymization by contrast, these same conditions do not pose a challenge. In fact, for example the high number of identities makes anonymization much easier, because the more identities we have, the more likely it is that for each identity there is another similar identity in the dataset. This makes it harder to distinguish between identities, which makes anonymization easier. We conclude that anonymization performance will not be accurately evaluated by methods designed to evaluate the performance of recognition systems.

Our analysis shows that the reusing of evaluation methods from recognition and anonymization causes three main problems.

The first problem we identified is that reuse of the scenario for the evaluation of recognition makes for an unrealistically weak adversary model for the evaluation of anonymization. Since in most papers the recognition system is trained on clear data and not on anonymized data (e.g. [27, 38, 67]), obviously the implicit assumption being made is that the adversary is unaware of the anonymization in place. However, an adversary which is aware of the anonymization can adapt to the anonymization and thus will present a greater threat. Consider, for example, an anonymization that performs a deterministic block permutation on a face image. The modification of the data would most likely cause the trained recognition model to break down, and therefore report a high performance. That report, however, will be based on flawed premises and is false.

The second problem we identified is that most evaluation methods assume that the recognition model which works best on clear data will also be the best model for recognizing people in anonymized data (e.g. [5, 13, 78]). We challenge this assumption. Recognition models are developed on clear data. No consideration is given to tampering with the data. Therefore, we doubt whether the recognition model which works best on the clear data is also the best for anonymized data.

The third problem we identified is that the same datasets are used to evaluate anonymization as are used to evaluate recognition (e.g. [38, 46, 67]). Consequently, anonymization techniques are evaluated almost exclusively on large numbers of identities. We argue that it is more challenging for anonymization techniques when there are low numbers of identities in the dataset. Furthermore, a low number of identities is more realistic because biometric data seldom exists alone and additional individuating information (e.g. device ids, soft biometrics, etc.) can be used to further reduce the number of identities in the group.

4.3 Our Improvements to State-of-the-Art Evaluation Methods

We use closed-set recognition for our general scenario to have a stronger attacker. Our adversary possesses a list of identities and consequently may simply test samples against the list to select the most likely identity for a given sample.

We use two different biometric recognition system architectures for the gait and face recognition systems. For our gait recognition systems, we use an architecture which only uses data specific to the target identities, and for our face recognition systems, we use an architecture that uses additional background data not specific to the target identities (see Fig. 1). Both architectures split the samples of each identity contained in the evaluation dataset into *train set* and *test set.* The train set is used to learn a representation for each identity which is then used to infer the identity of the samples in the test set. In addition to this, the face recognition systems are *pre-trained* prior to training on the train set. During pre-training, an additional background dataset representative of the general population is used to learn the features which can be used to differentiate between identities.

4.3.1 Training Recognition Systems with Anonymized Data.

In line with previous work [47, 56, 74, 84], we propose that recognition systems be trained on anonymized data so that a more reliable anonymization performance is achieved. The idea of retraining recognition systems was first proposed for face recognition by Newton

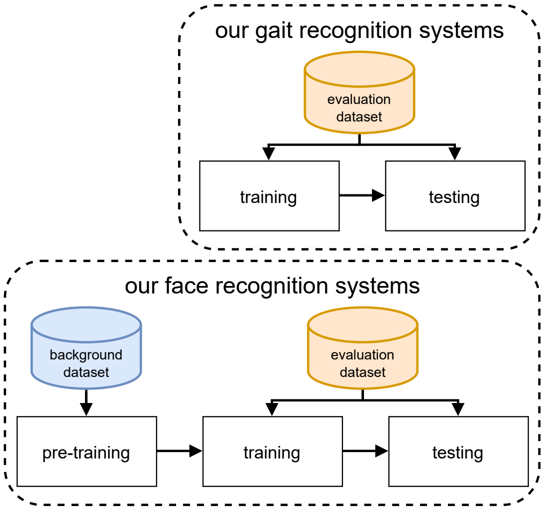

Figure 1: Dataset use through the phases of our recognition systems for face and gait recognition.

et. al. [56]. Their model is trained with anonymized data and then tested on anonymized data. The authors call this scenario parrot recognition, as opposed to training with clear data, which they call naive recognition. The authors report much better performance for parrot recognition compared to naive recognition.

Parrot recognition is another term for an informed attacker, as defined by Srivastava et. al. [74]. In the evaluation of voice anonymization, Srivastava et. al. [74] propose three attackers who differ in their awareness of the anonymization. The ignorant attacker is unaware of the anonymization (as in black-box assumptions), the semi-informed attacker knows the anonymization algorithm (as in gray-box assumptions), and the informed attacker knows the algorithm plus the given parameters (as in white-box assumptions).

The VoicePrivacy challenge [83, 84] used anonymized data to train a speaker verification system. The system was then tested against anonymized voice samples. It was found that training with anonymized data already improved recognition performance; however, performance improvement was greater when the recognition was pre-trained with anonymized data. The results of the VoicePrivacy challenge show that (pre-)training the recognition system with anonymized data leads to a much stronger evaluation of the privacy performance of a technique. Therefore, we recommend training and (where in use) also pre-training recognition systems with anonymized data. But even when a complete pre-training of the model is not possible, just training with anonymized data can already pose a more difficult challenge to an anonymization.

4.3.2 Test Against Different Recognition Systems. Most evaluation methods rely on the state-of-the-art recognition system currently available for the targeted biometric trait. However, during the design and development of recognition systems, anonymization is not considered. Consequently, recognition systems are not optimized to operate on anonymized data. For this reason, we challenge the assumption that the state-of-the-art recognition systems will also be the one that performs best on the anonymized data. Obviously, for practical reasons, not all types of recognition systems can be used in an evaluation. However, at least a few conceptually different recognition systems should be tested in order to assess which techniques work best on the anonymized data. The aim here is to approximate worst-case performance of the anonymization.

4.3.3 Use a More Challenging Evaluation Dataset. The datasets currently being used for the evaluation of biometric recognition are, as explained, recorded and designed to pose challenging recognition problem. It is our proposition, though, that evaluators of anonymization use an easy recognition problem in order to create a challenging anonymization scenario. Since the recording of biometric datasets is time-consuming and expensive (not to mention complicated by legal regulations like GDPR), we propose that existing recognition datasets be adapted so that the easy recognition problem becomes a hard anonymization problem. In particular, instead of using the entire dataset, we propose that the identities in the dataset be reduced in number. Further, we propose that the selection of identities be based on the criterion of easy distinguishability. For the reduced dataset, our strategies for identity selection are as follows:

- Random: As our baseline selection strategy, we use a random selection of identities. We repeat the selection multiple times to account for the variability of the selection.
- Classification: We use a biometric recognition system on the anonymized data to select the identities which have the highest identification accuracy.
- Metadata: We operationalize the fact that most biometric datasets also contain metadata about the identities, such as age and sex. Such metadata will typically be extractable via a recognition system. Our rationale is that identities with diverse attributes can be distinguished more easily when images are anonymized. Our procedure runs in three steps. First, we normalize each point of metadata information between 0 and 1, and then we calculate the pair-wise Euclidean distance between the points. Second, we obtain a subset of identities by locating pairs of identities at the greatest distances from one another. And third, we calculate the average of distances between the identities in our subset, and then we consistently select the identity located at the maximum distance to the average.
- Feature-space: Many recognition systems work by projecting the biometric data into a feature space and then calculating distances between the feature vectors. The rationale is that the recognition system is trained to project datapoints from the same identity onto similar features and as well, to project datapoints from different identities onto contrasting features. However, misclassification occurs when the feature of a datapoint belonging to one identity is farther from the correct feature and closer to a feature belonging to another identity. Therefore, we propose that recognition performance be improved by the intentional selection of identities whose feature spaces are distant from one another on anonymized data. In other words, we choose identities who are very different to one another when anonymized. We use this idea to develop two selection strategies:

– Distinctive: Inspired by the Biometric Menagerie [89], we calculate for each identity a genuine score and an imposter score (illustrated in Fig. 2). The genuine score of an identity is the furthest Euclidean distance of any feature vector of this identity to the average of all feature vectors of this identity. The imposter score is the shortest Euclidean distance of the average of all feature vectors of this identity to a feature vector of any other identity. Thus the genuine score is effectively an intra-class distance; conversely, the imposter score is effectively an inter-class distance. If the inter-class distance is high and the intra-class distance low, then the identity is less likely to be misclassified because the features of other identities lie farther away. In sum, we select identities that have the best average of genuine and imposter scores.

– Center: Our purpose is to create a subset of identities lying at the greatest distances from one another. As with the metadata vector above, we begin by selecting the two identities whose average feature vectors have the largest Euclidean distance. Then we consistently select the identities whose average feature vectors lie at maximum distances from the average feature vector of our subset of identities.

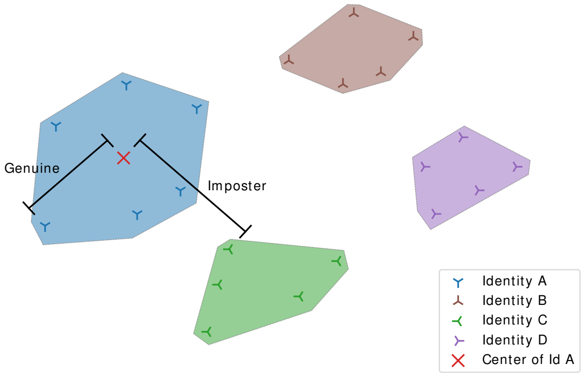

Figure 2: Simplified example for Genuine and Imposter scores of an identity A in a 2D projection of the feature space.

5 EXPERIMENTS

Our evaluation is based on the physiological biometric face and behavioral biometric gait. We begin by stating our hypotheses, and then we describe the experiments and present the results.

5.1 Hypotheses

Our aim in the evaluation is to test our three methodological proposals for improvements to the evaluation of biometric anonymization. We have proposed, first, that recognition systems also be trained on anonymized data; second, that multiple recognition systems be used; and third, that a more challenging dataset be used.

We begin our testing by formulating five hypotheses:

H1 Training the recognition system on anonymized data achieves more reliable anonymization performance than training on clear data.

H2 Training the recognition system on data in which a part of the samples is anonymized achieves more reliable anonymization performance than training on clear data.

H3 No single recognition system simulates worst-case performance on all anonymizations.

H4 A reduction in the number of identities in the evaluation dataset more robustly challenges the privacy protection of the anonymization.

H5 The identities selected by our selection strategies are a more robust challenge to the privacy protection of anonymization.

Our Hypotheses H1 and H2 hold that training recognition systems on anonymized data will achieve higher recognition performance. For our **H1**, we expect that (pre-)training recognition systems with anonymized data of the respective anonymization will result in higher recognition accuracies compared to (pre-)training on clear data. Further, for **H2**, we expect also that (pre-)training on partial anonymized datasets will perform better compared to (pre-)training on clear data. Further, we expect that increasing the amount of anonymized data in the train set will increase the recognition performance. We reason that the models we test must necessarily generalize more suitably to data that are noisier.

Our Hypothesis **H3** holds that no single recognition system will achieve the best performance on every anonymization. Our prediction for H3 is that, independent of results on clear data, some recognition systems will outperform others when using anonymized data. We reason that some recognition systems will better learn features from the anonymized data.

Our Hypothesis **H4** holds that reducing the number of identities in the evaluation dataset will present a more robust challenge to the performance of the anonymization. Our H5 builds on H4. For **H5**, we expect that selecting an evaluation dataset with our proposed selection strategies will pose a bigger challenge to the anonymization, and hence result in higher recognition performance then selecting random identities.

5.2 Experiments

We set an optimal performance bound by using chance-level performance of the anonymization as our baseline. We reason that perfect anonymization would leave adversaries with such a negligible advantage that their most effective strategy would be to guess identities at random. To approximate worst-case performance of the anonymization, we use the performance of clear level recognition, that is, the performance of the recognition system on clear data.

To test H1, we follow the same procedure for each anonymization technique: the recognition system is trained on the respective anonymized training data, and where possible, the system is also pre-trained on the anonymized data. To test H2, we (pre-)train the recognition system on different compositions of anonymized and clear training data using 25%, 50%, and 75% anonymized training data. Hence, we assess our H1 and H2 each with parrot and naive recognition.

For our H3, we use different recognition systems and perform parrot recognition for each anonymization.

For our H4, we again perform parrot recognition. However, instead of using the full evaluation dataset, we use only a random subset of identities of 50%, 25%, 12.5%, ..., until three of the original identities remain. For each number of identities, the sampling is repeated ten times to account for the variability of the random selection. Finally, in our last experiment for H5, we use the same numbers of identities as in the experiments for H4, but instead of randomly selecting, we choose identities according to the strategies described above in our methodology: Random, Classification, Metadata, Distinctive, and Center (see Subsection 4.3.3). We repeat the classification of the reduced dataset ten times to account for the randomness of the test/train split.

5.3 Datasets

For the face recognition, we use the CelebA [41] dataset because it is popular for face recognition and for anonymization evaluation, and we use the WebFace260M [91] dataset because its images are realistic. From both datasets we randomly select 1,000 identities as evaluation set and another 9,000 identities as background dataset for retraining. We only select identities with at least eight images, and we limit the maximum number of images per identity to 20. We crop all images to the face region, with images containing multiple faces cropped to the largest face. We resize all images to 224x224 pixel and rotate them until the eyes are level.

For gait we use the dataset of the gait patterns of 57 identities by Horst et. al. [26]. The dataset represents the most comprehensive publicly available dataset that contains multiple gait samples per identity, and this, in particular, recommends the dataset to the evaluation of anonymization performance. For each identity in the dataset there are 20 gait sequences, and we resample these to be 100 frames long. The dataset has used optical markers to capture motion. The motion capture covers 52 tracked points, each given as absolute 3D position (see Fig. 3).

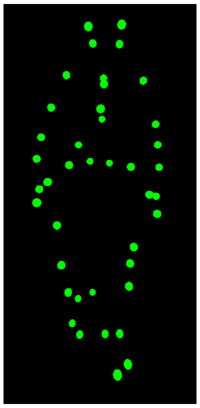

Figure 3: Sample pose of motion-captured gait information, represented as point-light walker.

5.4 Evaluation Framework

In order to run our experiments, we implemented the evaluation framework depicted in Fig. 4.

First, the clear dataset is copied and anonymized with a specific anonymization technique. Second, the selector performs a selection strategy to reduce the dataset to the configured numbers of identities. Third, the splitter splits the samples per identity into two sets, with 75% of samples going into the train set and 25% going into the test set. Depending on the configuration, either the clear samples or the anonymized samples go into the respective datasets.

Fourth and last, the recognition system is trained with the train set and evaluated with the test dataset. The resulting likelihood for a given test sample is recorded and saved for each identity.

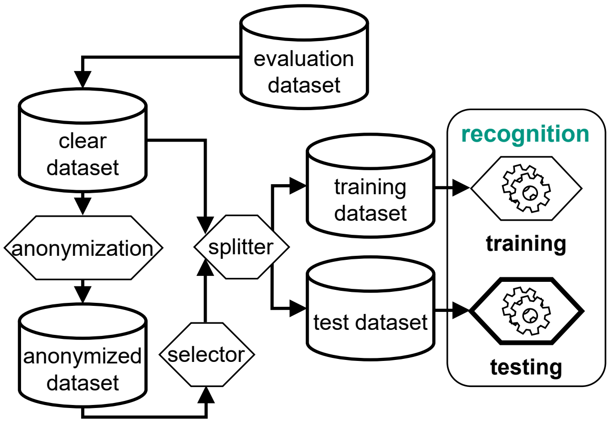

Figure 4: Schematic overview of the evaluation framework architecture, excluding pre-training for simplicity

5.5 Recognition Systems

For face recognition, we use the DeepFace [72] library because it covers the entire face recognition pipeline and includes pre-trained models for ArcFace [11], Facenet [71], and VGG-Face [60]. Additionally, we use the face recognition model (frknn) [14], which uses a pre-trained feature extractor and k-nearest neighbors for classification. In order to also test non-deep-learning approaches, we use a scalar, principal component analysis (PCA) and support vector machines (SVM) pipeline as described in a scikit tutorial[1] and a recognition method that uses Google AI's mediapipe[2] to extract 478 3-dimensional face landmarks before using a scalar, PCA and SVM pipeline on their coordinates. We also pre-train multiple models of ArcFace, which thereafter we referred to as Retrained ArcFace. For ArcFace pre-training, we used the remaining identities in CelebA or WebFace260M with the respective anonymization technique under evaluation applied to the samples. For %-parrot recognition approaches, we anonymized only the corresponding percentage of the samples in the background dataset. We validated Retrained ArcFace on clear data and achieved similar identification accuracy as the regular pre-trained ArcFace.

For gait recognition, we use two types of feature vectors. The flatten feature vector simply flattens all poses of a gait sequence into

[1] https://scikit-learn.org/stable/auto_examples/applications/plot_face_recognition.html

[2] https://developers.google.com/mediapipe/solutions/vision/face_landmarker

a single vector, as proposed by Horst et al. [26]. The simple feature vector does a PCA over all poses of a walking sequence and then concatenates the 4 first components of the PCA with an average over all poses of the sequence. For classification, we use SVM, random forest, and k-nearest neighbors. Unless stated otherwise, we used the combination SVM+flatten for gait recognition.

5.6 Anonymization Techniques

In the following, we present the anonymization techniques we use for our evaluation. For face anonymization, we select simple anonymization techniques such as blurring and state-of-the-art machine learning anonymizations such as CIAGAN [46]. For gait anonymization, we use a subset of the anonymizations used by Hanisch et al. [22]. If the anonymization is parameterized, we select the parameters in such a way that initially a low level of recognition accuracy is achieved. In this way, we can observe how our methodological improvements increase the recognition accuracy. Note that since we are investigating the efficiency of our methodological improvements, our selection of parameters does not allow a fair comparison of the anonymizations.

5.6.1 Face Anonymization. We consider the following techniques for face anonymization in our evaluation (see Fig. 5). The *Eye Masking* anonymization uses a black strip with 140 pixels height to cover the eye area of the face. *Gaussian Blur* applies a gaussian blur with a kernel size of 101. The anonymization k-randomized transparent overlays (*k-RTIO*) ($\alpha = 0.4$, $blocksize = 18$, $k = 3$) by Rajabi et al.[67] add a block-permuted semi-transparent overlay to the face image. The three methods *DP Pix* [12] ($\epsilon = 2$, $b = 12$, $m = 16$), *DP Snow* [30] ($d = 0.01$), and *DP Samp* [88] ($\epsilon = 5$, $k = 24$, $m = 12$) use differential privacy (DP) to provide formal privacy guarantees. We adapted these three methods from Reilly et al. [69] for RGB images. Our adaptation to RGB images prevents us from providing the formal guarantees given for grayscale images. Another formal privacy framework is k-anonymity, as used in the anonymization *k-Same-Pixel* ($k = 10$) by Newton et al. [57]. k-Same-Pixel expects a static dataset with a single image per identity. This does not apply to our scenario because we anonymize image by image and have multiple images per identity. Therefore, we use a separate background dataset with 200 identities. This means that the formal guarantees do not apply to our implementation. In Fawkes [73] ($mode = high$), adversarial machine learning is used to poison face recognition training data and thereby protect the identity in the picture. Both *DeepPrivacy* [27] and *CIAGAN* [46] anonymize faces by replacing them with new synthetic ones and then fitting them into the original background.

5.6.2 Gait Anonymization. For our gait experiments, we use simple anonymization techniques to select precisely the information to be perturbed in the samples. First, we suppress parts of the samples: *Keep(legs)* and *Keep(head)* both keep only the captured points for legs or head, respectively, while all other points are set to zero. Second, we perturb the samples: *Noise(x)* applies to each captured point normal ($\mu = 0$, $\sigma = 1$) distributed noise, which is scaled by 3, 10, or 100. Third, we generalize: *Motion Extraction* captures the differences between each next pose in order to extract only the dynamic parts of the data. The structure of the walkers is, then, effectively removed.

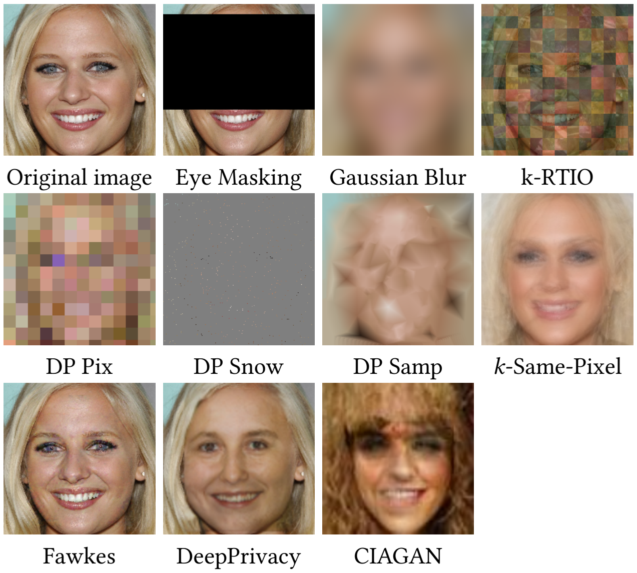

Figure 5: Example image for each of the face anonymization techniques we assess.

5.7 Selection Strategies

For our selections of face data using the Classification strategy, we use ArcFace to calculate the identification accuracy for each identity. We also use ArcFace to extract the feature vectors for the Center and Distinctive strategies. For gait, we use SVM+flatten for the Classification strategy and a PCA with four components over all samples as feature vector for Center and Distinctive.

5.8 Framework Implementation

Our evaluation framework was implemented using python (version 3.8) with numpy (1.19.5), scikit-learn (0.23), and DeepFace[72] (0.0.65) libraries.

6 RESULTS

We report here the results of our experiments. We assess, in turn, the validity of each of our hypotheses: whether recognition systems trained on anonymized data improve evaluation performance (H1, H2), whether no single recognition system performs best on every anonymization (H3), and lastly whether a reduction in the number of identities (H4) and whether a selection of identities in the evaluation dataset actually pose real challenges to the privacy protection of the anonymization (H5). In the Appendix B, there are additional results using the WebFace260M dataset and the selection on clear data.

6.1 Recognition Systems Trained on Anonymized Data Improve Evaluation Performance

In Fig. 6 and Fig. 7, we present the results of our experiments for H1 and H2 on the anonymization of face data and for gait data.

For face images, we find that, except for CIAGAN and *k*-Same-Pixel, all parrot recognition systems perform better than naive recognition. For *k*-Same-Pixel, all recognition systems have nearly the same performance, while for CIAGAN, naive recognition performs best. This anomaly in CIAGAN makes sense when we consider how CIAGAN performs the anonymization: every face is replaced by another face which shares the same soft biometrics. Therefore, we assume that CIAGAN's replacement of the face on each training image makes it harder for ArcFace Retrained to learn useful feature vectors.

We find significant results for parrot recognition of face anonymization. The performance of full parrot recognition and of all %-parrot recognition cluster close together for most anonymizations. In fact, %-parrot recognition often achieves the same performance as the full parrot recognition, and for DP Snow, the 75%-parrot recognition even outperforms the full parrot recognition.

In contrast to our results for face anonymization, the results for gait anonymization show full parrot recognition outperforming %-parrot recognition, with the exception of all Noise anonymization (cf. Fig. 7). For all gait anonymizations, naive recognition performs only at the chance-level. The %-parrot results for Noise(3) and Noise(10) are interesting because 25% performs best, 50% performs second best, 75% performs third best, and full parrot performs worst.

In our results for both face and gait anonymization, one thing defied our predictions. In the face and gait anonymization of DP Snow, Noise(3), and Noise(10) anonymization performance improves when the model is trained solely on a portion of anonymized images rather than on the full anonymized training set. We draw attention to the fact that all three anonymizations perform noise injection either by adding noise to each datapoint or by randomly removing pixels from the image. That portion of noisy data samples in the training set enables the recognition systems to adapt to DP Snow, Noise(3), and Noise(10) while still learning the features required for the classification from the clear data. We conclude, therefore, that there is a tipping point where more noisy data no longer improves training performance but begins impairing it.

6.2 No Single Recognition System Performs Best on All Anonymizations

We present the results of our experiments for H3 for the anonymization of face data in Fig. 8 and for the anonymization of gait data in Fig. 9.

All face anonymizations, except Fawkes, achieve a performance below 30% for all recognition systems except ArcFace Retrained. Fawkes achieves between 30% and 60% (except with Eigenfaces). The results for ArcFace Retrained differ significantly. With ArcFace Retrained, most anonymization techniques achieve much higher recognition rates. Only CIAGAN, DP Samp, and *k*-Same-Pixel are still below 30%, while Blur, DP Snow, and Fawkes are even above 60%. An interesting observation is that while Eigenfaces performs worst on clear data it performs better on DP Pix and Blur than most other recognition systems.

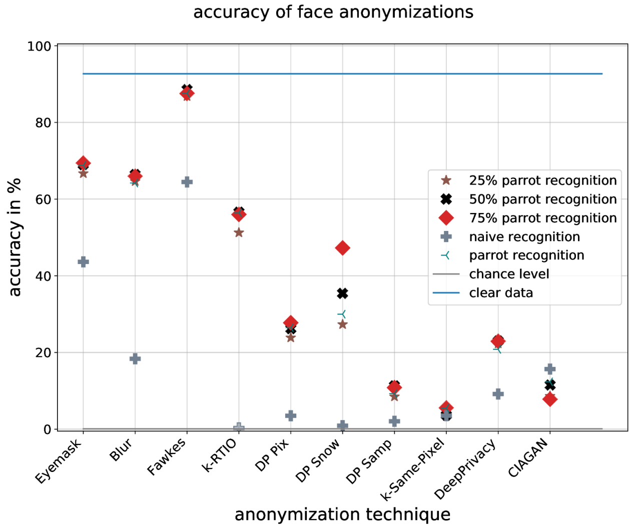

Figure 6: Accuracy for face anonymizations using ArcFace retrained on the CelebA dataset with naive, %-parrot, parrot recognition. A lower accuracy means better privacy protection.

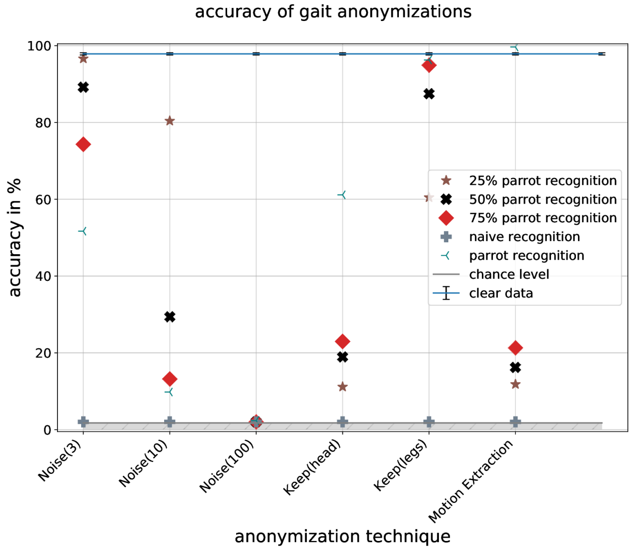

Figure 7: Accuracy for gait anonymizations using SVM+simple with naive, %-parrot, parrot recognition. A lower accuracy means better privacy protection.

For the gait data, all combinations of techniques perform between 80% and 98% on clear data, with SVM+flatten performing best on the clear data. The gait anonymization techniques across recognition systems perform in the same order, that is, we find the worst performance for Noise(100) and we find the best performance for either Keep(legs) or Motion Extraction.

We note that the differences between the gait anonymization techniques across the recognition systems can be quite large. For example, SVM+simple Noise(100), Noise(10), and Noise(3) score much higher when compared to the other recognition systems. However, among the anonymizations that do not use noise injection, SVM+simple scores lower than SVM+flatten. In sum, we observe that no single gait recognition system outperforms the others.

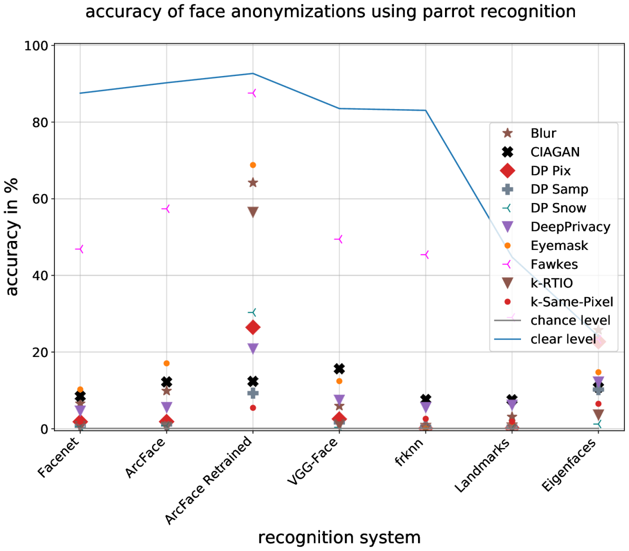

Figure 8: Accuracy of face anonymization over different recognition systems using parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.

6.3 Reducing the Number of Identities in the Evaluation Dataset Increases the Challenge for the Anonymization

We present the results of our experiments for H4 for the anonymization of face data in Fig. 10 and for the anonymization of gait data in Fig. 11.

For the face data, we assess the accuracy of our H4 by comparing the performances of parrot recognition on different numbers of identities in the evaluation dataset (see Fig. 10). For each number of identities (except the number of the full dataset), we selected 10 random subsets and calculated average performance and standard deviation. Every decrease in the number of identities increases the chance-level performance for the recognition systems. In short, the decreases make it easier for the recognition system to randomly guess an identity. We observe this increase in performance for all anonymization techniques. In particular, Fawkes attains the same performance plateau as initially on the clear data. Eyemask, Blur, and k-RTIO also start at high performance, but need longer to approach clear-level performance. *k*-Same-Pixel is the best performing anonymization. *k*-Same-Pixel stays close to the chance-level while mimicking the same increase in accuracy. In sum, we observe that decreases in numbers of identities increase the standard deviation of accuracy. From this, we reason that the selection of identities for the evaluation group is an decisive factor in evaluation accuracy.

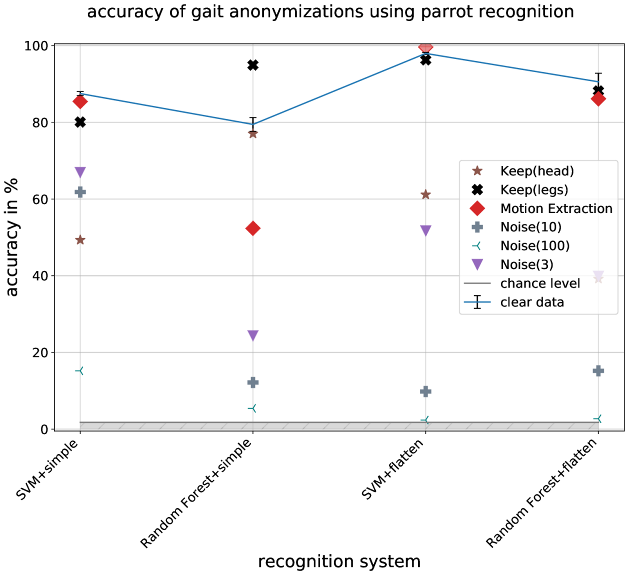

Figure 9: Accuracy of gait anonymization over different recognition systems using parrot recognition. A lower accuracy means better privacy protection.

For the gait data (Fig. 11), we observe a similar increase in recognition performance, except for the anonymization techniques Noise(10) and Noise(100), which stay close to the chance-level . The techniques Noise(10) and Noise(100) increase the standard deviation of the performance as the number of identities decreases. For the other gait anonymizations, we do not observe the same relation in the standard deviation.

We present the results of our experiments for H5 for the anonymization of face data in Fig. 12 and for the anonymization of gait data in Fig. 13.

Our selection strategies compare to random selection as follows: our strategies outperform when the number of identities is greater than 62, and under 62 Metadata starts performing worse than the best random selections, while the remaining techniques continue outperforming the best random selections down to 3 identities. Our Center and Classification strategies perform best across all numbers of identities, even matching the performance of random selection for 3 identities. What is more, for 500 to 15 identities, our Center and Classification strategies increases over 10% in performance compared to the best random selection.

For the gait data (Fig. 13), our results are not as good as for the face data. In general, we find that none of our selection strategies outperforms the best random selections. The strategy that performs consistently best is Classification. It always scores close to the best random selections. The strategy Metadata performs worst, as it does too in the face results. The strategies Center and Distinctive show varying results for different numbers of identities. Our explanation for the contrast between face and gait runs as follows: It is probable that the significant difference between the number of identities in

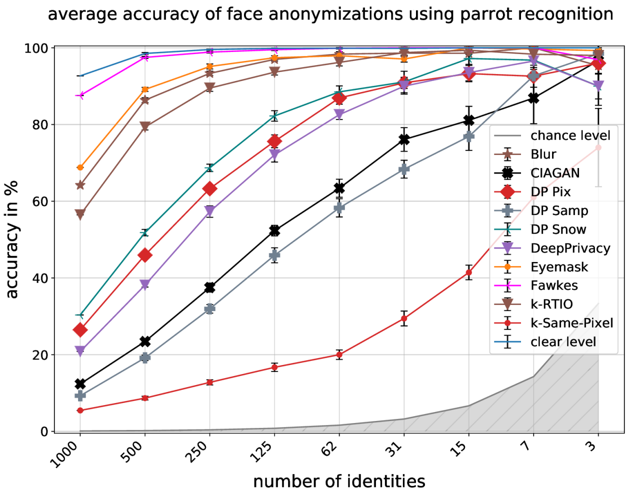

Figure 10: Mean accuracy of face recognition over ten random selections (excluding 1000 identities) from decreasing numbers of identities. The standard deviation of the random selection is given as error bars. ArcFace Retrained is used with parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.

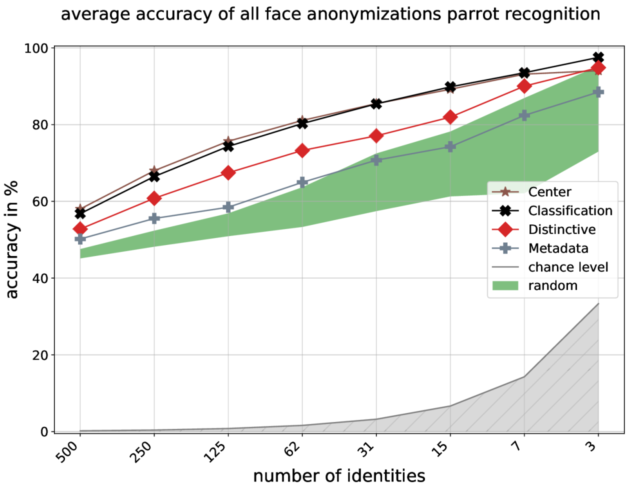

Figure 12: All accuracies are given as the average across all face anonymization techniques. The green area indicates the accuracy range of the previous ten random selections of identities. ArcFace Retrained is used with parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.

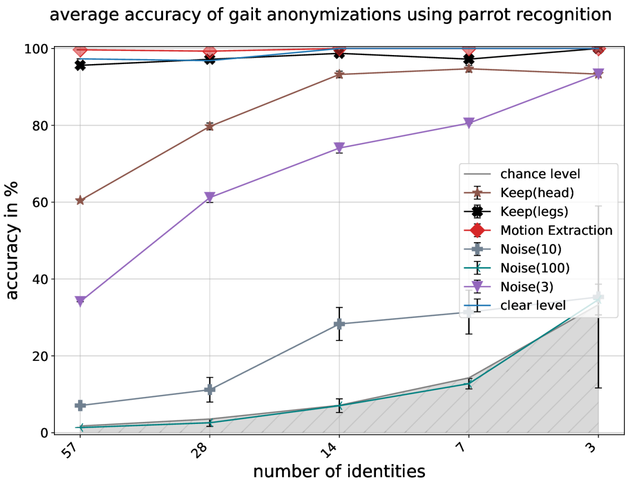

Figure 11: Mean accuracy of gait recognition over ten random selections (excluding 57 identities) for decreasing numbers of identities. The standard deviation of the random selection is given as error bars. SVM+flatten is used with parrot recognition. A lower accuracy means better privacy protection.

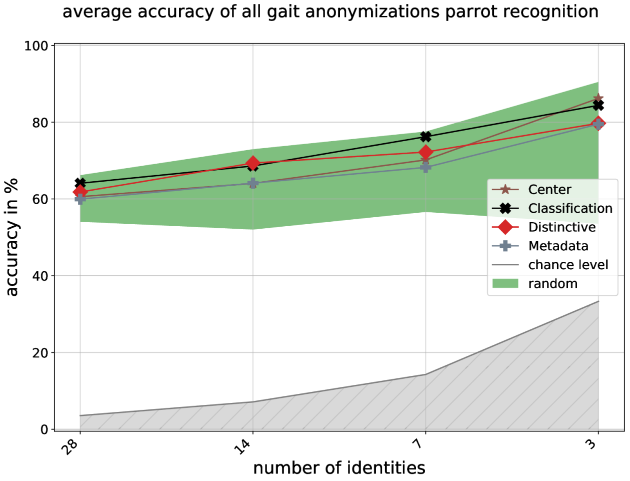

Figure 13: All accuracies are given as the average across all gait anonymization techniques. The green area indicates the accuracy range of the previous ten random selections of identities. SVM+flatten is used with parrot recognition. A lower accuracy means better privacy protection.

the full face dataset (n = 1,000) and the number in the full gait dataset (n = 57) results in less identities to pick from.

The accuracy we achieve with our Classification selection strategy deserves further attention here, because it performs best across anonymizations for both face and gait. We will examine Classification more closely by comparing it to the initial results for our decreases in numbers of identities.

For the face data (see Fig. 14), we observe that clear and Fawkes reach an early plateau close to 100% and that Eyemask, Blur, and k-RTIO begin scoring near the 80% mark and not near the 60% mark. For 125 identities, Eyemask, Blur, and k-RTIO also plateau earlier. DP Samp increases in accuracy steadily from 500 identities to 31 identities, and from there DP Samp accelerates in performance ultimately to achieve 100% at 3 identities. *k*-Same-Pixel achieves the lowest accuracies compared to the other anonymization techniques. However, *k*-Same-Pixel follows the same trend as the other techniques by steadily increasing as the identities decrease in number. When we compare to the random selection (see Fig. 10), we see an increase from around 60% to 90% for 3 identities. Similar increases can also be found for the other anonymization techniques. We conclude that the Classification strategy is effective in selecting identities that are hard for the anonymization techniques to anonymize.

For the gait data (see Fig. 15), we again find results similar to face. All anonymizations, except Noise(100), score higher. We consider this to be additional evidence that our Classification strategy is highly successful. Furthermore, we find that the Noise(100) results show that anonymization techniques exist which can achieve near perfect anonymization even in this challenging scenario.

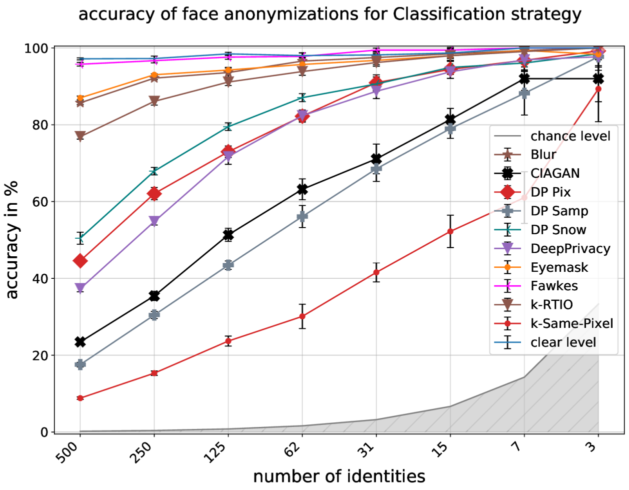

Figure 14: Accuracy of face anonymizations across decreasing numbers of identities. The strategy Center was used to select the identities. The error bars give the standard deviation over 10 test-train-splits. ArcFace Retrained is used with parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.

6.4 Summary of Results

- Recognition performance increases when the system is trained or especially pre-trained on anonymized data.
- Recognition performance increases when a reduction is made in the number of identities in the evaluation dataset.
- Our Classification selection strategy provides reliable evaluation of anonymization. When, however, the number of identities in the evaluation dataset is very small, Classification might be outperformed by best-case random selections.
- Anonymization techniques perform differently across recognition systems. As a direct consequence, it remains unclear which anonymization technique performs best in conjunction with which recognition system.
- For some anonymization techniques which use noise injection, it is crucial to determine the optimal proportion of anonymized data for both training and pre-training.

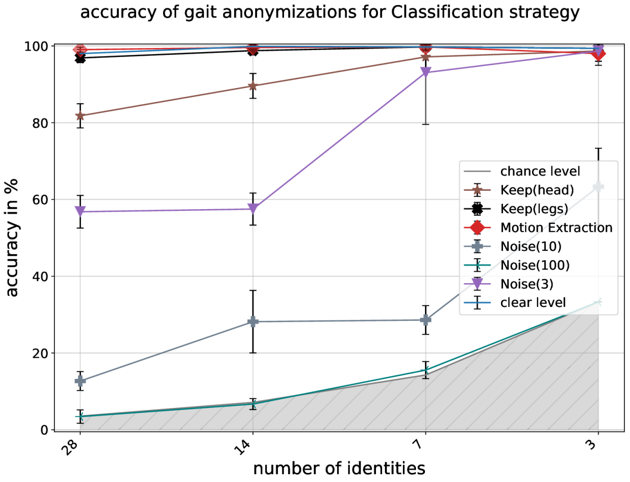

Figure 15: Accuracy of gait anonymizations across decreasing numbers of identities. The strategy Classification was used to select the identities. The error bars give the standard deviation over 10 test-train-splits. SVM+flatten is used with parrot recognition. A lower accuracy means better privacy protection.

7 DISCUSSION, LIMITATIONS, AND FUTURE WORK

The results of our three experiments confirm all five of our hypotheses. We see ourselves justified in drawing the overall conclusion that our methodological recommendations will improve the state-of-the-art in the evaluation of the anonymization of biometric data.

Our results for the Hypotheses H1 and H2 clearly show that training and also pre-training with anonymized data significantly improves the performance of the recognition and thus opens the door to improved evaluation of face and gait anonymization. As demonstrated for face anonymization, even a small amount of anonymized data greatly improves the training process. However, our results also indicate that an excess of noisy training data may decrease the performance. Therefore, for anonymization by noise injection (e.g. Laplace mechanism), we conclude that care should be taken to determine the right amount of anonymized training data. Nonetheless, we draw the final conclusion that training with anonymized data significantly improves the validity of the evaluation methodology. Without anonymized data in the train set, the performance of the anonymization is bound to be overestimated.

Our results for the Hypothesis H3 show that the recognition systems which perform comparably to one another on clear data may perform differently from one another on anonymized data. Since the performance on clear data is not a good predictor of performance on anonymized data, we conclude that the recognition system which seems to perform at the state-of-the-art on clear data might not accurately evaluate anonymization performance. This holds especially for anonymizations which use noise injection, as demonstrated by our results for gait anonymization. Therefore, we consider it the minimum that multiple recognition systems be used with different model architectures. Furthermore, we recommend designing recognition systems to be more resistant to anonymization. Our reason is clear: there is no single recognition system that performs best in all cases, not for face anonymization and not for gait anonymization. Understanding which recognition system architecture works best for which anonymization together with training the system on anonymized data will help to achieve more reliable evaluation results.

Our results for the Hypothesis H4 confirm that for most anonymization techniques, a reduced number of identities in the evaluation dataset increases the recognition performance more than what the increase in chance-level can explain. This reduction in the number of identities presents a more challenging scenario for the anonymization. Our results for H4 also show that, as the number of identities decreases, the run-to-run variation of possible results increases. We conclude that the selection of identities for the subset is a significant task in the evaluation of anonymization performance.

Our results for the Hypothesis H5 clearly indicate that a more challenging dataset is generated when our Classification selection strategy is used to select the identities for a reduced evaluation dataset. However, it appears that for very small datasets, multiple random selections can still outperform our Classification selection strategy. Hence, we recommend performing Classification and additionally the random selections in order to determine which performs best at identity selection for the evaluation dataset.

All in all, our proposed improvements will evaluate biometric anonymization techniques much more convincingly than these techniques are currently being evaluated. Further research, however, is clearly necessary. For example, our methodological improvements will need to be validated on other biometric traits. In addition, it remains an open research question precisely which types of recognition systems perform best on which types of anonymization. Answers here will help decide whether, in fact, a systematic approach exists for building recognition systems that perform well on specific anonymizations.

8 CONCLUSION

Biometric recognition technologies, such as face recognition systems, pose a real threat to privacy. Therefore, a crucial technique for privacy protection is anonymization, and likewise, evaluation is crucial to anonymization. This paper assesses the state-of-the-art methodologies used for the evaluation of anonymization techniques, finds flaws in those methodologies, and proposes how the methodologies can be improved.

We find several major flaws in the state-of-the-art methodologies for the evaluation of biometric anonymization. The state-of-the-art evaluation is based on weak and unrealistic assumptions about the adversary. These adversaries act in ignorance of the anonymization in place and are accordingly unable to adapt their recognition systems. These are not realistic adversaries of anonymization techniques. Therefore, the state-of-the-art methodologies largely fail to assess accurately the performance of the recognition.

To begin the work of correcting such flaws, we have proposed to improve the evaluation methodology for the anonymization of biometric data. It is our recommendation that recognition systems which are trained not only on clear data but also on anonymized data be used to evaluate anonymization performance. Furthermore, we argue that the use of a variety of different recognition systems will improve the rigor of the evaluation. The use of merely a single classifier trained only on clear data might result in unreliable, overoptimistic estimates of anonymization performance. Hence, we recommend using multiple recognition systems trained on anonymized data. And lastly, we recommend using a more challenging evaluation dataset to approximate worst-case performance. Our results indicate that such a dataset can be constructed by reducing the number of identities and selecting the easy-to-distinguish identities with our proposed Classification strategy. We have proposed improvements to the state-of-the-art in evaluation methodologies that will pre-empt overestimations of biometric anonymization performance. We have backed this finding with strong experimental evidence. Thus, we conclude that our proposed improvements lay the cornerstone of a more reliable evaluation methodology for the anonymization of biometric data.

ACKNOWLEDGMENTS

We would like to thank our in-house textician for his helpful feedback and valuable contributions in editing our manuscript, which have improved the clarity and scientific quality of our research. This work has been funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, and by funding of the Helmholtz Association through the KASTEL Security Research Labs (HGF Topic 46.23).

REFERENCES

[1] ISO/IEC JTC 1. 2017. *Information technology — Vocabulary — Part 37: Biometrics.* Standard. International Organization for Standardization, Geneva, CH.

[2] Mohamed Abou-Zleikha, Zheng-Hua Tan, Mads Graesboll Christensen, and Soren Holdt Jensen. 2015. A discriminative approach for speaker selection in speaker de-identification systems. In *European Signal Processing Conference.*

[3] Prachi Agrawal and P. J. Narayanan. 2011. Person De-Identification in Videos. *IEEE Trans. Circuits Syst. Video Technol.* 21 (2011).

[4] Ranya Aloufi, Hamed Haddadi, and David Boyle. 2019. *Emotionless: Privacy-Preserving Speech Analysis for Voice Assistants.* Technical Report arXiv:1908.03632.

[5] Fahimeh Bahmaninezhad, Chunlei Zhang, and John Hansen. 2018. Convolutional Neural Network Based Speaker De-Identification. In *The Speaker and Language Recognition Workshop* (2018-06-26). ISCA, 255–260. https://doi.org/10.21437/Odyssey.2018-36

[6] Jean-François Bonastre, Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Paul-Gauthier Noé, Jose Patino, Md Sahidullah, Brij Mohan Lal Srivastava, Massimiliano Todisco, Natalia Tomashenko, Emmanuel Vincent, Xin Wang, and Junichi Yamagishi. 2021. *Benchmarking and challenges in security and privacy for voice biometrics.* Technical Report arXiv:2109.00281 [cs, eess]. ISCA. https://doi.org/10.21437/spsc.2021-11

[7] Michael Boyle, Christopher Edwards, and Saul Greenberg. 2000. The effects of filtered video on awareness and privacy. In *ACM Conference on Computer supported cooperative work.*

[8] Efe Bozkir, Onur Günlü, Wolfgang Fuhl, Rafael F. Schaefer, and Enkelejda Kasneci. 2020. Differential Privacy for Eye Tracking with Temporal Correlations. _eprint: 2002.08972.

[9] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry. *ACM Trans. Inf. Syst. Secur.* 15, 3 (2012), 1–22. https://doi.org/10.1145/2382448.2382450

[10] Alice Cohen-Hadria, Mark Cartwright, Brian McFee, and Juan Pablo Bello. 2019. Voice Anonymization in Urban Sound Recordings. In *IEEE Workshop on Machine Learning for Signal Processing*.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019-06). IEEE, 4685–4694. https://doi.org/10.1109/cvpr.2019.00482

[12] Liyue Fan. 2018. Image Pixelization with Differential Privacy. In *Data and Applications Security and Privacy XXXII* (2018). Springer International Publishing, 148–162. https://doi.org/10.1007/978-3-319-95729-6_10

[13] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-Francois Bonastre. 2019. *Speaker Anonymization Using X-vector and Neural Waveform Models.* Technical Report arXiv:1905.13561 [cs, eess, stat].

[14] Adam Geitgey. 2021. Face Recognition. https://github.com/ageitgey/face_recognition.

[15] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. 2015. On the Reliability of Profile Matching Across Large Online Social Networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015-08). ACM. https://doi.org/10.1145/2783258.2788601

[16] Eric Granger and Dmitry Gorodnichy. 2014. *Evaluation methodology for face recognition technology in video surveillance applications.* Defence R & D Canada.

[17] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. 2005. Integrating utility into face de-identification. In *ACM Workshop on Privacy Enhancing Technologies*.

[18] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando De la Torre, and Simon Baker. 2009. Face de-identification. In *Protecting privacy in video surveillance*. Springer.

[19] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. 2006. Model-based face de-identification. In *IEEE Computer vision and pattern recognition workshop*.

[20] Jihun Hamm. 2017. Enhancing utility and privacy with noisy minimax filters. In *IEEE Acoustics, Speech and Signal Processing*.

[21] Simon Hanisch, Patricia Arias-Cabarcos, Javier Parra-Arnau, and Thorsten Strufe. 2021. *Privacy-Protecting Techniques for Behavioral Data: A Survey.* Technical Report 2109.04120. arXiv. http://arxiv.org/abs/2109.04120

[22] Simon Hanisch, Evelyn Muschter, Adamantini Chatzipanagioti, Shu-Chen Li, and Thorsten Strufe. 2022. Understanding person identification via gait. *arXiv preprint arXiv:2203.04179* (2022).

[23] Hanxiang Hao, David Guera, Janos Horvath, Amy R. Reibman, and Edward J. Delp. 2020. Robustness Analysis of Face Obscuration. In *IEEE Conference on Automatic Face and Gesture Recognition* (Buenos Aires, Argentina, 2020-11). IEEE, 176–183. https://doi.org/10.1109/FG47880.2020.00021

[24] Kei Hashimoto, Junichi Yamagishi, and Isao Echizen. 2016. Privacy-preserving sound to degrade automatic speaker verification performance. In *IEEE Conference on Acoustics, Speech and Signal Processing*.

[25] Yuki Hirose, Kazuaki Nakamura, Naoko Nitta, and Noboru Babaguchi. 2019. Anonymization of Gait Silhouette Video by Perturbing Its Phase and Shape Components. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

[26] Fabian Horst, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller, and Wolfgang I. Schöllhorn. 2019. Explaining the unique nature of individual gait patterns with deep learning. *Nature Scientific reports* 9, 1 (2019). https://doi.org/10.1038/s41598-019-38748-8

[27] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. *Advances in Visual Computing* (Sept. 2019), 565–578. https://doi.org/10.1007/978-3-030-33720-9_44 arXiv: 1909.04538. Implementation: https://github.com/hukkelas/DeepPrivacy.

[28] M. Ivasic-Kos, A. Iosifidis, A. Tefas, and I. Pitas. 2014. Person de-identification in activity videos. In *IEEE Convention on Information and Communication Technology, Electronics and Microelectronics*.

[29] Qin Jin, Arthur R. Toth, Tanja Schultz, and Alan W. Black. 2019. Voice convergin: Speaker de-identification by voice transformation. In *IEEE Conference on Acoustics, Speech and Signal Processing*.

[30] Brendan John, Ao Liu, Lirong Xia, Sanjeev Koppal, and Eakta Jain. 2020. Let It Snow: Adding pixel noise to protect the user's identity. In *Symposium on Eye Tracking Research and Applications*. ACM, Stuttgart Germany, 1–3. https://doi.org/10.1145/3379157.3390512

[31] Théo Jourdan, Antoine Boutet, and Carole Frindel. 2018. Toward privacy in IoT mobile devices for activity recognition. In *EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*.

[32] Tadej Justin, Vitomir Struc, Simon Dobrisek, Bostjan Vesnicer, Ivo Ipsic, and France Mihelic. 2015. Speaker de-identification using diphone recognition and speech synthesis. In *IEEE Automatic Face and Gesture Recognition*.

[33] Gokce Keskin, Tyler Lee, Cory Stephenson, and Oguz H. Elibol. 2019. *Measuring the Effectiveness of Voice Conversion on Speaker Identification and Automatic Speech Recognition Systems.* Technical Report 1905.12531. arXiv.

[34] Nigar Sen Koktas, Nese Yalabik, and Gunes Yavuzer. 2006. Ensemble Classifiers for Medical Diagnosis of Knee Osteoarthritis Using Gait Data. In *2006 5th International Conference on Machine Learning and Applications (ICMLA'06)*. 225–230. https://doi.org/10.1109/ICMLA.2006.22

[35] Pavel Korshunov and Touradj Ebrahimi. 2013. Using face morphing to protect privacy. In *IEEE Advanced Video and Signal Based Surveillance*.

[36] Pavel Korshunov and Touradj Ebrahimi. 2013. Using warping for privacy protection in video surveillance. In *IEEE Conference on Digital Signal Processing*.

[37] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Florian Müller. 2020. What does your gaze reveal about you? On the privacy implications of eye tracking. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 226–241.

[38] Minh-Ha Le and Niklas Carlsson. 2023. StyleID: Identity Disentanglement for Anonymizing Faces. *Proceedings on Privacy Enhancing Technologies* 2023, 1 (2023), 264—-278. https://doi.org/10.56553/popets-2023-0016

[39] Minh-Ha Le, Md Sakib Nizam Khan, Georgia Tsaloli, Niklas Carlsson, and Sonja Buchegger. 2020. AnonFACES: Anonymizing Faces Adjusted to Constraints on Efficacy and Security. In *Workshop on Privacy in the Electronic Society* (Virtual Event USA, 2020-11). ACM. https://doi.org/10.1145/3411497.3420220

[40] Juho Leinonen, Petri Ihantola, and Arto Hellas. 2017. Preventing Keystroke Based Identification in Open Data Sets. In *ACM Conference on Learning @ Scale*.

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Computer Vision* (2015-12). IEEE. https://doi.org/10.1109/iccv.2015.425

[42] Paula Lopez-Otero, Carmen Magariños, Laura Docio-Fernandez, Eduardo Rodriguez-Banga, Daniel Erro, and Carmen Garcia-Mateo. 2017. Influence of speaker de-identification in depression detection. *IET signal process.* 11 (2017).

[43] Emanuele Maiorana, Patrizio Campisi, and Alessandro Neri. 2011. Bioconvolving: Cancelable templates for a multi-biometrics signature recognition system. In *IEEE International Systems Conference*.

[44] Richard Matovu and Abdul Serwadda. 2016. Your substance abuse disorder is an open secret! Gleaning sensitive personal information from templates in an EEG-based authentication system. In *IEEE Conference on Biometrics Theory, Applications and Systems*.

[45] Richard Matovu, Abdul Serwadda, David Irakiza, and Isaac Griswold-Steiner. 2018. Jekyll and Hyde: On The Double-Faced Nature of Smart-Phone Sensor Noise Injection. In *IEEE Conference of the Biometrics Special Interest Group*.

[46] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020-06). IEEE, 5446–5455. https://doi.org/10.1109/CVPR42600.2020.00549 arXiv:2005.09544 [cs]. Implemenation: https://github.com/dvl-tum/ciagan.

[47] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. *Defeating Image Obfuscation with Deep Learning.* Technical Report 1609.00408. arXiV. http://arxiv.org/abs/1609.00408

[48] Lily Meng and Zongji Sun. 2014. Face de-identification with perfect privacy protection. In *IEEE Information and Communication Technology, Electronics and Microelectronics*.

[49] Lily Meng, Zongji Sun, Aladdin Ariyaeeinia, and Ken L Bennett. 2014. Retaining expressions on de-identified faces. In *IEEE Information and Communication Technology, Electronics and Microelectronics*.

[50] Denis Migdal and Christophe Rosenberger. 2019. Keystroke Dynamics Anonymization System. In *Joint Conference on e-Business and Telecommunications*.

[51] John V. Monaco and Charles C. Tappert. 2017. *Obfuscating Keystroke Time Intervals to Avoid Identification and Impersonation.* Technical Report arXiv:1609.07612 [cs].

[52] Andreas Nautsch, Jose Patino, Natalia Tomashenko, Junichi Yamagishi, Paul-Gauthier Noe, Jean-Francois Bonastre, Massimiliano Todisco, and Nicholas Evans. 2020. The Privacy ZEBRA: Zero Evidence Biometric Recognition Assessment. In *Interspeech* (2020-10). ISCA. https://doi.org/10.21437/interspeech.2020-1815

[53] Alexandru Nelus and Rainer Martin. 2018. Gender Discrimination Versus Speaker Identification Through Privacy-Aware Adversarial Feature Extraction. In *ITG Symposium on Speech Communication*.

[54] Alexandru Nelus and Rainer Martin. 2019. Privacy-aware Feature Extraction for Gender Discrimination versus Speaker Identification. In *IEEE Acoustics, Speech and Signal Processing*.

[55] Carman Neustaedter, Saul Greenberg, and Michael Boyle. 2006. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction* 13 (2006).

[56] Elaine Newton, Latanya Sweeney, and Bradley Malin. 2003. *Preserving Privacy by De-identifying Facial Images.* Technical Report CMU-CS-03-119. CMU.

[57] Elaine M Newton, Latanya Sweeney, and Bradley Malin. 2005. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering* 17, 2 (2005), 232–243. https://doi.org/10.1109/tkde.2005.32

[58] Paul-Gauthier Noé, Jean-François Bonastre, Driss Matrouf, N. Tomashenko, Andreas Nautsch, and Nicholas Evans. 2020. Speech Pseudonymisation Assessment Using Voice Similarity Matrices. In *Interspeech* (2020-10). ISCA. https://doi.org/10.21437/interspeech.2020-2720

[59] Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Jose Patino, Jean-François Bonastre, Natalia Tomashenko, and Driss Matrouf. 2022. Towards a unified assessment framework of speech pseudonymisation. *Computer Speech & Language* 72 (2022), 101299. https://doi.org/10.1016/j.csl.2021.101299

[60] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. Publisher: British Machine Vision Association.

[61] Sree Hari Krishnan Parthasarathi, Hervé Bourlard, and Daniel Gatica-Perez. 2011. LP residual features for robust, privacy-sensitive speaker diarization. In *Interspeech*.

[62] M. Pobar and I. Ipsic. 2014. Online speaker de-identification using voice transformation. In *IEEE Convention on Information and Communication Technology, Electronics and Microelectronics*.

[63] Jose Portelo, Bhiksha Raj, Alberto Abad, and Isabel Trancoso. 2014. Privacy-preserving speaker verification using garbled GMMS. In *European Signal Processing Conference*.

[64] Jiří Přibil, Anna Přibilová, and Jindřich Matoušek. 2018. Evaluation of speaker de-identification based on voice gender and age conversion. *Journal of Electrical Engineering* 69 (2018).

[65] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiangyang Li. 2021. Speech Sanitizer: Speech Content Desensitization and Voice Anonymization. *IEEE Transactions on Dependable and Secure Computing* (2021).

[66] Jianwei Qian, Haohua Du, Jiahui Hou, Linlin Chen, Taeho Jung, and Xiang-Yang Li. 2018. Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity. In *ACM Conference on Embedded Networked Sensor Systems*.

[67] Arezoo Rajabi, Rakesh B. Bobba, Mike Rosulek, Charles V. Wright, and Wu-chi Feng. 2021. On the ImPracticality of Adversarial Perturbation for Image Privacy. *Proceedings on Privacy Enhancing Technologies* 2021, 1 (2021), 85–106. https://doi.org/10.2478/popets-2021-0006

[68] Shantanu Rane. 2014. Standardization of Biometric Template Protection. *IEEE MultiMedia* 21, 4 (2014), 94–99. https://doi.org/10.1109/mmul.2014.65

[69] Dominick Reilly and Liyue Fan. 2021. A Comparative Evaluation of Differentially Private Image Obfuscation. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, Atlanta, GA, USA, 80–89. https://doi.org/10.1109/TPSISA52974.2021.00009

[70] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication* 47 (2016), 131–151. https://doi.org/10.1016/j.image.2016.05.020

[71] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Computer Vision and Pattern Recognition* (2015-06). IEEE. https://doi.org/10.1109/cvpr.2015.7298682

[72] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *IEEE Intelligent Systems and Applications Conference*.

[73] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. In *29th USENIX security symposium (USENIX Security 20)*. USENIX, 1589–1604.

[74] Lal Srivastava, Brij Mohan, Nathalie Vauquier, Md Sahidullah, Aurelien Bellet, Marc Tommasi, and Emmanuel Vincent. 2020. Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona, Spain, 2020-05). IEEE, 2802–2806. https://doi.org/10.1109/ICASSP40776.2020.9053868

[75] Julian Steil, Inken Hagestedt, Michael Xuelin Huang, and Andreas Bulling. 2019. Privacy-Aware Eye Tracking Using Differential Privacy. In *ACM Eye Tracking Research & Applications*.

[76] Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2013. *Classify, but Verify: Breaking the Closed-World Assumption in Stylometric Authorship Attribution*. Technical Report. Drexel University. 17 pages.

[77] Jimmy Tekli, Bechara al Bouna, Raphael Couturier, Gilbert Tekli, Zeinab al Zein, and Marc Kamradt. 2019. A Framework for Evaluating Image Obfuscation under Deep Learning-Assisted Privacy Attacks. In *2019 17th International Conference on Privacy, Security and Trust (PST)* (Fredericton, NB, Canada, 2019-08). IEEE, 1–10. https://doi.org/10.1109/PST47121.2019.8949040

[78] Ngoc-Dung T. Tieu, Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. An RGB Gait Anonymization Model for Low-Quality Silhouettes. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

[79] Ngoc-Dung T. Tieu, Huy H. Nguyen, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen. 2017. An approach for gait anonymization using deep learning. In *IEEE Workshop on Information Forensics and Security*.

[80] Ngoc-Dung T. Tieu, Huy H. Nguyen, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen. 2019. Spatio-temporal generative adversarial network for gait anonymization. *Journal of Information Security and Applications* 46 (2019).

[81] Julian Todt, Simon Hanisch, and Thorsten Strufe. 2022. Fantômas: Evaluating Reversibility of Face Anonymizations Using a General Deep Learning Attacker. https://doi.org/10.48550/ARXIV.2210.10651

[82] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Interspeech 2020* (2020-10). ISCA. https://doi.org/10.21437/interspeech.2020-1333

[83] N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. *Post-evaluation analysis for the VoicePrivacy 2020 Challenge: Using anonymized speech data to train attack models and ASR.* Technical Report. VoicePrivacy Challenge 2020.

[84] Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, Anaïs Chanclu, Jean-François Bonastre, Massimiliano Todisco, and Mohamed Maouche. 2021. *The VoicePrivacy 2020 Challenge: Results and findings.* Technical Report 2109.00648. arXiv. arXiv:2109.00648 http://arxiv.org/abs/2109.00648

[85] Tavish Vaidya and Micah Sherr. 2019. You Talk Too Much: Limiting Privacy Exposure Via Voice Input. In *IEEE Security and Privacy Workshops*.

[86] Gabriele Vassallo, Tim Van hamme, Davy Preuveneers, and Wouter Joosen. 2017. Privacy-Preserving Behavioral Authentication on Smartphones. In *Workshop on Human-centered Sensing, Networking, and Systems*.

[87] Changsheng Wan, Li Wang, and Vir V. Phoha. 2018. A Survey on Gait Recognition. *ACM Comput. Surv.* 51, 5, Article 89 (aug 2018), 35 pages. https://doi.org/10.1145/3230633

[88] Han Wang, Shangyu Xie, and Yuan Hong. 2019. VideoDP: A Universal Platform for Video Analytics with Differential Privacy. http://arxiv.org/abs/1909.08729 arXiv:1909.08729 [cs].

[89] Neil Yager and Ted Dunstone. 2010. The Biometric Menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 2 (2010), 220–230. https://doi.org/10.1109/TPAMI.2008.291

[90] Yue Yao, Josephine Plested, Tom Gedeon, Yuchi Liu, and Zhengjie Wang. 2019. Improved Techniques for Building EEG Feature Filters. In *IEEE Conference on Neural Networks*.

[91] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. 2021. WebFace260M: A Benchmark Unveiling the Power of Million-scale Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A SURVEY OF STATE-OF-THE-ART EVALUATION METHODOLOGY FOR THE ANONYMIZATION OF BIOMETRIC DATA

In order to learn about the current state-of-the-art for evaluating biometric anonymization, we perform a survey study on 49 papers (see Table 1). The majority of papers are from Hanisch et. al. [21], which collected papers that perform behavioral data anonymization and include traits like voice, gait, and brain activity. Additionally, we use the corpus of face anonymization papers from a survey by Ribaric et. al. [70], which focuses on anonymization in media content. We filtered the papers to match our scenario.

Our first category for separating the corpus is the *biometric trait* which the anonymization tries to protect and also the *protection goal*. The protection goal may be either to prevent identity disclosure or attribute disclosure. Since the anonymization approaches are tested against a biometric recognition system, we note whether the evaluations rely on a single approach or test *multiple recognition systems*. Further, we examine whether *multiple parameters* for the anonymization technique are evaluated. Our main interest in this survey was to learn which kind of attacker model the evaluations employed. For this, we compare whether an *open-set* or *closed-set* model was applied and with which kind of *training data* (clear or anonymized) the recognition system was trained. Further, we check if the *reversibility* of the anonymization approach was tested. Moreover, we compare the different *metrics* employed to measure the anonymization performance.

Taking together all the reviewed papers, we find that most focus on anonymizing voice data, then face, gait, and hand. Only one

Trait	Papers (Count and Sources)
Voice	22 ([2], [5], [4], [20], [74], [64], [53], [66], [42], [61], [33], [62], [54], [24], [63], [13], [32], [65], [10], [29], [85])
Face	10 ([55], [48], [17], [19], [57], [18], [49], [7], [35], [36])
Gait	8 ([3], [25], [28], [31], [45], [78], [79], [80])
Brain Activity	2 ([90], [44])
Eye-gaze	2 ([75], [8])
Hand	5 ([40], [43], [50], [51], [86])

Table 1: Publications included in the state-of-the-art survey with corresponding trait

Trait	Voice	Face	Gait	Hand	Brain	Eye
	22	10	8	5	2	2
Protection Goal	Identity		Attribute		Both	
	38		6		5	
Metric	Accuracy		EER		Other	
	36		10		3	

Table 2: Publication count for biometric trait, protection goal, and metric to evaluate the technique

	Yes	No
anonymized training data	8	41
test reversibility	1	48
closed-set assumption	38	11
multiple parameters	28	21
multiple recognition systems	12	37

Table 3: The number of papers for the remaining categories

paper tackles brain activity and one other tackles eye-gaze (see Table 2). Most papers try to protect against identity inference, while six paper try to protect against attribute inference, and five against both identity inference and attribute inference. Regarding metrics for the measurement of privacy protection, we find that accuracy (also including metrics closely based on accuracy e.g. $1 - accuracy$) is the most commonly used metric, followed by the equal error rate (EER). Some uncommon metrics we observed were the usage of F1-Score [45], and half total error rates (HTER) [44].

As seen in Table 3 slightly more than half of the papers evaluate different parameter configurations for their anonymization technique, while only about one in four papers uses more than one recognition system for its evaluation. For the recognition scenario, we find that most papers use a closed-set approach. When it comes to training the recognition system, all papers use clear data for training, only a minority also trains the recognition system with anonymized data. For the test whether the anonymization technique can be reversed, we find only one paper [66] that considers this for the evaluation, although it only performed a theoretical analysis.

B ADDITIONAL RESULTS

In the following we report additional results of our experiments, this includes the reproduction of H4 and H5 on the WebFace260M dataset (see Fig. 16 and Fig. 17), the Metadata strategy has been excluded as no soft biometric information of the identities was available. We also report the performance of our selection strategies on clear, instead of anonymized, data on the CelebA dataset for all strategies (see Fig. 18) and the Distinctive strategy (see Fig. 19) in particular.

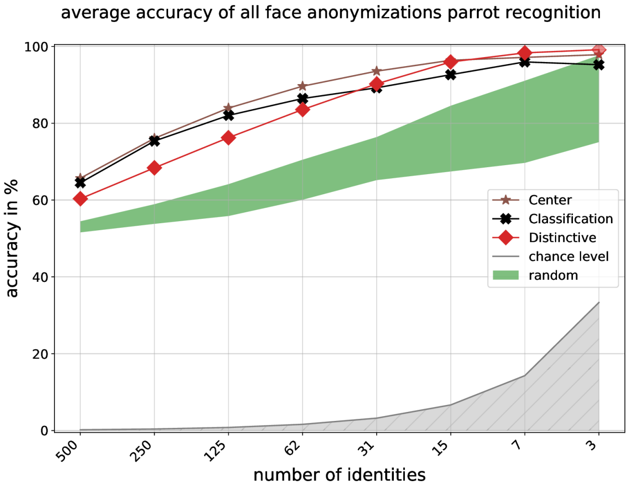

Figure 16: All accuracies are given as the average across all face anonymization techniques. The green area indicates the accuracy range of the previous ten random selections of identities. ArcFace Retrained is used with parrot recognition on the WebFace260M dataset. A lower accuracy means better privacy protection.

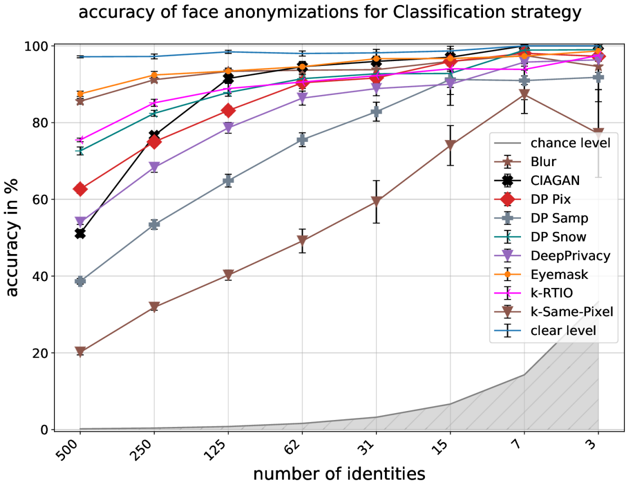

Figure 17: Accuracy of face anonymizations across decreasing numbers of identities. The strategy Center was used to select the identities. The error bars give the standard deviation over 10 test-train-splits. ArcFace Retrained is used with parrot recognition on the WebFace260M dataset. A lower accuracy means better privacy protection.

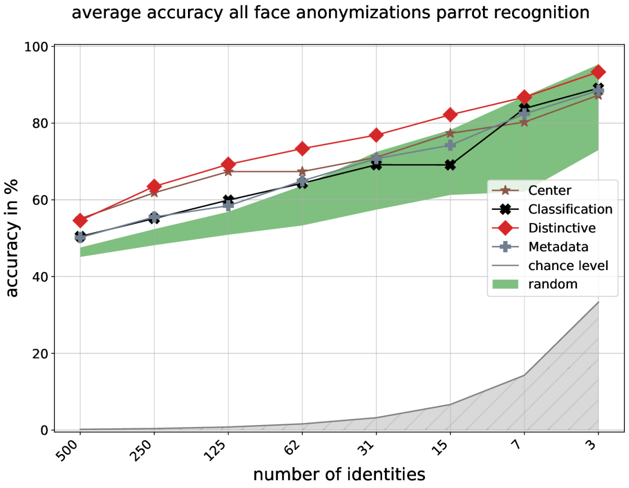

Figure 18: All accuracies are given as the average across all face anonymization techniques. The green area indicates the accuracy range of the previous ten random selections of identities. The selection have been performed on clear data. ArcFace Retrained is used with parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.

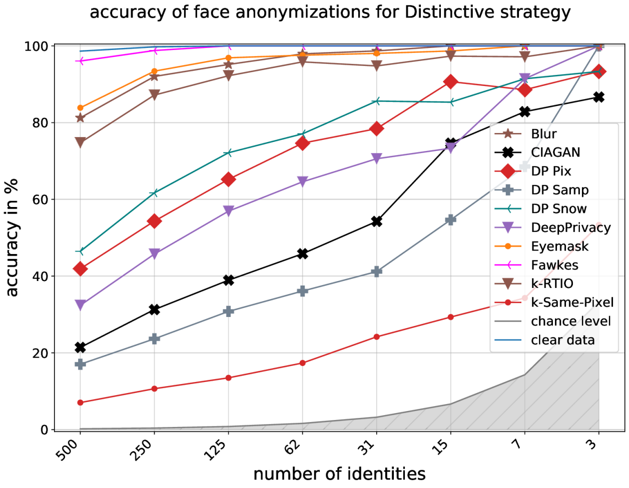

Figure 19: Accuracy of face anonymizations across decreasing numbers of identities. The strategy Distinctive was used to select the identities on clear data. ArcFace Retrained is used with parrot recognition on the CelebA dataset. A lower accuracy means better privacy protection.