Differentially Private Obfuscation of Facial Images

William L. Croft(✉), Jörg-Rüdiger Sack, and Wei Shi

Carleton University, Ottawa, ON, Canada
leecroft@cmail.carleton.ca

Abstract. The pervasiveness of camera technology in every-day life begets a modern reality in which images of individuals are routinely captured on a daily basis. Although this has enabled many benefits, it also infringes on personal privacy. To mitigate the loss of privacy, researchers have investigated methods of facial obfuscation in images. A promising direction has been the work in the k-same family of methods which employ the concept of k-anonymity from database privacy. However, there are a number of deficiencies of k-anonymity which carry over to the k-same methods, detracting from their usefulness in practice. In this paper, we first outline several of these deficiencies and discuss their implications in the context of facial obfuscation. We then develop the first framework to apply the formal privacy guarantee of differential privacy to facial obfuscation in generative machine learning models for images. Next, we discuss the theoretical improvements in the privacy guarantee which make this approach more appropriate for practical usage. Our approach provides a provable privacy guarantee which is not susceptible to the outlined deficiencies of k-same obfuscation and produces photo-realistic obfuscated output. Finally, while our approach provides a stronger privacy guarantee, we demonstrate through experimental comparisons that it can achieve comparable utility to k-same approaches in the context of preservation of demographic information in the images. The preservation of such information is of particular importance for enabling effective data mining on the obfuscated images.

Keywords: Privacy protection · Facial obfuscation · Differential privacy · Neural networks

1 Introduction

With the ever expanding presence of devices used to capture photos and video, visual privacy has become increasingly important. Images and video frames containing faces are routinely captured, e.g., through cameras, closed-circuit television systems [5], visual sensor networks [40] and a host of other devices and

The authors gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grants No. RGPIN-2015-05390, No. RGPIN-2016-06253 and No. CGSD2-503941-2017.

Published by Springer Nature Switzerland AG 2019
A. Holzinger et al. (Eds.): CD-MAKE 2019, LNCS 11713, pp. 229–249, 2019.
https://doi.org/10.1007/978-3-030-29726-8_15

methods. These systems have many benefits including mitigation of crime [5,40], improved care in assisted-living [34], and useful services such as Google Street View [17]. However, despite the benefits of the legitimate applications, the potential for infringement on personal privacy must be taken seriously.

Although many systems require only visual monitoring of behaviour, identities are often captured as well [40]. In some areas, the degree of public surveillance is reaching levels where it becomes possible to profile and track much of the population [34]. In cases where visual information is disseminated to the public, such as with Google Street View, it is imperative to hide the identities of individuals before the images are published [15]. Failure to sufficiently protect privacy may allow undesirable inferences to be drawn about individuals or enable malicious activities such as voyeurism or stalking. Users of mobile devices have also expressed strong aversion to the collection of images from their mobile devices via the applications they use [31]. Even in scenarios where users willingly share images to online platforms, they have expressed concerns over who is able to view their images [22]. In a similar context, the privacy of individuals captured in the backgrounds of images uploaded to such platforms should be taken into consideration. Rich visual information from these sources combined with the great advances in machine learning approaches to facial recognition (e.g., VGGFace [35]) make the exploitation of unprotected visual data a relatively easy task. While such machine learning algorithms are no doubt beneficial in many contexts, it is essential for approaches of privacy protection to be resistant to them.

To protect the privacy of individuals, methods for hiding identity via manipulation of the data can be employed. Many methods focus on the face as it is often the most identifiable piece of information. Trivially, the face could be covered by a uniformly coloured rectangle. This destroys all information about the face, guaranteeing that it can no longer be exploited to reveal an identity. However, this also destroys a great deal of utility. In scenarios where images are shared in online platforms, users have expressed a strong aversion to the use of such rectangles with respect to the visual quality and information content of the images [28]. Less severe methods of obfuscation present trade-offs between the level of privacy attainable and the utility of the data. Preservation of utility is especially important for machine learning and data mining. Visual data can be used to learn about customers in retail environments [29] and to detect anomalous or illegal events [39]. It is therefore essential for a good method of obfuscation to preserve as much of the non-sensitive information as possible.

A number of research directions have been explored for the obfuscation of visual identity in images, e.g., pixelization, blurring, etc. [34]. While many approaches lack a formal privacy guarantee, the k-same [33] family of approaches has gained a great deal of traction, largely thanks to its guarantee that for a chosen privacy parameter k, obfuscated individuals are indistinguishable within a group of k potential true identities. While this privacy guarantee is appealing, it suffers from susceptibilities (e.g., composition attacks [16]) carried over from the disclosure control method of k-anonymity on which it is based. In this paper, we

outline these susceptibilities in the context of privacy in images and propose an alternative, based on differential privacy, which addresses these susceptibilities.

1.1 Contributions and Paper Outline

Our contributions in this work are as follows:

- We examine susceptibilities of k-same obfuscation to composition attacks and background knowledge. We demonstrate how the privacy guarantee can be violated and discuss the implications this has on privacy in images.
- To address the deficiencies of k-same obfuscation carried over from k-anonymity, we propose as an alternative, the formal guarantee of differential privacy. We develop the first framework to apply differential privacy for the obfuscation of facial identity in images via generative machine learning models.
- We conduct a series of experiments to compare the quality of differential privacy to k-same obfuscation on two well-known datasets. The results of our experiments suggest that differential privacy offers a comparable level of utility in the obfuscated images to k-same obfuscation even though the privacy guarantee is improved.

We provide a review of existing work on the obfuscation of facial images in Sect. 2. We then cover the deficiencies of k-same in Sect. 3 and lay out a framework for differentially private obfuscation of images in Sect. 4. Finally, we describe our experimental comparisons and their results in Sect. 5.

2 Literature Review

Perhaps the most well-known and earliest studied alterations to images for the prevention of human recognition of faces are pixelization [21] and blurring [20]. Pixelization decreases the information conveyed in an image by dividing the image into a grid of cells and setting all pixels within each cell to a common pixel intensity. Blurring involves the addition of, typically Gaussian, noise to the image. While these methods have been successful at foiling human recognition, they have been shown to be highly ineffective against machine recognition [33].

Other ad hoc methods of privacy protection involving variations on blurring [26], warping [24], morphing [23] and face swapping [3] have been studied, however, the methods which have gained the most momentum are those which offer a formal guarantee of privacy. This trend has been reinforced by the legal and legislative demands in the broader context of the release of sensitive data [4,38]. To this end, k-same approaches have been quite successful. These approaches use an adaptation of k-anonymity [36], a concept from the field of database privacy which guarantees that an anonymized database record is linkable to at least k possible identities. The first adaptation of this concept to image obfuscation worked by aligning a set of input images on their facial features, partitioning

the set into clusters of k or more similar images, and then averaging the pixels within each cluster to produce an averaged face which would replace each of the original faces in the cluster [33]. By releasing only the averaged faces, it could be guaranteed that neither human nor machine recognition could do better than identifying the cluster of identities which produced the image, thus limiting the probability of successful re-identification by an upper bound of $\frac{1}{k}$.

One issue with the original k-same averaging of pixels was poor visual quality due to inexact alignment of facial features, leading to superimposed features. The k-same-m [18] approach improved upon this by using an active appearance model (AAM) [8] to obfuscate faces. AAMs are generative machine learning models for the approximation of visual representations of a particular class of objects (e.g., human faces). A model is trained on a set of images in order to learn about visual patterns and minimize differences with respect to shape and texture between the original images and the generated output of the model. The k-same-m approach first trains an AAM and then performs the clustering and averaging process within the parameter space of the model representations of faces to be obfuscated, thus eliminating the issue of superimposed features.

More recently, generative neural networks (GNNs) have been applied for k-same obfuscation [7,32]. GNNs are machine learning models which have shown great success in the generation of visual representations of input class labels [9]. A GNN passes the input labels through a sequence of convolutional layers, transforming them into features of finer granularity at each layer until reaching a pixel-space output. A training process adjusts weights used by filters in each convolutional layer in order to learn feature representations which minimize a loss function measuring the quality of the output. When trained on a set of images using identities as class labels, a GNN is able to produce a visual approximation of an identity based on an input class vector. By providing input vectors in which k identities are specified, the GNN produces k-anonymous output.

Efforts have also been devoted to the preservation of utility in the obfuscated images. The k-same-select approach [19] proposed partitioning the input images into classes based on the information to be preserved (e.g., male and female identities) before clustering such that the images within each cluster would share the same class, thus preserving this information in the averaged version. This idea has been extended to the k-same-m model by training a different AAM for each combination over the demographic attributes of age, gender and race [10]. By using the appropriately trained AAM for obfuscation, the attributes for which it was trained can be preserved in the output. In the context of GNNs, preservation of information has been considered by designing the network architecture to allow for multiple input vectors over different types of classes [9]. This has been applied to produce obfuscated images with specific facial expressions [32].

We note that differentially private obfuscation of images has been studied in the context of noise applied to pixel intensities [13]. While achieving a strong privacy guarantee, this leads to poor visual quality in the output as the obfuscated images no longer resemble the original class of the object (e.g., a human

face). To the best of our knowledge, our work is the first to study differential privacy applied to generative models for the obfuscation of facial images.

3 Weaknesses in Existing Facial Obfuscation

Given the importance of preserving privacy in images, a good method of obfuscation must assert a meaningful guarantee about the level of privacy it provides. Without such a guarantee, it is impossible to formally assess the effectiveness of the obfuscation. Empirical results may help to gain intuition on which approaches appear promising, however, without a formal guarantee to back up the results, it is impossible to assert that privacy will remain protected in untested scenarios against unknown attacks. For this reason, we focus our attention only on methods of obfuscation which offer a formal privacy guarantee.

The necessity of this restriction is underscored by the concept of parrot attacks [33]. A parrot attack uses a neural network to classify identities using labeled instances of obfuscated images as the training set. Having learned about patterns in the obfuscation during training, the network is made much more effective at defeating the obfuscation. Despite pixelization being reasonably effective against human recognition and even naive machine recognition, it can be completely defeated by a parrot attack. This formed a strong basis for the need of a formal privacy guarantee such as that provided by the k-same family.

The k-same approaches employ a privacy guarantee derived from k-anonymity [36] which asserts that the original identity for any obfuscated image is indistinguishable from at least $k-1$ others. This guarantee is a result of the obfuscation process which draws upon clusters of k or more images to produce averaged instances as replacements for all images in each cluster. This makes it impossible for any software to achieve a better probability of re-identification than $\frac{1}{k}$.

However, the k-same guarantee relies on assumptions about the nature of the attack. In this section, we discuss these assumptions. We show why they are often unrealistic in practice, making the guarantee weaker than it appears to be.

3.1 Background Information

A well-known deficiency of k-anonymity is its susceptibility to attacks which employ background information [2]. This refers to cases where the attacker uses prior knowledge about the sensitive information to draw inferences which violate the privacy guarantee. This concept carries directly over to the k-same privacy guarantee. If, via prior knowledge, the attacker knows with certainty that some of the k individuals could not be in the obfuscated image, they can discount them from the set of k identities. An attacker could come by this knowledge in a number of ways: personal knowledge about friends and family, information scrapped from other data sources such as social media, etc. The simple combination of knowledge about the time at which an photo was taken and the approximate

locations of some of the k individuals at that time can be enough to derive a proper subset of the k individuals which violates the privacy guarantee.

Contextual information in an image can often enable these types of inferences. Using signs, architecture or landscapes in an image, an attacker might recognize the location or employ software to determine it. Knowledge about locations that individuals frequent may greatly increase the probability of some possibilities over others. Similarly, if some of the k identities are known to live in different cities than where the photo was taken or worse yet, different continents, these identities become much less probable. Other cues such as accessories or clothing on obfuscated individuals may also greatly impact the probabilities accorded to the k possible identities. Since the privacy guarantee asserts that each of the k identities are equally probable, this is also in violation of the guarantee.

We note that the original k-same paper does acknowledge this vulnerability to contextual information and asserts that the privacy guarantee applies strictly to the information contained within the face, not to the image as a whole [33]. While this important distinction allows for the privacy guarantee to be upheld, it is a major restriction on the practical applicability of the k-same guarantee. Most contexts in which facial obfuscation is applied will be rich with contextual information, making the privacy guarantee much less meaningful.

3.2 Composition Attacks

Another deficiency of k-anonymity is a susceptibility to composition attacks [16]. This is a class of attacks which exploit information from multiple, potentially uncoordinated, obfuscated releases to violate the privacy guarantee. A simple instance of this is the intersection attack. An attacker first identifies the clusters in which a particular individual exists from two different releases. If the releases were uncoordinated, the clusters likely differ, allowing the attacker to take their intersection to achieve a new set with a cardinality less than k.

This attack again carries directly over to the k-same approach. Consider a scenario where an individual takes a photo which they wish to upload to social media. Privacy protection might be applied to the individual or perhaps to bystanders who were captured in the background of the photo. Should the individual decide to upload the same photo to two or more social media platforms, the issue of uncoordinated obfuscation immediately arises. An attacker needs only scrape these platforms for similar photos to apply an intersection attack.

Intersection attacks may even be effective for multiple releases from the same organization if care is not taken. For example, an individual may take consecutive photos and then upload all of them. Algorithms for k-same determine clusters based on the similarity of faces but many factors beyond facial identity (e.g., pose, angle and lighting) could impact similarity. It is therefore not unlikely that multiple images of the same individual will result in different clusters. Sequences of images uploaded in this way would be an ideal target for intersection attacks.

Most k-same approaches require each individual to appear only once in the gallery of images to be obfuscated. This prevents intersection attacks for releases

from the same organization but does not protect against uncoordinated releases across multiple organizations. Furthermore, enforcing this restriction may be very challenging in practice. While the primary subject in a photo might be determined based on the account used to upload the photo, other individuals in the photo cannot be correctly identified 100% of the time. Face recognition software has not yet reached this level of accuracy. Without manual labeling, such a policy cannot be enforced. Beyond this, the restriction of one image per identity is very severe and does not match typical use cases for image sharing.

3.3 Other Difficulties

We discuss here two other difficulties that arise when using k-same obfuscation in practice. Although these difficulties do not violate the privacy guarantee, they hinder meaningful applications for k-same obfuscation in some contexts.

The first problem arises from the requirement of an input gallery of images. This may be appropriate for scenarios where batches of images are obfuscated but it is awkward to apply to cases where images are sporadically uploaded (e.g., in social media platforms). One might consider the use of a preloaded static gallery or even a dynamic gallery that gets updated as new images are uploaded. This, however, is not a good solution since identities can then participate in more than one cluster. Furthermore, if an attacker records information about identities known to be in the gallery, those identities can be discounted when an image is uploaded for a new identity. An alternative solution could rely on buffering uploaded images to form a gallery which can eventually be used to release a batch of obfuscated images. However, this necessitates a trade-off between the size of the gallery (and thus the quality of the output) and ability to deliver a timely service. In an era where users expect images to be uploaded instantly, this is not likely to be a manageable trade-off. The release of multiple batches also increases the chances of enabling composition attacks.

The second problem relates to the preservation of utility in the obfuscated output. Approaches which partition the gallery according to classes to be preserved (e.g., combinations of age and gender) place an even greater strain on the input gallery requirement. Working separately with the subset of images from each class greatly reduces the number of images available for clustering. Such an approach is not scalable for large numbers of classes that would be needed for finely grained attention to utility. In the worst case, some classes may be outliers in the overall distribution and could lack sufficient images to form a cluster. These classes would have to be merged with others in order to achieve the k-same guarantee, thus failing to achieve the desired granularity of classes.

4 Differential Privacy for Generative Models

Due to the deficiencies of the k-same privacy guarantee in practical applications of facial obfuscation, we argue that a more robust privacy guarantee is required. Following the advances in the field of database privacy, we consider the potential

of differential privacy to provide a stronger privacy guarantee. In this section, we first review basic theory of differential privacy. We then adapt the privacy guarantee to fit the context of generative machine learning models for images and we formalize a framework to apply differential privacy to facial images. We discuss how the derived privacy guarantee addresses the issues identified with the k-same approach. Finally, we apply our framework to implement differentially private facial obfuscation using an AAM and a GNN.

4.1 Differential Privacy for Databases

A privacy guarantee which offers an absolute bound on re-identification risk necessitates restrictive assumptions about the attacker. This is due to the fact that it is impossible to prevent an attacker from learning about the sensitive information through means other than the obfuscated release [11]. Differential privacy recognizes this difficulty and instead adopts a privacy guarantee which limits the increase in an attacker's knowledge about the sensitive information. In the context of databases, the goal is to release aggregate information about the database while preventing that information from being exploited to derive sensitive details about the individual records. Differential privacy functions by using a *randomization mechanism* to add controlled noise to database query responses in order to release useful responses while achieving a desired level of indistinguishability between potential configurations of the database contents.

Two databases are considered to be *adjacent* if they differ by a single record. Informally, the privacy guarantee enforces that any pair of adjacent databases must be bounded within a multiplicative factor of e^{ϵ} (where ϵ is the *privacy parameter*) in their probabilities of producing the same noisy query response. This is often interpreted as a ratio of e^{ϵ} between these probabilities. With a sufficiently small ratio, similar databases have similar probability distributions over their noisy query responses, causing them to behave similarly with respect to the noisy query responses they produce. This limits the usefulness of the noisy responses as a means to distinguish between potential configurations of the database. The privacy guarantee [12] in Formula 1 formally states this requirement in terms of any pair of adjacent databases $D_1, D_2 \in \mathbb{D}$, where $\mathbb{D}$ is the set of valid database configurations, and a randomization mechanism $K : \mathbb{D} \to \mathbb{R}^n$, where $n \in \mathbb{Z}^+$.

$$\Pr\left(K\left(D_1\right) = R\right) \leq e^{\epsilon} \Pr\left(K\left(D_2\right) = R\right) \qquad \forall R \in \mathbb{R}^n. \tag{1}$$

To achieve this privacy guarantee, the mechanism K must take into account the value of ϵ and the *query sensitivity*. The sensitivity ΔF of a query $f : \mathbb{D} \to \mathbb{R}^n$ is defined as the maximum possible L_1 distance between the query responses for any pair of adjacent databases. The guarantee can be achieved by adding to the query response a vector of n continuous random variables, each drawn independently from a Laplace distribution with $\frac{\Delta F}{\epsilon}$ as its scaling parameter [12]. The exponential decay of probability density in the Laplace distribution benefits the utility of the mechanism by limiting the expected perturbation of the query

responses. Through the selection of an appropriate value for ϵ, a data custodian can control how much information is revealed about the contents of the database.

4.2 Framework for Generative Models

We now consider how differential privacy can be applied to generative models for images. A generative model can represent images of instances from specific classes (e.g., human faces) using a numeric representation which abstracts from pixel intensities. Our goal is to protect the privacy of individuals in images by modifying these numeric representations to prevent facial identification while maintaining utility and visual quality. Differential privacy is ideal for this purpose as it provides a robust guarantee against the accuracy of the inferences an attacker can make about the original data. The application of noise to the numeric representation of the model allows for the generation of photo-realistic instances of novel human faces. This avoids the significant degradation in visual quality which results from the addition of noise to pixel intensities.

When moving from the domain of databases to that of generative model representations, the concepts of adjacency and query sensitivity can no longer be applied for the configuration of a mechanism. In place of a database where each record is an individual, we have a numeric representation of a single individual. To protect sensitive data in this form, one can apply a generalization of differential privacy to arbitrary *secrets* [6], where a secret is any numeric representation of data. In our case, the secret is the generative model representation of an individual. This generalization substitutes the notion of adjacency between databases with distance between secrets. By controlling noise according to an appropriate distance metric, the privacy guarantee is adapted to ensure that similar secrets are highly indistinguishable while very different secrets remain distinguishable. For a pair of databases, the distance between them is the number of records by which they differ. For other types of secrets, the distance metric must be carefully chosen in order to provide an appropriate privacy guarantee.

The notion of distance between secrets is appropriate for the representation of images within a generative model. Any model which employs a numeric representation of images allows for the calculation of distance between images. While the exact representation of an image differs from model to model, they can generally mapped to a vector of fixed length with little difficulty. We provide details on how this concept can be applied to both AAMs and GNNs in Sect. 4.4. To develop a general framework here, we consider the representation of an image to be a vector $X \in \mathbb{R}^n$ and the randomization mechanism to be a function $K : \mathbb{R}^n \to \mathbb{R}^n$ used to produce an obfuscated instance of the image. Although the differential privacy generalization only deals explicitly with one and two-dimensional secrets [6], its generalization to an n-dimensional vector is straightforward. We therefore adapt the privacy guarantee to suit this purpose in Formula 2, using a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

$$\Pr\left(K\left(X_1\right) = R\right) \leq e^{\epsilon d(X_1, X_2)} \Pr\left(K\left(X_2\right) = R\right) \qquad \forall X_1, X_2, R \in \mathbb{R}^n. \tag{2}$$

Comparing this to Formula (1), the databases D_1 and D_2 have been replaced by secrets X_1 and X_2 and the distance function now appears in the exponent of the multiplicative factor e^{ϵ}. The distance between any pair of secrets acts as a coefficient to ϵ when interpreting the ratio of their probabilities. Intuitively, the meaning is that the more similar a pair of images are to each other, the harder is it to determine which of them led to a given obfuscated instance. This hampers the accuracy with which attempts at re-identification can be made. To achieve this guarantee, we must first determine an appropriate distance metric to measure the distinguishability of the numeric representations of images.

A natural choice for the distance metric is L_1 distance, however, we must be wary of the meaning of each element in the vectors. Should certain elements have differently sized ranges, they should be obfuscated using different magnitudes of noise. If one element has a much larger range than the others, the addition of noise configured to the smaller range would do little to prevent an inference of high accuracy on the original value of the element. We therefore apply normalization such that the distance between any pair of elements in the i^{th} position of a pair of vectors falls within the range $[0, 1]$. Letting $R_i = [i_{min}, i_{max}]$ be the range of elements in the i^{th} position of a model representation vector, we define a normalized, element-wise distance metric as follows:

$$d_e(x_1, x_2) = \frac{|x_1 - x_2|}{i_{max} - i_{min}} \qquad \forall x_1, x_2 \in R_i. \tag{3}$$

A distance metric for vectors defined as the sum of the element-wise distances for each position would be appropriate for images represented by the same model. However, a more useful framework would allow for reasoning about the level of privacy across different models. Ideally, the meaning of a privacy parameter ϵ applied to one model should have a similar meaning for a different model. For this, we require another normalization to account for models having vectors of different lengths. We therefore define the distance metric for vectors as follows:

$$d(X_1, X_2) = \frac{\sum_{i=1}^{n} d_e(X_{1i}, X_{2i})}{n} \qquad \forall X_1, X_2 \in \mathbb{R}^n. \tag{4}$$

By using this distance metric in combination with Formula 2, we obtain a meaningful privacy guarantee for the model representations of images. Although this type of metric is not novel, its use in this context is. We must therefore address how to configure a mechanism to satisfy this instantiation of the privacy guarantee. This leads to our main result in the development of a framework for the application of differential privacy to generative models for images.

Theorem 1. *Any image $X \in \mathbb{R}^n$ can be protected by ϵ-differential privacy through the addition of a vector $(Y_1, ..., Y_n) \in \mathbb{R}^n$ where each Y_i is a random variable independently drawn from a Laplace distribution using a scaling parameter $\sigma_i = \frac{n(i_{max} - i_{min})}{\epsilon}$.*

Proof. We must satisfy the privacy guarantee (Formula 2) using our proposed distance metric (Formula 4). The form this privacy guarantee takes is our starting point in Formula 5. Through manipulation of this inequality and the substitution of mechanism probabilities with a Laplace distribution, we prove that the selection of an appropriate scaling parameter for each instance of the Laplace distribution allows for the privacy guarantee to be satisfied.

$$\prod_{i=1}^{n} \Pr\left(K\left(X_{1i}\right) = R_i\right) \leq e^{\frac{\epsilon \sum_{i=1}^{n} d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^{n} \Pr\left(K\left(X_{2i}\right) = R_i\right) \forall X_1, X_2, R \in \mathbb{R}^n. \quad (5)$$

$$\prod_{i=1}^{n} \Pr\left(K\left(X_{1i}\right) = R_i\right) \leq \prod_{i=1}^{n} e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^{n} \Pr\left(K\left(X_{2i}\right) = R_i\right) \forall X_1, X_2, R \in \mathbb{R}^n. \quad (6)$$

$$\prod_{i=1}^{n} \frac{e^{-\frac{|X_{1i}, R_i|}{\sigma}}}{2\sigma} \leq \prod_{i=1}^{n} e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} \prod_{i=1}^{n} \frac{e^{-\frac{|X_{2i}, R_i|}{\sigma}}}{2\sigma} \qquad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (7)$$

$$\prod_{i=1}^{n} e^{\frac{|X_{2i}, R_i| - |X_{1i}, R_i|}{\sigma}} \leq \prod_{i=1}^{n} e^{\frac{|X_{2i} - X_{2i}|}{\sigma}} \leq \prod_{i=1}^{n} e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} \qquad \forall X_1, X_2, R \in \mathbb{R}^n. \quad (8)$$

$$\prod_{i=1}^{n} e^{\frac{\epsilon d_e(X_{1i}, X_{2i})}{n}} = \prod_{i=1}^{n} e^{\frac{\epsilon |X_{2i} - X_{2i}|}{n(i_{max} - i_{min})}} \qquad \forall X_1, X_2. \quad (9)$$

From Formula 9, it becomes clear that the inequality holds when using an independent Laplace distribution for each pair of elements X_{1i}, X_{2i}, substituting the scaling parameter σ with a corresponding value $\sigma_i = \frac{n(i_{max} - i_{min})}{\epsilon}$. □

Using the generalization of differential privacy, the notion of query sensitivity is implicitly captured in the distance metric. Since the distance metric of Formula (4) has a range of $[0, 1]$, the ratio of probabilities for a pair of maximally dissimilar images to produce the same obfuscated output is e^{ϵ}. This is akin to the meaning of the privacy guarantee for a pair of databases which differ on every record. In order to select an appropriate value of ϵ, a data custodian must keep in mind that similar images will have a very small distance between them, requiring much larger values of ϵ to provide a reasonable ratio. In Sect. 5, we demonstrate the implications of the choice of ϵ on the levels of privacy and utility.

4.3 Benefits of Differentially Private Facial Obfuscation

We now describe the improvements we obtain from the use of differential privacy for each of the problems identified in Sect. 3.

Background Information. By removing dependence of the attack model on an absolute level of re-identification risk, we are able to reason about the level of privacy in the presence of attackers with background knowledge. If the location in a photo is identified as a particular city, no facial obfuscation can prevent the inference that individuals living in the identified city have a higher probability of being the obfuscated identity than individuals living elsewhere. Yet, the

differential privacy guarantee continues to hold as the background information does not impact the conditional probability distribution used by the randomization mechanism. Since the privacy guarantee concerns only the change in the attacker's knowledge when presented with the obfuscated data (e.g., the face), it is unaffected by other sources of information the attacker may gain access to.

Composition Attacks. Another very important property of differential privacy is its resilience to composition attacks. The composition theorem [12] states that for two differentially private releases using privacy parameters ϵ_1 and ϵ_2 respectively, the privacy guarantee holds for a privacy parameter $\epsilon = \epsilon_1 + \epsilon_2$. Thus, even in the case of uncoordinated releases, we still have a valid privacy guarantee. Furthermore, this removes the restriction on the same individual appearing only once in the release of obfuscated images.

Input Image Gallery. Differentially private image obfuscation has no need for a gallery of images in order to perform obfuscation. Since noise is added on a per-image basis, there is no computation of clusters required. Given a trained model, obfuscation of a single image or a batch of images can be performed with ease. This makes the obfuscation process much more versatile.

4.4 Implementation Details

AAMs and GNNs have the very useful property of producing photo-realistic images. We now describe how our framework can be applied to these models. Provided that the addition of noise is properly controlled, the output will be a photo-realistic image of any newly created identity.

The AAM representation of an image consists of a shape vector and a texture vector. We take the concatenation of these vectors as the overall model vector. It is important to note that this gives a representation of the identity which is strictly contained within the contour of the face, leaving features such as hair and ears as contextual information which is untouched by the obfuscation. This is not ideal for the goal of hiding an identity since this contextual information can greatly facilitate inferences about the identity. Although in theory an AAM could be designed to incorporate the hair and ears, we are unaware of any research in which this has been done. Thus, although we include AAMs in our experimental comparisons, we recommend the use of GNNs instead.

For GNNs, we consider architectures which take one or more class vectors as input and employ up-convolution to transform the input into a visual representation in pixel space [9]. By considering each identity to be a different class, an input vector can specify the individual to be generated. The identity class vector is an obvious choice as the model vector to be obfuscated. However, this leads to some form of interpolation between the identities. To apply a finer degree of modification to the identity, we propose the application of obfuscation at the second layer of the network. Typically, the second layer applies convolution to the class vector and transforms it into a vector of numeric representations of high

level facial features. By applying obfuscation to these features instead, we can achieve a richer variety in the potential modifications to the face. We therefore apply obfuscation to the output of the first convolutional layer of the network and pass the obfuscated feature vector on as the input to the next layer of the network. A sample architecture is shown in Fig. 1.

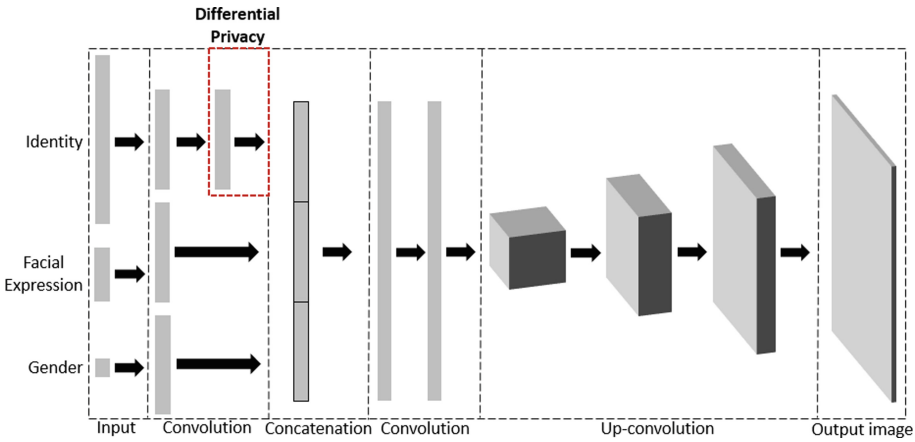

Fig. 1. Visualization of the layer architecture in an up-convolutional neural network using differential privacy. Noise is applied to the output of the second identity layer. The numbers and shapes of the convolutional layers shown here are not exact and represent only the general structure of such a network.

Information about the range of each model vector element can be used as a means to preserve the visual quality of the obfuscated output. Noisy elements which have gone too far beyond the valid range may lead to visual artifacts or distortions in the output image. To prevent this, we snap any out-of-bounds noisy value back to the nearest valid value. Since differential privacy is resistant to any form of post-processing [12] and the ranges of the elements are non-sensitive information, this step cannot violate the privacy guarantee.

5 Experiments

In this section, we run an experimental comparison between our proposed implementation of Sect. 4.4 and k-same implementations following the designs of k-same-m [18] and k-same-net [32]. We employ these experiments to gain insight into the relative performances of differential privacy and k-same obfuscation in the context of a trade-off between re-identification risk and utility.

5.1 Experimental Design and Results

We have implemented AAM obfuscation using AAM-API [37] for model training and generation of output. For GNN obfuscation, we have built on top of the

DeconvFaces [14] network which implements the concept of up-convolution for the generation of images of input classes [9]. For both models, we have applied differential privacy as described in Sects. 4.2 and 4.4. For the k-same implementation in the AAM, we have followed the process of k-same-m [18]. For the GNN, we have followed the approach of k-same-net [32]. In both cases, we have implemented clustering as described for k-same-m. This deviates from the use of a proxy gallery as described for k-same-net. It is important to note that, while a proxy gallery can reduce re-identification risk, it involves a step which is external to the k-same privacy guarantee. Thus, in the absence of a privacy guarantee which incorporates this detail, we omit the use of a proxy gallery in order to focus our experiments on the formalized aspects of the privacy guarantees.

We apply each method of obfuscation to two different datasets - RAFD [25] and KDEF [30]. These datasets provide frontal facial images of subjects wearing same coloured shirts. The use of same coloured shirts prevents bias in re-identification from the exploitation of information in unique clothing. The RAFD and KDEF datasets contain images of 67 and 70 subjects, respectively, and provide a variety of facial expressions. Due to apparent issues with lens exposure in the KDEF dataset, we have removed two of the subjects from our experiments.

The GNN architecture accepts class vectors for identity and facial expression as input. The RAFD and KDEF datasets are therefore highly suitable for this network. We have trained the network for 1000 epochs on each of the datasets to obtain models capable of reproducing these identities. An example of obfuscated output is shown in Fig. 2. The AAM has the advantage of being able to approximate previously unseen identities. To use this in our experiments, we have trained a model for each dataset using the other dataset (e.g., RAFD as training data for KDEF) with pre-processing to adjust the colour saturation of the training data in order to better match the target data. Since the training requires annotations of facial landmarks, we have employed OpenFace [1] to compute high accuracy approximations of the landmarks.

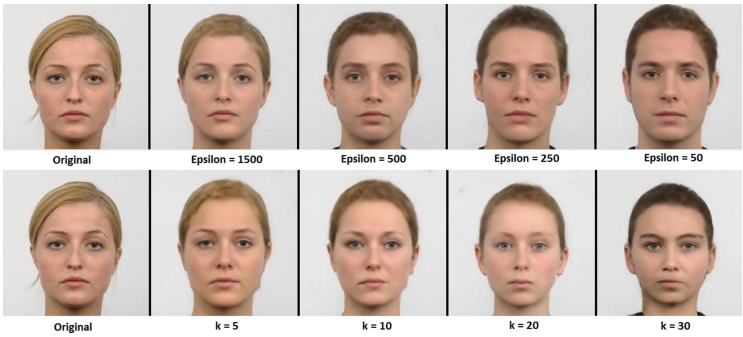

Fig. 2. Obfuscation via the GNN on the RAFD dataset. The top row employs differential privacy and the bottom row employs k-same obfuscation.

To measure re-identification risk, we have employed VGGFace D [35], a deep convolutional neural network which has been shown to achieve excellent facial identity classification accuracy. This simulates how an attacker might leverage machine learning models to launch an attack on obfuscated images. We have trained a separate model for each dataset, using the neutral and sad expressions for each identity for validation and the remaining expressions for training. To improve the robustness of the models, we have also augmented the datasets by creating two additional versions of each image - one with increased contrast and one with decreased contrast.

In our experiments, we generate obfuscated images having a neutral facial expression. We measure re-identification risk based on the accuracy of the top 1 guesses of the VGGFace network. Given that differential privacy is a stochastic process, for each combination of a privacy parameter and an identity to be protected, we have generated 10 obfuscated instances over which we take the average of the re-identification risk. We measure overall re-identification risk for a given privacy parameter as the average risk over all individuals in the dataset. Since the k-same approaches are deterministic, we produce only a single output image per identity and then take the average re-identification risk over the whole dataset (Fig. 3). In contrast to the typical ϵ values applied to differentially private mechanisms for databases, the values used in our experiments may appear unusually high. The larger magnitude is simply a side-effect of the normalization for the model vector, resulting in the interpretation of ϵ on a different scale.

To compare the methods of obfuscation in terms of utility, we have focused on gender classification in the obfuscated output. As forms of demographic classification may be desirable for data mining purposes, we consider high classification accuracy to reflect good utility. To this end, we employ a convolutional neural network for the classification of gender in facial images [27]. Since we wish to compare differential privacy to k-same obfuscation, we plot the data as a function of identity classification error in order to abstract away from the proprietary privacy parameters (Fig. 4). Given the poor obfuscation achieved by the AAMs, we omit them from this comparison. To highlight the ability of GNNs to incorporate properties relevant to image utility into the network architecture, we have also created a modified version of the architecture which preserves gender in the obfuscated output. To do so, we have created an input layer having two classes which specify the gender in the image. By training a model with gender labels, it learns to separate features relevant to gender from those relevant to identity. This enables us to focus obfuscation only on the features relevant to identity while leaving the gender feature vector untouched. An example of gender-preserving obfuscation is shown in Fig. 5.

5.2 Discussion

We first consider the results on re-identification risk. It is immediately notable that the AAMs are ineffective at privacy protection, even under severe privacy settings. This is due to the contextual information outside of the facial contour such as the hair, ears and neck. In an alternative setting, the obfuscated face

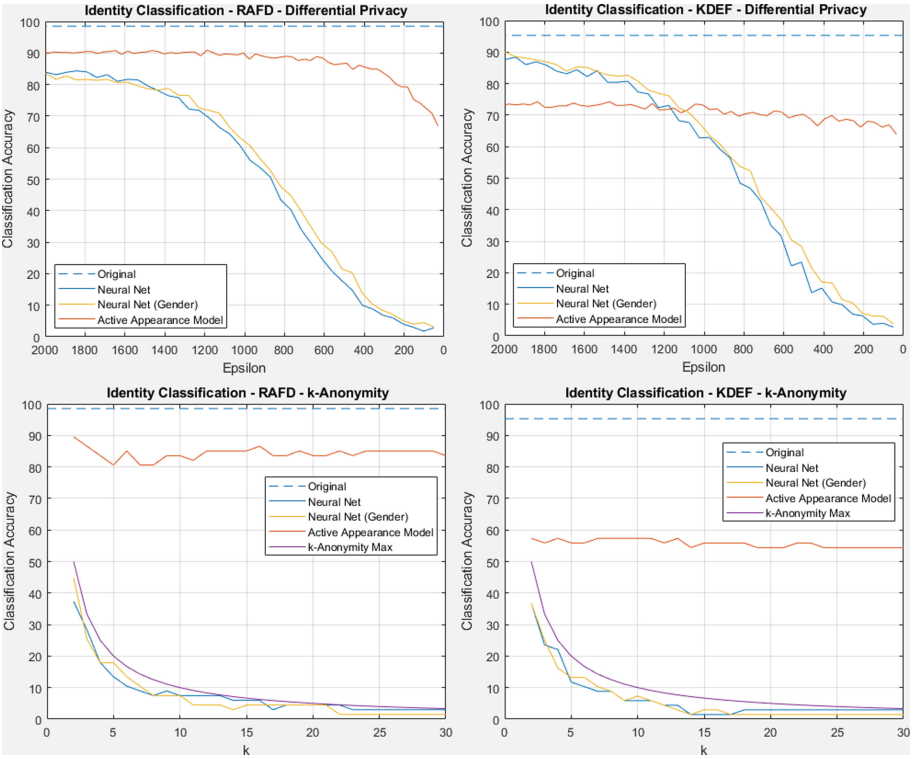

Fig. 3. Identity classification accuracy for the methods of obfuscation

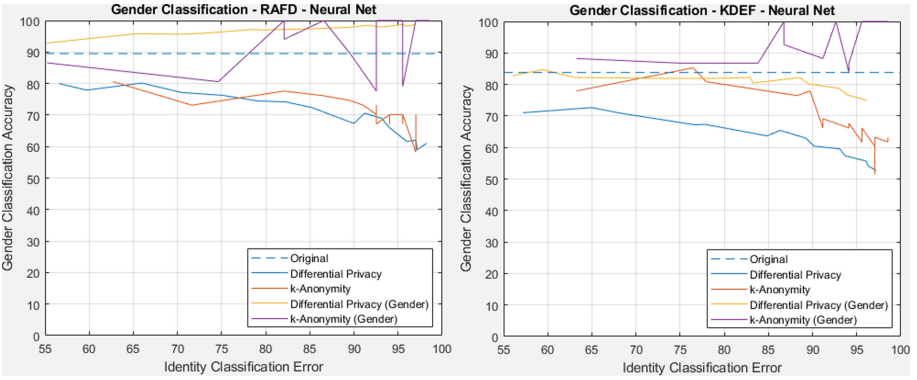

Fig. 4. Gender classification accuracy for the methods of obfuscation

could be released on its own (e.g., against a black background) to prevent leaking this contextual information. While this would greatly improve the privacy protection, it would omit a great deal of useful information in the image and would lead to an output which is not very visually pleasing. We expect that in most practical situations, it is desirable to release a full version of the image. We therefore recommend against the use of AAMs for obfuscation, at least in a form which does not obfuscate the full head.

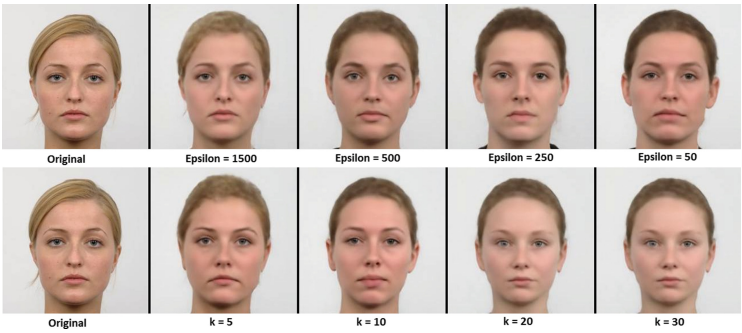

Fig. 5. Gender-preserving obfuscation via the GNN on the RAFD dataset. The top row employs differential privacy and the bottom row employs k-same obfuscation.

The k-same AAM results clearly demonstrate the violation of the k-same privacy guarantee since the re-identification risk is well above the theoretical maximum. We again note that if we only consider information within the contour of the face, the guarantee is not violated, however, such a guarantee is not useful in most practical situations. With differential privacy, we maintain a meaningful privacy guarantee for images which contain such information. Although our experiments do not illustrate the susceptibility of k-same obfuscation to composition attacks, we plan to demonstrate this in future work.

In the gender classification comparisons, we see that the basic models for differential privacy and k-same obfuscation suffer a degradation in classification accuracy at high levels of privacy protection. Comparing the gender-preserved models to their basic counterparts, we see a large improvement in the classification accuracy, suggesting that this is an effective approach for the preservation of specific properties in the obfuscated output. In some cases, the classification accuracy of the obfuscated images has surpassed that of the original data. This is a result of the explicit specification of gender labels in the network input which can lead to obfuscated identities that more prominently display these features.

The overall comparison between the utility preserved in differential privacy and k-same obfuscation appears to be inconclusive in these results. Some experiments show better results for differential privacy while other experiments show better results for k-same obfuscation. The k-same results are also more difficult to assess given the sporadic nature of the plots. This is likely due to changes

in clusters between each level of obfuscation which can greatly impact classification accuracy, both for identity and for gender. It is clear that the utility is also data-dependent given the variations in the results seen on the two datasets. Notably, many subjects in the KDEF dataset, including some males, have long hair whereas all subjects in the RAFD dataset have short hair. The males with long hair in KDEF may have contributed to the lower gender classification accuracy.

Given that differential privacy offers great improvements in the privacy guarantee over k-same obfuscation and that the levels of utility in our experiments appear to be similar between the two approaches of obfuscation, we consider differential privacy to be a preferable choice for the obfuscation of facial images.

6 Conclusions

We have studied how to obtain a formalized privacy guarantee for the obfuscation of facial images in practice. We have identified shortcomings of the k-same privacy guarantee including susceptibilities to background information and composition attacks as well as the awkwardness in the requirement for a gallery of input images. To improve upon this, we have proposed the use of differential privacy in the context of obfuscation applied to generative models for images. We have developed a framework which provides a meaningful privacy guarantee for such models and we have derived the configuration of Laplace mechanism which can achieve this privacy guarantee. Our approach preserves the privacy guarantee in the presence of attackers with background information, provides resistance to composition attacks and removes the requirement for a gallery of input images. We have implemented both our proposed framework as well as k-same obfuscation in order to run experimental comparisons. Through our comparisons, we have shown that this application of differential privacy can achieve comparable utility to k-same obfuscation. We conclude that the key improvements in the privacy guarantee combined with comparable levels of utility make differential privacy a much more appropriate choice for the obfuscation of facial images.

References

1. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: OpenFace: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science (2016)
2. Basu, A., Nakamura, T., Hidano, S., Kiyomoto, S.: k-anonymity: risks and the reality. In: 2015 IEEE Trustcom/BigDataSE/ISPA, vol. 1, pp. 983–989 (2015). https://doi.org/10.1109/Trustcom.2015.473
3. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. ACM Trans. Graph. **27**(3), 39:1–39:8 (2008). https://doi.org/10.1145/1360612.1360638
4. Brand, A., Lal, J.A.: European best practice for quality assurance, provision and use of genome-based information and technologies. Drug Metab. Drug Interact. **27**, 177–182 (2012). https://doi.org/10.1515/dmdi-2012-0026

5. Cavailaro, A.: Privacy in video surveillance [In the Spotlight]. IEEE Signal Process. Mag. **24**(2), 166–168 (2007). https://doi.org/10.1109/MSP.2007.323270
6. Chatzikokolakis, K., Andrés, M.E., Bordenabe, N.E., Palamidessi, C.: Broadening the scope of differential privacy using metrics. In: Privacy Enhancing Technologies, pp. 82–102 (2013). https://doi.org/10.1007/978-3-642-39077-7_5
7. Chi, H., Hu, Y.H.: Face de-identification using facial identity preserving features. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 586–590 (2015). https://doi.org/10.1109/GlobalSIP.2015.7418263
8. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001). https://doi.org/10.1109/34.927467
9. Dosovitskiy, A., Springenberg, J.T., Tatarchenko, M., Brox, T.: Learning to generate chairs, tables and cars with convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 692–705 (2017). https://doi.org/10.1109/TPAMI.2016.2567384
10. Du, L., Yi, M., Blasch, E., Ling, H.: GARP-face: balancing privacy protection and utility preservation in face de-identification. In: IEEE International Joint Conference on Biometrics, pp. 1–8 (2014). https://doi.org/10.1109/BTAS.2014.6996249
11. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
12. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014). https://doi.org/10.1561/0400000042
13. Fan, L.: Image pixelization with differential privacy. In: Kerschbaum, F., Paraboschi, S. (eds.) DBSec 2018. LNCS, vol. 10980, pp. 148–162. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95729-6_10
14. Flynn, M.: Generating faces with deconvolution networks (2016). https://github.com/zo7/deconvfaces
15. Frome, A., et al.: Large-scale privacy protection in google street view. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2373–2380 (2009). https://doi.org/10.1109/ICCV.2009.5459413
16. Ganta, S.R., Kasiviswanathan, S.P., Smith, A.: Composition attacks and auxiliary information in data privacy. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 265–273 (2008). https://doi.org/10.1145/1401890.1401926
17. Google: Google Maps. https://www.google.be/maps. Accessed 27 Feb 2019
18. Gross, R., Sweeney, L., de la Torre, F., Baker, S.: Model-based face de-identification. In: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2006), pp. 161–161 (2006). https://doi.org/10.1109/CVPRW.2006.125
19. Gross, R., Airoldi, E., Malin, B., Sweeney, L.: Integrating utility into face de-identification. In: Privacy Enhancing Technologies, pp. 227–242 (2006). https://doi.org/10.1007/11767831_15
20. Harmon, L.: The recognition of faces. Sci. Am. **229**(5), 71–82 (1973)
21. Harmon, L., Julesz, B.: Masking in visual recognition: effects of two-dimensional filtered noise. Science **180**(4091), 1194–1197 (1973). https://doi.org/10.1126/science.180.4091.1194
22. Hu, X., et al.: How people share digital images in social networks: a questionnaire-based study of privacy decisions and access control. Multimed. Tools Appl. **77**(14), 18163–18185 (2018). https://doi.org/10.1007/s11042-017-4402-x

23. Korshunov, P., Ebrahimi, T.: Using face morphing to protect privacy. In: 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 208–213 (2013). https://doi.org/10.1109/AVSS.2013.6636641
24. Korshunov, P., Ebrahimi, T.: Using warping for privacy protection in video surveillance. In: 2013 18th International Conference on Digital Signal Processing (DSP), pp. 1–6 (2013). https://doi.org/10.1109/ICDSP.2013.6622791
25. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D., Hawk, S., van Knippenberg, A.: Presentation and validation of the radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010). https://doi.org/10.1080/02699930903485076
26. Letournel, G., Bugeau, A., Ta, V., Domenger, J.: Face de-identification with expressions preservation. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 4366–4370 (2015). https://doi.org/10.1109/ICIP.2015.7351631
27. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 34–42 (2015). https://doi.org/10.1109/CVPRW.2015.7301352
28. Li, Y., Vishwamitra, N., Knijnenburg, B.P., Hu, H., Caine, K.: Effectiveness and Users' Experience of Obfuscation As a Privacy-Enhancing Technology for Sharing Photos. Proc. ACM Hum.-Comput. Interact. 1, 1–24 (2017). https://doi.org/10.1145/3134702
29. Liu, X., Krahnstoever, N., Yu, T., Tu, P.: What are customers looking at? In: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 405–410 (2007). https://doi.org/10.1109/AVSS.2007.4425345
30. Lundqvist, D., Flykt, A., Öhman, A.: The Karolinska Directed Emotional Faces – KDEF (1998). ISBN 91-630-7164-9
31. Martin, K., Shilton, K.: Putting mobile application privacy in context: an empirical study of user privacy expectations for mobile devices. Inf. Soc. **32**(3), 200–216 (2016). https://doi.org/10.1080/01972243.2016.1153012
32. Meden, B., Emersic, Z., Struc, V., Peer, P.: k-same-net: neural-network-based face deidentification. In: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), pp. 1–7 (2017). https://doi.org/10.1109/IWOBI.2017.7985521
33. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE Trans. Knowl. Data Eng. **17**(2), 232–243 (2005). https://doi.org/10.1109/TKDE.2005.32
34. Padilla-López, J.R., Chaaraoui, A.A., Flórez-Revuelta, F.: Visual privacy protection methods. Expert Syst. Appl. **42**(9), 4177–4195 (2015). https://doi.org/10.1016/j.eswa.2015.01.041
35. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)
36. Samarati, P., Sweeney, L.: Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression. Technical report (1998). http://www.csl.sri.com/papers/sritr-98-04/
37. Stegmann, M.B.: The AAM-API (2003). http://www.imm.dtu.dk/~aam/aamapi/, platform: MS Windows
38. U.S. Department of Health & Human Services: Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (2015). https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html. Accessed 09 Feb 2018

39. Venetianer, P.L., Zhang, Z., Scanlon, A., Hu, Y., Lipton, A.J.: Video verification of point of sale transactions. In: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 411–416 (2007). https://doi.org/10.1109/AVSS.2007.4425346
40. Winkler, T., Rinner, B.: Security and privacy protection in visual sensor networks: a survey. ACM Comput. Surv. **47**(1), 1–42 (2014). https://doi.org/10.1145/2545883