Soft biometric privacy: Retaining biometric utility of face images while perturbing gender

Conference Paper · October 2017

DOI: 10.1109/BTAS.2017.8272743

2 authors:

Vahid Mirjalili
Michigan State University
32 PUBLICATIONS 666 CITATIONS
SEE PROFILE

Arun Ross
Michigan State University
360 PUBLICATIONS 23,547 CITATIONS
SEE PROFILE

V. Mirjalili and A. Ross, "Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender,"
Proc. of International Joint Conference on Biometrics (IJCB), (Denver, USA), October 2017

Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender

Vahid Mirjalili and Arun Ross
Department of Computer Science and Engineering
Michigan State University
{mirjalil,rossarun}@cse.msu.edu

Abstract

While the primary purpose for collecting biometric data (such as face images, iris, fingerprints, etc.) is for person recognition, yet recent advances in machine learning has shown the possibility of extracting auxiliary information from biometric data such as age, gender, health attributes, etc. These auxiliary attributes are sometimes referred to as soft biometrics. This automatic extraction of soft biometric attributes can happen without the user's agreement, thereby raising several privacy concerns. In this work, we design a technique that modifies a face image such that its gender as assessed by a gender classifier is perturbed, while its biometric utility as assessed by a face matcher is retained. Given an arbitrary biometric matcher and an attribute classifier, the proposed method systematically perturbs the input image such that the output of the attribute classifier is confounded, while the output of the biometric matcher is not significantly impacted. Experimental analysis convey the efficacy of the scheme in imparting gender privacy to face images.

1. Introduction

Biometrics is the science of recognizing individuals based on their physical or behavioral characteristics such as face, fingerprints, iris, gait, *etc*. A typical biometric system acquires biometric data from a subject (*e.g*., a 2D face image), extracts a feature set and compares the feature set to templates in a database (*e.g*., face images labeled with an identifier) in order to verify a person's claimed identity (biometric verification) or to determine the person's identity (biometric identification) [15]. While the biometric data acquired from a subject is expected to be used for recognition purposes only, recent research has established the possibility of deducing additional attributes of a person from their biometric data [6, 18, 20]. For instance, attributes such as age [19, 8, 9], gender [22, 26, 21, 25, 34, 37, 19] and ethnicity [13] can be automatically deduced from a 2D face image obtained in the visible spectrum. These attributes may not be distinctive enough for recognition purposes and are, therefore, called *soft* biometric attributes [6]. However, they have valuable applications [6] such as narrowing down the search space for biometric identification; gender or age-specific advertisement; improving the recognition accuracy of a biometric system; *etc*.

While extracting such auxiliary information from biometric data may be useful, there are legitimate concerns about users' privacy [1]. For example, while the user may have provided their biometric data willingly for *recognition* purposes, they may not have agreed to release other information such as their age, gender, or ethnicity. Therefore, protecting the privacy of the biometric data of an individual becomes crucial for the following reasons:

- Extracting soft biometric information and combining them with other publicly available data of an individual can lead to identity theft [11, 39];
- There could be the possibility of a link-attack [1], where the biometric data of an individual present in one database (*e.g*., driver's license database) could be linked with another non-biometric database (*e.g*., a demographic database), thereby divulging more information and leading to data accretion;
- The extracted information can be misused for profiling users based on gender/race;
- There are ethical and privacy concerns due to extracting information without consent of the users [2].

Therefore, from a privacy standpoint, due to the unexpected consequences of privacy breaches [17], it is crucial to protect the privacy of users and ensure that the stored biometric data cannot be utilized to deduce any additional information beyond that which the users expect.

In this paper, we focus on transforming a face image such that it can be used for recognition purposes by a biometric

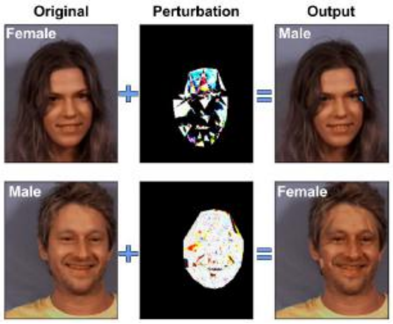

Figure 1: Objective of our work: perturb a face image such that the gender attribute is flipped as assessed by an automated gender classifier.

matcher, but information such as gender and race cannot be reliably estimated by a soft biometric classifier. Specifically, we consider flipping the gender attribute of a given face image, as shown in Fig. 1. Given the original input image (left column), applying the perturbations in the middle column results in an output image (right column) whose gender, as assessed by an automated gender classifier, is flipped to male (respectively, female) in the top (respectively, bottom) row. Our goal is to find these perturbations for an input face image.

1.1. Related work

1.1.1 Biometric privacy

Extensive work has been done on protecting the privacy of biometric data [31, 27]. Natgunanathan *et al.* [27] surveyed a comprehensive list of privacy-preserving biometric schemes and identified their drawbacks. One aspect of biometric privacy is face de-identification, where a face image is modified such that it cannot be used to determine the identity of an individual. The ad-hoc solution for face de-identification is to naively distort images by blurring or pixelation [4]; however, this technique obscures facial details which reduces the *utility* of the biometric data. Other de-identifying techniques that preserve much of the facial details have been proposed in the literature. Newton *et al.* [28] proposed the *k-same* technique in which the facial color and texture of k faces are averaged, thereby reducing the chance of uniquely identifying a face image. Later, Gross *et al.* [12] improved this technique and proposed the *k-same-M* approach by incorporating the Active Appearance Model (AAM) [5]. Face swapping was proposed by Bitouk *et al.* [3], where a source face is seamlessly blended with candidate images similar in appearance and pose, resulting in new de-identified faces. Jourabloo *et al.* [16] adopted the *k-same* algorithm and proposed an optimization scheme to find the optimal set of weights for the k face images to preserve facial attributes such as gender, age-group, ethnicity and other details like eyewear.

Inspired by the work of Othman and Ross [29], Sim and Zhang [35] promoted a new concept in face privacy, in which certain attributes are selectively suppressed or retained during face de-identification. While most of the previously proposed face de-identification methods perturb facial attributes, they developed a scheme that is able to preserve such attributes, thereby ensuring data utility in visual analytic applications where such facial details are needed. For this purpose, they applied multimodal discriminant analysis [36] to project images onto three subspaces pertaining to the age, gender, and ethnicity attributes, plus a residual space that accounts for identity. Altering or retaining the vectors in these subspaces results in a new face image with the corresponding facial attribute either modified or preserved.

Another aspect of privacy is to protect the identity of the faces while selectively suppressing other deducible attributes such as age, gender, *etc*. This has several privacy applications; since the most common reason for storing biometric data is for recognition purposes, preventing the automatic extraction of other facial attributes becomes important. Therefore, this aspect of privacy will ensure that the perturbed biometric data is still usable for biometric recognition, while it precludes the possibility of automatically inferring auxiliary information from the data.

One of the earliest attempts in this regard was by Rowland and Perrett [32] for gender conversion of facial images. In their work, they proposed finding prototypes of male and female faces, and used these prototypes to define a gender conversion axis as a vector difference between the two prototypes. An input face image is then modified by adding a fraction α of this vector to it. Later, Suo *et al.* [38] proposed a component-based approach in which face images are decomposed into several facial components. For a given face image, they replace each facial component with that of the closest match from the opposite gender group. The resulting image could successfully be used for recognition purposes, although its gender information was perturbed. Othman and Ross [29] developed a privacy preserving scheme where the gender of an input face image is progressively suppressed by mixing it with candidate face images from the opposite gender. The resulting face images, however, could be successfully used by a biometric matcher, thereby retaining their intended utility.

1.1.2 Adversarial images

Szegedy *et al.* [41] discovered that neural networks are vulnerable to adversarial examples. They defined adversarial examples as input data that are perturbed slightly in such a way that the network will misclassify them. They proposed

Table 1: Overview of gender perturbation methods.

Authors	**Proposed Method**	**Comments**
Rowland and Perrett [32]	Prototypes of male/female faces	Ghosting artifacts Pronounced perceptual changes
Tiddeman *et al*. [42]	Prototypes in the wavelet domain	Improved ghosting effects
Suo *et al*. [38]	Component-based	Seamless; No ghosting effect Pronounced perceptual changes
Othman and Ross [29]	Mixing faces of opposite gender	Generating multiple outputs Ghosting artifacts Pronounced perceptual changes
Sim and Li [35]	Multimodal Discriminant Analysis	Pronounced perceptual changes

a box-constrained optimization problem to find the smallest perturbations required to modify the input such that the output target label is changed. The perturbations applied to input images are barely perceptible to the human eye. Further, they showed that these perturbations are robust in that the same perturbations can cause misclassification on different networks with the same topology but trained on different subsets of data. Later, Goodfellow *et al*. [10] proposed a fast-gradient sign method for generating adversarial perturbations, and observed that adversarial examples generalize well to different neural network models with different architectures or trained using disjoint training sets. Rozsa *et al*. [33] proposed a fast flipping attribute (FFA) algorithm for generating adversarial examples, which leverages backpropagation and the negative gradients of the decision layer of a neural network to perturb input examples. They found that input examples that were misclassified naturally (referred to as natural adversarial examples) could be correctly classified after perturbations using the fast flipping algorithm.

Focusing on the gender attribute of a face image, besides the adversarial technique that leverages neural networks, the previously mentioned attribute conversion methods relied on either using prototypes for different classes or fusion of facial components. Considering the fact that the primary objective of these methods was to modify the apparent gender of an input face image as assessed by a human observer, the output contains perceptual changes compared to the input face image. As a result, a human observer is potentially fooled into assigning an incorrect gender label to the modified image.[1] These methods modify the face and texture of the input face image, without explicitly determining the specific features of the face that is being exploited by the attribute classifier. Therefore, these methods induce unnecessary changes to the input face image which may not be directly affecting attribute conversion. As opposed to previous methods, in this work, our goal is to apply changes that *specifically* target a particular attribute (in this case, flipping the gender attribute). Given a specific face matcher and a specific face attribute classifier, we propose an attribute flipping algorithm to iteratively perturb face images and show that it is possible to generate adversarial images which are misclassified by a robust attribute classifier (*e.g.*, gender classifier). We show that, in most cases, slightly perturbing a few pixels in the input can confuse the gender classifier, while retaining the utility of a biometric matcher. Note that we target a specific gender classifier; in other words, we do not intend to flip the gender attribute as assessed by a human observer. As a result, if a human is monitoring the images, they may be able to detect the correct gender. In summary, the contributions of the proposed work compared to previous methods are as follows:

- The proposed method is generalizable to work with any biometric matcher and facial attribute classifier;
- The proposed method finds perturbations that specifically target flipping an attribute, resulting in imperceptible changes (in most cases).

2. Proposed Method

2.1. Problem formulation

We assume that we are given a binary[2] attribute classifier, f, that outputs a classification score for an input image $X \in R^n$, *i.e.*, $f : R^n \to R$. The class label is computed as $\text{sign}(f(X))$. Furthermore, we denote $M(X_i, X_j)$ as a biometric matcher that computes the match score between face images X_i and X_j. Our goal is to efficiently find a minimally perturbed image $X' = \phi(X)$ such that $g(X)f(X') < 0$, where $g(X) \in \{-1, 1\}$ is the ground-truth attribute label (*e.g.*, female=-1; male=1). Here, ϕ is

[1] It is worthwhile to note that the focus of our paper is on fooling an *automated classifier* as opposed to a *human observer*.

[2] In this paper we assume gender has 2 labels; however, it must be noted that societal and personal interpretation of gender can result in many more classes. Facebook, for example, suggests over 58 gender classes: https://goo.gl/lwTJhr.

the transformation function that modifies the image. Therefore, we define the following cost function for optimization:

$$J(X', X) = f(X') \operatorname{sign}(g(X)). \tag{1}$$

This cost function is designed to give a positive value if the estimated attribute score of the perturbed image X' has the same sign (positive for male, negative for female) as its ground-truth label.

Optimizing Attribute Perturbation: Based on the objective function given above, the optimization problem for attribute perturbation can be stated as follows:

$$\min_{\phi} J(X', X) \text{ where } X' = \phi(X). \tag{2}$$

For this optimization, we apply small incremental perturbations to an input image as described next.

2.2. Finding perturbation direction

One way of perturbing an input image is through modifying one pixel at a time and computing the cost function. However, this method is not efficient given the large search space; also, the attribute classification output may not be useful due to low sensitivity after changing only one pixel. Therefore, we use a warping technique to simultaneously modify a group of pixels. The group of pixels to be modified are determined via Delaunay triangulation based on facial landmark points [24]. Also, in order to find the "direction" of perturbing a group of pixels in one triangle, we first select a candidate face image from a gallery set that has the highest correlation of facial landmark points with the input face image. The proposed method for finding the perturbations that would flip the gender attribute of a face image is illustrated in Fig. 2.

Given a source face image X_0^S, a set of 77 facial landmark points L^S are extracted using the Stasm software (see Fig. 3 for an example). Then, a candidate image X^C that has the highest correlation of landmark points L^C with those of X_0^S is chosen from a gallery set of faces. The correlations are calculated by averaging over the correlations of x and y coordinates of corresponding landmark points. Next, Delaunay triangulation is performed on points in L^S. For each triangle T^S, the corresponding triangle points T^C are found from the candidate image X^C. The iterative perturbation procedure starts by initializing $X^S \leftarrow X_0^S$. For each triangle T^S and its corresponding T^C, the affine transformation matrix A_t is estimated that maps T^C onto T^S (*i.e.*, $T^S = A_t\, T^C$). Next, a binary mask $Mask_T$ is defined, that has a value of 1 corresponding to image pixels inside triangle T^S. Finally, the source image is perturbed as,

$$\begin{aligned} X_{T,\alpha}^{S'} &= (1 - Mask_T)X^S + \\ &Mask_T\left((1-\alpha)X^S + \alpha A_t X^C\right), \end{aligned} \tag{3}$$

Algorithm 1: Attribute Perturbations

```
Input: a 2D face image $X_0^S$, a gallery of face images,
       a face attribute classifier $f$, threshold $\eta$
Output: perturbed image $X'$ whose attribute is flipped
Find landmark points $L^S$ on $X_0^S$
Find a candidate in gallery, $X^C$, that has highest
  correlation of landmark points, $L^C$, with those of $X_0^S$
Apply Delaunay triangulation to $L^S$ and find the
  corresponding triangles in $L^C$
Initialize $X^S \leftarrow X_0^S$
repeat
    For each triangle $T$ in $L^S$:
    – create matrix $Mask_T$ with ones for pixels inside
      triangle $T$ and zeros everywhere else
    – estimate the affine transformation matrix $A_t$:
          $T^S = A_t\, T^C$
    – define $\alpha = (\epsilon^+, \epsilon^-)$
    – apply perturbations in two directions:
          $X_{T,\alpha}^{S'} = (1 - Mask_T)X^S + Mask_T\left((1-\alpha)X^S + \alpha A_t X^C\right) \quad \forall\alpha$
    – calculate the cost function $J$ associated with
      perturbed images: $J(X_{T,\epsilon^+}^{S'}, X_0^S)$, $J(X_{T,\epsilon^-}^{S'}, X_0^S)$
    – compute numerical gradients for cost function:
          $\nabla_T J = \frac{J(X_{T,\epsilon^+}^{S'}, X_0^S) - J(X_{T,\epsilon^-}^{S'}, X_0^S)}{2 \times \epsilon^+}$
    – choose the perturbed new image for the next
      step:
          $X^{S^{new}} = \begin{cases} X_{T,\epsilon^+}^{S'}, & \text{if } \nabla_T J < 0, \\ X_{T,\epsilon^-}^{S'}, & \text{otherwise.} \end{cases}$
until cost function $J(X^S, X_0^S)$ goes below threshold $\eta$
```

where, coefficient α determines perturbing pixels either towards the candidate image (when $\alpha = \epsilon^+ > 0$) or away from the candidate image (when $\alpha = \epsilon^- < 0$), and $|\alpha|$ determines the magnitude of the perturbations. $Mask_T$ ensures that the perturbations are only applied to the triangle T^S, while face pixels outside T^S stay unmodified.

Since we do not have the closed mathematical form of attribute classifier $f(X)$, we use the central-difference to compute the gradient of cost function $\nabla_{T,\alpha} J$ numerically:

$$\nabla_T J(X_T^{S'}, X_0) = \frac{J(X_{T,\epsilon^+}^{S'}, X_0) - J(X_{T,\epsilon^-}^{S'}, X_0)}{2\epsilon^+}. \tag{4}$$

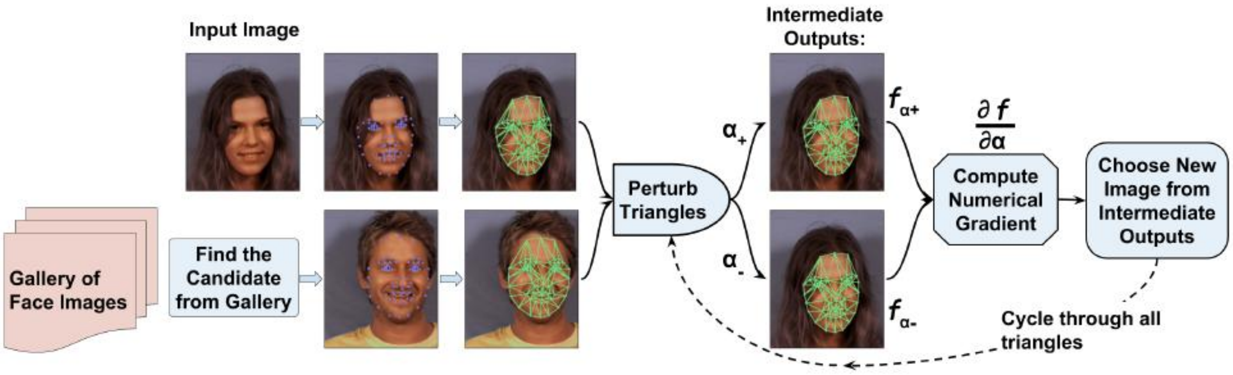
Input Image
Intermediate Outputs:
Gallery of Face Images
Find the Candidate from Gallery
Perturb Triangles
α+
α-
$f_{\alpha+}$
$f_{\alpha-}$
$\frac{\partial f}{\partial \alpha}$
Compute Numerical Gradient
Choose New Image from Intermediate Outputs
Cycle through all triangles

Figure 2: Workflow of the proposed method for finding per image perturbations in order to flip the gender attribute of a face image.

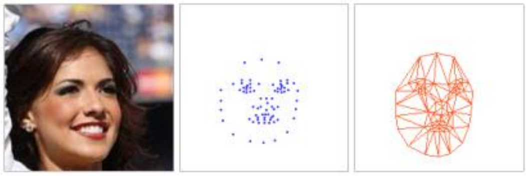

Figure 3: Example of Delaunay triangulation on landmark points extracted from an input face image.

Based on the numerical gradient computed above, the perturbation which results in decreasing the cost function J is accepted according to the following rule:

$$X^{S^{new}} = \begin{cases} X^{S'}_{T,\epsilon^+}, & \text{if} \quad \nabla_T J < 0, \\ X^{S'}_{T,\epsilon^-}, & \text{otherwise.} \end{cases} \tag{5}$$

The next iteration starts with $X^S = X^{S^{new}}$. This procedure is repeatedly performed for all triangles until the cost function goes below a predefined threshold η.

The entire process is summarized in Algorithm 1.

3. Results

We used two face datasets for evaluating the proposed method of perturbing face attributes: the MUCT dataset [23] which has 276 subjects – 131 male subjects, 145 female subjects – with 10 or 15 samples per subject and the LFW dataset [14] which has 5740 subjects – 1461 female and 4274 male subjects – and a total of 13227 face samples. In this section, we present the results of our approach to perturb the gender attribute of input images. For this purpose, we designed two experiments as shown in Table 2, where perturbations are generated based on two gender classifiers, IntraFace [7] and a Commercial-of-The-Shelf (GCOTS) software. For computing the numerical gradient as mentioned in the previous section, we used $\epsilon^+ = 0.05$ and $\epsilon^- = -0.05$. Our algorithm is run iteratively until the cost function reaches $\eta = -0.1$ or less. A secondary stopping criterion is invoked when the number of iterations exceeds a maximum user set value; in our experiments, this value is set to 40.

Table 2: Summary of designed experiments.

Experiments	Perturbations guided by
Exp1	IntraFace
Exp2	GCOTS
Exp3	None (Ref. [29])

Analysis 1: Assessing how gender prediction is affected.

Dataset	Gender classifier	
Original (before)	IntraFace	GCOTS
Exp1 output	IntraFace	GCOTS
Exp2 output	IntraFace	GCOTS
Exp3 output	IntraFace	GCOTS

Analysis 2: Assessing how identity matching is affected through computing genuine/impostor match scores.

Dataset	Match score estimator	
Original (before)	VGG	MCOTS
Exp1 output	VGG	MCOTS
Exp2 output	VGG	MCOTS

3.1. Gender perturbation

Two examples from the MUCT dataset are shown in Fig. 4 where the gender score is progressively suppressed as assessed by the IntraFace gender classifier. Figure 4(a) shows a face image whose gender score is initially negative (*i.e.*, female), which after 31 triangle update steps becomes positive (*i.e.*, male). A similar trend is observed in Fig. 4(b), where the initial positive gender score (suggesting a male face) is successfully flipped to a negative value (suggesting a female face).

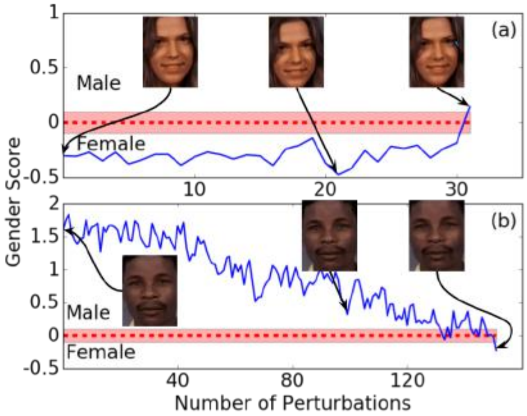

Figure 4: Two examples showing the progress of incremental gender perturbation based on IntraFace gender classifier. (a) Input image initially classified as female (gender score=-0.4), gradually perturbed until classified as male (gender score=0.1). (b) Input image initially classified as male (gender score=1.7), gradually perturbed until classified as female (gender score=-0.1).

While some of the previous gender perturbation methods rely on utilizing a candidate face image of the opposite gender [29, 32, 38], our method does not stipulate this condition. Given an input source image, our method was tested using candidates from the same gender as well as candidates from the opposite gender; it was observed that our method successfully works with both types of candidates. However, in some cases, utilizing candidates from opposite gender required fewer iterations, although the difference was not statistically significant.

The histograms of male and female scores of images in the MUCT dataset before and after gender perturbation are displayed in Fig. 5. In Fig. 5, panels (a) and (b) show the distribution of gender scores of images in the original dataset as computed using IntraFace and GCOTS, respectively. Panels (c) and (d) show the distribution of gender scores on the images after they have been perturbed using the proposed method, and panels (e) and (f) shows the scores for output images generated using the algorithm in [29]. In the original dataset (Fig. 5(a)), the distribution of gender score for male subjects is shown in white, and that of female subjects is shown in green. The histogram of gender scores after applying our gender-perturbation method and using the IntraFace gender classifier to guide the process, is shown in Fig. 5(c). This analysis indicates that gender scores are flipped, *i.e.*, those which were originally classified as male now have negative scores, and vice versa. Similar results are obtained when GCOTS is used to guide the perturbation process (see Fig. 5(b) and (d)). Note that in this case, although the distribution of gender scores for ground-truth males and females are very well separated in the original dataset, yet our method can successfully perturb the gender attribute. The histograms of gender scores computed for output images from [29] are completely overlapped. This is expected according to the K-anonymity [40] principle, since two subjects from opposite genders are mapped to a single mixed face. Quantitatively, the confusion matrices before and after gender perturbation (see Table 4) indicate that a 14.9% misclassification rate in the original dataset has increased to 76.6% after gender perturbation based on the proposed method and guided by IntraFace.

Table 3: Gender prediction errors $(\%)$ computed using IntraFace and GCOTS on the MUCT and LFW datasets.

Dataset		IntraFace	GCOTS
MUCT	Original	14.9	5.1
	Perturbed by IntraFace	76.6	5.4
	Perturbed by GCOTS	24.1	90.1
	Ref. [29]	50.2	51.9
LFW	Original	10.5	2.7
	Perturbed by IntraFace	90.0	2.5
	Perturbed by GCOTS	24.3	68.6
	Ref. [29]	55.5	45.0

Table 4: Confusion matrices for gender prediction using IntraFace, on the original MUCT dataset (top) and after perturbations guided by IntraFace (bottom).

		Predictions	
		Male	Female
Ground Truth	Male	1762	17
	Female	521	1300

		Predictions	
		Male	Female
Ground Truth	Male	276	1503
	Female	1255	566

While the objectives of our current work are similar to that of [29], there are some important differences. In our work, we intend to flip the gender attribute as assessed by a specific gender classifier, while in [29], two face images from opposite genders are mixed without taking into account any specific gender classifier. Furthermore, in their

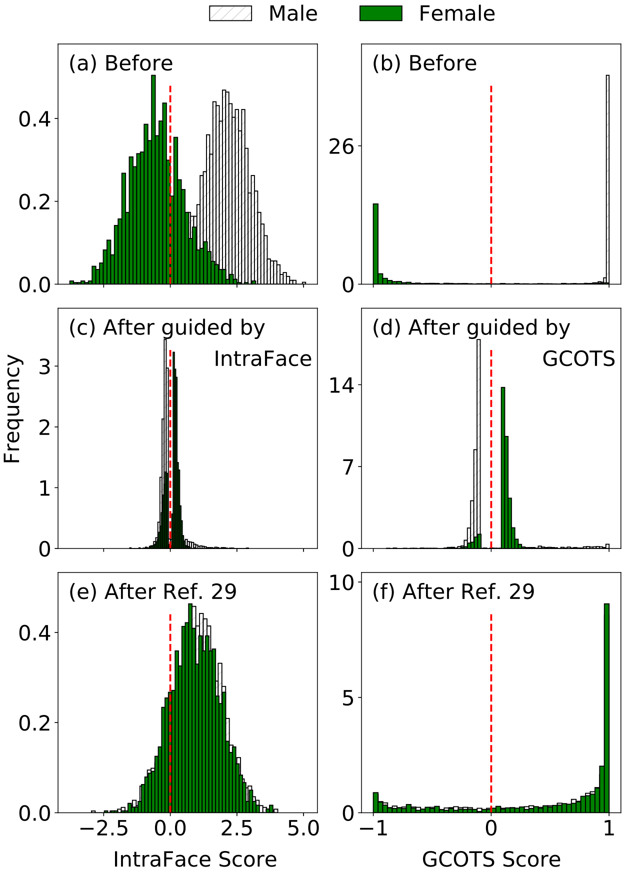

Figure 5: Histogram of gender scores obtained by IntraFace [7] ((a),(c), and (e)) and GCOTS ((b),(d), and (f)) on the MUCT dataset. Top row shows the histogram of gender scores in the original data set before perturbations, and middle shows histograms after perturbation. For comparison, the histograms of gender scores for the method proposed by Othman and Ross [29] is shown in (e) and (f). Note that the proposed algorithm is successfully flipping the gender attribute as assessed by both gender classifiers.

work, both shape and texture are modified, while in our work, only the texture has changed and the shape of the source face image stays unchanged.

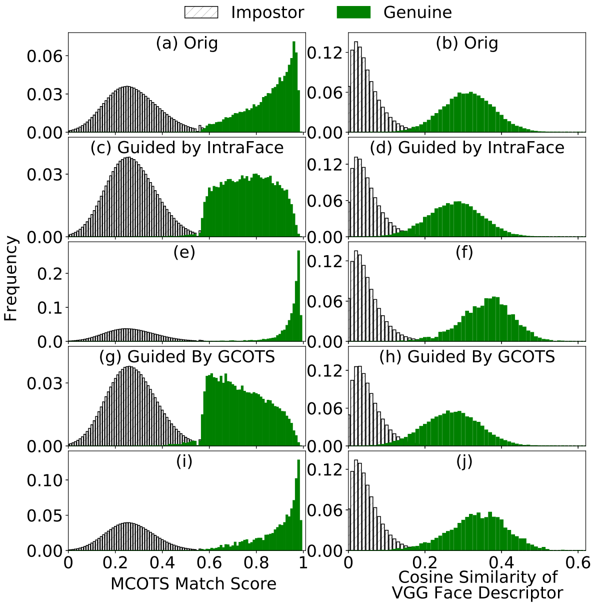

Figure 6: Distribution of genuine and impostor match scores obtained via MCOTS software (left column) and using the c VGG face descriptor [30] (right column) on the MUCT dataset; results from original dataset ((a),(b)); after gender perturbations as guided by the IntraFace gender classifier ((c),(d)); cross comparison between original and perturbed images, where perturbation was guided by IntraFace ((e),(f)); after gender perturbations as guided by the GCOTS gender classifier ((g),(h)); and cross comparison between original and perturbed images, where perturbation is guided by GCOTS ((i),(j)).

3.2. Match scores

In order to determine if the match scores are affected by the proposed gender perturbation method, we computed genuine and impostor match scores on the original MUCT and LFW datasets (before perturbation), as well as genuine and impostor scores on the perturbed datasets (guided by IntraFace gender classifier and GCOTS gender classifier). Furthermore, we also computed cross-genuine and cross-impostor scores, where face images in the original dataset are compared against those in the perturbed datasets. These experiments are conducted using two face matchers: MCOTS[3] and VGG face descriptor [30]. To obtain match scores using the VGG face descriptor, we used the cosine similarity to compare feature descriptors corresponding to a pair of images. Comparing the genuine and impostor histograms for the original dataset and the perturbed datasets (Fig. 6) shows that the distributions of genuine and impostor match scores are still well separated.

Furthermore, Receiver Operating Characteristic (ROC) curves for all three cases (before perturbation, after perturbation, and cross-comparison (before/after)) is shown in Fig. 7. The ROC curves for all these three cases show little divergence from each other, which provides further evidence that the matching accuracy is not adversely affected

[3] The face matcher in this case is a state-of-the-art COTS software that demonstrates excellent performance in challenging face datasets.

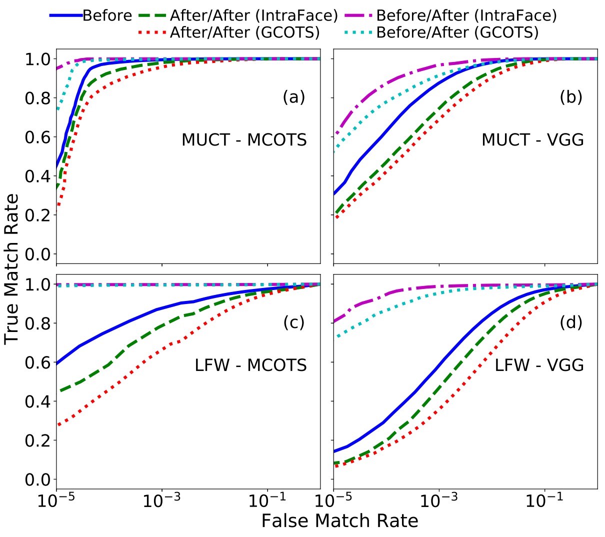

Figure 7: ROC curves for face matching obtained using the MCOTS software ((a), (c)) and the VGG face descriptor [30] ((b), (d)). Top row shows the results obtained on the MUCT dataset, and the bottom row on the LFW dataset. Note that the recognition performance is not significantly impacted in most cases.

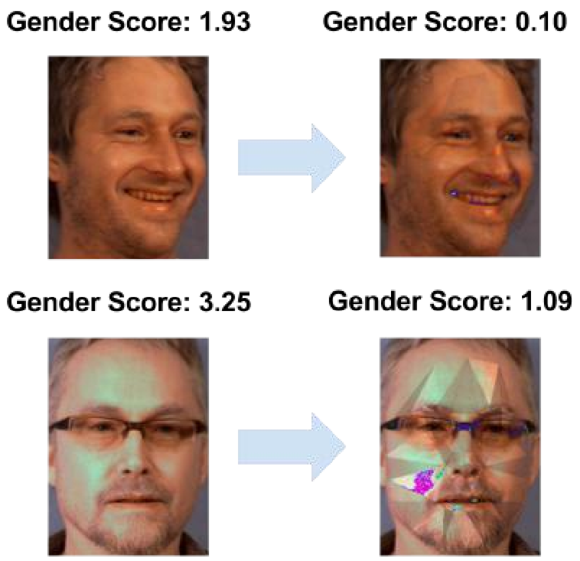

Figure 8: Two examples of unsuccessful cases where our method fails to completely flip the gender attribute as assessed by IntraFace [7].

by the perturbations.

Two unsuccessful cases are shown in Fig. 8, where the gender scores of the original images and perturbed images are both in the positive region thereby indicating the male class. We observed that the average number of perturbations in successful cases was 1084.5 (± 20) for male faces, and 667.2 (±18) for female faces, when IntraFace is used to guide the perturbation process. These numbers were found to be slightly higher when the GCOTS software is used to guide the perturbation process: 1592 (± 32) for male faces, and 782 (± 22) for female faces.

In a practical application, perturbing the gender attribute of *all* stored face images would not be prudent since the output of the attribute classifier can be trivially flipped by the user in order to obtain the true attribute value. In order to avoid this, we can apply the perturbation randomly on a certain proportion of the stored images and leave the rest unchanged. As a result, the certainty of the correct gender label will be reduced.

4. Conclusion and Future Work

While biometric data is solely expected to be used for recognizing an individual, advances in machine learning has made it possible to extract additional information such as age, gender, ethnicity, and health indicators from biometric data. These auxiliary attributes are referred to as soft biometrics. Extracting such attributes from the biometric data of an individual, without their knowledge, has raised several privacy concerns. In this work, we focused on extending privacy to face images. In particular, we designed a technique to modify a face image such that gender information cannot be easily extracted from it, while the image can still be used for biometric recognition purposes. The proposed method entails iteratively perturbing a given face image such that the performance of the face matcher is not adversely affected, but that of the soft biometric classifier is confounded. The perturbation is accomplished using a gradient descent technique. Experiments involving 2 face matchers and 2 gender classifiers convey the efficacy of the proposed method.

In the current work, we did not include any term pertaining to the biometric matcher in the cost function. In our future work, we plan to incorporate this in our cost function in order to prevent the matching accuracy from being heavily compromised. Furthermore, we plan to extend the technique to other attributes such as age, race, *etc*.

5. Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant Number 1618518.

References

[1] A. Acquisti and R. Gross. Predicting social security numbers from public data. *Proceedings of the National Academy of*

Sciences, 106(27):10975–10980, 2009. 1

[2] A. Acquisti, L. K. John, and G. Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013. 1

[3] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. *ACM Transactions on Graphics (TOG)*, 27(3):39, 2008. 2

[4] M. Boyle, C. Edwards, and S. Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 1–10, 2000. 2

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995. 2

[6] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016. 1

[7] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn. Intraface. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8, 2015. 5, 7, 8

[8] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010. 1

[9] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007. 1

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[11] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, pages 71–80, 2005. 1

[12] R. Gross, L. Sweeney, F. De la Torre, and S. Baker. Model-based face de-identification. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2006. 2

[13] S. Hosoi, E. Takikawa, and M. Kawade. Ethnicity estimation with facial images. In *6th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 195–200, 2004. 1

[14] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5

[15] A. Jain, A. A. Ross, and K. Nandakumar. *Introduction to biometrics*. Springer Science & Business Media, 2011. 1

[16] A. Jourabloo, X. Yin, and X. Liu. Attribute preserved face de-identification. In *International Conference on Biometrics (ICB)*, pages 278–285, 2015. 2

[17] E. J. Kindt. *Privacy and data protection issues of biometric applications*. Springer. 1

[18] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *12th IEEE International Conference on Computer Vision*, pages 365–372, 2009. 1

[19] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015. 1

[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 1

[21] H. Lu, Y. Huang, Y. Chen, and D. Yang. Automatic gender recognition based on pixel-pattern-based texture feature. *Journal of Real-Time Image Processing*, 3(1):109–116, 2008. 1

[22] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008. 1

[23] S. Milborrow, J. Morkel, and F. Nicolls. The MUCT landmarked face database. *Pattern Recognition Association of South Africa*, 2010. 5

[24] S. Milborrow and F. Nicolls. Active shape models with SIFT descriptors and MARS. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 380–387. IEEE, 2014. 4

[25] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002. 1

[26] E. Mkinen and R. Raisamo. An experimental comparison of gender classification methods. *Pattern Recognition Letters*, 29(10):1544–1556, 2008. 1

[27] I. Natgunanathan, A. Mehmood, Y. Xiang, G. Beliakov, and J. Yearwood. Protection of privacy in biometric data. *IEEE Access*, 4:880–892, 2016. 2

[28] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. 2

[29] A. Othman and A. Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European Conference on Computer Vision Workshop*, pages 682–696. Springer, 2014. 2, 3, 5, 6, 7

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015. 7, 8

[31] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001. 2

[32] D. A. Rowland and D. I. Perrett. Manipulating facial appearance through shape and color. *IEEE Computer Graphics and Applications*, 15(5):70–76, 1995. 2, 3, 6

[33] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boult. Are facial attributes adversarially robust? *arXiv preprint arXiv:1605.05411*, 2016. 3

[34] A. Samal, V. Subramani, and D. Marx. Analysis of sexual dimorphism in human face. *Journal of Visual Communication and Image Representation*, 18(6):453–463, 2007. 1

[35] T. Sim and L. Zhang. Controllable face privacy. In *11th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–8, 2015. 2, 3

[36] T. Sim, S. Zhang, J. Li, and Y. Chen. Simultaneous and orthogonal decomposition of data using multimodal discriminant analysis. In *12th IEEE International Conference on Computer Vision*, pages 452–459, 2009. 2

[37] Z. Sun, G. Bebis, X. Yuan, and S. J. Louis. Genetic feature subset selection for gender classification: A comparison study. In *6th IEEE Workshop on Applications of Computer Vision*, pages 165–170, 2002. 1

[38] J. Suo, L. Lin, S. Shan, X. Chen, and W. Gao. High-resolution face fusion for gender conversion. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(2):226–237, 2011. 2, 3, 6

[39] L. Sweeney. Uniqueness of simple demographics in the US population. *Technical Report, Carnegie Mellon University*, 2000. 1

[40] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002. 6

[41] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[42] B. P. Tiddeman, M. R. Stirrat, and D. I. Perrett. Towards realism in facial prototyping: results of a wavelet MRF method. In *Proc. Theory and Practice of Computer Graphics*, volume 1, pages 20–30, 2006. 3