Anonymizing k-Facial Attributes via Adversarial Perturbations

Saheb Chhabra[1], **Richa Singh**[1], **Mayank Vatsa**[1] and **Gaurav Gupta**[2]
[1] IIIT Delhi, New Delhi, India
[2] Ministry of Electronics and Information Technology, New Delhi, India
{sahebc, rsingh, mayank@iiitd.ac.in}, gauravg@gov.in

Abstract

A face image not only provides details about the identity of a subject but also reveals several attributes such as gender, race, sexual orientation, and age. Advancements in machine learning algorithms and popularity of sharing images on the World Wide Web, including social media websites, have increased the scope of data analytics and information profiling from photo collections. This poses a serious privacy threat for individuals who do not want to be profiled. This research presents a novel algorithm for anonymizing selective attributes which an individual does not want to share without affecting the visual quality of images. Using the proposed algorithm, a user can select single or multiple attributes to be surpassed while preserving identity information and visual content. The proposed adversarial perturbation based algorithm embeds imperceptible noise in an image such that attribute prediction algorithm for the selected attribute yields incorrect classification result, thereby preserving the information according to user's choice. Experiments on three popular databases i.e. MUCT, LFWcrop, and CelebA show that the proposed algorithm not only anonymizes k-attributes, but also preserves image quality and identity information.

1 Introduction

"The face is the mirror of the mind, and eyes without speaking confess the secrets of the heart." - Letter 54, St. Jerome.

Face analysis has been an area of interest for several decades in which researchers have attempted to answer important questions ranging from identity prediction [Zhou *et al.*, 2018], to emotion recognition [Li *et al.*, 2016], and attribute prediction [Abdulnabi *et al.*, 2015], [Sethi *et al.*, 2018]. Focused research efforts and use of deep learning models have led to a high performance for tasks involved in face analysis. For instance, state-of-the-art results on YouTube Faces [Wolf *et al.*, 2011] and Point and Shoot Challenge [Beveridge *et al.*, 2013] databases are over 95%, and 97% [Goswami *et al.*, 2017], and attribute recognition on the Celeb-A database is more than 90% [Sethi *et al.*, 2018].

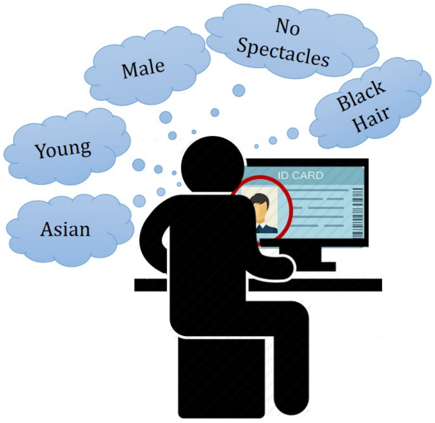

Figure 1: Profiling of a person using his face image in ID card for malicious purpose.

While there are several advantages of these high accuracies, they also pose a threat to the privacy of individuals. As shown in Figure 1, several facial attributes such as age, gender, and race can be predicted from one's profile or social media images. In a recent research, Wang and Kosinksi predicted the "secrets of the heart", such as predicting *sexual orientation* from face images [Wang and Kosinski, 2017]. They reported 81% accuracy for differentiating between gay and heterosexual men with a single image and 74% accuracy is achieved for women. Similarly, targeted advertisements by predicting the gender and age from the profile and social media photographs have been a topic of research for the last few years.

Motivated by these observations, in this research, we raise an important question: "Can we anonymize certain attributes for privacy preservation and *fool* automated methods for attribute predictions?" The answer to this question relates to k-anonymity literature where information of each individual cannot be differentiated from the $k-1$ individuals [Sweeney, 2002]. It differs from the fact that in attribute anonymization, certain attributes are to be suppressed/anonymized while remaining ones are retained. For instance, if the images are uploaded to the driving license database to ascertain identity, it should not be used for any other facial analysis, except identity matching.

Author	Method	No. of Attributes	Dataset	Controlling Attributes	Visual Appearance
Othman and Ross, 2014	Face Morphing and fusion	One	MUCT	No	Partially Preserved
Mirjalili and Ross, 2017	Delaunay Triangulation and fusion	One	MUCT, LFW	No	Partially Preserved
Mirjalili *et al.*, 2017	Fusion using Convolutional Autoencoder	One	MUCT, LFW, Celeb-A, AR-Face	No	Preserved
Rozsa *et al.*, 2016, 2017	Fast Flipping Attribute	Multiple	CelebA	No	Preserved
Proposed	Adversarial Perturbations	Multiple	MUCT, LFWcrop, Celeb-A	Yes	Preserved

Table 1: Summarizing the attribute suppression/anonymization algorithms in the literature.

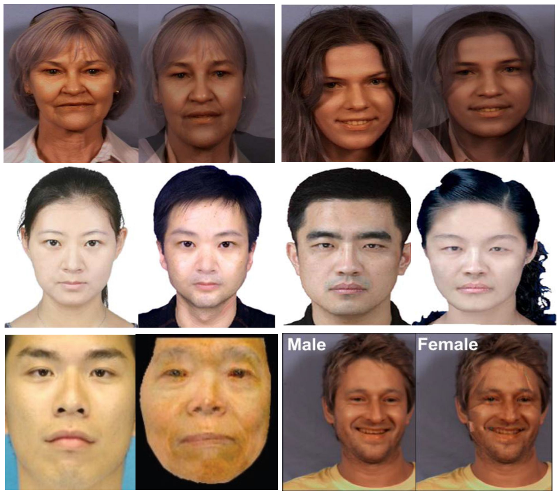

Figure 2: Demonstrating the attribute anonymized images generated by existing algorithms. First and third column images are original images while second and fourth column images are gender anonymized images [Sim and Zhang, 2015], [Suo *et al.*, 2011], [Othman and Ross, 2014], [Mirjalili and Ross, 2017].

1.1 Literature Review

In the literature, privacy preservation in face images has been studied from two perspectives. Researchers have studied this problem in terms of privacy-preserving biometrics while others have termed it as attribute suppression. As mentioned above, face images reveal a lot of ancillary information for which the user may not have consented. In order to protect such ancillary information (soft biometrics), researchers have proposed several different methodologies. [Boyle *et al.*, 2000] developed an algorithm to blur and pixelate the images in the video. [Newton *et al.*, 2005] have developed an algorithm for face de-identification in video surveillance such that the face recognition fails while preserving other facial details. [Gross *et al.*, 2006] have shown that distorting image via blurring and pixelation method gives poor results. To improve the results, they proposed model based face de-identification method for privacy protection. Some researchers have also worked on the privacy of soft biometrics. [Othman and Ross, 2014] have proposed attribute privacy preserving technique, in which the soft biometrics attribute such as gender is "flipped" while preserving the identity for face recognition. In order to flip the gender, face morphing scheme is used in which the face of other opposite gender is morphed with the input image. As an extension of this work, [Mirjalili and Ross, 2017] and [Mirjalili *et al.*, 2017] have proposed Delaunay triangulation, and convolutional autoencoders based methods to flip gender information while preserving face identity. [Suo *et al.*, 2011] have presented an image fusion framework in which the template of opposite gender face image is taken for fusion with the candidate image while preserving face identity. [Jourabloo *et al.*, 2015] have developed an algorithm for de-identification of face image while preserving other attributes. For attribute preservation, it uses k images (motivated by k-Same) which shares the same attributes for fusion. [Sim and Zhang, 2015] have proposed a method which independently controls the identity alteration and preserves the other facial attributes. It decomposes the facial attribute information into different subspaces to independently control these attributes.

To anonymize the facial attributes, [Rozsa *et al.*, 2017] have proposed deep learning based model for facial attribute prediction. A deep convolutional neural network is trained for each attribute separately and in order to test the robustness of the trained model, adversarial images are generated using fast flipping attribute (FFA) technique. As an extension of their work [Rozsa *et al.*, 2016] have used FFA and adversarial images are generated in which a facial attribute is flipped. They have observed that the few attributes are effected while flipping the targeted attribute. For instance, while changing the 'wearing lipstick' attribute, other attributes such as 'attractive' and 'heavy makeup' are also flipped.

Based on the literature review, we observe that there are algorithms for single attribute anonymization. However, there are three major challenges in anonymizing multiple attributes.

1. While anonymizing facial attributes, there should be no visual difference between original and anonymized images.
2. Selectively anonymizing few and retaining some attributes require a "control" mechanism. For example, gender and expression can be anonymized while retaining race and eye color and other attributes such as attractiveness and hair color may be in "do not care" condition.
3. In applications involving matching faces for recognition, identity should be preserved while anonymizing attributes.

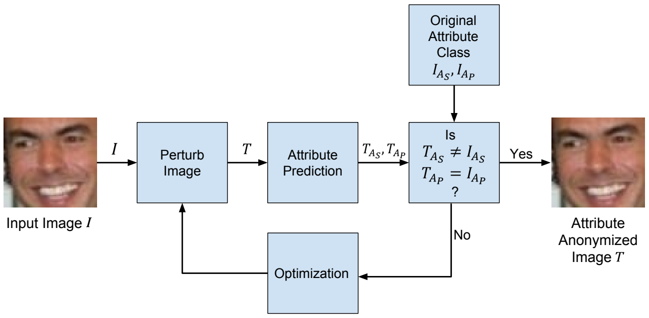

Figure 3: Illustrating the steps involved in the proposed algorithm.

The major limitation of these methods is that they do not address the first two challenges mentioned above. Existing algorithms primarily depend on a candidate image which is fused with the original image. This privacy preserving fusion leads to major transform and loss in visual appearance. [Rozsa *et al.*, 2016] have addressed this issue but due to lack of any control mechanism, the other attributes are also suppressed, while suppressing one attribute.

1.2 Research Contributions

This research proposes a novel algorithm for privacy-preserving k-attribute anonymization using adversarial perturbation. The proposed algorithm jointly anonymizes and preserves multiple facial attributes without affecting the visual appearance of the image. The proposed algorithm is also extended for identity preserving k-attributes anonymization. Experiments on three databases and comparison with existing techniques showcase the efficacy of the proposed algorithm.

2 Proposed Approach

The problem statement can formally be defined as: "create an image such that a set of pre-defined attributes are preserved while another set of pre-defined attributes are suppressed". As shown in Figure 3, the proposed algorithm can change the prediction output of certain attributes from the true class to a different target class. The detailed description of the proposed algorithm is given below:

Let $\mathbf{I}$ be the original face image with k number of attributes in the attribute set $\mathbf{A}$. For each attribute $\mathbf{A_i}$, there are C_i number of classes. For instance, in attribute 'Gender', two classes are $\{$Male, Female$\}$ and 'Expression' attribute has five classes, namely $\{$Happy, Sad, Smiling, Anger, Fearful$\}$. Mathematically, it is written as:

$$\mathbf{A} = \{\mathbf{A_1}(C_1), \mathbf{A_2}(C_2), ... \ \mathbf{A_k}(C_k)\} \tag{1}$$

Let $\mathbf{T}$ be the attribute anonymized image generated by adding perturbation $\mathbf{w}$ in the original image $\mathbf{I}$ where range of $\mathbf{T}$ is between 0 to 1. Mathematically, it is written as:

$$\mathbf{T} = \mathbf{I} + \mathbf{w} \tag{2}$$

$$\text{such that} \quad \mathbf{T} \in [0, 1]$$

To satisfy the above constraint, $tanh$ function is applied on $\mathbf{I} + \mathbf{w}$ as follows:

$$\mathbf{T} = \frac{1}{2}(tanh(\mathbf{I} + \mathbf{w}) + 1) \tag{3}$$

In an attribute suppressing/preserving application, let $\mathbf{A_S}$ and $\mathbf{A_P}$ be the sets of attributes to be suppressed and preserved where, $\mathbf{A} \geq (\mathbf{A_S} \cup \mathbf{A_P})$ and $\mathbf{A_S} \cap \mathbf{A_P} = \phi$. In $k-$ attribute anonymization task, it has to be ensured that the class of $\mathbf{A_S}$ attributes in $\mathbf{T}$ changes to some other class while preserving $\mathbf{A_P}$ attributes in image $\mathbf{T}$.

$$\mathbf{T_{A_S}} \neq \mathbf{I_{A_S}}, \mathbf{T_{A_P}} = \mathbf{I_{A_P}} \tag{4}$$

In order to suppress and preserve the sets of attributes $\mathbf{A_S}$ and $\mathbf{A_P}$ respectively, the distance between attributes of $\mathbf{A_P}$ in $\mathbf{I}$ and $\mathbf{T}$ is minimized, while the distance between attributes $\mathbf{A_S}$ is maximized. The objective function is thus represented as:

$$\min \ [D(\mathbf{I_{A_P}}, \mathbf{T_{A_P}}) - D(\mathbf{I_{A_S}}, \mathbf{T_{A_S}})] \tag{5}$$

$$\text{such that} \quad \mathbf{T_{A_S}} \neq \mathbf{I_{A_S}}, \mathbf{T_{A_P}} = \mathbf{I_{A_P}}$$

where, D is the distance metric. To preserve the visual appearance of the image, the distance between $\mathbf{I}$ and $\mathbf{T}$ is also minimized. Experimentally, we found that the ℓ_2 distance metric is most suitable for preserving the visual appearance of the image. Equation 5 is updated as:

$$\min \ \left\{D(\mathbf{I_{A_P}}, \mathbf{T_{A_P}}) - D(\mathbf{I_{A_S}}, \mathbf{T_{A_S}}) + ||\mathbf{I} - \mathbf{T}||_2^2\right\} \tag{6}$$

$$\text{such that} \quad \mathbf{T_{A_S}} \neq \mathbf{I_{A_S}}, \mathbf{T_{A_P}} = \mathbf{I_{A_P}}$$

The first two constraints i.e. $\mathbf{T_{A_S}} \neq \mathbf{I_{A_S}}, \mathbf{T_{A_P}} = \mathbf{I_{A_P}}$ and the term $[D(\mathbf{I_{A_P}}, \mathbf{T_{A_P}}) - D(\mathbf{I_{A_S}}, \mathbf{T_{A_S}})]$ in the above Equation is non linear. Therefore, to solve the same, an alternative function $f(\mathbf{T})$, inspired from [Carlini and Wagner, 2017], is used. The above Equation is thus written as:

$$\min \left\{f(\mathbf{T}) + ||\mathbf{I} - \mathbf{T}||_2^2\right\} \tag{7}$$

Here, $f(\mathbf{T})$ attempts to preserve the attributes $\mathbf{A_P}$ and suppress attributes $\mathbf{A_S}$. There can be multiple cases for facial attribute anonymization, the objective function for each case is discussed as follows:

Case I - Single Attribute Anonymization: This case formulates the scenario pertaining to changing the class of a single attribute i.e., the set $\mathbf{A_S}$ contains only one attribute, $\mathbf{U}$. In order to change the class of a single attribute $\mathbf{U}$ with class i to any other class j where $(j \neq i)$, the objective function $f(\mathbf{T})$ is formulated as:

$$f(\mathbf{T}) = max \left\{0, max(P(\mathbf{U}|\mathbf{T})) - P(U^j|\mathbf{T})\right\} \tag{8}$$

where, $P(x)$ denotes a function giving the probability value of x. For our case, we have used the Softmax output score (discussed in Section 4). The term $max(P(\mathbf{U}|\mathbf{T}))$ used in Equation 8 outputs the maximum class score of attribute $\mathbf{U}$ and term $P(U^j|\mathbf{T})$ outputs the score of each class of attribute $\mathbf{U}$ except the i^{th} class. It is important to note that Equation 8 can also provide the target attribute class by giving a class label to the variable j.

Experiment	Dataset	# Attributes Anonymized	Attribute Anonymized	
			Suppressed	Preserved
Single Attribute	MUCT, Celeb-A, LFWcrop	1	Gender	-
Multiple Attributes	Celeb-A	3, 5	Gender, Attractive, Smiling	Heavy makeup, High cheekbones
Identity Preservation	MUCT, LFWcrop	1+1	Gender	Identity

Table 2: The experiments are performed pertaining to three cases to showcase the effectiveness of the proposed algorithm.

Case II - Multiple Attribute Preservation and Suppression: This case formulates the scenario when the prediction output of multiple attributes, e.g. 'Gender' and 'Attractiveness' are suppressed while preserving other attributes e.g. 'Ethnicity', 'Wearing glasses', and 'Heavy makeup'. In this specific example, the set $\mathbf{A_P}$ contains a list of three attributes whereas, $\mathbf{A_S}$ contains list of two attributes. For such scenarios of multiple attributes, the function $f(\mathbf{T})$ will be the summation of Equation 8 for each attribute. Mathematically, it is expressed as:

$$f(\mathbf{T}) = \sum_{\mathbf{U} \in \{\mathbf{A_S}, \mathbf{A_P}\}} max\{0, max(P(\mathbf{U}|\mathbf{T})) - P(U^j|\mathbf{T})\} \quad (9)$$

In order to preserve the attributes, the target attribute class will be same as input attribute class. The optimization method in Equation 9 can control the separation between target class score and next maximum class score by providing value '$-c$' in place of 0 where $c \in [0, 1]$, i.e.

$$f(\mathbf{T}) = \sum_{\mathbf{U} \in \{\mathbf{A_S}, \mathbf{A_P}\}} max\{(-c, max(P(\mathbf{U}|\mathbf{T})) - P(U^j|\mathbf{T}))\} \quad (10)$$

Case III - Identity Preservation: The third and the most relevant case with respect to face recognition applications, is preserving the identity $\mathbf{Id}$ of a person while suppressing an attribute. The objective function in this case can be written as:

$$\min \left\{ f(\mathbf{T}) + ||\mathbf{I} - \mathbf{T}||_2^2 + \mathcal{D}(\mathbf{Id_I}, \mathbf{Id_T}) \right\} \quad (11)$$

Here, $\mathbf{Id_I}$ and $\mathbf{Id_T}$ is the identity of the face images $\mathbf{I}$, and $\mathbf{T}$ respectively obtained from any face recognition algorithm, and $\mathcal{D}$ is the distance metric to match two face features.

The proposed formulation of k-anonymizing attributes can be viewed as adding adversarial noise such that some attributes are suppressed while preserving selected attributes and identity information. Figure 4, shows some examples of attribute anonymization using the proposed algorithm.

3 Datasets and Experimental Details

The proposed algorithm is evaluated on three datasets: MUCT [Milborrow *et al.*, 2010], LFWcrop [Huang *et al.*, 2007], and CelabA[Liu *et al.*, 2015]. Comparison has been performed with the algorithm proposed by [Mirjalili and Ross, 2017]. The details of the databases are described as follows:

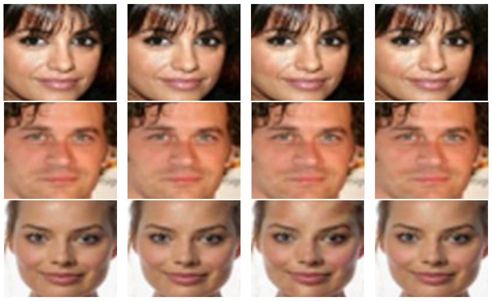

Figure 4: This figure shows the images before and after anonymizing attributes of Celeb-A dataset. First column represents the original images while second, third, and fourth columns show one, three, and five attributes anonymized images.

MUCT dataset contains 3,755 images of 276 subjects out of which 131 are male and 146 are female captured under varying illuminations using five webcams. Viola-Jones [Viola *et al.*, 2005] face detector is applied to all images and the detector fails to detect 49 face images. Therefore, only 3,706 images are used for further processing. **LFWcrop dataset** contains 13,233 face images of 5,749 subjects. In order to evaluate the performance, view 2 of the dataset which consist of 6,000 pairs of images, has been considered. Out of these 6,000 pairs of images, only one image from each pair has been selected for anonymization. **CelebA dataset** contains 202,599 face images of celebrities. From this dataset, only the test set of 19,962 images has been considered for attribute anonymization. The results of anonymizing a single attribute are shown on all three databases while the results of anonymizing multiple attributes are shown on CelebA dataset.

As shown in Table 2, three experiments are performed, one corresponding to each case discussed in Section 2. For attribute anonymization task, white box attacks are performed whereas for identity preservation, black box attack is performed. The attributes are anonymized corresponding to the attribute classification model while the identities are preserved according to face recognition algorithm. For attribute classification, we have selected fine-tuned VGGFace [Parkhi *et al.*, 2015] and for identity preservation, OpenFace [Amos *et al.*, 2016] model is used. The distance metrics used with both the models is Euclidean distance. The performance of face recognition is evaluated on both OpenFace [Amos *et al.*, 2016] and VGGFace (black box model) [Parkhi *et al.*, 2015].

Implementation details: The proposed algorithm is implemented in Tensorflow with 1080 Ti GPU. For learning the perturbation, L2 attack has been performed with Adam optimizer. The learning rate is set 0.01 and number of iterations used is 10000.

		Attribute Class	Prediction		Attribute Class	Prediction		Attribute Class	Prediction	
			Male	Not Male		Smiling	Not Smiling		Attractive	Not Attractive
Ground Truth	Before Anonymization	Male	87.70	12.30	Smiling	64.59	35.41	Attractive	89.31	10.69
		Not Male	19.64	80.36	Not Smiling	24.66	75.34	Not Attractive	28.41	71.59
	After Anonymization	Male	3.89	96.11	Smiling	0.02	99.98	Attractive	0.28	99.72
		Not Male	100	0	Not Smiling	99.90	0.10	Not Attractive	99.59	0.41

Table 3: Confusion matrix displaying classification accuracies (%) of before and after suppression of three attributes together on the CelebA dataset

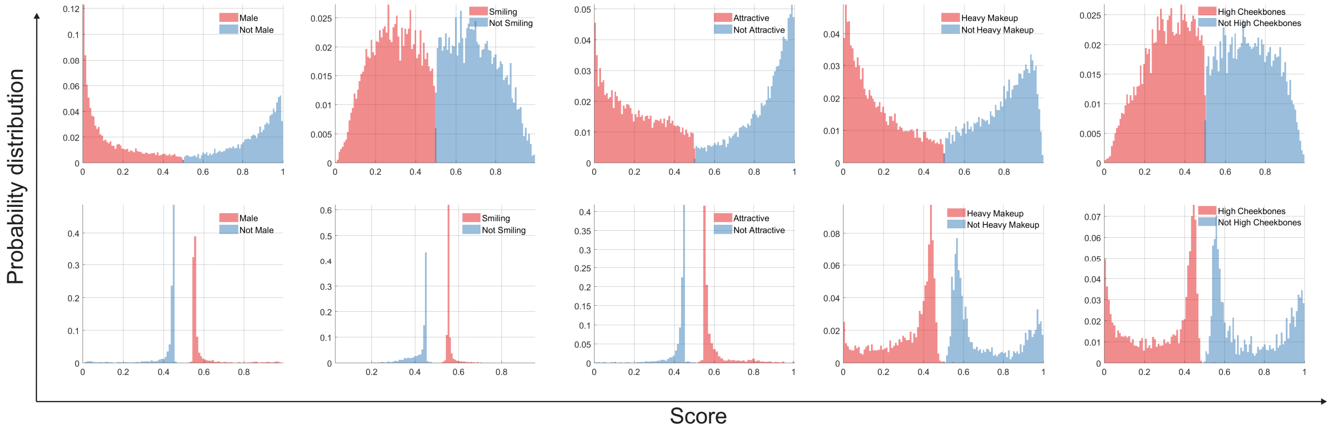

Figure 5: Comparing the attribute class score distributions of images before and after anonymization of 5 attributes together on the CelebA dataset. The first row distributions pertain to the original images whereas, second row distributions corresponds to the anonymized images.

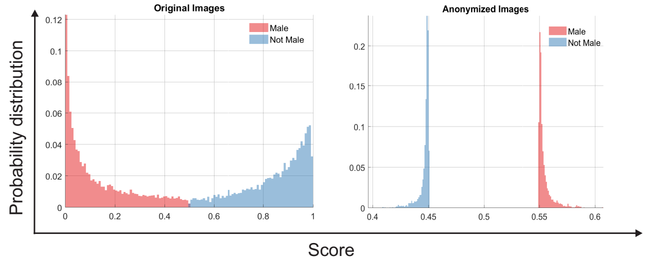

Figure 6: Comparing score distributions before and after anonymizing 'Gender' attribute on the CelebA dataset.

4 Performance Evaluation

The proposed model is evaluated for single and multiple attribute anonymization. For multiple attribute anonymization, two types of experiments are performed: attribute suppression, and attribute suppression + preservation. In case of single attribute anonymization, the performance of the proposed algorithm has been analyzed for the tasks of attribute suppression as well as identity preservation. To generate the attribute anonymized images, only those samples have been considered which are correctly classified by the attribute classification model.

4.1 Multiple Attribute Anonymization

The proposed algorithm for multiple attribute anonymization is evaluated on the CelebA dataset. Three attributes are considered for the task of attribute suppression, while attribute suppression and preservation is performed with five attributes.

Attribute Suppression: Experiments have been performed with three attributes i.e. 'Gender', 'Smiling', 'Attractive'. The confusion matrix in Table 3 shows that in over 96% of the images, attributes are correctly suppressed. After applying the proposed algorithm for the attribute 'Attractive', the classification accuracy dropped from 89.31% to 0.28%; thereby showcasing the efficacy of the proposed technique for the task of attribute suppression.

Attribute Suppression and Preservation: The results for five attributes 'Gender', 'Attractive', 'Smiling', 'Heavy Makeup' and 'High Cheekbones' are shown in Figure 5. Three attributes are suppressed i.e. 'Gender', 'Attractive' and 'Smiling', while remaining two are preserved. The confidence value ('c' value as defined in Equation 10) is set to 0.1. In Figure 5, the first and second row histograms show the attribute class score distributions before and after anonymization of images. It is observed that the attribute class score distributions of first three attributes 'Gender', 'Attractive' and 'Smiling' are flipped while the class score distributions of 'Heavy Makeup' and 'High Cheekbones' are preserved. This illustrates the utility of the proposed algorithm for multiple attributes suppression and preservation. Figure 4 presents the original and anonymized images for one, three, and five attributes. The similarity of visual appearance between original and modified images further strengthens the usage of the proposed algorithm.

Anonymization	Attribute Class	[Mirjalili and Ross, 2017]		Proposed	
		Male	Female	Male	Female
Before	Male	1762	17	1741	87
	Female	521	1300	252	1626
After	Male	276	1503	0	1828
	Female	1255	566	1878	0

Table 4: Comparison of confusion matrix of the proposed algorithm with [Mirjalili and Ross, 2017] on the MUCT dataset.

4.2 Attribute Suppression with Identity Preservation

To evaluate the performance of single attribute suppression with identity preservation, experiments are performed on MUCT, CelebA and LFWcrop datasets. The results of single attribute suppression and attribute suppression along with identity preservation are discussed below.

Single Attribute Suppression: Figure 6 shows the result of the 'Gender' attribute suppression on CelebA dataset. It can be observed that the score distribution of attributes are completely flipped before and after anonymization. The confusion matrix for MUCT dataset before and after anonymization is shown in Table 4. On comparing the results with the algorithm proposed by [Mirjalili and Ross, 2017], it can be observed that after anonymization, all samples are misclassified by the proposed method. It is important to note that the method proposed by [Mirjalili and Ross, 2017] is dependent on the fusion of the other candidate images which embeds visual distortion in the image. On the other hand, in the proposed method, the anonymization of attribute is independent of another candidate image, thus resulting in limited visual distortions. Some example images of a suppressed attribute on MUCT and LFWcrop datasets are demonstrated in Figure 7. These samples show that there are minimal effect on visual appearance of the image after attribute anonymization.

Identity Preservation: To evaluate identity preservation performance, while anonymizing an attribute, experiments are performed on both MUCT and LFWcrop datasets. For MUCT dataset, a single image gallery and two probe sets are used. The first probe set consist of original images while the second probe set contains the corresponding 'Gender' suppressed images of probe set 1. Two widely use face recognition models i.e. OpenFace [Amos *et al.*, 2016] and VGGFace [Parkhi *et al.*, 2015] are used for (i) original to original face matching (i.e., gallery to probe set 1) and (ii) original to anonymized face matching (i.e., gallery to probe set 2). The Cumulative Match Characteristic (CMC) curves summarizing the performance of face recognition before and after suppressing the attributes are show in Figure 8. The low variation observed in accuracy for recognition before and after anonymization motivates the applicability of the proposed algorithm for identity preservation as well.

Figure 9 shows the Receiver Operating Characteristic (ROC) curves for face verification on LFWcrop dataset. Experiments are performed on (i) 6000 original pair images, and (ii) 6000 original-anonymized pairs. Face verification is performed using OpenFace [Amos *et al.*, 2016] and VGGFace model [Parkhi *et al.*, 2015]. The results show minimal effect on the face verification performance before and after attribute anonymization.

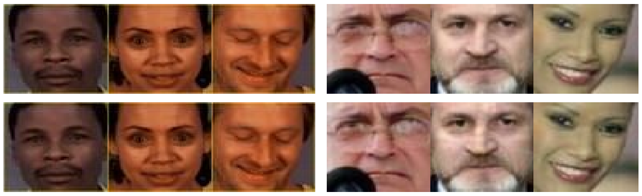

Figure 7: Examples of before and after anonymization of 'Gender' attribute on MUCT and LFWcrop datasets. First row images are original images, second row images are 'gender' anonymized images.

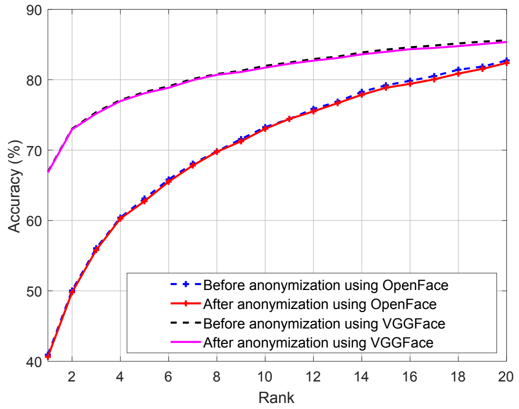

Figure 8: CMC curve showing original to original face matching vs original to anonymized face matching on the MUCT dataset.

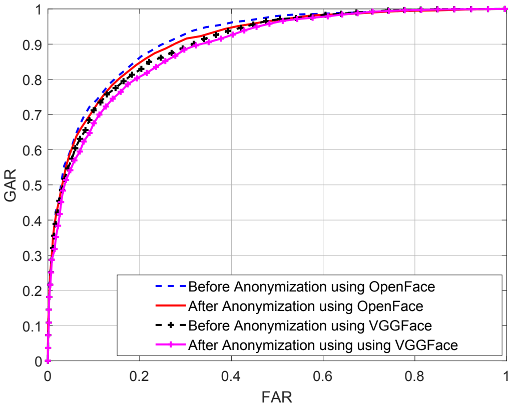

Figure 9: ROC curve showing face recognition performance on the LFWcrop Dataset

5 Conclusion

Attribute anonymization while preserving identity has several privacy preserving applications. This paper presents a novel algorithm based on adversarial noise addition concept such that selected attributes (or features) are anonymized and selected attributes (including identity information) are preserved for automated processing. Experiments are performed on CelebA, LFWcrop and MUCT databases with three different application scenarios. The results demonstrate that the proposed algorithm can handle multiple attribute anonymization process without affecting visual appearance and face recognition performance.

Acknowledgments

Vatsa and Singh are partially supported through Infosys Center for AI at IIIT-Delhi.

References

[Abdulnabi *et al.*, 2015] Abrar H Abdulnabi, Gang Wang, Jiwen Lu, and Kui Jia. Multi-task cnn model for attribute prediction. *IEEE TM*, 17(11):1949–1959, 2015.

[Amos *et al.*, 2016] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, 2016.

[Beveridge *et al.*, 2013] J Ross Beveridge, P Jonathon Phillips, David S Bolme, Bruce A Draper, Geof H Givens, Yui Man Lui, Mohammad Nayeem Teli, Hao Zhang, W Todd Scruggs, Kevin W Bowyer, et al. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE BTAS*, pages 1–8, 2013.

[Boyle *et al.*, 2000] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *CSCW*, pages 1–10. ACM, 2000.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE SP*, pages 39–57, 2017.

[Goswami *et al.*, 2017] Gaurav Goswami, Mayank Vatsa, and Richa Singh. Face verification via learned representation on feature-rich video frames. *IEEE TIFS*, 12(7):1686–1698, 2017.

[Gross *et al.*, 2006] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *IEEE CVPRW*, pages 161–161, 2006.

[Huang *et al.*, 2007] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik 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.

[Jourabloo *et al.*, 2015] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *IEEE ICB*, 2015.

[Li *et al.*, 2016] Chao Li, Chao Xu, and Zhiyong Feng. Analysis of physiological for emotion recognition with the irs model. *Neurocomputing*, 178:103–111, 2016.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[Milborrow *et al.*, 2010] Stephen Milborrow, John Morkel, and Fred Nicolls. The MUCT Landmarked Face Database. *PRASA*, 2010. `http://www.milbo.org/muct`.

[Mirjalili and Ross, 2017] Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. *IEEE IJCB*, 2017.

[Mirjalili *et al.*, 2017] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. *arXiv preprint arXiv:1712.00321*, 2017.

[Newton *et al.*, 2005] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE TKDE*, 17(2):232–243, 2005.

[Othman and Ross, 2014] Asem A Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *ECCV Workshops (2)*, pages 682–696, 2014.

[Parkhi *et al.*, 2015] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[Rozsa *et al.*, 2016] Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Are facial attributes adversarially robust? In *IEEE ICPR*, pages 3121–3127, 2016.

[Rozsa *et al.*, 2017] Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Facial attributes: Accuracy and adversarial robustness. *PRL https://doi.org/10.1016/j.patrec.2017.10.024*, 2017.

[Sethi *et al.*, 2018] Akshay Sethi, Maneet Singh, Richa Singh, and Mayank Vatsa. Residual codean autoencoder for facial attribute analysis. *PRL*, 2018.

[Sim and Zhang, 2015] Terence Sim and Li Zhang. Controllable face privacy. In *IEEE AFGR*, volume 4, pages 1–8, 2015.

[Suo *et al.*, 2011] Jinli Suo, Liang Lin, Shiguang Shan, Xilin Chen, and Wen Gao. High-resolution face fusion for gender conversion. *IEEE TSMC*, 41(2):226–237, 2011.

[Sweeney, 2002] Latanya Sweeney. k-anonymity: A model for protecting privacy. *IJUFKS*, 10(05):557–570, 2002.

[Viola *et al.*, 2005] Paul Viola, Michael J Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *IJCV*, 63(2):153–161, 2005.

[Wang and Kosinski, 2017] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *PsyArXiv preprint arXiv: 10.17605/OSF.IO/HV28A*, 2017.

[Wolf *et al.*, 2011] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE CVPR*, pages 529–534, 2011.

[Zhou *et al.*, 2018] Xiuzhuang Zhou, Kai Jin, Qian Chen, Min Xu, and Yuanyuan Shang. Multiple face tracking and recognition with identity-specific localized metric learning. *PR*, 75:41–50, 2018.