Gender Obfuscation through Face Morphing

Shunxin Wang
Faculty of Electrical Engineering, Mathematics & Computer Science
University of Twente
Supervised by:
Prof.Dr.Ir. R.N.J. Veldhuis (1^{st}), Dr.Ir. J. Goseling (2^{nd}), U.M. Kelly

***Abstract*—While facial biometric data has been widely adopted for person recognition, recent developments in machine learning show that soft biometrics such as gender, age and ethnicity can be automatically extracted from the facial photographs without permission, which raises privacy concerns. In this work, face morphing is applied to face images so that facial attributes such as gender, can no longer be deduced correctly by the corresponding attribute classifier. Meanwhile, the face images can still be used for identity verification. Experiments show that soft biometrics obfuscated through face morphing cannot be recovered or retrieved easily. It is concluded that face morphing is a good approach to protect soft biometric privacy in face images.**

1. Introduction

Physical and behavioral human characteristics such as fingerprint, face, iris, retina, palm-print and gait, can be used for identity verification, which can then control access to systems, devices or data [31]. Such characteristics are defined as biometrics. A typical face verification system is shown in Figure 1. The input of the system is a face image while the output is the decision made by the system. The features extracted from the input image are compared to the templates stored in the database. Through the decision module, it can either ensure or reject the validity of the claimed identity. However, a face also contains some ancillary information such as gender, age and ethnicity. Such characteristics that carry some information but lack the validity to distinguish between two different persons, are defined as soft biometrics [9].

For a typical face verification system, as shown in Figure 1, attributes such as gender, age and ethnicity can be automatically detected from the face images stored in the reference database by different facial attribute classifiers. This raises legitimate concerns about users' privacy if biometric data stored in the database were leaked, since users may not have agreed to share information other than identity. Therefore, protection of biometric data becomes critical for several reasons:

- Legitimate concern due to extracting other information without permission from users.
- Leakage of biometric data might cause identity theft.
- The extracted soft biometrics might be misused for the profiling of users.

According to General Data Protection Regulation (GDPR), it is essential to ensure the stored biometric data cannot be used for other purposes that are beyond the users' expectation [1]. For example, for the face verification system shown in Figure 1, the stored biometric data can only be utilized for identity verification, while other ancillary information cannot be deduced.

In this paper, face morphing is applied to face images in such a way the images can still be used for identity verification while the ancillary information is suppressed. Face morphing mixes face images through image warping and cross-dissolving. It has shown in [32] that the morphed face can be successfully used to verify the identity of the two contributing faces. Therefore, the idea is that morphing with an average face of the opposite gender would help to confuse a gender classifier, but the identity of the obfuscated face remains. Specifically, gender obfuscation ensures that a gender classifier is not able to distinguish between male and female faces. To achieve this, we morph a face (Non-obfuscated face) image with an average face of the opposite gender (Gender obfuscator), resulting in a new face (Obfuscated face) image that can confuse an arbitrary gender classifier. The goal is to investigate the utility and reversibility of face morphing in gender obfuscation. First, we determine roughly the morphing parameters suitable for gender obfuscation by comparing the degree of gender obfuscation and identity preservation. Second, we explore the influence of average faces in the performance of obfuscating gender from several aspects, such as the number of faces needed to create an average face, gender characteristics of the average face, similarity between the average face and the non-obfuscated face. Third, we investigate two possible gender retrieval approaches, which are face demorphing and gender classifier retraining. We demonstrate that face morphing is stable to obfuscate gender in face images while the gender information suppressed is hard to recover or retrieve.

The remainder of this paper is organized as follows. Section 2 introduces related work on face morphing and biometrics privacy protection. Section 3 describes the proposed gender obfuscation scheme and two potential retrieval approaches, while Section 4 gives the experimental results

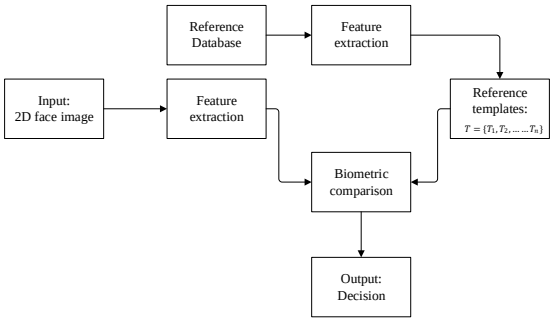


Figure 1. A typical face verification system

and analysis. Later then, some discussions and conclusions are drawn in Section 5 and 6 respectively.

2. Related work

2.1. Morphing

Morphing or metamorphosis, can be considered as a combination of multiple images through image warping and cross-dissolving [19]. Among those morphing techniques, feature-based approaches are used most frequently [3][4][18]. In [4], given two images, pairs of corresponding line segments are located for calculating the mapping functions in between. Then images are warped according to the mapping function and blended with specific weights. Later, instead of using line segments, researchers suggest using pairs of corresponding key points to improve quality of morphed images [3][18]. Early morphing approaches rely on human assistance to locate feature points while modern morphing techniques are able to create morphs automatically with the advancement in feature detection [21].

Applying morphing in face images is a well-studied topic. Traditional face morphing derives a morphed face from only two face images. Extending from this morphing technique, polymorph, introduced in [19], create a morphing framework to morph multiple faces uniformly or non-uniformly. With the development of facial landmark detection, automatic generation of morphed faces becomes possible. Nevertheless, automatic face morphing is still influenced by characteristics such as hair, glasses, pose variations, resulting in morphed images with artefacts if done without manually retouching [21].

Bui et al. [7] apply Radial Basis Functions (RBF) Networks to face morphing. The networks are able to transform the muscles on the prototype face to the source face in order to create realistic facial expressions. Not satisfied with morphing 2D images, researchers further develop realistic 3D face morphing based on dense point-to-point alignment [15].

With the advancement of morphing techniques, face morphing attack on Automated Border Control (ABC) systems has already raised concerns [34] while automatic face morphing allows researchers to create a database with numerous morphed faces, which are helpful in experiments on face morphing detection [21]. Unlike face morphing, face demorphing is the reverse process of it, which aims to discover whether a face is morphed with another face [12]. It can be considered as a special approach to detect face morphing.

Instead of morphing faces in pixel space, Generative Adversarial Networks (GANs) achieve face morphing in latent space and the resulted images are more realistic [16][36]. It can not only change one's gender, age, and ethnicity, but also de-identify one's face [37]. However, though the resulted images have better quality, training such a model can be tricky.

2.2. Biometrics privacy

Recently, biometrics privacy has raised many concerns, because some biometric systems, especially face verification systems, also collect auxiliary information like age, gender and ethnicity during acquisition. Furthermore, with advanced machine learning techniques, it has become easier to extract this auxiliary information from a face image automatically.

Privacy preservation has been studied roughly from two aspects:

1) Face de-identification
2) Facial attributes suppression

In terms of face de-identification, pixelation and blurring are the easiest ways to de-identify a face. Pixelation reduces information contained in a face image by subsampling [38], while blurring smooths a face image by filtering [6][26]. However, they cannot well retain facial attributes. Replacing a face in an image by another face which is known to the public is also a feasible approach [5][10]. The face is selected based on the similarity to the probe face. Instead of replacing the whole face, a component-based approach can be utilized [25]. In this approach, a source face image is decomposed into multiple facial components (eyes, nose, mouth) which are replaced by templates that are similar. Among these methods, there is a trade-off between identity verification and facial attributes preservation. Therefore, Wu et al. [37] further utilize GAN to generate realistic face images that preserve facial attributes but cannot be used for identification.

In terms of facial attributes suppression, Rowland and Perrett [33] achieve gender conversion by manipulating shape and color of face images based on difference between prototypes of male and female. We can also use a component-based approach, same as [25], in which a face image is decomposed into several components and these components are replaced with templates from opposite gender group that have highest similarity [35]. In this way, the identity is preserved while gender is flipped. The component-based approach results in unnatural images since templates from different images are spliced together.

Othman and Ross [27] perform gender suppression differently. Instead of face replacement, the source face image is mixed with another face image with opposite gender. The gender suppression level differs based on the face used for morphing. Different from our approach, a face instead of an average face from the opposite gender group is morphed with the non-obfuscated face. Therefore, the degree of gender suppression is hard to control because a single male face can have strong male characteristics or weak male characteristics, same for a female face. Furthermore, the performance of their gender suppression approach is evaluated by one gender classifier. Thus, it is unknown whether other gender classifiers would perform similarly or not. Additionally, it is unknown whether the gender suppressed in this way can be decompressed or retrieved.

Later on, Mirjalili et al.[23] develop Semi-Adversarial Networks (SANs) that can generate adversarial images which are robustly misclassified by gender classifiers. Continuing on the previous work, Mirjalili et al. [24][22] design FlowSAN which combines a set of SANs in order to remunerate the weakness of each model. Further investigating adversarial images, Chhabra et al. [8] studied an algorithm to embed imperceptible noise in images so that the corresponding attribute classifier outputs wrong prediction (not only gender, but also age and ethnicity). The complexity of [8][24] is much higher than using face morphing although they perform well in gender conversion.

3. Methods

In this section, we specify the gender obfuscation scheme and the evaluation methods for the degree of gender obfuscation and identity preservation.

3.1. Gender obfuscation

The average face can be considered as prototypes with specific gender characteristics. Generally, when a face is morphed with an average face of the opposite gender, its original gender characteristics are weakened. Therefore, the gender obfuscation scheme is as follows:

1) Male face + Female obfuscator (Average female face)
2) Female face + Male obfuscator (Average male face)

The '+' indicates the process of face morphing and an obfuscator is an average face with opposite gender (relative to the non-obfuscated face). Examples of obfuscators are shown in Figure 2.

The contour of the face and the region outside of it affect gender classification, such as hair style, wearing clothes. To avoid the influence of region outside of face in gender classification, we generate a neutral background, as shown in Figure 2c. It is created by morphing 10 female faces and 10 male faces, so the background is the average background of the 20 face images. After morphing with an obfuscator, the resulted face is pasted into the neutral background in the face region using Poisson blending [30][1]. The schema of our proposed gender obfuscation procedure is shown in Figure 3. The average background is generated by mixing background in non-obfuscated face and obfuscator.

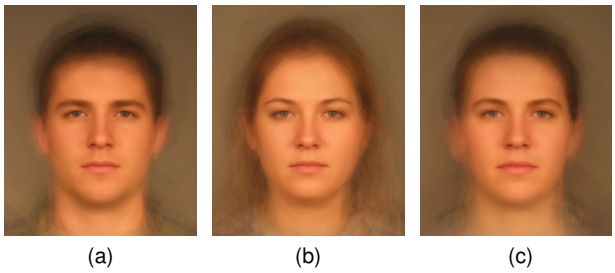


Figure 2. Obfuscators and neutral background (a: Male obfuscator, b: Female obfuscator, c: Neutral background)

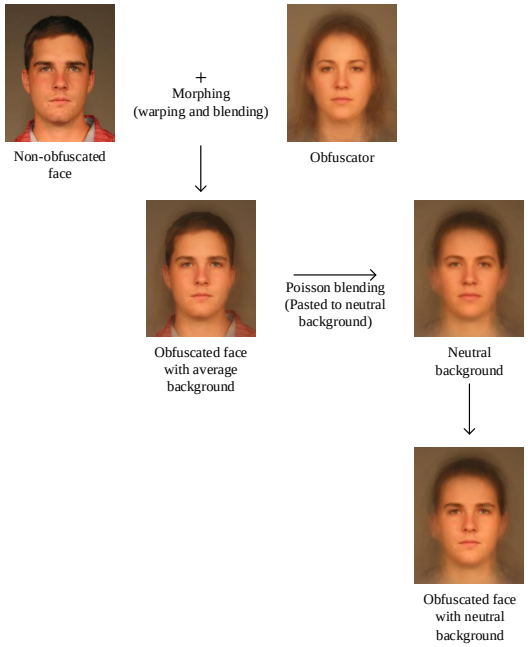
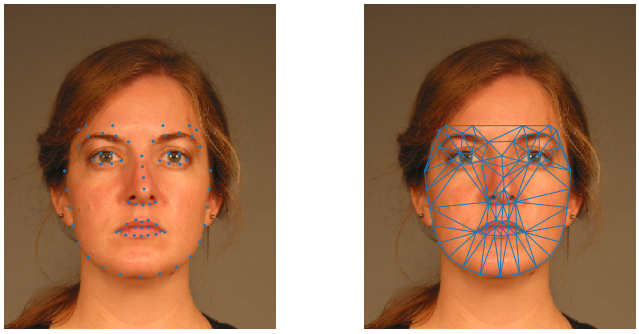


Figure 3. Gender obfuscation scheme

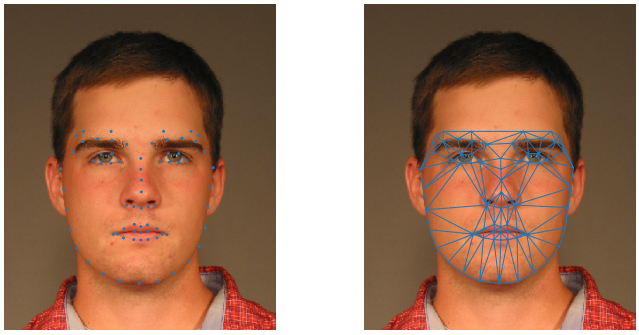
3.2. Face Morphing

Given two face images, the corresponding facial landmarks need to be located first. In total, 71 landmarks are used. 68 landmark points are localized by *shape_predictor_68_face_landmarks* from the *dlib* library (http://dlib.net) at first. The faces we use have a neutral expression, which means that the mouth should be closed. In that case, the three landmarks depicting the upper contour

1. Note that the function of Poisson blending is different from blending in face morphing. Poisson blending is used to copy the morphed face to the neutral background while blending in face morphing is to combine faces.



(a) (b)



(c) (d)

Figure 4. Delaunay triangles based on the detected facial landmarks

of the lower lip are identical to the landmarks on the lower contour of the upper lip. We remove these three duplicates and add six landmarks on the forehead. The positions of the six landmarks are determined by the width w and height h of the bounding rectangle of the detected facial landmark points. Suppose that (x, y) is the position of the up-right corner of the bounding rectangle. The positions of the extra three landmarks above the left (from the view of observers) eyebrow are: $(x+0.1\cdot w, y-0.02\cdot h)$, $(x+0.15\cdot w, y-0.05\cdot h)$, $(x + 0.33 \cdot w, y - 0.05 \cdot h)$ (pixel indices). The other three landmarks above the right eyebrow are located in: $(x+w-0.1\cdot w, y-0.02\cdot h)$, $(x+w-0.15\cdot w, y-0.05\cdot h)$, $(x + w - 0.33 \cdot w, y - 0.05 \cdot h)$.

After locating the corresponding points, as shown in Figure 4a and 4c, Delaunay Triangulation [29] is applied (Figure 4b, 4d), which can improve the quality of morphs since it separates face into multiple triangles, and for each pair of corresponding triangles, there is a warping function, instead of one warping function for the whole face. Warping functions are used to calculate the intermediate facial structure, as explained in the next paragraph.

Warping changes the facial structure, to what extent is controlled by the parameter α_w (warping parameter). As shown in Figure 5, landmark points in blue belongs to image I_0 and landmark points in orange belongs to image I_1. The locations of the brown points are computed based on the locations of blue and orange points using the following equation:

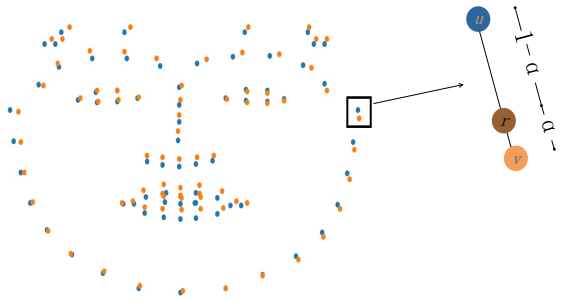


Figure 5. Facial landmarks of intermediate frame I_α (brown point)

$$P_\alpha = \{r_i | r_i = \alpha_w \cdot u_i + (1-\alpha_w)\cdot v_i, u_i \in P_0, v_i \in P_1\} \quad (1)$$

where P_0 is the set of landmarks in I_0, P_1 is the set of landmarks in I_1 and P_α is the set of landmarks in intermediate frame I_α. Warping functions from I_α to I_0 ($w_{P_\alpha \longrightarrow P_0}$) and from I_α to I_1 ($w_{P_\alpha \longrightarrow P_1}$) are further used in the following equation:

$$I_\alpha(p) = \alpha_b \cdot I_0(\omega_{P_\alpha \to P_0}(p)) + (1-\alpha_b)\cdot I_1(\omega_{P_\alpha \to P_1}(p)) \quad (2)$$

where p is the position of a point in I_α and $\omega_{P_\alpha \to P_0}(p)$ demonstrates its corresponding position in I_0. $I_\alpha(p)$ is the pixel value in position p in the intermediate frame I_α, computed from the pixel values in the corresponding positions in I_0 and I_1 with weights α_b (blending parameter) and $1-\alpha_b$.

α_w and α_b can be different. The morphed face images with different combinations of α_w and α_b are shown in Figure 6. In this figure, faces are not pasted into the neutral background, and the background is the weighted average of the background in the two images. Research has shown that the influence of α_b is larger than that of α_w in face recognition [11][17].

In our gender obfuscation scheme, average faces are needed, which are created by morphing multiple face images. The locations of corresponding points in I_α are computed by the following equation:

$$P_\alpha = \{r_i | r_i = \frac{u_{1,i} + u_{2,i} + ... + u_{N,i}}{N}, 1 \le n \le N\} \quad (3)$$

where $u_{n,i}$ is the position of the i^{th} point in image I_n and N is the number of faces for morphing. An example of computing an average position from three points is shown in Figure 7.

Deducing from equation (2), the pixel values in an average face are computed as follows:

$$I_\alpha(p) = \frac{I_0(\omega_{P_\alpha \to P_0}(p)) + ... + I_N(\omega_{P_\alpha \to P_N}(p))}{N} \quad (4)$$

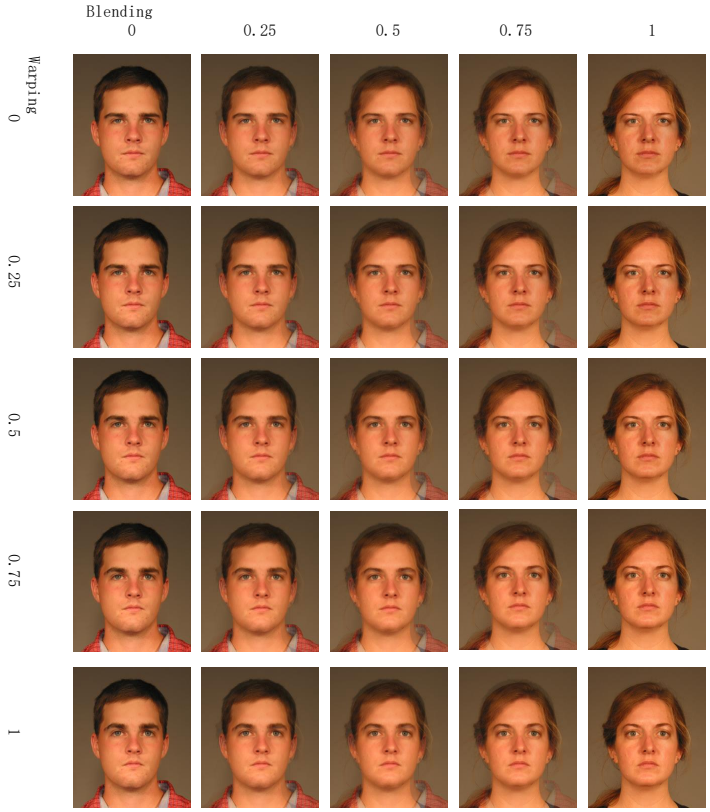


Figure 6. Morphed faces with different combinations of α_w and α_b

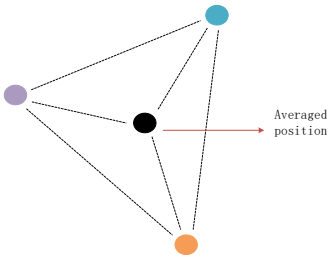


Figure 7. Average position from three points

3.3. Gender retrieval

From the aspect of privacy protection, it would be undesirable if it were possible to recover or retrieve the protected information. In this section, potential retrieval methods are explored. One method is face demorphing. Another is to train a gender classifier which can directly distinguish between obfuscated female and male faces.

Face demorphing is the reverse process of face morphing. Suppose that we have subjects A and B, and their combined face through face morphing is noted as C. Image C is the linear combination of A and B, which means $C = A + B$. In real scenario, we have only the combined face C and a live captured face image A' which belongs to subject A. Then the demorphed image $D = C - A'$, and it should be matched to subject B. In the case of recovering gender information in obfuscated faces, there is no such live captured image, but an average face. So the obfuscated face is demorphed with an average face, and its gender score is utilized further. An example is shown in Figure 8, where the '-' indicates the process of face demorphing.

Training a gender classifier to distinguish between obfuscated female and male faces is a potential way as well. Due to the computational cost of training, a pre-trained model in [13] is used, with ShuffleNet V2 architecture [20]. To distinguish from the pre-trained gender classifier used in evaluation, the re-trained gender classifier is called Re-trained ShuffleNet while the pre-trained gender classifier is called ShuffleNet.

3.4. Evaluation

The goal is to obfuscate gender characteristics in face images while maintaining identity information. Therefore, two aspects need to be evaluated. The first is the degree of gender obfuscation, and the second is the degree of identity preservation.

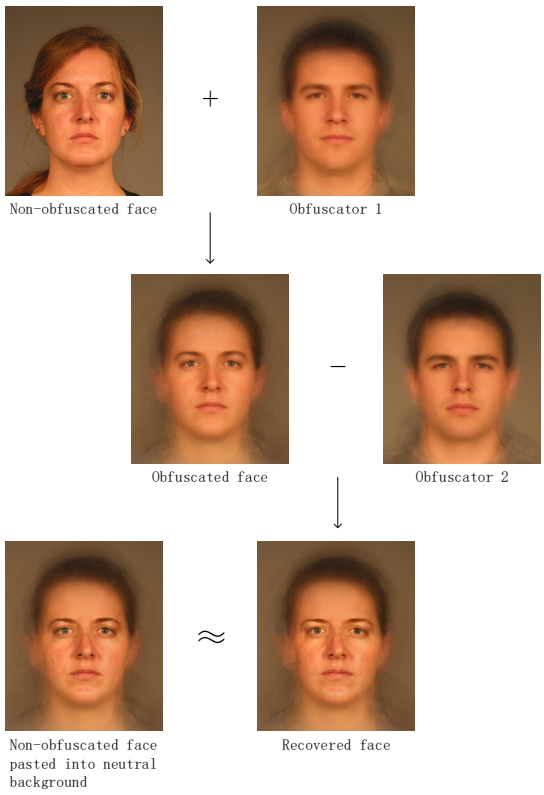


Figure 8. An example of gender information recovery by face demorphing

3.4.1. Degree of gender obfuscation. For a confused gender classifier, its corresponding ROC curve should lie near to the diagonal and the Area Under the Curve (AUC), which measures discrimination, is approximately equal to 0.5, indicating the gender classifier loses its ability to distinguish between male and female faces. Therefore, the degree of gender obfuscation is computed as:

$$O_{\text{gender}} = \int_0^1 |R(t) - t| dt \tag{5}$$

where $R(t)$ is the value of the ROC curve in False positive rate t, and O_{gender} indicates the area between the ROC curve and the diagonal. The lower the O_{gender} is, the better the gender classifier is confused.

The gender score distributions are analyzed as well. It is considered that if a gender classifier cannot distinguish between male and female faces, then the distributions of males' and females' gender scores will overlap with each other. To observe intuitively the distributions of gender scores, the Empirical Cumulative Distribution Function (ECDF) is needed. Given a set of *N* ordered scores $(S_1, S_2, S_3, ..., S_N)$, ECDF is computed using Equation 6, where *n(i)* is the number of scores that are smaller than t and $t \in \{S_1, S_2, S_3, ..., S_N\}$. Therefore, the computed ECDF is actually a step function.

$$E_N(t) = \frac{n(i)}{N} \tag{6}$$

Two gender classifiers are utilized and their performance are compared. One is from *FaceVacs SDK* (https://www.cognitec.com/) and another is provided by [13] with ShuffleNet architecture.

3.4.2. Degree of identity preservation. The identity information should be preserved mostly after obfuscating gender. Two face comparison modules are used. One is from *FaceVacs SDK*, and another is from python library *face_recognition* (https://github.com/ageitgey/face_recognition). Note that the python *face_recognition* module performs well in White faces, but its performance in other ethnicity groups cannot be guaranteed [2]. Meanwhile, *FaceVacs* face comparison module outputs a face similarity score while python *face_recognition* module outputs a face distance score.

Unlike confusing a gender classifier, the ROC curve for the face comparison modules are supposed to be as ideal as possible, which indicates the corresponding AUC should be closed to 1. We define the degree of identity preservation as P_{identity}, which is:

$$P_{\text{identity}} = AUC \cdot 100\% \tag{7}$$

4. Experiments and Results

The purpose of the following experiments is to investigate the utility of face morphing in gender obfuscation systematically from two aspects: (1) influence of morphing parameters in gender obfuscation, (2) influence of obfuscator in gender obfuscation, and to explore the reversibility of our gender obfuscation scheme by two possible methods: (1) face demorphing, (2) retraining a gender classifier.

4.1. Databases

Two databases are used for different tasks. One task is to evaluate the performance of gender obfuscation through face morphing and another is to investigate the potential to train a gender classifier that can distinguish between obfuscated male and female faces. These two databases are: FRGC [28] and CMU Multi-PIE [14]. FRGC [28] is used for gender obfuscation while CMU Multi-PIE [14] is used for training a gender classifier.

FRGC. This database consists of around 50,000 recordings captured from different sessions. For a subject in one session, 4 controlled still images, two uncontrolled still images and one 3D image are captured. *Controlled* means that the illumination condition is controlled (studio setting). Additionally, every subject has two facial expressions (smiling and neutral). To create high-quality obfuscated faces, only the controlled still images with neutral facial expression are used. The data distribution is shown in Table 1. Most of the subjects are White, and due to the demand of generating numerous morphed faces and the bad performance of python *face_recognition* module in faces of other races, we use White faces only in later experiments.

TABLE 1. DATA DISTRIBUTION(NEUTRAL FACIAL EXPRESSION)

Session	FN Fall 03	MN Fall 03	FN Spring 04	MN Spring 04
Recordings/Subjects	1847/359	1847/359	2063/333	2063/333
Recordings/Subjects(White: Females)	509/98	509/98	617/103	617/103
Recordings/Subjects(White: Males)	678/139	678/139	750/138	750/138
Recordings/Subjects(Asian: Females)	260/47	260/47	336/40	336/40
Recordings/Subjects(Asian: Males)	259/45	259/45	266/34	266/34
Recordings/Subjects(Other: Females)	43/9	43/9	27/5	27/5
Recordings/Subjects(Other: Males)	98/21	98/21	67/13	67/13

TABLE 2. INDIVIDUAL SESSION ATTENDANCE (REMOVING SUBJECTS WEARING GLASSES)

Session	1	2	3	4
White (Females/Males)	26/71	24/55	28/62	28/67
Asian (Females/Males)	17/16	12/9	15/10	16/9
Others (Females/Males)	7/16	5/15	7/15	7/14

CMU Multi-PIE. 755,370 images from 337 different subjects are contained in this database, and they are recorded in 4 sessions. Individual session attendance is shown in Table 2. 129 of all subjects appeared in all four sessions. Among those subjects, 69.7% are males while the rest are females. The subjects are predominantly European-Americans (60%) and Asian (35%), while the rest of them are African-American (3%) and from other ethnicity groups (2%). To be noticed, some of the subjects wear glasses, which affects the resulted morphed face a lot. Therefore, face images which contain glasses are removed manually. Furthermore, we only use face images from European-Americans not only because they are the vast majority in this database, but also gender obfuscation is only applied in White faces. Therefore, when training a gender classifier, influence of other factors such as ethnicity should be avoided. There are two kinds of images captured from the subjects:

1) Multi-view;
2) High-resolution.

The high-resolution face images are utilized since they are captured frontally, and have higher quality than those captured from multi-view. After manually cleaning the data, the number of utilizable face images (may come from the same subject) decreases to:

- 326 White male faces
- 132 White female faces

There are more male faces than female faces, thus, when generating data for training the gender classifier, the dataset needs to be balanced. Detailed process is described in Section 4.5.1.

4.2. Select appropriate morphing parameters

As shown in [11], α_b influences the performance of face recognition more than α_w does. But the influence of these parameters on gender classification is unknown. Therefore, faces are morphed with different combinations of α_w and α_b. The gender classifiers and face verification modules introduced in Section 3.4 are used to evaluate the degree of gender obfuscation and identity preservation.

In the experiment, a face is morphed with an obfuscator. When $\alpha_w = \alpha_b = 1$, face morphing is not applied to the subject face. When $\alpha_w = \alpha_b = 0$, the face is completely the obfuscator. Examples of obfuscated faces are shown in Figure 9. To better compare the difference among faces, the non-obfuscated faces are pasted into the neutral background as well.

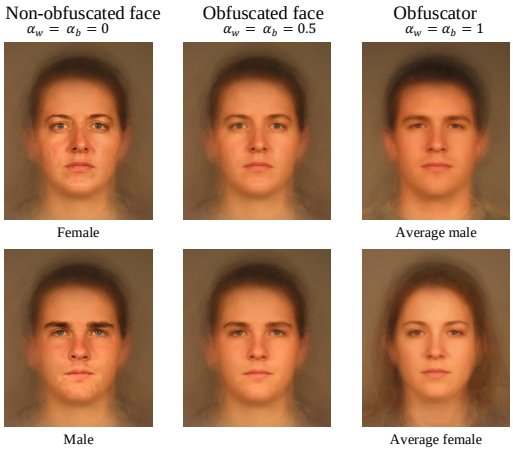
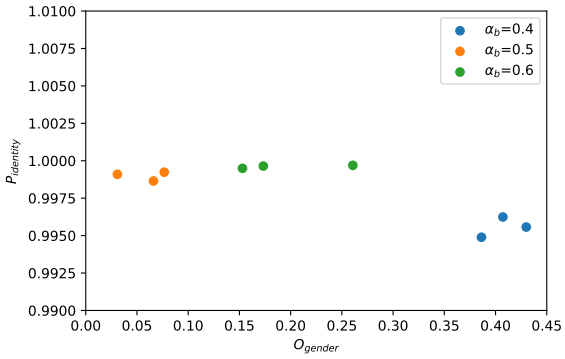


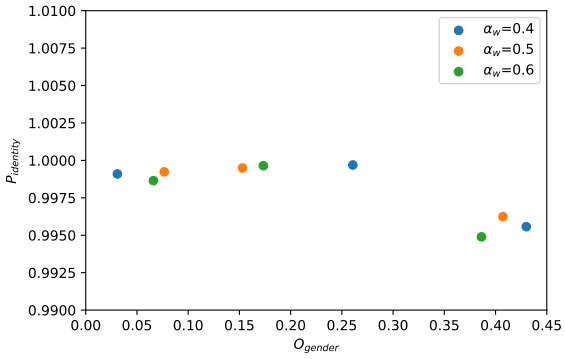
Figure 9. Examples of obfuscated face (middle column)

Figure 10 shows the trade-off between P_{identity} and O_{gender} under different combinations of α_w and α_b, which is evaluated by *FaceVacs* face comparison module and gender classifier. Figure 10a and 10b are plotted from same scores while the legends are labeled differently. It can be observed that α_w does not affect identity verification significantly, while P_{identity} drops with the decrease of α_b. However, it can also be observed that *FaceVacs* face comparison module is vulnerable to morphing attacks since the P_{identity} only decreases by at most 0.48%. The degree of identity preservation is shown in Table 3. For the non-obfuscated faces, P_{identity} equals to 100% for *FaceVacs* face comparison module and 99.90% for python *face_recognition* module.

In terms of gender obfuscation, the degree of obfuscation with different combinations of α_w and α_b is shown in Table 4. It shows that α_w has a slight influence on gender obfuscation while α_b matters more. However, the two gender classifiers are confused differently. For the *FaceVacs*



(a)



(b)

Figure 10. Scatter plots of P_{identity} vs. O_{gender} under different combinations of α_w and α_b (a and b are plotted from the same gender and similarity scores evaluated by *FaceVacs SDK* while labeled differently)

TABLE 3. DEGREE OF IDENTITY PRESERVATION UNDER DIFFERENT COMBINATIONS OF α_w AND α_b

Face comparison module		P_{identity}		
		α_w = 0.4	α_w = 0.5	α_w = 0.6
FaceVacs	$\alpha_b = 0.4$	99.56%	99.62%	99.49%
	$\alpha_b = 0.5$	99.91%	99.92%	99.86%
	$\alpha_b = 0.6$	99.97%	99.95%	99.96%
face_recognition	$\alpha_b = 0.4$	85.77%	89.67%	90.78%
	$\alpha_b = 0.5$	96.48%	96.48%	94.87%
	$\alpha_b = 0.6$	98.39%	98.24%	98.14%

TABLE 4. DEGREE OF GENDER OBFUSCATION UNDER DIFFERENT COMBINATIONS OF α_w AND α_b

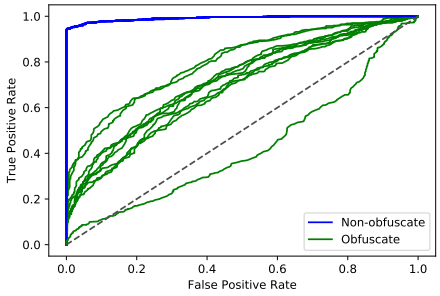
Gender classifier		O_{gender}		
		α_w = 0.4	α_w = 0.5	α_w = 0.6
FaceVacs	$\alpha_b = 0.4$	0.429	0.407	0.386
	$\alpha_b = 0.5$	0.031	0.077	0.066
	$\alpha_b = 0.6$	0.261	0.153	0.173
ShuffleNet	$\alpha_b = 0.4$	0.397	0.396	0.347
	$\alpha_b = 0.5$	0.193	0.179	0.157
	$\alpha_b = 0.6$	0.154	0.165	0.106

gender classifier, α_w = 0.4 and α_b = 0.5 leads to the best obfuscation, while for the ShuffleNet gender classifier, α_w = α_b = 0.6 leads to the best obfuscation.

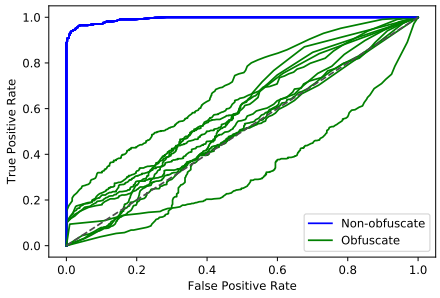
Combining the results in Table 3 and Table 4, the morphing parameters are set to: $\alpha_w = \alpha_b = 0.5$. Another reason to set $\alpha_w = 0.5$ is that some of the faces are not fully frontal, and setting $\alpha_w = 0.5$ can reduce pose variations while not influencing on identity verification.

4.3. Generating average faces

In experiment 4.2, the obfuscators are created from 20 randomly chosen subjects (20 males or 20 females). As shown in Figure 11, the ROC curves for 10-test (ten repetitions of the same experiment) gender obfuscation are plotted. For each test, the obfuscators are different while the selected subjects for gender obfuscation are the same. It can be observed that the characteristics of average face is closely related to the degree of gender obfuscation. In this Section, we investigate the influence of average faces on gender obfuscation.



(a) *FaceVacs* gender classifier

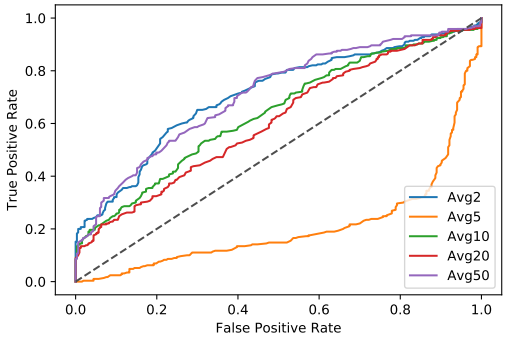


(b) ShuffleNet gender classifier

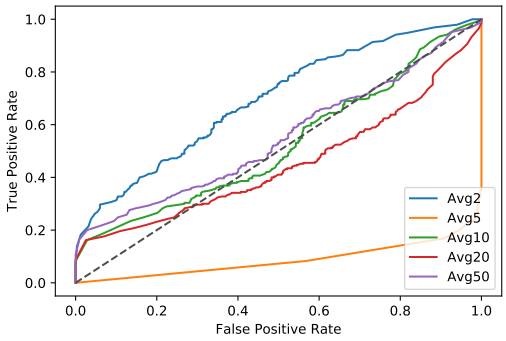
Figure 11. ROC curves for 10-test gender obfuscation (a,b: test on FRGC database, obfuscators generated from FRGC database)

4.3.1. Number of faces to create obfuscators. The more faces used to create the average face, the more blurry the average face is. The number of faces used to create obfuscators needs to be set appropriately. If too many faces are used, facial landmarks cannot be located on the blurry face accurately and the computation time to generate obfuscators increases; If too few faces are used, the resulted face cannot be considered as a prototype, i.e it's not "average enough". Therefore, the performance of gender obfuscation by obfuscators generated from different numbers of faces is

compared. Session FN Fall 03 and MN Spring 04 in FRGC database are used as Average-face gallery while MN Fall 03 is used as Reference gallery and FN Spring 04 is used as Probe gallery. As shown in Figure 12, the number of faces used to create obfuscators has a slight influence on gender obfuscation, except for the case using 5 faces to generate the obfuscator. In order to avoid too much randomness (averaging too few faces) and on the other hand unnecessary computation, the number of faces is set to 20, which enables more combinations of different faces at the same time.



(a) *FaceVacs* gender classifier + *FaceVacs* face comparison module (Corresponding values of P_{identity} are: 99.94%, 99.99%, 99.94%, 99.97%, 99.98%)



(b) ShuffleNet gender classifier + python *face_recognition* (Corresponding values of P_{identity} are: 98.84%, 98.42%, 99.45%, 99.37%, 99.57%)

Figure 12. ROC curves of gender classifiers (obfuscators are generated from different numbers of faces)

4.3.2. Gender characteristics of obfuscators. Instead of randomly choosing 20 subjects, we further investigate gender characteristics of the obfuscators and influence of the similarity between the Non-obfuscated face and the obfuscator in gender obfuscation.

For the first case, another gallery including 50 faces with high gender scores (50-face male/female gallery) is created. The gender scores of faces are computed by ShuffleNet gender classifier and galleries for males and females are separated. Then 20 faces are randomly chosen from the 50-face male/female gallery.

TABLE 5. INFLUENCE OF OBFUSCATORS WITH DIFFERENT CHARACTERISTICS IN O_{gender} AND P_{identity}

Obfuscator's characteristics	P_{identity}		O_{gender}	
	FaceVacs	*face_recognition*	*FaceVacs*	ShuffleNet
Similar	99.99%	99.71%	0.239	0.145
Strong	100.00%	99.60%	0.147	0.072
Trade-off	99.99%	99.64%	0.174	0.037

For the second case, the similarity scores between the Non-obfuscated face and faces with opposite gender in Average-face gallery are calculated first, and then the 20 most similar faces are chosen to generate the obfuscator. In this case, the obfuscator is more similar to the Non-obfuscated face while its gender characteristics might not be strong.

It shows in Table 5 that obfuscators with strong gender characteristics have a slightly better performance in gender obfuscation. When choosing only similar faces to create obfuscators, the identity information is preserved a bit better.

Trade-off between identity preservation and gender obfuscation is considered and thus we choose faces with high similarity scores from the 50-face gallery instead of Average-face gallery. From Table 5, it shows that it is possible to preserve slightly more identity information without affecting gender obfuscation much.

However, the gallery for generating obfuscators is not large enough, and using the above methods leads to invariant obfuscators for individuals. In a real scenario, we may not know the faces used to create the obfuscators. Therefore, to guarantee the randomness, in the following experiments, we use obfuscators that are created from 20 faces chosen randomly from the Average-face gallery.

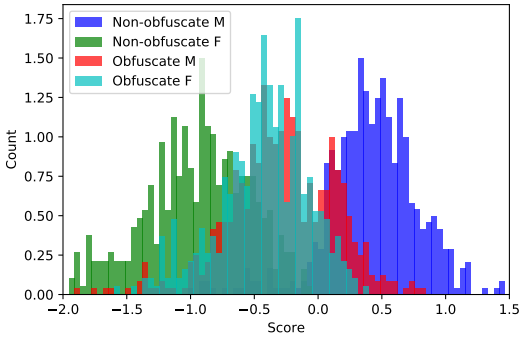
4.4. Obfuscate gender in face images

In previous experiments, when $\alpha_w = \alpha_b = 0.5$, and obfuscators are generated from 20 randomly chosen faces, gender information can be well obfuscated and identity information is mostly preserved. Due to the randomness of obfuscators, the experiment for gender obfuscation is repeated 10 times, each time with a newly created obfuscator. It can be observed from Figure 11a and 11b that ROC curves lie near to the diagonal but there are some deviations based on obfuscators we used. Selecting gender scores for one of the tests, the score distributions for male and female faces are shown in Figure 13a and 14a. Before face morphing, the gender scores for males and females are distinguishable. Because it is difficult to determine whether the gender scores evaluated by ShuffleNet overlap, ECDFs are used instead, as shown in Figure 15b. After applying face morphing, the gender score distributions for males and females overlap. It can be considered as confusing the gender classifiers. O_{gender} of the 10-test gender obfuscation is measured by Equation 5, which ranges from 0.077 to 0.298 (*FaceVacs*), and from 0.009 to 0.170 (*ShuffleNet*). The overlapped area of distributions of gender scores computed by *FaceVacs* ranges

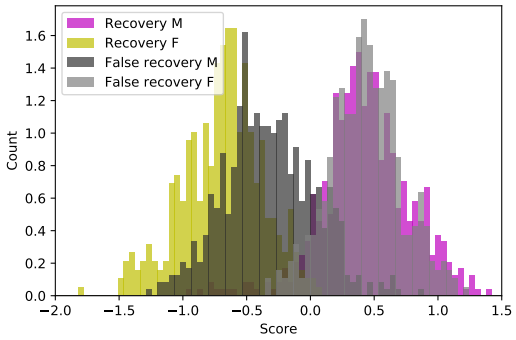
from 54.66% to 75.94%, while by *ShuffleNet*, it ranges from 61.94% to 81.61%.

4.5. Retrieve gender in obfuscated face

It is important that the hidden gender information cannot be recovered or retrieved easily. One possible way to retrieve gender information of the obfuscated faces is face demorphing, which is the reverse process of face morphing. We assume that the obfuscators used for obfuscation is not publicly available (nor the database used). Therefore, another database to create new obfuscators is needed. Suppose that we have an obfuscated face, and a new gallery of faces which is for creating new obfuscators, the obfuscated face is demorphed with the new obfuscator. Therefore, CMU Multi-PIE database is utilized in order to avoid reusing same subjects when generating obfuscators. When trying to retrieve gender information, it is no longer necessary to focus on identity preservation. Another way is to train a classifier that can classify gender in the obfuscated faces.



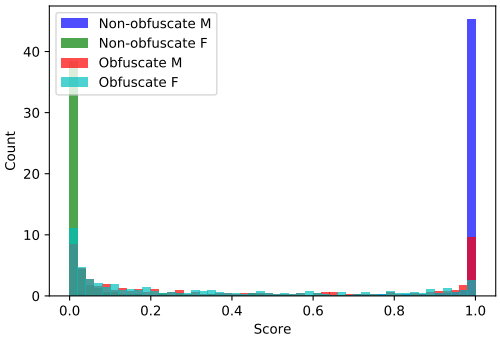
(a)



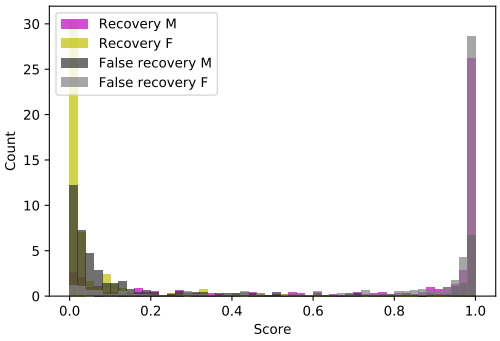
(b)

Figure 13. Comparison of gender score distributions (a: between non-obfuscated and obfuscated faces, b: between recovered and false recovered faces; gender scores are evaluated by *FaceVacs*)

Retrieve by face demorphing. We consider a scenario in which someone has been presented with an obfuscated face. However, it is unknown whether the original gender



(a)



(b)

Figure 14. Comparison of gender score distributions (a: between non-obfuscated and obfuscated faces, b: between recovered and false recovered faces; gender scores are evaluated by *ShuffleNet*)

corresponding to this face was male or female. Therefore it could have been morphed with a male obfuscator (if the original face was female), or with a female obfuscator (if the original face was male). Thus, the obfuscated face should be demorphed with both female obfuscator and male obfuscator.

In the experiments, we have the pre-known gender information of the obfuscated faces, and it enables to analyze whether the gender information can be retrieved in this way. Before the experiments, some essential terms are defined as follows:

- Recovered face: an obfuscated face is demorphed with a correct obfuscator;
- False recovered face: an obfuscated face is demorphed with an incorrect obfuscator.

Because the morphing parameters are set to $\alpha_w = \alpha_b = 0.5$, the demorphing parameters are also set to be the same. The gender score distributions before and after demorphing are shown in Figure 13 and 14. To better visualize the overlapped distributions, the corresponding ECDFs of gender scores are plotted in Figure 15.

It can be observed from Figure 15a and 15b that:

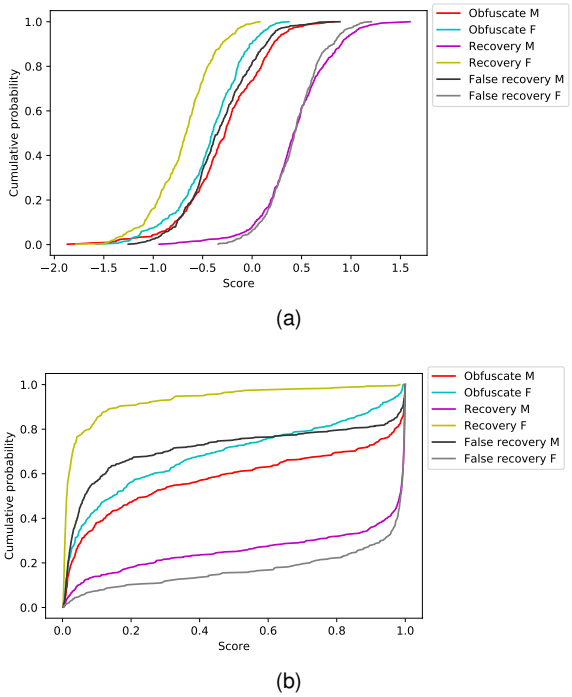


Figure 15. Comparison of ECDFs (a: *FaceVacs* gender classifier, b: ShuffleNet gender classifier)

- When demorphing with a female obfuscator,the gender score distributions of both obfuscated female faces and obfuscated male faces overlap.
- When demorphing with a male obfuscator, the gender score distributions of both obfuscated female faces and obfuscated male faces overlap.

Therefore, face demorphing is not a feasible method to retrieve gender in obfuscated faces since the gender score distributions are still overlapped after demorphing.

4.5.1. Retrain gender classifier. It has been shown that face demorphing is not a practical technique in retrieving gender information of an obfuscated face. Therefore, we would like to investigate whether it is possible to train a gender classifier that can detect gender in an obfuscated face.

First, we prepare a dataset that contains obfuscated faces only. The faces are generated from 71 White males and 26 White females in CMU Multi-PIE database session 1. α_w and α_b are set to be the same and range from 0.4 to 0.6 with step = $\frac{0.6-0.4}{61}$. In total, 5917 obfuscated faces are generated. Due to the imbalanced data distribution, undersampling is utilized. Face images of males are removed randomly. Later, Gaussian noise is added to each face with mean = 0 and variance = 0.02 for each color channel to avoid overfitting. Finally, there are 771 obfuscated male faces and 771 obfuscated female faces. To be noted, the obfuscator for each obfuscated face is generated singly from 20 randomly chosen subjects from CMU Multi-PIE database session 2, i.e. the obfuscator for every face is newly created.

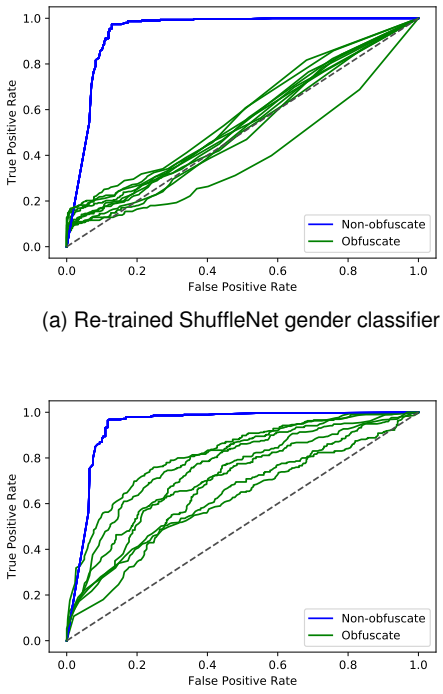


Figure 16. ROC curves for 10-test gender obfuscation (a: test on FRGC database, obfuscators generated from FRGC database, b: test on FRGC database, obfuscators generated from CMU Multi-PIE session 2)

80% of the generated dataset is used for training, and the rest is for validation. After training, the gender classifier obtains an accuracy of 98.95% in training dataset and 89.97% in validation dataset. However, when applying the model to FRGC database (Non-obfuscated faces and obfuscators are from FRGC), the performance is not consistent with the results obtained during training (as shown in Figure 16a, the gender classifier is still confused). Hence, instead of using obfuscators created from FRGC database, we use obfuscators generated from CMU Multi-PIE session 2, to see whether it is the obfuscator that influences the performance of the re-trained classifier. The results are shown in Figure 16b. The performance of gender classification improves a bit when the obfuscated faces contain obfuscators used during training. But the gender classifier is still confused. Thus, we conclude that it is complicated to train a gender classifier which can distinguish between obfuscated male and female faces.

5. Discussion

Our results have shown that face morphing is a useful approach to obfuscate gender information in face images. Compared with other related methods such as GAN [16][36], SAN [24], it is a much simpler method that requires less computations. It has also been proved that recovering or retrieving the gender information from obfuscated face images is a tricky task, which demonstrates that

the gender information is well secured. The shortcoming of the method is that the obfuscated faces carry some artefacts caused by hair and facial hair. Additionally, another facial attribute "age" is severely influenced. We conclude from visual inspection that most of the subjects used for creating the obfuscators are below 40 years old. Some examples of obfuscated faces are shown in Appendix A.

Identity preservation differs based on different face comparison module. A face verification system which is vulnerable to morphing attack is able to utilize face morphing to obfuscate gender information since its original performance is well maintained. In addition, the experiments are only done in the White faces only, thus, the performance of our gender obfuscation scheme for other ethnicity groups is unknown and further research is needed.

In the process of preparing dataset for training a gender classifier, face images are generated from a few subjects (1 subject with 61 obfuscated faces). It is possible that the gender classifier becomes overfitted due to too few subjects.

Morphing with an obfuscator can be considered as adding opposite gender characteristic to a face. But when a non-obfuscated face is directly demorphed with an average face with same gender, its original gender characteristics are not removed (examples are shown in Appendix B), and it even results in faces with severe artefacts such as big eyes, tilted nose. Therefore, a face directly demorphed with an average face which has the same gender cannot confuse gender classifiers. However, if applying face demorphing on the obfuscated faces (in the case of False recovery), the gender information is further suppressed and the texture on the resulted face makes the face look realistic (examples are shown in Appendix C).

It is suggested to create a benchmark dataset containing obfuscated faces generated from more subjects and further investigate on training gender classifiers for obfuscated faces specifically. Further experiments can explore the functionality of face morphing in multiple facial attributes obfuscation and it is assumed that there exists a trade-off between number of facial attributes and identity preservation.

6. Conclusion

In this paper, we explore the potential to obfuscate gender information in face images by face morphing and investigate the possibility to retrieve gender information in such obfuscated faces. The obfuscated faces are generated by morphing a face with an average face that has the opposite gender, which is called obfuscator. Experimental results indicate that face morphing is a feasible method to confuse gender classifiers while identity information is well preserved, especially when the identity verification system is sensitive to morphing attacks. It is found that morphing parameters have different influence on gender classification. When α_b = 0.5, the degree of gender obfuscation is satisfactory, while α_w has a slight influence on gender classification. The characteristics of the obfuscator influence the degree of gender obfuscation as well. To retrieve gender information hidden by face morphing, two methods are utilized. One is to retrieve gender information by applying face demorphing, the other is to retrieve gender information by training a gender classifier that can classify gender in obfuscated faces. Our results show that gender information in obfuscated faces is difficult to recover or retrieve with either of the two methods. Although face demorphing is exactly the reverse process of face morphing, it is not an effective approach, because it is too vulnerable to the choice of average faces. Directly training a gender classifier to distinguish between obfuscated male and female faces is a potential method, while experiments demonstrate the difficulty in training such a model.

Acknowledgments

I would like to express my deepest appreciation to my committee. Thank you to my supervisor, Raymond Veldhuis, for providing me guidance and feedback throughout this project. Thanks to Una Kelly who was always willing to assist me in any way. Thanks also to Jasper Goseling, who gave some valuable advice. Finally, thanks to Antonio Greco, for his contribution in gender classification experiments.

References

[1] "General data protection regulation," *Official Journal of the European Union*, vol. 59, p. 35–36, May 2016.

[2] Ageitgey, "ageitgey/face_recognition." [Online]. Available: https://github.com/ageitgey/face_recognition/wiki/Face-Recognition-Accuracy-Problems

[3] N. Arad, N. Dyn, D. Reisfeld, and Y. Yeshurun, "Image warping by radial basis functions: Application to facial expressions," *CVGIP: Graphical Models and Image Processing*, vol. 56, no. 2, p. 161–172, 1994.

[4] T. Beier and S. Neely, "Feature-based image metamorphosis," *Proceedings of the 19th annual conference on Computer graphics and interactive techniques - SIGGRAPH '92*, 1992.

[5] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping," *ACM SIGGRAPH 2008 papers on - SIGGRAPH '08*, 2008.

[6] M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, 2000.

[7] B. Bui Huu Trung, T. Bui, M. Poel, D. Heylen, and A. Nijholt, "Automatic face morphing for transferring facial animation," in *Proceedings of the 6th IASTED International Conference on Computers, Graphics, and Imaging (CGIM 2003)*, H. Hamza, Ed. Canada: ACTA Press, 8 2003, pp. 19–24.

[8] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing k facial attributes via adversarial perturbations," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

[9] A. Dantcheva, C. Velardo, A. D'Angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identifi-

cation," *Multimedia Tools and Applications*, vol. 51, no. 2, p. 739–777, 2010.

[10] L. Du, M. Yi, E. Blasch, and H. Ling, "Garp-face: Balancing privacy protection and utility preservation in face de-identification," *IEEE International Joint Conference on Biometrics*, 2014.

[11] M. Ferrara, A. Franco, and D. Maltoni, "Decoupling texture blending and shape warping in face morphing," in *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2019, pp. 1–5.

[12] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, p. 1008–1017, 2018.

[13] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, pp. 1–1, 2020.

[14] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *2008 8th IEEE International Conference on Automatic Face and amp; Gesture Recognition*, 2008.

[15] Y. Hu, M. Zhou, and Z. Wu, "A dense point-to-point alignment method for realistic 3d face morphing and animation," *International Journal of Computer Games Technology*, vol. 2009, p. 1–9, 2009.

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] P. Korshunov and T. Ebrahimi, "Using face morphing to protect privacy," *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013.

[18] S. Lee, G. Wolberg, K.-Y. Chwa, and S. Y. Shin, "Image metamorphosis with scattered feature constraints," *IEEE Transactions on Visualization and Computer Graphics*, vol. 2, no. 4, p. 337–354, 1996.

[19] S. Lee, G. Wolberg, and S. Y. Shin, "Polymorph: morphing among multiple images," *IEEE Computer Graphics and Applications*, vol. 18, no. 1, p. 58–71, 1998.

[20] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient cnn architecture design," *Computer Vision – ECCV 2018 Lecture Notes in Computer Science*, p. 122–138, 2018.

[21] A. Makrushin, T. Neubert, and J. Dittmann, "Automatic generation and detection of visually faultless facial morphs," *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2017.

[22] V. Mirjalili, S. Raschka, and A. Ross, "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, vol. 7, pp. 99 735–99 745, 2019.

[23] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial Networks: Convolutional autoencoders for imparting privacy to face images," *2018 International Conference on Biometrics (ICB)*, 2018.

[24] V. Mirjalili, S. Raschka, and A. Ross, "Gender privacy: An ensemble of Semi Adversarial Networks for confounding arbitrary gender classifiers," *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2018.

[25] S. Mosaddegh, L. Simon, and F. Jurie, "Photorealistic face de-identification by aggregating donors' face components," *Computer Vision – ACCV 2014 Lecture Notes in Computer Science*, p. 159–174, 2015.

[26] C. Neustaedter, S. Greenberg, and M. Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 1, p. 1–36, 2006.

[27] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," *Computer Vision - ECCV 2014 Workshops Lecture Notes in Computer Science*, p. 682–696, 2015.

[28] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*.

[29] F. P. Preparata and M. I. Shamos, *Proximity: Fundamental Algorithms*. Springer-Verlag, 1985, p. 185–223.

[30] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," *ACM SIGGRAPH 2003 Papers on - SIGGRAPH '03*, 2003.

[31] H. T. F. Rhodes, *Alphonse Bertillon, father of scientific detection*. Greenwood Press, 1968.

[32] D. J. Robertson, R. S. S. Kramer, and A. M. Burton, "Fraudulent id using face morphs: Experiments on human and automatic recognition," *Plos One*, vol. 12, no. 3, 2017.

[33] D. Rowland and D. Perrett, "Manipulating facial appearance through shape and color," *IEEE Computer Graphics and Applications*, vol. 15, no. 5, p. 70–76, 1995.

[34] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, p. 23012–23026, 2019.

[35] 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*, vol. 41, no. 2, p. 226–237, 2011.

[36] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Can GAN generated morphs threaten face recognition systems equally as landmark based morphs? - vulnerability and detection," *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 2020.

[37] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-Protective-GAN for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, no. 1, p. 47–60, 2019.

[38] Q. A. Zhao and J. T. Stasko, "Evaluating image filtering based techniques in media space applications," *Proceedings of the 1998 ACM conference on Computer supported cooperative work - CSCW '98*, 1998.

Appendix A.

Some examples of obfuscated faces compared with non-obfuscated faces are shown in figure 17. For a better comparison, the non-obfuscated faces are pasted into the neutral background as well. There are 5 males and 5 females in total. The quality of the resulted images is closely related to the accuracy of facial landmarks detection, hair in the forehead. As shown in the second obfuscated male face, the detected face region is much smaller than the face region in the neutral background. And for the last female face, the hair covered in the face region affects the resulted obfuscated face.

Appendix B.

In Figure B, there are some examples showing how face demorphing behaves in non-obfuscated faces. It can be observed that the resulted demorphed faces (second and third columns) usually have rougher skin and slightly bigger eyes. And the asymmetric facial structure is highlighted.

Appendix C.

As shown in Figure C, when comparing the non-obfuscated faces with recovered face (the first column and the third column), the gender information can be recovered if such information is pre-known. Meanwhile, comparing the obfuscated faces with false recovered faces, it is observed that demorphing further enhances the gender characteristics of the obfuscated faces, which can be an approach to flip the gender information.

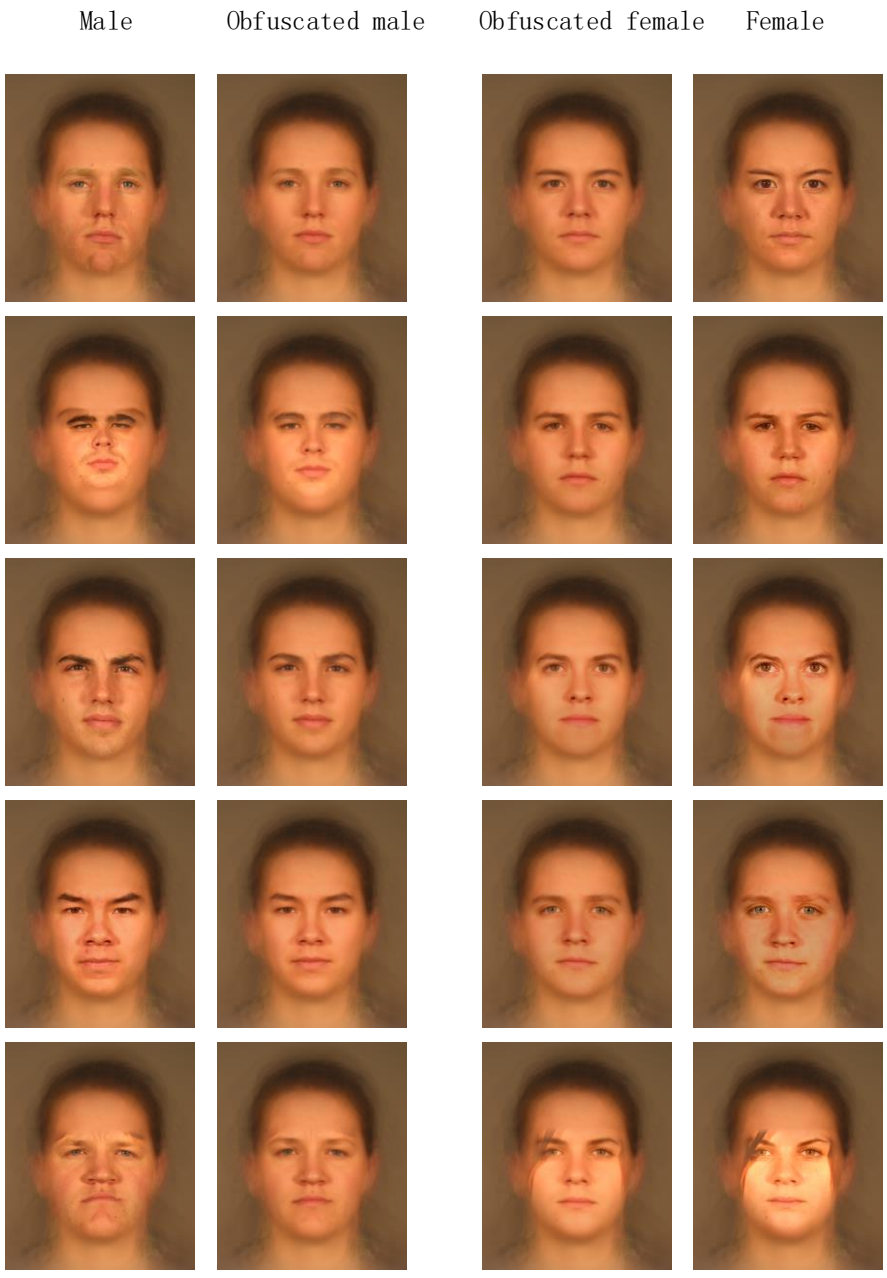


Figure 17. Obfuscated faces compared with non-obfuscated faces

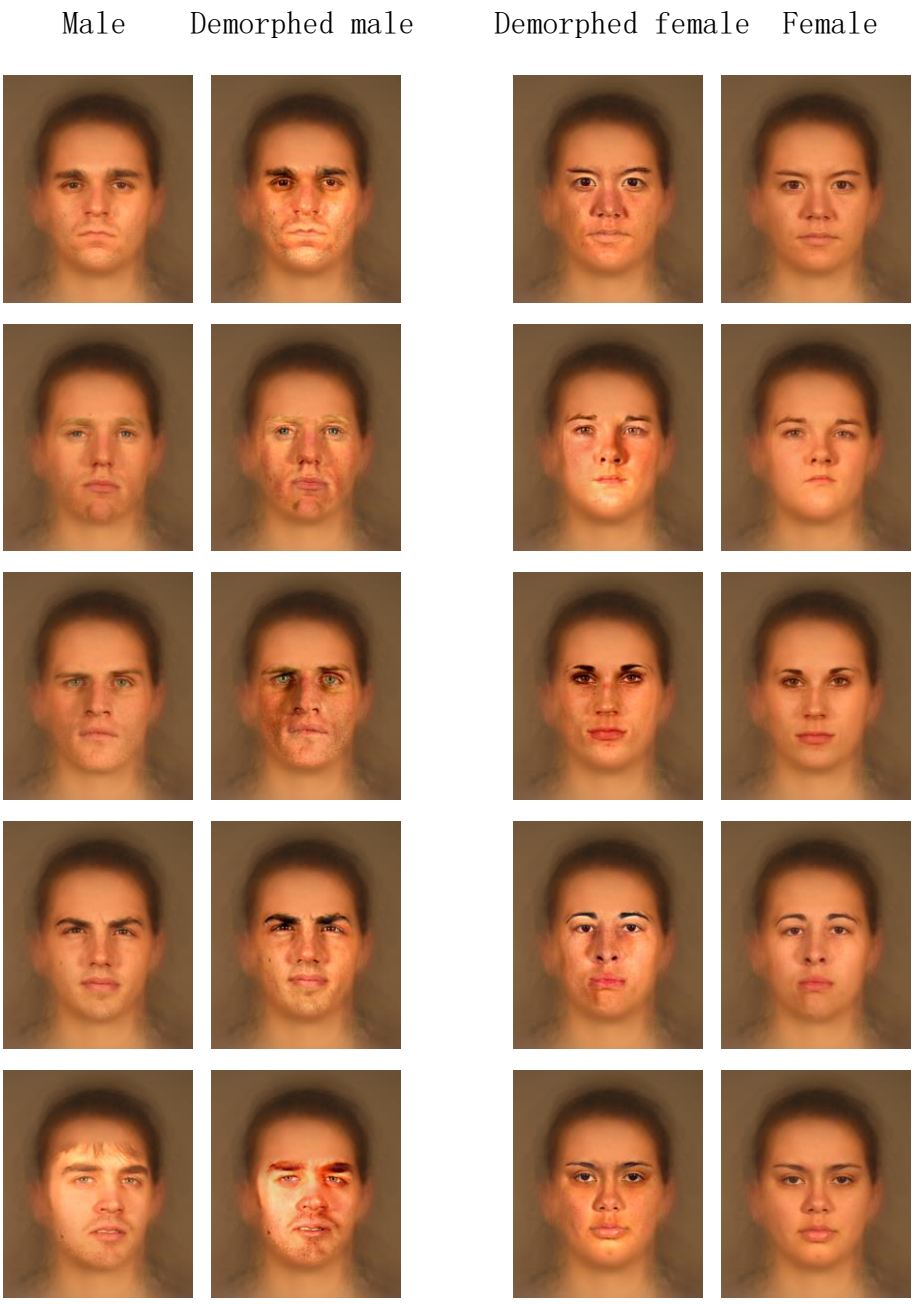


Figure 18. Examples of demorphed non-obfuscated faces

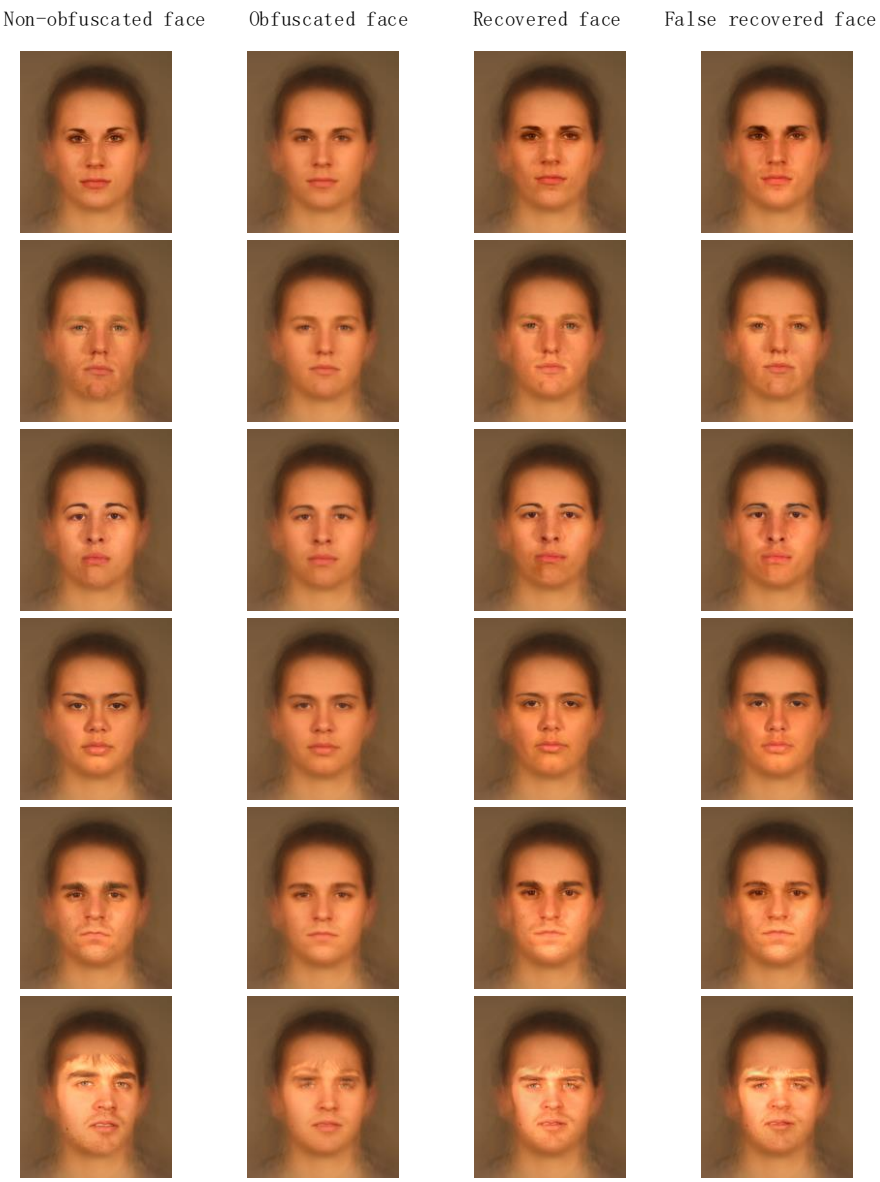


Figure 19. Examples of non-obfuscated face (first column), obfuscated faces (second column), Recovered faces (third column), False recovered faces (fourth column)