Attribute Preserved Face De-identification

Amin Jourabloo*, Xi Yin*, and Xiaoming Liu
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{jourablo, yinxi1, liuxm}@msu.edu

Abstract

In this paper, we recognize the need of de-identifying a face image while preserving a large set of facial attributes, which has not been explicitly studied before. We verify the underling assumption that different visual features are used for identification and attribute classification. As a result, the proposed approach jointly models face de-identification and attribute preservation in a unified optimization framework. Specifically, a face image is represented by the shape and appearance parameters of AAM. Motivated by k-Same, we select k images that share the most similar attributes with those of a test image. Instead of using the average of k images, adopted by k-Same methods, we formulate an objective function and use gradient descent to learn the optimal weights for fusing k images. Experimental results show that our proposed approach performs substantially better than the baseline method with a lower face recognition rate, while preserving more facial attributes.

1. Introduction

The human is the dominant object of interest in the huge network of surveillance cameras deployed in buildings, airports, streets, and so on. The captured images and videos enable various applications such as face recognition, person re-identification, crowd and behavior analysis, etc. Certain applications, such as face recognition, rely on the specific facial characteristics associated with identifiable information - typical for biometrics problems. On the other hand, for a wide range of other applications, it is not necessary to access identifiable information on faces, yet the recording of faces arouses potential privacy concern from the general public [27]. To address this concern, one possible solution is *face de-identification* [19], which aims to eliminate the identifiable information by modifying a face image while preserving data utility – a *counter* problem to biometrics.

In addition to the identifiable information, face also contains other information associated with various facial attributes [13], such as gender, age, race, glasses, expression, etc. For many real-world applications, these facial attributes carry useful information and are desired to be extracted. For example, age and race attributes can help analyze the demography of customers in a retail store. Therefore, it is important to *preserve* these facial attributes while performing the face de-identification. In other words, ideally face de-identification should operate *only* on the identifiable information while keeping a wide variety of attributes *intact*.

Figure 1. The proposed APFD method can change the identity of a face image while preserving facial attributes.

However, most state-of-the-art de-identification approaches are designed to thwart face recognition methods without explicitly modeling the preservation of facial attributes. Previous k-Same methods [19, 9] transform a face image in an one-shot operation to make it unrecognizable, and then test its utility such as expression classification. In these approaches, attribute preservation is a consequence after face de-identification. To the best of our knowledge, there is no prior work that *jointly* models de-identifying face images and preserves a *large set* of facial attributes.

In recognizing this problem, we propose an *Attribute-Preserved Face De-identification* (APFD) approach, which jointly models these two parts in a unified optimization framework. As shown in Fig. 1, given a test image, our method can change the identity of the subject while preserving as many attributes as possible. Specifically, this paper studies 17 attributes including gender, race, age, expression, glasses, image quality, etc. In contrast to the k-Same methods where the average of k images is treated as the de-identified image, we formulate an objective function to estimate the optimal weights to fuse the k images.

The objective function includes two terms defined using a set of pre-trained face attribute classifiers and a face verifi-

*denotes equal contribution by the authors.

cation classifier. The first term is the score of a face verification classifier between a test image and its de-identified image, where a lower score indicates *different* identities. The second term is the total difference of scores from the facial attribute classifiers of the test image and those of its de-identified image, where a smaller difference indicates the preservation of more facial attributes. During the gradient descent-based optimization, the supervised learned face verification classifier and attribute classifiers serve as the oracle to guide the estimation of the weighting parameters, i.e., the minimization leads to both the de-identification and attribute preservation. Experiments on a set of 200 images demonstrate the effectiveness of our proposed approach in de-identifying face images while preserving attributes.

In summary, this paper makes two key contributions:

$\diamond$ We identify the problem of face de-identification while preserving attributes, which emphasizes the importance of attribute preservation that lacks attention in prior work.

$\diamond$ We formulate an optimization problem to explicitly model face de-identification and attribute preservation. By optimizing the joint objective function, the de-identified face images will preserve as many attributes as possible.

2. Related Work

Face de-identification related topics have been addressed in data mining, graphics, and computer vision. For example, [14] studies the tradeoff between privacy protection and data utility, where some ideas are applied to face de-identification. Bitouk et al. [1] automatically replace face images in photographs, although its purpose is not for face de-identification. In computer vision, earlier work on face de-identification are ad-hoc methods using simple operations such as masking, blurring, and pixelation [2]. While these simple techniques are easily applicable to any image, there is no guarantee in the de-identification performance.

To overcome this problem, Newton et al. [19] propose the k-Same algorithm based on the k-anonymity concept. It uses the averaged face of k images from the gallery as the de-identified image. Therefore, the performance of face recognition is theoretically limited to $\frac{1}{k}$. Although the privacy protection can be guaranteed, there is no guarantee in preserving data utility and the de-identified images usually suffer from blurring and ghosting artifacts. Some variants of k-Same are proposed to solve these problems. k-Same-Select algorithm [8] divides the image set to mutual exclusive subsets and applies the k-Same algorithm to each subset. This algorithm attempts to preserve attributes of each subset. The k-Same-M algorithm [10] relies on an Active Appearance Model (AAM) [18] where an image is represented by its shape and appearance parameters. It applies the k-Same algorithm to the appearance parameters of k similar images. Using the AAM model solves the misalignment problem and remedies the undesirable artifacts.

Recently, there are more work that consider facial attributes in face de-identification. Gross et al. [9] propose multi-factor models to factorize test images into identity and non-identity factors and apply de-identification to the former factors. However, only expression is considered in this paper. Du et al. [6] explicitly preserve race, gender, age attributes in face de-identification. Given a test image, it first computes the attributes and select the corresponding attribute-specific AAMs for k-Same. However, for any possible combination of the interested attributes, one specific AAM model is needed. Therefore, this method is not scalable to a *large* set of attributes. Another related work is [22] where identity is preserved while gender is changed.

Our work is an extension of k-Same-M. Instead of using an averaged face, we propose to estimate the best weights to combine k images by explicitly modeling attribute preservation and face de-identification in an optimization process. In addtion, a large set of 17 attributes is studied in this paper.

In all k-Same-based methods, there is a gallery set for selecting k similar images. The test set in their experiments may either overlap [9, 19] or disjoint [8, 6] with the gallery set. The overlap scenario is more challenging because the same test image will be one of k images and retain identity information in the de-identified image. Further, this also mimics the worst case scenario since for an arbitrary test subject, there might exist a *visually similar* gallery subject. Therefore, our experiments employ the overlap scenario.

3. The Proposed APFD Approach

As shown in Fig. 2, the proposed method mainly consists of two parts. First, an AAM model, a set of facial attribute classifiers, and a face verification classifier are learned. We apply these classifiers and model on the gallery set to compute the attributes, shape and appearance parameters of all images. Second, given a test image from the same set, we find the top k images that share the most number of similar attributes to those of the test image and save the corresponding shape and appearance parameters. Similar to k-Same-M algorithm, we update the image by linearly weighting the shape and appearance parameters of k images. Instead of applying a constant weight, we formulate an objective function to estimate the optimal weights such that the de-identified image and the original image will have as many common attributes as possible while being classified as two different subjects. We present each part in the following.

3.1. Attribute Classifiers

In order to ensure the facial attributes are preserved, we learn a set of N_t attribute classifiers, each for detecting the presence of one individual attribute. Given the success of gist [21] feature in scene classification and attribute studies [23], we adopt it as the feature for our attribute classifiers. Gist feature extraction includes Gabor filtering and

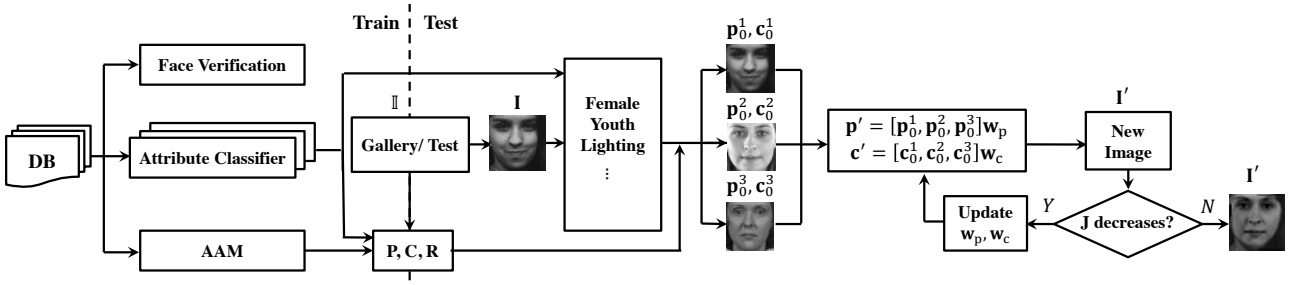

Figure 2. Overview of the proposed Attribute-Preserved Face De-identification (APFD) algorithm.

grid averaging. First, a set of B Gabor filters with different scales and orientations are applied to an image. Second, the filtered image is normally divided into a 4×4 grid and the values are averaged in each cell to form a 16-dim vector for one filtered image. In our work, we use a denser grid with the cell size of s and the overlap of $\frac{s}{2}$. For a test image $\mathbf{I}$, we denote the extracted gist feature as $\mathbf{f} = F(\mathbf{I})$.

Given the high-dimensional gist feature, we use Boosting for both feature selection and classifier learning due to its proven success [29]. We denote an attribute classifier as,

$$G(\mathbf{f}) = \sum_{i=1}^{N_w} \lambda_i g_i(f_i), \tag{1}$$

where f_i is the ith selected gist feature - one element among $\mathbf{f}$, λ_i is the weight, and $g_i(f_i)$ is a weak classifier. Similar to [15], we use atan() in order to make $g_i(f_i)$ differentiable,

$$g_i(f_i) = \frac{2}{\pi}\text{atan}(p_i f_i - p_i \theta_i), \tag{2}$$

where θ_i is a threshold and p_i is a parity term (1 or -1). p_i is assigned such that on average the positive class has a larger $p_i f_i$ than $p_i \theta_i$. For the detailed training process, the reader may refer to [15]. During classification, the sign of $G(\mathbf{f})$ indicates the binary attribute value of an image.

3.2. Face Verification

Face verification is a well-studied topic [17]. We leverage the work of [4] to learn a joint Bayesian face verification classifier via the gist feature. Specifically, the gist features of one face image pair, $\mathbf{f}$ and $\mathbf{f}'$, is assumed to form a Gaussian distribution. The appearance of a face is influenced by two factors: the identity and the intra-personal variations. That is, we have the probability of $\mathbf{f}$ and $\mathbf{f}'$ belonging to the same subject as $P(\mathbf{f}, \mathbf{f}'|H_I) = N(0, \Sigma_I)$, and different subjects as $P(\mathbf{f}, \mathbf{f}'|H_E) = N(0, \Sigma_E)$, where Σ_I and Σ_E are two covariance matrices that can be estimated from training data. The log likelihood ratio $R(\mathbf{f}, \mathbf{f}')$ is computed as,

$$R = \log \frac{P(\mathbf{f}, \mathbf{f}'|H_I)}{P(\mathbf{f}, \mathbf{f}'|H_E)} = \mathbf{f}^\intercal \mathbf{A} \mathbf{f} + \mathbf{f}'^\intercal \mathbf{A} \mathbf{f}' - 2\mathbf{f}^\intercal \mathbf{G} \mathbf{f}', \tag{3}$$

where $\mathbf{A}$ and $\mathbf{G}$ are two matrices derived from Σ_I and Σ_E. More details can be found in [4]. Finally, the smaller R is, the more likely that $\mathbf{f}$ and $\mathbf{f}'$ belong to different subjects.

Chen et al. [5] suggests face recognition to use the inner face region rather than the background or hair, while the hair has shown contribution to face verification [25]. For simplicity, given the face image pair, we apply a fixed oval-shaped mask to eliminate the background information before feeding them to Bayesian face verification.

3.3. Face Image Synthesis

Given a test image $\mathbf{I}$, face de-identification attempts to synthesize a new image $\mathbf{I}'$ that still visually appears as a face with a similar image quality, i.e., maintaining *data usability* [6]. Hence we describe how to represent $\mathbf{I}$ and synthesize a face image. Conventional k-Same method generates blurred images $\mathbf{I}'$ with ghosting effects due to the misalignment among k images. Similar to the k-Same-M [10] algorithm, we represent an image $\mathbf{I}$ by its shape parameter $\mathbf{p}$ and appearance parameter $\mathbf{c}$ in a pre-trained AAM model,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{N_s} p_i \mathbf{s}_i, \quad \mathbf{a} = \mathbf{a}_0 + \sum_{i=1}^{N_a} c_i \mathbf{a}_i, \tag{4}$$

where $\mathbf{s}$ is a $2N_l$-dim vector containing the 2D coordinates of N_l landmarks on a face image. $\mathbf{s}_0$ and $\mathbf{a}_0$ are the mean shape and appearance, and $\mathbf{s}_i$ and $\mathbf{a}_i$ are the basis functions of shape and appearance, respectively. The AAM model $\mathbb{M} = \{\{\mathbf{s}_i\}_{i=0}^{N_s}, \{\mathbf{a}_i\}_{i=0}^{N_a}\}$ is learned from a set of face images with manually labeled landmarks [18].

In order to compute $\mathbf{p}$ for $\mathbf{I}$, we employ the state-of-the-art linear regression-based face alignment approach [30] to estimate N_l landmarks ($\mathbf{s}$). As for $\mathbf{c}$, the key is to compute the warped test image in the mean shape space $\mathbf{a} = \mathbf{I}(W(\mathbf{x}_0; \mathbf{s}))$, where $W(\mathbf{x}_0; \mathbf{s})$ represents the piecewise affine warping from $\mathbf{s}_0$ to $\mathbf{s}$, and $\mathbf{x}_0$ is the collection of all coordinates in $\mathbf{s}_0$.

With the face representation of $\mathbf{p}$ and $\mathbf{c}$, our de-identification algorithm can synthesize a new image $\mathbf{I}'$ by using the estimated $\mathbf{p}'$ and $\mathbf{c}'$. This can be achieved by using the inverse warping in the AAM model, which, however, only generates the inner part of the face. In contrast, we need to synthesize a rectangular-shaped face image for the purpose of evaluating attribute classification. We address this problem by generating both the inner and outer

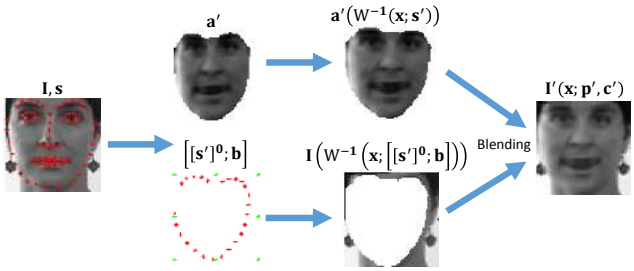

Figure 3. Synthesizing a face image $\mathbf{I}'$ with the updated shape and appearance parameters, $\mathbf{p}'$ and $\mathbf{c}'$.

parts of a face. Specifically, we add 8 points on the image boundary (green points in Fig. 3), denoted as a 16-dim vector $\mathbf{b}$, together with the boundary of the face shape to form triangles for the outer part. By using the test image and the updated parameters ($\mathbf{p}'$, $\mathbf{c}'$), the synthesized image $\mathbf{I}'$ can be generated by,

$$\mathbf{I}'(\mathbf{x};\mathbf{p}',\mathbf{c}') = \begin{cases} \mathbf{a}'(W^{-1}(\mathbf{x};\mathbf{s}')), & \text{if} \quad \mathbf{x} \in \mathbf{s}', \\ \mathbf{I}(W^{-1}(\mathbf{x};[[\mathbf{s}']^{\mathrm{O}};\mathbf{b}])), & \text{otherwise}, \end{cases}$$

where $\mathbf{s}'$ and $\mathbf{a}'$ are computed via Eqn. 4 with $\mathbf{p}'$ and $\mathbf{c}'$, $W^{-1}(\mathbf{x};\mathbf{s}')$ represents the inverse warping from the coordinates $\mathbf{x}$ in $\mathbf{I}'$ to the coordinates in $\mathbf{s}_0$, when $\mathbf{x}$ resides within the boundary of $\mathbf{s}'$, or otherwise to the coordinates in the test image space. The operator $[\mathbf{s}']^{\mathrm{O}}$ returns the subset of $\mathbf{s}'$ corresponding to the landmarks on the boundry. The whole process is illustrated in Fig. 3. To combine the inner and outer parts, we apply image blending by replacing the pixels near the boundary as the average of their neighborhood.

3.4. Optimization-based De-identification

In this section, we describe the proposed method for face de-identification. Our goal is to change the identity of a test image while preserving a large set of facial attributes. In order to do this, we combine the attribute classifiers and face verification classifier in a joint objective function. Given a test image, we first find k images from the gallery set that share the most similar attributes as those of $\mathbf{I}$. Our optimization process estimates the optimal weights for fusing k images to generate the de-identified image.

We denote the gallery set $\mathbb{I} = \{\mathbf{I}_i\}_{i=0}^{N_g}$. To create a more challenging scenario for de-identification, we assume that the test image is from $\mathbb{I}$. For the efficiency of testing, we pre-compute the representations and attributes for all images in $\mathbb{I}$. Specifically, we employ the face alignment [30] to estimate landmarks on each image and the collections of its shape and appearance representation are denoted as $\mathbf{P}$ and $\mathbf{C}$ respectively. The learned attribute classifiers are used to estimate the N_t facial attributes of all images in $\mathbb{I}$, denoted as a $N_t \times N_g$ matrix $\mathbf{R}$ where $\mathbf{R}(i,j) \in [-1,1]$.

Given a test image, we apply the attribute classifiers to estimate the N_t attributes $\mathbf{r}$. Based on the similarity between $\mathbf{r}$ and each column of $\mathbf{R}$, we find the top k images with the highest similarities, and denote their corresponding shape and appearance parameters as $\mathbf{P}_0 = [\mathbf{p}_0^1, \cdots, \mathbf{p}_0^k]$ and $\mathbf{C}_0 = [\mathbf{c}_0^1, \cdots, \mathbf{c}_0^k]$. Following the k-Same theory [19], it is assumed that the shape and appearance parameters ($\mathbf{p}$, $\mathbf{c}$) of the de-identified image can be generated by,

$$\mathbf{p} = \mathbf{P}_0\mathbf{w}_p, \quad \mathbf{c} = \mathbf{C}_0\mathbf{w}_c, \tag{5}$$

where $\mathbf{w}_p$ and $\mathbf{w}_c$ are k-dim vectors representing the weights of k images.

Rather than assigning $\frac{1}{k}$ to $\mathbf{w}_p$ and $\mathbf{w}_c$, which is universally adopted in all prior k-Same methods, this paper formulates an objective function $J(\mathbf{w_p}, \mathbf{w_c})$, whose minimization leads to the optimal $\mathbf{w}_p$ and $\mathbf{w}_c$,

$$J = \sum_{i=1}^{N_t} ||\frac{2}{\pi}\text{atan}(G_i(\mathbf{f}')) - \frac{2}{\pi}\text{atan}(G_i(\mathbf{f}))||_2^2 + \lambda R(\mathbf{f}', \mathbf{f}), \tag{6}$$

where $\{G_i()\}_{i=1}^{N_t}$ are the attribute classifiers, R is the face verification classifier, and $\mathbf{f}' = F(\mathbf{I}'(\mathbf{x};\mathbf{p},\mathbf{c}))$ is the extracted gist feature from the synthesized image $\mathbf{I}'$.

The first term in Eqn. 6 measures the difference between the attributes of $\mathbf{I}'$ and those of $\mathbf{I}$. The atan() function is used to approximate the sign function since we are only interested in the polarity of the attribute classifier, i.e., whether the face image has the positive or negative attribute value, rather than the continuous classifier score. The second term aims to classify $\mathbf{I}$ and $\mathbf{I}'$ as different subjects by minimizing the score of the face verification classifier.

Since image warping is involved, this is a non-linear optimization problem. We decide to use gradient descent to minimize the objective function by iteratively estimating $\mathbf{w}_p$ and $\mathbf{w}_c$. We use the chain rule to compute the derivative

Algorithm 1: Attribute-preserved face de-identification.

Input: Test image $\mathbf{I}$, k, attribute classifiers $\{G_i()\}_{i=1}^{N_t}$, face verifier $R()$, AAM model $\mathbb{M}$, $\mathbf{P}$, $\mathbf{C}$, and $\mathbf{R}$.
Output: De-identified image $\mathbf{I}'$.
1. Compute $\mathbf{r} = [G_1(\mathbf{I}); \cdots; G_{N_t}(\mathbf{I})]$;
2. Find k images by comparing $\mathbf{R}$ and $\mathbf{r}$. Save their corresponding parameters as $\mathbf{P}_0$ and $\mathbf{C}_0$;
3. Initialization:
$\mathbf{w}_p^0 = \mathbf{w}_c^0 = [\frac{1}{k}, \frac{1}{k}, ..., \frac{1}{k}]$;
$\mathbf{p}_0 = \mathbf{P}_0\mathbf{w}_p^0$, $\mathbf{c}_0 = \mathbf{C}_0\mathbf{w}_c^0$, $t = 0$;
4.**while** $J_t() < J_{t-1}()$ **do**
  $t = t + 1$;
  Compute $\frac{\partial J}{\partial \mathbf{w}_p}$ according to Eqn. 7;
  $\mathbf{w}_p^t = \mathbf{w}_p^{t-1} - \eta_p\frac{\partial J}{\partial \mathbf{w}_p}$, $\mathbf{p}_t = \mathbf{P}_0\mathbf{w}_p^{t-1}$;
  Compute $\frac{\partial J}{\partial \mathbf{w}_c}$ according to Eqn. 7;
  $\mathbf{w}_c^t = \mathbf{w}_c^{t-1} - \eta_c\frac{\partial J}{\partial \mathbf{w}_c}$, $\mathbf{c}_t = \mathbf{C}_0\mathbf{w}_c^{t-1}$;
Return: $\mathbf{I}'(\mathbf{x};\mathbf{p}_t,\mathbf{c}_t)$.

of J w.r.t. $\mathbf{w}_p$ and $\mathbf{w}_c$,

$$\frac{\partial J}{\partial \mathbf{w}_p} = \frac{\partial J}{\partial \mathbf{f}'}\frac{\partial \mathbf{f}'}{\partial \mathbf{w}_p}, \quad \frac{\partial J}{\partial \mathbf{w}_c} = \frac{\partial J}{\partial \mathbf{f}'}\frac{\partial \mathbf{f}'}{\partial \mathbf{w}_c}, \tag{7}$$

where $\frac{\partial J}{\partial \mathbf{f}'}$ can be easily computed. For the second part, due to non-linear image warping, we use the slope of the tangent line of $\mathbf{f}'$ for approximation. For example, the derivative of $\mathbf{f}'$ w.r.t. the jth element of $\mathbf{w}_p$ is,

$$\frac{\partial \mathbf{f}'}{\partial \mathbf{w}_p(j)} = \frac{F(\mathbf{I}'(\mathbf{x};\mathbf{p}+\epsilon\mathbf{p}_0^j,\mathbf{c})) - F(\mathbf{I}'(\mathbf{x};\mathbf{p},\mathbf{c}))}{\epsilon}, \tag{8}$$

where ϵ is a small perturbation, and $\mathbf{p}_0^j$ is the jth column of $\mathbf{P}_0$. For the iterative gradient descent, all elements in $\mathbf{w}_p$ and $\mathbf{w}_c$ are initialized as $\frac{1}{k}$, i.e., we assign equal weights for all k images. The iteration continues as long as the objective function keeps decreasing. Algorithm 1 summarizes the de-identification process.

4. Experiments

This section presents the details of our experimental setup, the results, and comparison with prior work.

4.1. Experimental Setup

Attribute Classifiers We use FaceTracer dataset [12] to train and test our attribute classifiers. This is one of the few publicly available dataset providing ground truth labels for a large number of facial attributes. There are $5,000$ attribute labels for 10 different attributes including *gender* (female, male), *race* (Asian, White, Black), *age* (baby, child, youth, middle-aged, senior), *hair color* (blond, no blond), *eye wear* (none, eyeglasses, sunglasses), *mustache* (true, false), *expression* (smiling, not smiling), *blurry* (true, false), *lighting* (harsh, flash), and *environment* (outdoor, indoor). We are able to download an image subset with $3,268$ labels, where 80% of them are used for training and 20% for testing. We consider all attributes except the *environment*. To enjoy the simplicity of binary classification, we treat attributes with more than two choices separately by defining the detection of each choice as a binary classification problem, similar to [13], which leads to a total of $N_t = 17$ attributes.

Face Verification It is shown in [4] that the performance of Bayesian face verification increases when more subjects are used for training. We use FaceScrub dataset [20], which includes a large set of face images for 530 subjects. We randomly select five images for each subject and manually label the eye centers to rectify all images. These data are used to compute the two covariance matrices Σ_I and Σ_E, and subsequently $\mathbf{A}$ and $\mathbf{G}$ in Eqn. 3. We use the testing set in LFW [11] to evaluate the performance.

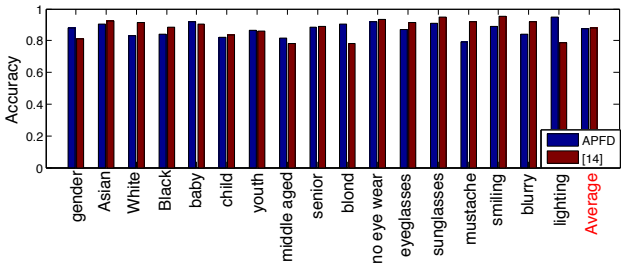

Figure 4. The accuracies of 17 attributes and their average.

Face De-identification We use the training set of 300W dataset [26], which includes $3,148$ images with ground truth labels of $N_l = 68$ facial landmarks per image, to learn the AAM model $\mathbb{M}$. For the test set $\mathbb{I}$, it is desirable that it contains face images with diverse combinations of facial attributes. However, conventional face recognition datasets normally have non-uniform distribution in their attributes due to the biased population of volunteers. Therefore, we manually select $N_g = 200$ images[1] from four datasets, ND1 [3], FERET [24] , CAS-PEAL [7], and BioID [28], with the goal of minimal bias in any attribute. Furthermore, this testing set $\mathbb{I}$ also serves as the gallery set for selecting k images. Given a test image, including the same image in the gallery leads to a more challenging problem, because this same image is guaranteed to be selected as the top of k images. Therefore the de-identified image will have a higher probability to maintain the same identify as the original test image, compared to the case of *not* including the same image in the gallery set. For each de-identified image, we use Bayesian face verification to perform closed-set face recognition on the gallery set and evaluate the performance based on the face recognition rate and the number of preserved attributes. Since no prior work has explicitly performed face de-identification while preserving attributes, we choose to use k-Same-M, one of the state-of-the-art de-identification methods, as the baseline method for comparison.

Note that all datasets discribed above are used for traing except the gallery set. The trained models can be directly applied to a test image. In real-world applications, the only data needed is the gallery set, which can be a few hundred of face images with diverse attributes.

Parameter Setting All images in our experiments are rectified with two eye centers and cropped to the size of 64×64. In the gist feature extraction, we set $B = 32$ and $s = 8$, which results in a $7,200$-dim feature $\mathbf{f}$. We apply PCA to reduce the high feature dimension to 400 before learning the face verification classifier. Each attribute classifiers consists of $N_w = 300$ weak classifiers. In the AAM model, by preserving 99% of energy, we use $N_s = 101$ and $N_a = 1,334$ basis functions for shape and appearance, re-

[1] http://www.cse.msu.edu/~liuxm/faceDeID

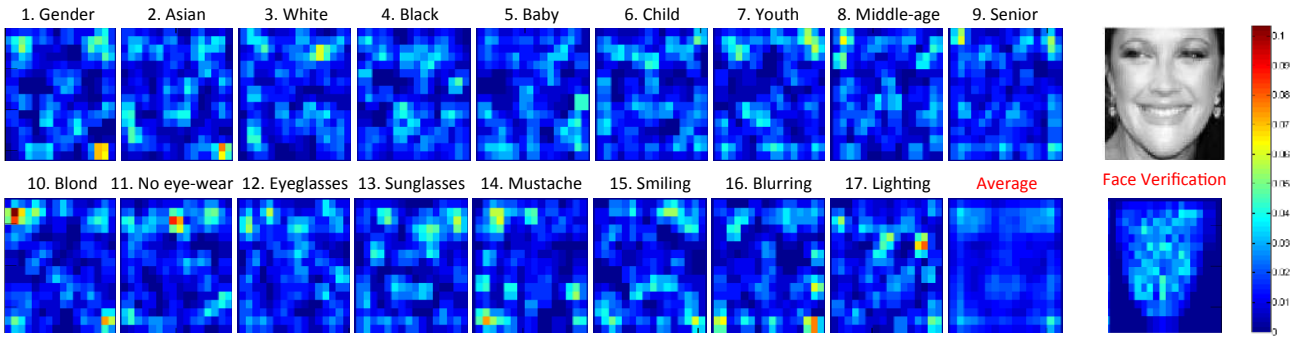

Figure 5. Feature maps of attribute classifiers and face verification classifier. Higher intensities indicate more contribution to classification.

spectively. Finally, we experimentally determine other parameters as $\lambda = 0.2$, $\eta_p = 0.001$, $\eta_c = 0.05$, and $\epsilon = 0.4$.

4.2. Experimental Results

Since the proposed method relies on the supervised learned classifiers to guide the optimization, it is important that these classifiers can achieve good classification performances in their own tasks. The performances of all attribute classifiers are shown in Fig. 4, with the average accuracy of 87%. We compare it with the results reported in [13] since both methods are evaluated on the same dataset. The overall performance is comparable, which allows us to integrate reliable attribute classifiers into the proposed face de-identification framework. The Bayesian classifier achieves the face verification rate of 70%. As discussed in Sec. 3.2, the performance of Bayesian face verification increases as the number of training subjects increases. With only 530 subjects, we achieve a reasonable good performance.

The underlying assumption of the proposed method is that the visual features accounting for facial identity and facial attributes are distinct and separable. In order to verify this assumption, we plot the feature maps of each attribute classifier (summation of λ_i at all N_w feature locations) and the Bayesian classifier to show the spatial distributions of selected features. For the feature map of Bayesian classifier, we use the approximation similar to Eqn. 8 to compute the difference of classifier output w.r.t. the change of a specific feature. As shown in Fig. 5, the selected feature locations for each attribute are consistent with intuition. For example, most features are selected from the boundary of the face for the *blond hair* attribute. For *eye-wear* attribute, most features are from the upper part of the face. However, some features from the top of the face are selected for classifying the *mustache* attribute. The reason could be that this attribute is relevant to male and the top-left corner is salient for the gender classifier. Overall, the average of all 17 feature maps are separable with the feature map of Bayesian classifier. Therefore, the assumption holds, and it is possible to de-identify the face while preserving the attributes.

Figure 6 shows an example of the gradient descent process in the proposed face de-identification algorithm. The

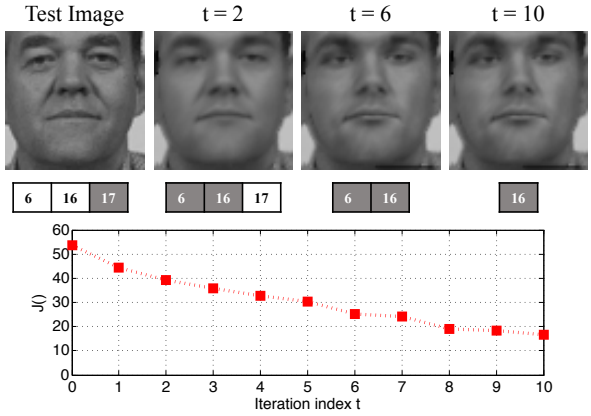

Figure 6. An example of iterative gradient descent process. The box under each image indicates the attribute values, where the white box represents the positive value and gray the negative. The attribute indexes can be referred to Fig. 5. The number of boxes under each de-identified image is equal to the number of non-preserved attributes compared to the test image.

face images are gradually updated such that the Bayesian classifier score is decreasing and the number of non-preserved attributes keeps decreasing. Therefore, the joint formulation of our objective function serves the purpose of both face de-identification and attribute preservation.

We compare the proposed method with k-Same-M [10], which can be considered as the initialization of our proposed method. We apply both methods to all 200 images. For each de-identified image, we use the attribute classifiers to compute the number of preserved attributes, and apply the Bayesian classifier to calculate the rank-1 face recognition rate. The performances of both methods under different k are shown in Fig. 7. The face recognition rate for both methods decreases as k increases, which is consistent with the k-anonymity theory. And it almost saturates when $k = 8$. Comparing with the baseline, the improvement in de-identification is substantial. Across different k, the face recognition rate of APFD is on average only 37% of the baseline method, i.e., we reduce the recognition rate by more than half. In contrast, the improvement in attribute

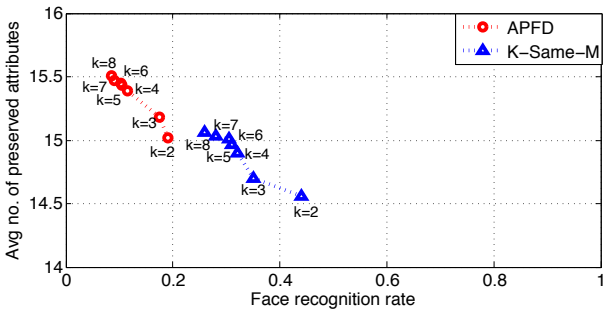

Figure 7. Comparison of the face recognition rates and the average number of preserved attributes for k-Same-M and APFD.

preservation is relatively minor, around 0.5 more attributes for all k. This is partially due to the limited number of attributes in our experiments. With potentially larger number of attributes in real-world applications, it is less likely to find k images with all attributes being the same, and therefore the optimization has more advantages in preserving the maximum number of attributes. Note that the Bayesian face verification is one of the state-of-the-art recognizers. Although our implementation does not achieve comparable results due to the lack of training subjects. The trained recognizer is tested on both methods, which is a fair comparision.

Figure 8 shows the de-identified images at different k. In general, we see that a larger k value leads to lower face verification scores (R), and hence a better de-identification performance. It is also able to preserve more attributes.

Figure 9 shows exemplar de-identification results of seven images with $k = 3$. The proposed method is superior in terms of the number of non-preserved attributes and the amount of identity difference w.r.t. the test image. The simple averaging scheme can result in distorted face shapes (Col. #1) and undesirable attributes (Col. #7). On the contrary, the proposed method can generate images of better quality and preserve more attributes. Noted that the accuracy of the attribute classifiers influences the final performance. E.g., our attribute classifier wrongly classify the test image in Col. #6 as having eye-wear, which encourages the de-identified image to preserve eye-wear, as seen in Fig. 9 (e) at Col. #6. Therefore, as research progresses, the better (attribute and verification) classifiers would be beneficial to, and can be easily incorporated into, our framework.

5. Conclusions

This paper studies the problem of attribute preservation in face de-identification. We first learn a set of attribute classifiers and a face verification classifier that achieve reasonablely good performance. These classifiers are used in the joint modeling of face de-identification and attribute preservation. An image is de-identified by combining the shape and appearance parameters of k images with similar attributes. The main technical novelty of our work is to estimate the optimal weights for k images instead of using the average weights. Experimental results show the effectiveness of the proposed method compared to the baseline. In the future, we will apply our method to face videos assuming video-based face alignment is conducted [16].

References

[1] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar. Face swapping: automatically replacing faces in photographs. In *SIGGRAPH*, page 39. ACM, 2008.

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

[3] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *Proc. ACM Workshop on Multimodal User Authentication*, pages 25–32. ACM, 2003.

[4] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian face revisited: A joint formulation. In *ECCV*, pages 566–579. Springer, 2012.

[5] L.-F. Chen, H.-Y. M. Liao, J.-C. Lin, and C.-C. Han. Why recognition in a statistics-based face recognition system should be based on the pure face portion: a probabilistic decision-based proof. *Pattern Recognition*, 34(7):1393–1403, 2001.

[6] L. Du, M. Yi, E. Blasch, and H. Ling. Garp-face: Balancing privacy protection and utility preservation in face de-identification. In *IJCB*. IEEE, 2014.

[7] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE T-SMC A, Syst. Humans*, 38(1):149–161, 2008.

[8] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies*, pages 227–242. Springer, 2006.

[9] R. Gross, L. Sweeney, J. Cohn, F. de la Torre, and S. Baker. Face de-identification. In *Protecting Privacy in Video Surveillance*, pages 129–146. Springer, 2009.

[10] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *CVPRW*, pages 161–161. IEEE, 2006.

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

[12] N. Kumar, P. N. Belhumeur, and S. K. Nayar. FaceTracer: A Search Engine for Large Collections of Images with Faces. In *ECCV*, pages 340–353. Springer, 2008.

[13] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372. IEEE, 2009.

[14] T. Li and N. Li. On the tradeoff between privacy and utility in data publishing. In *SIGKDD*, pages 517–526. ACM, 2009.

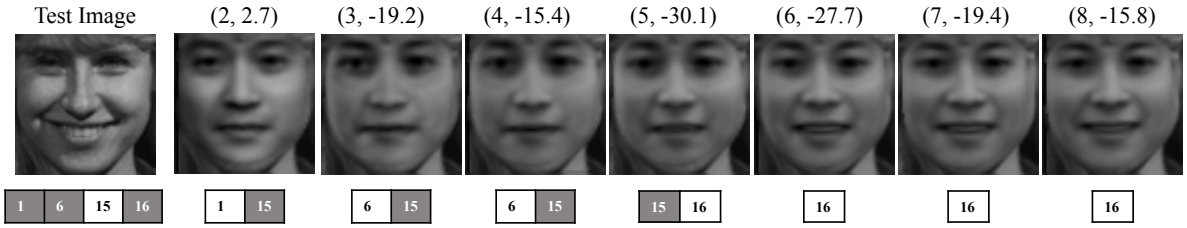

Figure 8. Results of a test image with different numbers of similar images (k). The numbers above are (k, R) values.

Figure 9. (a) Test images; (b) original attribute values; (c) results of k-Same-M; (d) wrongly preserved attributes of k-Same-M; (e) results of the AFPD; (f) wrongly preserved attributes of AFPD.

[15] X. Liu. Discriminative face alignment. *IEEE T-PAMI*, 31(11):1941–1954, 2009.

[16] X. Liu. Video-based face model fitting using adaptive active appearance model. *Image and Vision Computing*, 28(7):1162–1172, 2010.

[17] X. Liu, T. Chen, and B. V. K. V. Kumar. On modeling variations for face authentication. In *FG*, pages 384–389, 2002.

[18] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[19] E. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *T-KDE*, pages 232–243, 2005.

[20] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *ICIP*, pages 343 – 347. IEEE, 2014.

[21] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, pages 145–175, 2001.

[22] A. Othman and A. Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *Proc. of ECCV Workshop on Soft Biometrics*, 2014.

[23] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510. IEEE, 2011.

[24] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE T-PAMI*, 22(10):1090–1104, 2000.

[25] J. Roth and X. Liu. On hair recognition in the wild. In *AAAI*, 2014.

[26] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPRW*, pages 896–903. IEEE, 2013.

[27] A. Senior. Privacy protection in a video surveillance system. In *Protecting Privacy in Video Surveillance*, pages 35–47. Springer, 2009.

[28] M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - A flexible appearance modeling environment. *IEEE T-MI*, 22(10):1319–1331, 2003.

[29] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[30] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *ICCVW*, pages 392–396. IEEE, 2013.