Controllable Face Privacy

Terence Sim and Li Zhang
School of Computing, National University of Singapore, Singapore

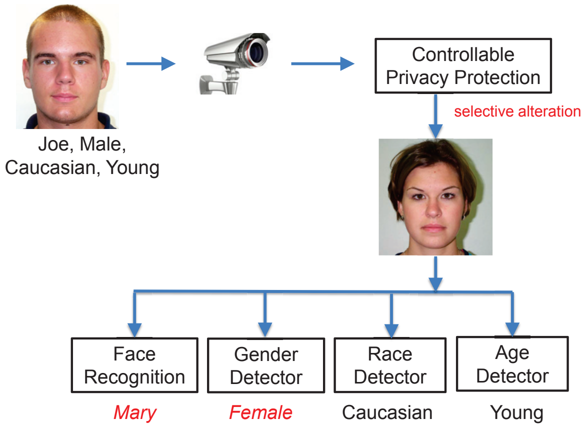

Fig. 1. A controllable privacy protection system can selectively alter some facial attributes, *e.g.* identity and gender, while retaining others.

***Abstract*— We present the novel concept of Controllable Face Privacy. Existing methods that alter face images to conceal identity inadvertently also destroy other facial attributes such as gender, race or age. This all-or-nothing approach is too harsh. Instead, we propose a flexible method that can independently control the amount of identity alteration while keeping unchanged other facial attributes. To achieve this flexibility, we apply a subspace decomposition onto our face encoding scheme, effectively decoupling facial attributes such as gender, race, age, and identity into mutually orthogonal subspaces, which in turn enables independent control of these attributes. Our method is thus useful for nuanced face de-identification, in which only facial identity is altered, but others, such gender, race and age, are retained. These altered face images protect identity privacy, and yet allow other computer vision analyses, such as gender detection, to proceed unimpeded. Controllable Face Privacy is therefore useful for reaping the benefits of surveillance cameras while preventing privacy abuse. Our proposal also permits privacy to be applied not just to identity, but also to other facial attributes as well. Furthermore, privacy-protection mechanisms, such as k-anonymity, L-diversity, and t-closeness, may be readily incorporated into our method. Extensive experiments with a commercial facial analysis software show that our alteration method is indeed effective.**

I. Introduction

Big Brother is already watching you: over 3,000 eyes (cameras) watch the streets of Lower Manhattan in New York City [9], while about 13,000 cameras blanket the subway system of London [20]. This scales up nationally as well, with an estimated six million cameras, or one CCTV per 11 people, in the United Kingdom [20]. On the Web, people routinely upload faces of themselves onto social websites like Facebook. In turn, this has engendered websites that aggregate face photos, such as `www.thefacesoffacebook.com`, which purports to house 1.2 billion "faces of Facebook users together". Web faces have also prompted the US National Security Administration to scour the net, at the rate of "55,000 facial recognition quality images" per day, to boost its crime-fighting intelligence [17]. These statistics and websites alarm privacy advocates, and ought to be a grave concern for ordinary citizens as well.

To be sure, there are benefits to having surveillance cameras, such as increased crime deterrence and public safety in common areas, and speedier police investigations should something bad happen (as in the case of the 2013 Boston Marathon bombing [2]). A more recent trend, called *visual analytics*, involves the usage of in-store cameras by retailers to analyze customer behavior. The goal is to assess customers' moods, determine product preferences based on age or gender, gauge staff responsiveness, as well as to streamline the layout of store shelves [23]. Such clever use of computer vision techniques may yet bring about greater convenience to customers and better sales for businesses. The mere presence of cameras does not always doom privacy.

Indeed, privacy protection need not be at the expense of visual analytics. The tricky part is to balance the benefits of surveillance cameras against the drawbacks of privacy abuse. In this paper, we propose a novel concept, called Controllable Face Privacy (Figure 1), that is helpful for protecting privacy in face images while allowing visual analytics to function normally. The key idea is to decouple facial attributes, such as gender, race (ethnicity), age and identity, into parameters that can be independently controlled. More precisely, we apply MMDA [18], a subspace decomposition technique we had previously developed, onto our face encoding scheme to selectively alter some facial attributes (*e.g.* gender, race), while retaining others (*e.g.* age). Figure 2 illustrates the changing of identity without changing gender, age or race. Figure 3 shows one example each of altering age, gender, and race attributes, respectively. The key to privacy protection lies in altering the identity of faces images (called *de-identification* or *anonymization* in the literature); while the key to visual analytics lies in analyzing non-identity facial attributes. Our method thus permits a more nuanced privacy protection: identity can be altered while non-identity attributes can be left unchanged.

What sets our paper apart from previous works is the *selectivity* with which we can alter faces. Previous face de-identification methods, such as Newton et al.[16], Neustaedter and Greenberg [15], Berger [4], alter facial identity

This research was partially carried out at the SeSaMe Centre. It is supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

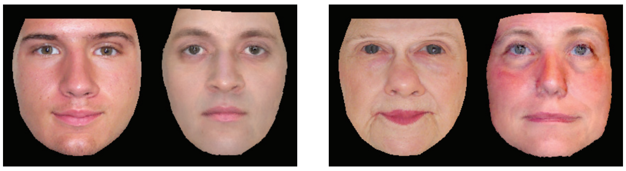

Fig. 2. Two examples of altering facial identity while retaining gender, age and race attributes. In each pair, the left image is the original, while the right is the altered face. Such an alteration protects privacy without thwarting visual analytics.

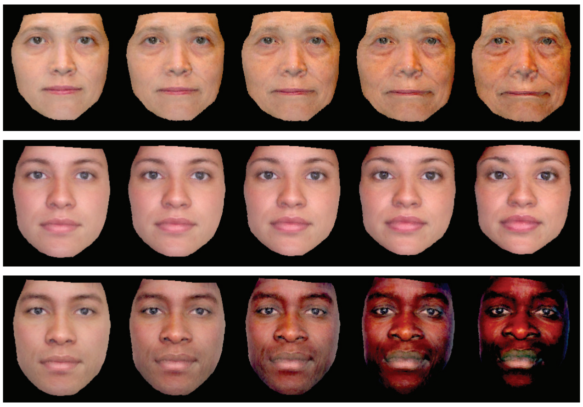

Fig. 3. (Top row) Altering the age attribute. (Left to right) Intensity, σ, is increased from 0.5 to 2.5 in steps of 0.5. The face appears older as intensity is increased. (2nd row) Increasing Femininity. (3rd row) Making the face more African.

but inadvertently destroy other facial attributes. For example, Newton et al. perform a k-to-1 replacement of faces, thereby making the altered face indistinguishable from $k-1$ other faces. Unfortunately, non-identity attributes are also rendered indistinguishable in the process. By comparison, Neustaedter and Greenberg blur all detected faces. This destroys identity information, *and* to a great extent, gender, age and race information as well. Berger employs pixelation, which preserves skin color (and may thus preserve racial information), but alas is not effective against modern face recognition algorithms, as demonstrated by Newton et al. In other words, while existing methods succeed in foiling a face identification system, they also thwart gender detectors, race detectors and age detectors. Alas, these detectors form the building blocks of visual analytics, which gather useful information about consumer preference based on non-identity facial attribues. Previous methods may thus be characterized as all-or-nothing: they protect privacy at the expense of visual analytics. Not so with our method.

To be fair, Gross et al. [10] attempted to inject *utility* into their de-identification method. By this the authors meant the retention of facial attributes not related to identity, such as facial expression. They showed that, compared to naive methods, their k-Same-Select algorithm was superior at retaining facial smiles while removing identity information from a face. Thus, a facial-expression classifier would perform well on their de-identified images but a face identification system would fail. However, it is unclear how their algorithm could be generalized to simultaneously retain multiple facial attributes such as gender, race and age. Their experiments retain only one type of facial expression, namely, smiles. Indeed, it appears that a different k-Same-Select algorithm is needed to preserve different facial attributes. Furthermore, it is doubtful that all these algorithms can work in harmony; their combined operation is not guaranteed to preserve all the desired attributes. In contrast, our method, which exploits the orthogonality property of a subspace decomposition, allows us to alter facial attributes independently and simutaneously.

Playing the devil's advocate for a moment, we could argue for a simpler method to achieve the same goal of nuanced privacy, as follows. First, prepare beforehand a fixed set of M template faces consisting of all possible combinations of gender, race and age. Then, given an input face to be de-identified, run it through a series of gender, race and age detectors to determine these facial attributes. Finally, replace the given face with the appropriate template face. It is clear that this Simple Method can also alter any desired attribute while retaining others. Indeed, it can even achieve $k-$Sameness. Why, then, bother with the method proposed in this paper?

The answer is twofold: uniqueness and diversity. The Simple Method cannot replace a face with another unique face. Indeed, k-Sameness prevents this. Instead, it is obvious that there are only M possible faces in the output of the Simple Method. This makes tracking impossible. Note that one common visual analytics task is to track the same person as she moves within a retail store. When Simple Method replaces two customers with the same face, tracking them becomes impossible (at least, when tracking with faces). Our method however, can in principle guarantee unique replacement faces because, as described in Section III-B, identity is encoded in a subspace of infinite extent. Thus an infinite number of identities is possible.[1]

Second, Simple Method lacks diversity. In fact, diversity is exactly M, the size of the template faces, which must be prepared beforehand. In contrast, our method can synthesize *at run time* as many races or genders or identities as machine precision allows. For instance, we can synthesize different degrees of Masculinity or "Caucasian-ness". We can also create mixed races, or androgynous faces, see Figures 6(b) and 7. Diversity is important in visual analytics applications that require the altered set of images to mimic the natural diversity found in the original input videos.

Our contribution: This paper pioneers the notion of Controllable Face Privacy for the protection of privacy in face images. The key idea is to *selectively* alter some facial attributes while retaining others. To this end we employ a subspace decomposition technique to decouple the parameters that control different facial attributes. In each subspace, we may then independently vary the said parameters and then synthesize faces with new attributes. This not only permits the privacy protection of facial identity (which is the

[1] In practice, due to the machine's 64-bit precision, we can synthesize 2^{768} ($\sim 10^{231}$) unique identities. While finite, this is still very large.

sole concern of all existing work), but also of gender, race and age as well. Furthermore, we show that we can easily incorporate the mechanisms of k-anonymity, L-diversity, and t-closeness [13] (pioneered by the data mining research community) to provide *provable* privacy guarantees on the altered faces. We run extensive experiments — we tested our altered images on Face++, a commercial face analysis software[2] that can classify gender, age and race — to show that our alteration is indeed effective.

II. Related Work

As mentioned in the Introduction, the proliferation of surveillance cameras in public places has been a boon for visual analytics but a bane for privacy protection. In recent years, numerous research have attempted to protect the identity of persons in surveillance videos. Broadly speaking, these works may be divided into two categories: (i) those that de-identify faces, verus (ii) those that de-identify the whole body. The two are complementary: if a face appears large enough in a video to be de-identified, the whole body is usually not seen; on the other hand, if the full body is visible, then the face region is too small to be recognized.

For category (i), the group at Carnegie Mellon University (CMU) has published a number of papers [16], [11], [10]. These authors championed formal mechanisms that can provably guarantee the protection of privacy, introducing notions such as k-Same, and k-Same-Select. The guarantee is that a face is sufficiently de-identified if no face recognition algorithm, human or machine, can distinguish it from at least $k-1$ other faces (people). The authors showed that ad-hoc methods for de-identification, such as masking out the eyes, pixelation, and adding random noise, do not sufficiently protect identity. In fact, modern face recognition software can easily overcome such ad-hoc methods.

Another notion introduced by the CMU group is that of *utility*: a de-identified face image is deemed to have preserved utility if other non-identity image analyses can still succeed on it, for example, facial expression recognition. The authors proposed their own algorithm that sufficiently alters facial identity to thwart face recognition software, but yet permits an expression classifier to work. This is not the case for ad-hoc methods such as blurring [15] or pixelation [4]. Indeed, utility is a good property, since many surveillance videos are now routinely being analyzed by computer vision algorithms for purposes like people-counting, gender detection, age classification. Unfortunately, the authors appear to have demonstrated utility only for expression and gender analysis. In this regard, our current work may be considered a generalization of theirs: we show that gender, race and age detectors all work on our de-identified images.

For category (ii), full-body de-identification, we may list [1], [7], [24], [14]. All these methods apply simple image distortions, such as pixelation, blurring, or silhouetting, to mask the identity of the person. While effective, these methods appear too harsh, since they also thwart computer vision algorithms that count people, or track human motion. None of the above authors have yet argued for, nor proposed, a more nuanced privacy protection scheme.

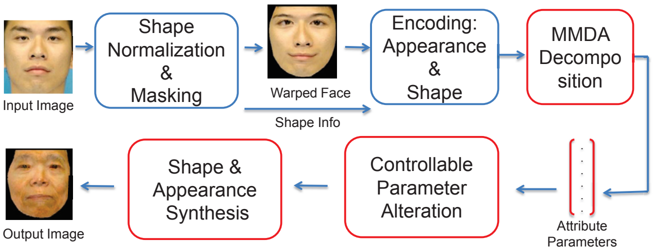
Input Image
Shape Normalization & Masking
Warped Face
Shape Info
Encoding: Appearance & Shape
MMDA Decompo sition
Attribute Parameters
Controllable Parameter Alteration
Shape & Appearance Synthesis
Output Image

Fig. 4. The main steps in our method.

III. Methodology

We first learn subspaces from a training set that is appropriately labeled with gender, age and race attributes. Since the attribute of a face is determined by both appearance and shape (for instance, Africans usually have black skin with flared noses while the Asians have yellow skin with small mouths), we need to encode both the shape and appearance into a single vector before applying multimodal analysis.

In the literature, Tensorface [22] is a popular multilinear analysis method. However, we instead prefer Multimodal Discriminant Analysis (MMDA) [18] because its zero intra-class variance property means that it can capture the essence of gender, race, or age in a single constant vector parameter (see Section III-B.1), unlike Tensorface. Altering facial attributes would thus be more difficult with Tensorface.

After applying MMDA to obtain the required orthogonal subspaces, a new face image can be easily parameterized by projecting it onto these subspaces. We can then selectively alter the parameters in the subspace corresponding to the facial attribute we wish to change. Moreover, we may control the "intensity" of the alteration by simply scaling the norm of the said parameters. Finally, because MMDA is an invertible transform, we can synthesize new faces that will exhibit the desired attributes. We illustrate our method in Figure 4.

A. Pre-processing: Normalization, Masking & Encoding

(This section describes the blue boxes in Figure 4). Given a face image, we first locate facial landmarks using AAM [8] and align the image to a reference face via the eye positions. Our images are 385×343 in size. After alignment, we normalize the shape by non-linearly warping the image to our reference face. Any non-linear warping method may be used; for convenience, we chose Thin Plate Spline [6]. A mask is then applied to remove hair and background. After this, each face image is encoded as the concatenation of two parts: a shape-normalized appearance vector and a shape vector. The appearance vector is obtained by simple column-scanning the image. To handle color, we experimented with various color spaces and found that the XYZ color space produced the best results. The shape vector contains the (x, y) coordinates, before warping, of the 63 landmarks. Let D_a denote the

[2] www.faceplusplus.com

dimension of the appearance vector, and D_s the dimension of the shape vector. Our encoded vector thus has $D = D_a + D_s$ dimensions. The blue boxes in Figure 4 illustrates the main idea of this section.

B. Decomposition, Alteration & Synthesis

(This section describes the red boxes in Figure 4.) At the heart of our method, we employ MMDA, a linear algebra subspace decomposition technique based on the Whitened Fisher Linear Discriminant. Essentially, MMDA decomposes a set of data vectors containing multiple modes into mutually orthogonal subspaces, as shown in Figure 5. A face image X, suitably encoded as a column vector, is projected onto three subspaces, one for each mode of gender, race and age. Figure 5 shows each subspace as a blue axis. In each subspace, MMDA computes an orthonormal basis by which parameters (coefficients) of X for that mode may be derived. These parameters can then be altered, and a new face image synthesized. Altering the parameters in one subspace affects only the mode captured by that subspace. For instance, in Figure 5, altering the gender parameters moves X to A, the resulting image at A exhibits only a gender change, while race and age is unchanged.

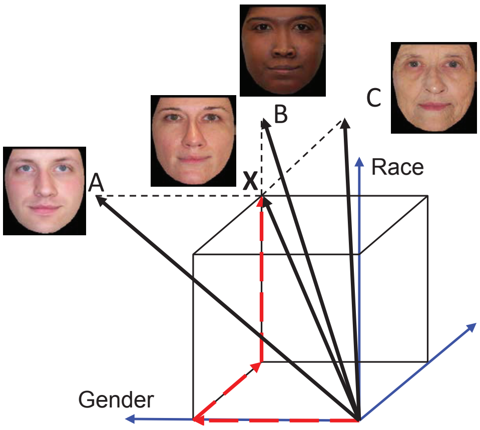

Fig. 5. MMDA projects a vector X into three mutually orthogonal subspaces (represented by the blue axes) that encode gender, race and age, respectively. Changing the parameters of X in the race subspace alters only the race, not the gender or age. This corresponds to moving X to B. Reconstructing the vector B reveals a new face image exhibiting only a change in race, not of age or gender.

1) MMDA theory: Mathematically, let $\mathbf{X}$ denote a $D \times N$ training matrix whose columns are $\mathbf{x}_i, i = 1, \ldots, N$ with K attributes. Each training vector $\mathbf{x}_i$ is multiply labeled, one label per mode. For each attribute i, there are C_i classes, $\mathbf{L}_1^i, ..., \mathbf{L}_{C_i}^i$. In our case, there are three attributes, gender, race and age ($K = 3$). For gender, there are two class labels, Male and Female; For race, there are three class labels, Caucasian, African and Oriental; For age, there are also three class labels, Young, Middle-aged and Old.[3] Therefore, in our case we have $N = 2 \times 3 \times 3 = 18$ training images. Let $\mathbf{m}_k^i$ denote the mean of class $\mathbf{L}_k$ for attribute i. Without loss of generality, we assume that the global mean of $\mathbf{X}$ is zero. If not, we may simply subtract the mean from each element. Note that the global mean is the same for all attributes.

To begin, we whiten the data $\mathbf{X}$ for all modes. For each mode i, we compute the total scatter matrix $\mathbf{S}_t^i = \mathbf{X}\mathbf{X}^\top$, then eigen-decompose it to get $\mathbf{S}_t^i = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, retaining only non-zero eigenvalues in the diagonal matrix $\mathbf{D}$ and their corresponding eigenvectors in $\mathbf{U}$. Now, compute the $(N-1) \times D$ matrix $\mathbf{P} = \mathbf{U}\mathbf{D}^{-1/2}$ and apply it to the data to get $(N-1) \times (N-1)$ matrix: $\tilde{\mathbf{X}} = \mathbf{P}^\top\mathbf{X}$. Note that we only need to perform this step once, as all modes come from the same training set. The data is now whitened because the scatter matrix of $\tilde{\mathbf{X}}$ equals $\mathbf{I}$, the identity matrix.

After data whitening, we maximize the Fisher Criterion for each mode i,

$$\mathbf{J}_F(\mathbf{V}^i) = trace\{((\mathbf{V}^i)^\top \tilde{\mathbf{S}}_w^i (\mathbf{V}^i))^{-1}((\mathbf{V}^i)^\top \tilde{\mathbf{S}}_b^i (\mathbf{V}^i))\} \quad (1)$$

where $\mathbf{V}^i$ contains the basis for each subspace, $\tilde{\mathbf{S}}_b^i$ is the inter-class scatter matrix, and $\tilde{\mathbf{S}}_w^i$ is the intra-class scatter matrix. We repeat the same steps for all the modes until we get $\mathbf{V}^1$ to $\mathbf{V}^K$. The last step is to compute the Residual Space, $\mathbf{V}^0$, via the Gram-Schmidt algorithm. The Residual Space captures any residual discriminant information present in the training vectors. We combine all these to obtain:

$$\mathbf{V} = [\mathbf{V}^1\, \mathbf{V}^2 \cdots \mathbf{V}^K\, \mathbf{V}^0]. \quad (2)$$

We now describe some useful properties of MMDA.

P1. For mode i, $\mathbf{J}_F(\mathbf{V}^i)$ is equal to $\frac{\lambda_b}{\lambda_w}$, where λ_b and λ_w are the eigenvalues of $\tilde{\mathbf{S}}_b^i$ and $\tilde{\mathbf{S}}_w^i$. Moreover, $\lambda_b + \lambda_w = 1$. $\mathbf{V}^i$ can be obtained when $\frac{\lambda_b}{\lambda_w} = \frac{1}{0} = +\infty$. The dimension of $\mathbf{V}^i$ is $C_i - 1$ and

$$(\mathbf{V}^i)^\top \tilde{x}^i = (\mathbf{V}^i)^\top \tilde{m}_k^i, \quad \forall \tilde{x}^i \in \mathbf{L}_k^i. \quad (3)$$

P2. For $\mathbf{V}^i$ and $\mathbf{V}^j$, $(\mathbf{V}^i)^\top \mathbf{V}^j = 0$, if $i \neq j$.

The proofs of these properties may be found in [18]. From these, we may deduce several important facts:

(a) From P1, the projection of any vector onto $\mathbf{V}^i$ has only $C_i - 1$ coefficients. These are the parameters that control the facial attributes.
(b) Also from P1, after projection, the intra-class variance is zero because $\lambda_w = 0$. This property allows us to capture the essence of “Asian-ness” or “Female-ness” in a single vector of $C_i - 1$ coefficients.
(c) From P2, the $\mathbf{V}$ matrix is orthogonal and thus invertible. This makes MMDA suitable for both analysis and synthesis of face images.

2) Decomposition: Given a new vector $\mathbf{x}$ and trained orthogonal vector $\mathbf{V}$, it may be decomposed using (4) to yield the parameter vector $\mathbf{y}$.

$$\mathbf{y} = \mathbf{V}^\top \mathbf{P}^\top \mathbf{x} \quad (4)$$

Recall from Section III-B.1 that we have $M = 3$ modes (= facial attributes), *i.e.* gender, age and race. Gender can take

[3] We concede that these label choices are somewhat arbitrary. MMDA requires the discretization of each mode into class labels, and more or fewer ages or races is certainly permissible; we chose said labels merely for convenience. Note, however, that MMDA does not currently handle labels with continuous values, so that we cannot treat age as a continuous number.

one of two labels: Male or Female; race has three labels: Caucasian, African, or Oriental; and age is either Young, Middle-aged, or Old.

Decomposing any training face $\mathbf{x}_i$ using (4) to get its parameter vector $\mathbf{y}_i$ gives:

$$\mathbf{y}_i^\top = [\ \underbrace{g_i}_{\text{1 param}}\quad \underbrace{\mathbf{r}_i^\top}_{\text{2 params}}\quad \underbrace{\mathbf{a}_i^\top}_{\text{2 params}}\quad \underbrace{\mathbf{s}_i^\top}_{\text{params}}\ ] \tag{5}$$

From the above properties and deduced facts, we further conclude:

i) The scalar g_i can only take on two constant values: G_1 or G_2, representing Male and Female, respectively.
ii) The 2D vector $\mathbf{r}_i$ can only take on three constants: $\mathbf{R}_1, \mathbf{R}_2$, or $\mathbf{R}_3$, representing Caucasian, African, and Oriental, respectively.
iii) Likewise, the 2D vector $\mathbf{a}_i$ can only take on three constants: $\mathbf{A}_1, \mathbf{A}_2$, or $\mathbf{A}_3$, representing Young, Middle-aged, and Old, respectively.
iv) The remaining vector $\mathbf{s}_i$ controls Residual Space.

In essence, these constant scalars and vectors capture the average appearance of Male, Female, Young, Old, etc. Therefore, to alter the facial attribute of a novel face which is not in the training set $\tilde{\mathbf{x}}$, we simply perform the following:

a) Decompose $\tilde{\mathbf{x}}$ using (4) to get parameter vector $\tilde{\mathbf{y}}$.
b) To change gender, set its gender parameter to either constant: G_1 or G_2.
c) To change age, set its age parameter to one of these constants: $\mathbf{A}_1, \mathbf{A}_2$ or $\mathbf{A}_3$.
d) To change race, set its race parameter to one of these constants: $\mathbf{R}_1, \mathbf{R}_2$ or $\mathbf{R}_3$.
e) Reconstruct using (6) to produce $\tilde{\mathbf{x}}$.

Furthermore, we can control the intensity of the change. For instance, we can set the age parameter to $\sigma\mathbf{A}_3$. By varying σ, we will vary the appearance of Old.

3) Synthesis: Given an altered parameter vector, synthesis is achieved using

$$\tilde{\mathbf{x}} = \mathbf{P}_r\mathbf{V}\mathbf{y} \tag{6}$$

where $\mathbf{P}_r$ reverses the whitening and PCA operation of $\mathbf{P}$. The results in a face encoded vector $\tilde{\mathbf{x}}$: the first D_a values contain the new appearance of the face, while the remaining D_s values are the new coordinates of the facial landmarks. We now re-arrange the first D_a values back into a 2D image, and then apply Thin Plate Spline using the new coordinates to unwarp the face into its new shape.

IV. Visualization

For more insight, let's visualize the column vectors in the MMDA decomposition. Let $\mathbf{Q} = \mathbf{P}_r\mathbf{V}$, then (6) becomes:

$$\tilde{\mathbf{x}} = \underbrace{\begin{bmatrix} | & | & & | \\ \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_{17} \\ | & | & & | \end{bmatrix}}_{\mathbf{Q}} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_{17} \end{bmatrix}}_{\mathbf{y}} \tag{7}$$

Figure 6(a) displays these 17 column vectors as images. As described in Section III-B.2, the first 5 parameters control the gender, race, and age attributes, respectively. Therefore, the corresponding columns $\mathbf{q}_1, \ldots, \mathbf{q}_5$ should somehow encode these attributes as well. This is clearly evident in the first row of Figure 6(a). The remaining faces encode information in the Residual Space, containing identity, and any remaining image variation.

Hence, we term these vectors *Semantic Faces*, since they capture the semantics of facial attributes. Furthermore, these vectors form a *Semantic basis* for the set of face images: taking arbitrary linear combinations of these vectors will generate faces with different amounts of gender, race, and age attributes. To further illustrate this, consider the vector $G_1\mathbf{q}_1$. This creates the average Male face (Figure 6(b)). Changing this to $G_2\mathbf{q}_1$ creates an average Female face. Increasing the scalar to $1.5G_1$ or $1.5G_2$ makes the face more Masculine or Feminine, respectively.

For race, the three races, Caucasian, African and Oriental, are encoded by two Semantic Faces: $\mathbf{q}_2, \mathbf{q}_3$. By taking linear combinations of these, we may generate the three said races, or increase the racial "intensity" (*e.g.* appear more African). We may also create mixed races (Figure 7).

Finally, we may visualize the subspace for age in a similar way. Like race, age is controlled by two Semantic Faces: $\mathbf{q}_4, \mathbf{q}_5$. Taking appropriate linear combinations will synthesize faces with different age appearances. We omit this figure due to space constraint.

To summarize, applying MMDA onto a set of labeled face images yields a *semantic basis* with which we may decompose a face into its gender, race and age attributes. This basis also allows us to synthesize new faces with desired new attributes. We thus have a discriminability and flexibility not found in other subspace decomposition techniques, such as AAM [8], 3D Morphable model [5], Eigenfaces[19], Fisherfaces [3], Laplacianfaces [12] and Tensorface [21].

V. Privacy Protection

A. Identity Alteration

How do we alter identity? There does not seem to be any explicit identity parameter in (5). The answer lies in $\mathbf{s}_i$, the parameter that controls Residual Space. Identity is a tricky concept. It includes gender and race, but not age. For instance, we can talk about an older or younger Michael Jackson, but there is no such thing as a Female Michael Jackson, or an Oriental Michael Jackson. In terms of MMDA, identity is a label that cannot be applied on a vector independently of race or gender. Thus there is no identity parameter in MMDA. This is not a flaw of MMDA; rather, it is a reflection of the fact that identity is not independent of race or gender. Thus the only parameter left for alteration is $\mathbf{s}_i$. MMDA theory assures us that Residual Space captures any remaining discriminant information that is not captured in the other subspaces. This would include identity, illumination, facial expression, and so on. Since our training images are all frontally illuminated, frontal faces with neutral expression, the main thing captured in Residual Space must be identity. Hence, to alter identity, we may simply replace $\mathbf{s}_i$ with,

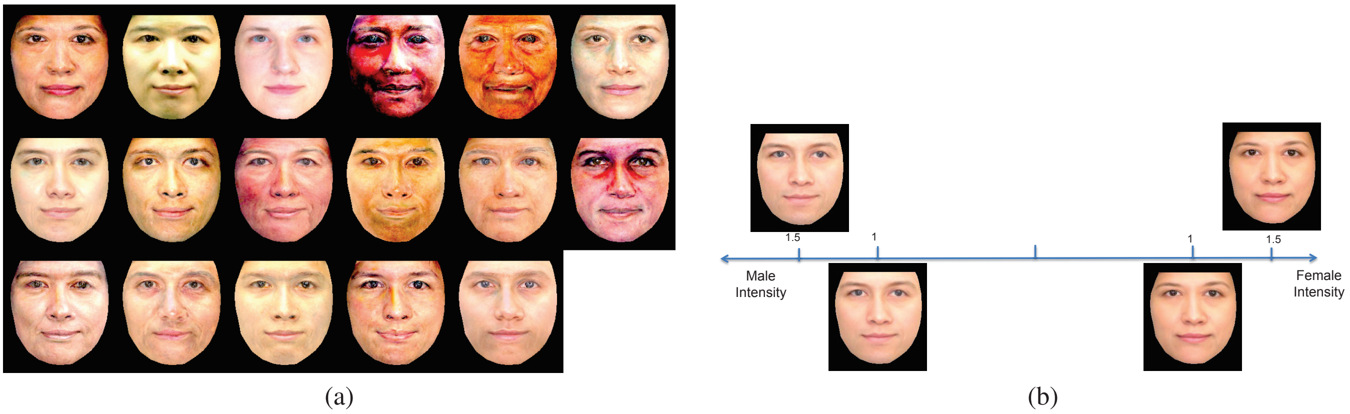

Fig. 6. (a) Semantic faces: the columns $\mathbf{q}_1, \ldots, \mathbf{q}_{17}$ in (7) visualized as images. The first row images appear to encode gender (1^{st} face), race (2^{nd} and 3^{rd} faces), and age (4^{th} and 5^{th} faces), respectively. The other 12 encode identity and any residual image variation. (b) By changing the sole parameter of $\mathbf{q}_1$ (the gender Semantic Face), Male (left) and Female (right) faces are synthesized. Varying this parameter also makes the face appear more or less Masculine (or Feminine).

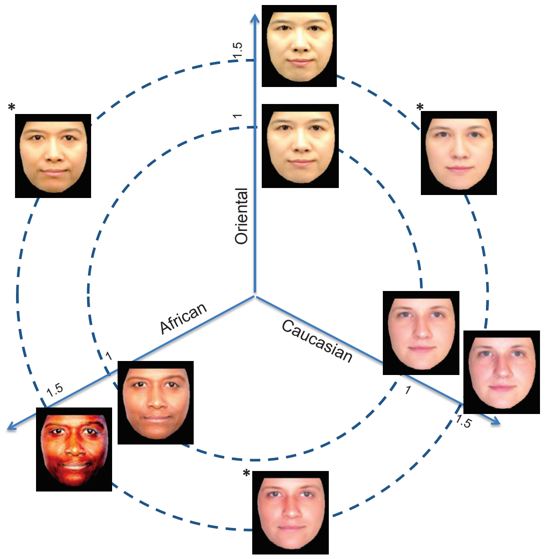

Fig. 7. The three races of Caucasian, African and Oriental are encoded along the three blue axes in a 2D subspace spanned by $\mathbf{q}_2, \mathbf{q}_3$ (the Semantic Faces for race). By taking suitable linear combinations of these vectors, we may intensify the racial appearance, and even synthesize new, mixed races.

say, $-\sigma \mathbf{s}_i$. This "reverses" identity, with σ controlling the intensity of reversal. This is how Figure 2 was produced.

Finally, note that altering identity need not be an all-or-nothing affair. We could exaggerate or diminish one's identity simply by increasing or decreasing the norm of $\mathbf{s}_i$. This allows for more subtle control of identity.

B. Privacy Guarantees

It is clear that the above method to alter identity by negating the Residual Space parameter, s, is easily reversible. This defeats our privacy protection. To truly protect privacy, we can employ any of the following schemes:

1. Add a random vector $\mathbf{w}$ to s. This makes the identity alteration non-deterministic and thus irreversible.

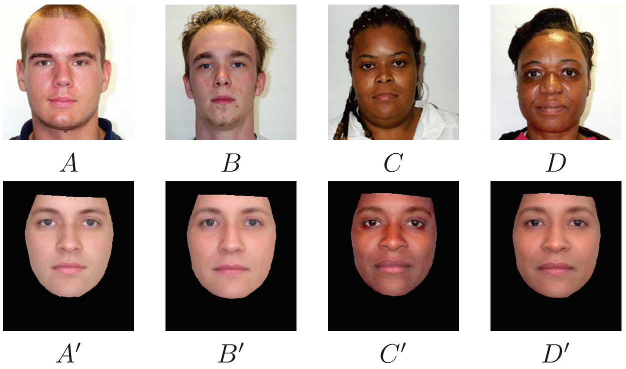

Fig. 8. Irreversible identity alteration: A and B are clearly 2 different persons with the same sex, race and age. Their identity parameter is set to $\mathbf{0}$, yielding A' and B' respectively, which look alike. Similarly for C, D and C', D'. Face++ confirms that A' and B' are the same person, as are C' and D'; however, A' and C' are different people.

2. Always set s to $\mathbf{0}$. Figure 8 shows this procedure, demonstrating that we can irreversibly alter identity but retain other attributes. In effect, this achieves ∞-anonymity, an extreme form of k-anonymity. In general, to k-anonymize a set of M faces: first, group them into $L = \lfloor M/k \rfloor$ groups; in each group, compute the average identity parameter of all faces in that group; then replace each face's identity parameter with its group average.
3. Set s to one of L pre-defined vectors in a uniformly random manner. This achieves L-diversity. These L vectors may be chosen by sampling Residual Space.

Indeed, by viewing the parameter vector $\mathbf{y}$ of (5) as a row in a relational table, the protection mechanisms of k-anonymity and L-diversity may be readily applied to *all* other facial attributes as well. Likewise, any desired probability distribution on the facial attributes can also be easily enforced, *e.g.* ensuring that 10% of synthesized faces be African, 15% be Oriental, *etc..*. This fulfills the t-closeness criterion [13]. In other words, our method can provably guarantee the privacy of altered faces.

VI. Experiments & Discussion

How good is our face alteration? More precisely,

Q1. When we alter a facial attribute, say, gender, is it effective?

Q2. When we alter one facial attribute but retain others, are the unchanged attributes perceived as such?

Q3. Does increasing the intensity of a parameter manifest in a corresponding increase in the attribute?

Q4. When we alter identity, is it effective?

To answer these questions, we will use a Change Detector (CD), *i.e.* a vision algorithm, to compare an original face image with its altered image. This is in line with our motivation to protect privacy while allowing visual analytics (*i.e.* other computer vision algorithms) to function normally. In all our experiments, we use a set of test images that are different from our training set.

A. Evaluation metric

We build several CDs, one each for identity, gender, race and age. Each CD accepts two inputs, an original face image plus its altered version, and outputs "Changed" if it judges that the two images differ in that particular attribute; otherwise it outputs "Unchanged". All our CDs are built using Face++, a commerical face attribute classifier. Once we have our CDs, it would be a simple matter to run them on our test cases. This would give the *raw performance* of our method. However, our CDs are not perfect; they make errors. How then can we be sure that the judgment of our CD is a true statement of our method, and not an inherent error? We proceed as follows. Let β be the probability that our method correctly alters a facial attribute. Also define:

1. t_p to be the true positive probability, *i.e.* when the input image pair correctly differ in an attribute and the CD says "Changed".
2. f_p to be the false positive probability, *i.e.* when the input image pair do not differ in an attribute but the CD says "Changed".
3. α to be the observed changed rate,*i.e.* the fraction of time the CD outputs "Changed".

From these definitions, we may derive:

$$\beta = \frac{\alpha - f_p}{t_p - f_p} \tag{8}$$

Equation (8) allows us to compensate the raw performance of our method (α) from the inherent errors of the CD (t_p, f_p). To estimate the inherent errors, we generated a separate ground-truth set of about $130K$ test images. The true positive rates (t_p) for gender, race and age CDs are estimated to be 0.90, 0.93 and 0.83, respectively; while the false positive rates (f_p) are 0.11, 0.16 and 0.30, respectively. This shows that all CDs have good performance.

B. Experiments on single attribute change

To answer questions Q1 and Q2, we changed one facial attribute while retaining the other two. We generated between $18K$ and $71K$ test image pairs. All three CDs were then used to measure α, after which we use (8) to derive β.

TABLE I

Actual "Changed" Rate (β) values for single-attribute change. These show that our method is effective in changing an attribute (bold values), and also in retaining an attribute (values in normal font).

	Intensity σ	0.5	1.0	1.5	2.0	2.5
Gender Change	Gender CD	**0.50**	**0.63**	**0.75**	**0.88**	**1.00**
	Race CD	0.31	0.31	0.31	0.31	0.31
	Age CD	0.56	0.19	0.38	0.38	0.19
Race Change	Gender CD	0.38	0.31	0.44	0.38	0.31
	Race CD	**0.57**	**0.64**	**0.70**	**0.76**	**0.89**
	Age CD	0.19	0.19	0.47	0.66	0.28
Age Change	Gender CD	0.31	0.38	0.31	0.38	0.19
	Race CD	0.25	0.25	0.12	0.12	0.25
	Age CD	**0.66**	**0.84**	**1.00**	**1.00**	**1.00**

Looking at the "Gender Change" rows in Table I, we can see that when gender (only) is changed, the Gender CD reported consistently higher β values at all intensity levels (in bold) than the other CDs. This means that the perceived gender in the face is indeed changed, while its race and age remain unchanged. The other rows show the performance for race-only, and age-only changes. In other words, our method is effective in changing facial attributes, as well as in retaining attributes. The β values that ought to be high are indeed high, while those that ought to be low are low.

C. Experiments on multiple attribute change

We next examine the effect of changing two or more facial attributes. Table II summarizes the β values. Again, the conclusion is that altered attributes are effectively manifested in the image, and detected as such by the Change Detectors. Likewise, any unaltered attribute is usually detected as unchanged by the corresponding CD. Finally, increasing the intensity does increase the β values, as expected.

The results are not perfect, however. The entry in italics (0.68) show that changing both gender and race at a high intensity ($\sigma = 2.0$) appear to cause a change in age as well.

D. Experiments on identity change

In fact, identity change can be easily observed in our experimental results. To validate this, we asked 5 volunteers to compare the identities in an original image and its altered image. All volunteers confirmed the identity change for all test images. This is expected. Besides human evaluation, we also conducted an experiment by using Face++ matcher to test the identity change. We randomly altered several hundred samples at four intensity levels. Figure 2 shows two examples. In all these tests, the identity CD output "Changed". That is, $\alpha = 100\%$. If we scrutinize the confidence values returned by Face++, we see that at all levels of intensity, these values are very high (Table III). This shows that our method succeeded in changing facial identity.

E. Discussion

1. From all these experiments, we conclude that our method is effective in altering the facial attributes of gender, race, age, and identity, whether singly or in different

TABLE II

ACTUAL "CHANGED" RATE (β) FOR MULTI-ATTRIBUTE CHANGE.

	Gender+Race			Gender+Age			Race+Age			All 3 attributes		
Intensity σ	Gender	Race	Age	Gender	Race	Age	Gender	Race	Age	Gender	Race	Age
1.0	0.45	0.52	0.23	0.55	0.26	0.71	0.39	0.54	0.77	0.36	0.47	0.72
2.0	0.60	0.60	*0.68*	0.53	0.25	0.82	0.33	0.72	0.88	0.50	0.71	0.82

TABLE III

AVERAGE CONFIDENCE VALUES RETURNED BY FACE++ AT DIFFERENT INTENSITIES OF IDENTITY CHANGE.

Intensity σ	0.5	1.0	1.5	2.0
Average Confidence	0.9137	0.943	0.951	0.966

combinations. Question Q4 is answered in the affirmative by Table III; while Q1, Q2 and Q3 are all answered in the affirmative by Tables I and II.

2. We could not compare with existing works because ours is the first to selectively alter some facial attributes while retaining other attributes. There is no prior work to compare to.
3. We cannot claim to have solved the problem of protecting privacy in videos. This is because in real videos, faces may not always be frontal. Our method currently works on frontal faces only. Moreover, there is the problem of hair (head and facial hair), clothes, and accessories such as jewelry, handbags, etc. All these contrive to reveal the identity of the person, even if the face is unrecognizable. We acknowledge that there are still numerous challenges ahead, and ours is but a first step.

VII. CONCLUDING REMARKS

We are pleased to present the novel concept of Controllable Face Privacy for the nuanced protection of face images. Applying multimodal discriminating analysis on our face encoding scheme results in a *Semantic basis* with which we may decompose a face into its gender, race and age attributes. In turn, this permits the synthesis of novel faces with new, desired attributes. Moreover, privacy protection mechanisms, such as k-anonymity, L-diversity, t-closeness, are easily incorporated into our method, thereby providing *provable* guarantees on our altered faces. In the near future, we intend to get human volunteers to assess the quality of our altered images.

We thank Dr. Shih-Tung Ngiam for his insightful advice on how to estimate the actual change rate (β) from the observed rate (α), and Dr. Ee-Chien Chang for the initial impetus to embark on this research, and for his on-going engagement with us.

REFERENCES

[1] P. Agrawal and P. Narayanan. Person de-identification in videos. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21(3):299–310, 2011.

[2] T. Atlas and G. Stohr. Surveillance Cameras Sought by Citites After Boston Bombs. *Bloomberg News*, 2013.

[3] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[4] A. M. Berger. Privacy mode for acquisition cameras and camcorders, May 23 2000. US Patent 6,067,399.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[6] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6), 1989.

[7] D. Chen, Y. Chang, R. Yan, and J. Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[8] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[9] C. Francescani. NYPD expands surveillance net to fight crime as well as terrorism. *Reuters News*, 2013.

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

[11] R. Gross, L. Sweeney, F. De La Torre, and S. Baker. Semi-supervised learning of multi-factor models for face de-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

[12] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):328–340, 2005.

[13] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, volume 7, pages 106–115, 2007.

[14] M. Mrityunjay and P. Narayanan. The de-identification camera. In *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2011 Third National Conference on*, pages 192–195. IEEE, 2011.

[15] C. Neustaedter and S. Greenberg. The design of a context-aware home media space for balancing privacy and awareness. In *UbiComp 2003: Ubiquitous Computing*, pages 297–314. Springer, 2003.

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

[17] Risen, James and Poitras, Laura. N.S.A. Collecting Millions of Faces From Web Images. *New York Times*, May 2014.

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

[19] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[20] Unknown. Spy Britain: six million CCTV cameras - and most are in private hands. *London Evening Standard*, 2013.

[21] M. A. O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Computer VisionECCV 2002*, pages 447–460. Springer, 2002.

[22] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433. ACM, 2005.

[23] J. Williams. In-Store Cameras: From Security Aids to Sales Tools. *Insead Knowledge*, 2013.

[24] X. Yu, K. Chinomi, T. Koshimizu, N. Nitta, Y. Ito, and N. Babaguchi. Privacy protecting visual processing for secure video surveillance. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 1672–1675. IEEE, 2008.