Retaining Expression on De-identified Faces

Li Meng[1] and Aruna Shenoy[2]

[1] University of Hertfordshire, Hatfield, AL10 9AB, United Kingdom
[2] National Physical Laboratory, Teddington, United Kingdom
[1] L.1.Meng@herts.ac.uk
[2] aruna.shenoy@npl.co.uk

Abstract. The extensive use of video surveillance along with advances in face recognition has ignited concerns about the privacy of the people identifiable in the recorded documents. A face de-identification algorithm, named k-Same, has been proposed by prior research and guarantees to thwart face recognition software. However, like many previous attempts in face de-identification, k-Same fails to preserve the utility such as gender and expression of the original data. To overcome this, a new algorithm is proposed here to preserve data utility as well as protect privacy. In terms of utility preservation, this new algorithm is capable of preserving not only the category of the facial expression (e.g., happy or sad) but also the intensity of the expression. This new algorithm for face de-identification possesses a great potential especially with real-world images and videos as each facial expression in real life is a continuous motion consisting of images of the same expression with various degrees of intensity.

Keywords: Privacy Protection, Face De-Identification, Facial Expression Preservation, Linear Discriminant Analysis, k-Anonymity.

1 Introduction

Recent advances in both camera technology and computing hardware have highly facilitated the effectiveness and efficiency of image and video acquisition. This capability is now widely used in a variety of scenarios to capture images of people in target environments, either for immediate inspection or for storage and subsequent analysis/sharing [1]. These improved recording capabilities, however, has ignited concerns about the privacy of people identifiable in the scenes. The Council of Europe Convention of 1950 formally declared privacy protection as a human right. This was later embodied in the 1995 Data Protection Directive of the European Union (Directive 95/46/EC) and the 2016 General Data Protection Regulation (GDPR Regulation (EU) 2016/679). Both regulations demand the deployment of appropriate technical and organizational measures to protect private information in the course of transferring or processing such data. This legal requirement along with ethical responsibilities has restricted data sharing and utilization while various organizations may require the use of such data for research, business, academic, security and many other purposes. To comply with the regulations, de-identification has become the focus of attention by many organizations with the ultimate goal of removing all personal identifying information while protecting the utility of the data.

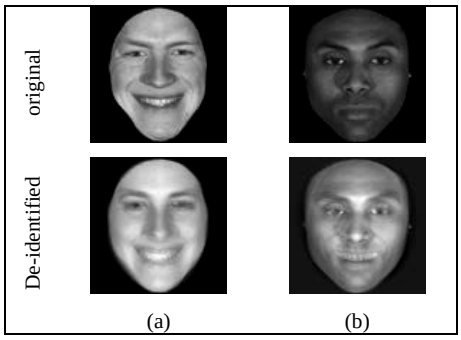


Fig. 1. Risk of losing data utility with the *k*-Same algorithm: (a) loss of gender and (b) loss of expression.

Various methods have been proposed for the de-identification of faces in still images. These methods can be put into two categories: the ad hoc methods (such as masking, pixelation and blurring [2-4]) and the *k*-anonymity based methods (such as *k*-Same [5]). The ad hoc methods are usually simple to implement. However, these methods significantly distort the integrity of the image data. Imagine the eye area being blacked out by masking methods or the resolution of the image being sacrificed by pixelation or blurring. Even worse, ad hoc methods fail to serve their purpose as they are unable to thwart the existing face recognition software [5, 6]. To achieve privacy protection, the concept of *k*-anonymity was introduced by Sweeney in 2002 [7]. All *k*-anonymity based methods de-identify a face image by replacing it with the average of *k* faces from a gallery and hence achieve privacy protection by guaranteeing a recognition risk lower than 1/*k*. Among the *k*-anonymity methods, the most widely used method is *k*-Same [5]. However, *k*-Same was not designed for preserving data utility. As a result, the de-identified version of a male face might look feminine (Fig. 1(a)) and a neutral face might put on a smile (Fig. 1(b)). The work presented in this paper is an extension to the *k*-Same method. In addition to privacy protection, consideration has been taken for retaining the facial expression of the original image and the intensity of the expression.

The next section introduces the benchmark algorithms that support the method proposed in this work. Section 3 defines our method. Section 4 describes the face image database used in this work, gives an overview of the approaches used in the experiments and presents the results. Finally, the findings of this work are concluded in Section 5, with a discussion on the general applicability of this work and some proposals for further work.

2 Benchmark Methods

2.1 Principal Component Analysis and Face Recognition

Principal Component Analysis (PCA) is a benchmark method for the unsupervised reduction of dimensionality [8]. It has been widely used in the global approach to face recognition [9], where *D*-dimensional pixel-based face images are projected into a *d*-dimensional PCA subspace called the facespace (typically $d << D$). The goal of PCA is to reduce the dimensionality of the face images while retaining as much as possible of

the variation present in them. In face recognition, the PCA projection along with the face space are established through training a set of face images and PCA achieves its goal by projecting these training images along the directions where they vary the most. Fig. 2 presents the general PCA-based training process of a face recognition system. Typically, the face space is defined based on the eigenvectors of the covariance matrix corresponding to the largest eigenvalues. The magnitude of the eigenvalues corresponds to the variance of the data along the eigenvector directions. In this work, all eigenvectors with a nonzero eigenvalue are kept to avoid losing information on data utility.

In face recognition, each PCA eigenvector is a face image. These eigenvectors are therefore named Eigenfaces. Fig. 3 displays the top two and the last two Eigenfaces used in this work. The image set used for computing/training these Eigenfaces contains both neutral and smiley faces.

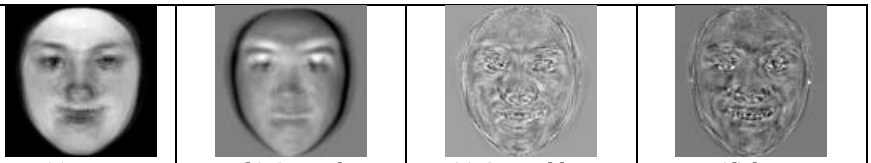
PCA-based face recognition, also known as the Eigenfaces technique [9], projects a probe face image Γ into the Eigen face space using (1) and matches faces there based on the Euclidean distance.

$$\text{projected probe face: } \Omega = \mathbf{V}^T(\Gamma - \bar{\Gamma}) \tag{1}$$

Let $\mathbf{H}$ be the training set of M face images and every image Γ_i in $\mathbf{H}$ be a $N \times 1$ vector. Perform the following steps:

1) $\bar{\Gamma} = \frac{1}{\mathrm{M}}\sum_{i=1}^{M} \Gamma_i$
2) For each $\Gamma_i \in \mathbf{H}$, $\Phi_\mathrm{i} = \Gamma_\mathrm{i} - \bar{\Gamma}$
3) Form the matrix $\mathbf{A} = [\Phi_1, \dots, \Phi_\mathrm{M}]$, then compute the covariance matrix $\mathbf{C} = \mathbf{A}\,\mathbf{A}^T$ (covariance matrix $\mathbf{C}$ characterizes the scatter of the face images in $\mathbf{H}$)
4) Compute the eigenvalues $\lambda_1, \dots, \lambda_N$ of $\mathbf{C}$ such that $|\mathbf{C} - \lambda_i| = 0$ for $i = 1, \dots, N$
5) Sort eigenvalues in descending order, i.e. $\lambda_1 \geq \cdots \geq \lambda_N$ for $i = 1, \dots, N$
6) Compute the eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_N$ of $\mathbf{C}$ such that $\mathbf{C}\mathbf{v}_i = \lambda_\mathrm{i}\mathbf{v}_i$ for $i = 1, \dots, N$
7) Select the top N' eigenvectors such that
$$\lambda_i \neq 0 \text{ for } i = 1, \dots, N'$$
8) Form the matrix $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_{N'}]$. Construct the eigen face space by projecting the training images to the PCA subspace defined by $\mathbf{V}$:
$$\mathbf{facespace}_i = \mathbf{V}^T(\Gamma_i - \bar{\Gamma}) \text{ for } i = 1, \dots, M$$

Fig. 2. Training process of a PCA-based face recognition system.



(a) First (b) Second (c) Second last (d) last

Fig. 3. The first two and the last two Eigenfaces used in this work.

2.2 Linear Discriminant Analysis and Classification of Facial Expression

Face data have multiple attributes and individual face images can be grouped into classes according to attributes such as age or gender. Although PCA is effective in terms of maximizing the scatter among individual face images, it ignores the underlying class structure. As a result, the projection axes chosen by PCA might not provide good discrimination power for classification purposes.

To this problem, Linear Discriminant Analysis (LDA) or Fisher Linear Discriminant (FLD) analysis [10, 11] seems to be the perfect solution as it maximizes the scatter between image classes while minimizing the scatter within the classes. The steps involved in the LDA process are presented in Fig. 4. With face data, $\mathbf{S}_{\mathrm{W}}$ is often singular since image vectors are of large dimensionality while the size of the data set is much smaller. To alleviate this problem, typically the original face images are first projected into the PCA space to reduce dimensionality. LDA is then applied to find the most discriminative directions. In this work, $\mathbf{x}_i$ is the PCA projection of face image Γ_i. The eigenvectors obtained from LDA are called Fisherfaces. In the cases with two classes, for example our work, the corresponding eigenvalues will have only one nonzero value and therefore only the top Fisherface is kept and used for projecting data into the Fisher face space.

LDA can be used to estimate various attributes of the face, for example expression, gender, age, identity and race, etc. In this work, LDA has been used to identify the expression on a face as either 'neutral' or 'smiley' and evaluate the intensity of the expression identified. Next section presents more detail on how LDA is utilized in the proposed algorithm.

2.3 *k*-Same for Face De-identification

Introduced for preserving privacy [5], *k*-Same is based on the *k*-anonymity framework of Sweeney's [7]. It guarantees that each de-identified face image could be representative of *k* faces and therefore limit the recognition risk of the de-identified faces to $1/k$.

1) For M $N \times 1$ samples $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, C classes $\{\mathbf{X}_1, \dots, \mathbf{X}_C\}$, calculate the average $\boldsymbol{\mu}_i$ for each class i along with the total average $\boldsymbol{\mu}$.
2) Calculate the scatter $\mathbf{S}_i$ for each class i as
$$\mathbf{S}_i = \sum_{\mathbf{x}_k \in \mathbf{X}_i} (\mathbf{x}_k - \boldsymbol{\mu}_i)(\mathbf{x}_k - \boldsymbol{\mu}_i)^T$$
3) Calculate the within-class scatter as
$$\mathbf{S}_{\mathrm{W}} = \sum_{i=1}^{C} \mathbf{S}_i$$
4) Calculate the between-class scatter as:
$$\mathbf{S}_{\mathrm{B}} = \sum_{i=1}^{C} |\mathbf{X}_i| (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$
5) Compute the matrix $\mathbf{C} = \mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}$.
6) Compute the eigenvalues $\lambda_1, \dots, \lambda_N$ of $\mathbf{C}$ such that $|\mathbf{C} - \lambda_i| = 0$ for $i = 1, \dots, N$
7) Compute the eigenvectors (Fisherfaces) $\mathbf{v}_1, \dots, \mathbf{v}_N$ of $\mathbf{C}$ such that $\mathbf{C}\mathbf{v}_i = \lambda_{\mathrm{i}}\mathbf{v}_i$ for $i = 1, \dots, N$.

Fig. 4. General procedure of LDA.

In [5], Newton introduced two versions of the *k*-Same algorithm, namely *k*-Same-Pixel and *k*-Same-Eigen. Both versions find the *k* closest faces to the probe in the PCA face space, while the former returns the pixel-wise average of the *k* closest and the later performs the averaging in the PCA facespace. Compared to *k*-Same-Pixel, *k*-Same-Eigen brings an extra blurring effect which contributes to the reduction of ghost artifacts in the de-identified face. Considering this, this work follows the approach of *k*-Same-Eigen. For more in-depth details of the *k*-Same algorithm, refer to [5].

3 The New Algorithm – *k*-SameClass-Eigen

The *k*-Same algorithms ignores utility features of the face images (such as gender, age, and expression) and searches for the *k* closest faces merely based on the appearance. As displayed in Fig. 1, the utility information are often lost. To address this problem, consideration has been taken in this work for retaining the utility of the original face. Although this work focuses merely on facial expressions, the same principal can be applied to the preservation of other utilities of face images.

Inspired by [12], the algorithm proposed here is an extension to *k*-Same, where the *k* closest faces are selected only among the gallery faces with the same expression to the probe image. However, unlike [12] where the classification of facial expression is achieved using a support vector machine, our algorithm uses LDA. One advantage of LDA is that it is able to not only classify the expression but also evaluate the intensity of the expression. The work in [12] focused on the preservation of merely the class label of a given face image (e.g. male or female, young or old). While we aim to preserve both class label and intensity here. In this work, the LDA Fisher space is trained to classify two expression classes {neutral, smiley}. As shown in Fig. 5, the LDA projection of all the gallery neutral faces used in this work has an average of 4.2 with a small variance of 0.004; while the average LDA projection of the smiley faces is -4.2 with a small variance of 0.011. Results in Fig. 5 suggest that the LDA projection can be used as the classifier of facial expressions and the measure of expression intensity.

Furthermore, through changing the value of the LDA projection, fine adjustment of facial expression has been achieved in this work. As mentioned in the previous section, for a two-class problem only the top Fisherface is available for the LDA projection. Fig. 6 displays the Fisherface used in this work. The number of components in the Fisherface equals the number of Eigenfaces used in our PCA projection. As the facial expression might be encoded by the last few Eigenfaces (refer to Fig. 3), all Eigenfaces with a non-zero Eigenvalue are kept in this work. As shown in Fig. 6, the Fisherface has a dominant component (the sixth component). In other words, the expression on a face is mainly determined by the value of this component. For this reason, this dominant component is named the 'expression index'. In this work, the expression on a face or the intensity of the expression is adjusted through changing the value of the expression index while keeping the value of other Fisherface components unchanged. Given a target expression intensity d, the value of the expression index $v(6)$ is changed to:

$$v(6)_{\text{new}} = d - \sum_{i=1}^{N} v(i) + v(6)_{\text{current}} \quad (2)$$

As the algorithm proposed in this work selects the closest faces from a specific class in the gallery, we name it as *k*-SameClass-Eigen.

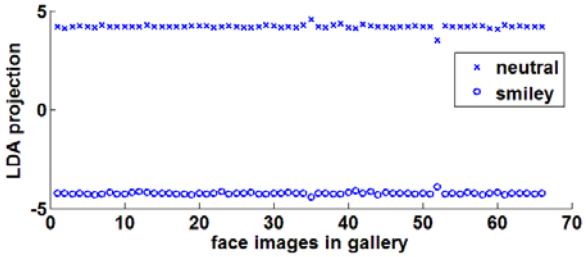


Fig. 5. LDA projection of the gallery images used in this work.

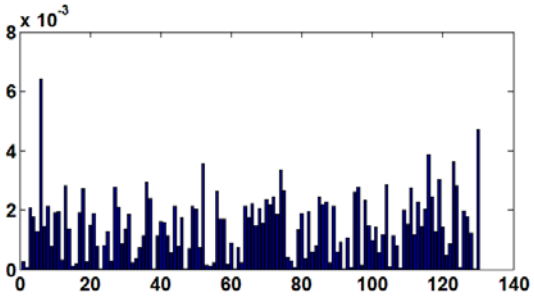


Fig. 6. The Fisherface used in this work.

4 Experiments

4.1 Image Database: Binghamton

Our experiments have been carried out with the 95×95 gray scale images from the Binghamton database [13]. The face gallery in our experiments contains 132 Binghamton images (33 images from each of the following four classes: Female Neutral, Female Smiley, Male Neutral, and Male Smiley. All images in the gallery are used for training both the PCA and the LDA spaces.

4.2 Evaluation of Privacy

There are two types of probe image sets, 'seen' and 'unseen', with each set containing 20 images randomly selected from the Binghamton database. The 'seen' probe image set follows the closed universe model meaning that every face image in this probe set is also a member of the gallery whereas the face images in the 'unseen' probe set are taken from outside the gallery. Visual results of our *k*-SameClass-Eigen algorithm are displayed in Fig. 7 for k = 2, 5, 10, and 20 from left to right. Fig. 7(I)

displays the results for examples of seen probe faces and Fig. 7(II) displays the results for examples of unseen probe faces.

To evaluate the privacy protection power of the proposed algorithm *k*-SameClass-Eigen, the de-identified probe images are matched to the original images in the gallery. The privacy protetion performance is measured in terms of the average recognition risk of the de-identified face images. For the seen probe images, the recognition risk is the percentage at which a de-identified face is recognized as its own original. For the unseen probe images, it is the percentage of a de-identified face being recognised as the gallery image that is closest to the original probe. For both probe sets, close-set identification has been performed, i.e., the closest face from the gallery is always returned as the best match despite how large the closest distance is. To exam the impact of *k*, the level of *k* has been varied between 1 and 20. Figs. 8(a) and 8(b) show the recognition risk for the de-identified seen and unseen probe faces, respectively.

In both Fig. 8(a) and 8(b), the recognition risk decreases with the increase of *k*-level and the recognition risk tends to remain lower than the theoretical maximum of *1/k*. The zig-zags presented in Figs. 8(a) and 8(b) are due to the relatively small size of each probe set, which can be improved by introducing more probe images. Nevertheless, the recognition risk converges to a value below 1/*k* in both Figs. 8(a) and 8(b) despite the fact of a small probe set.

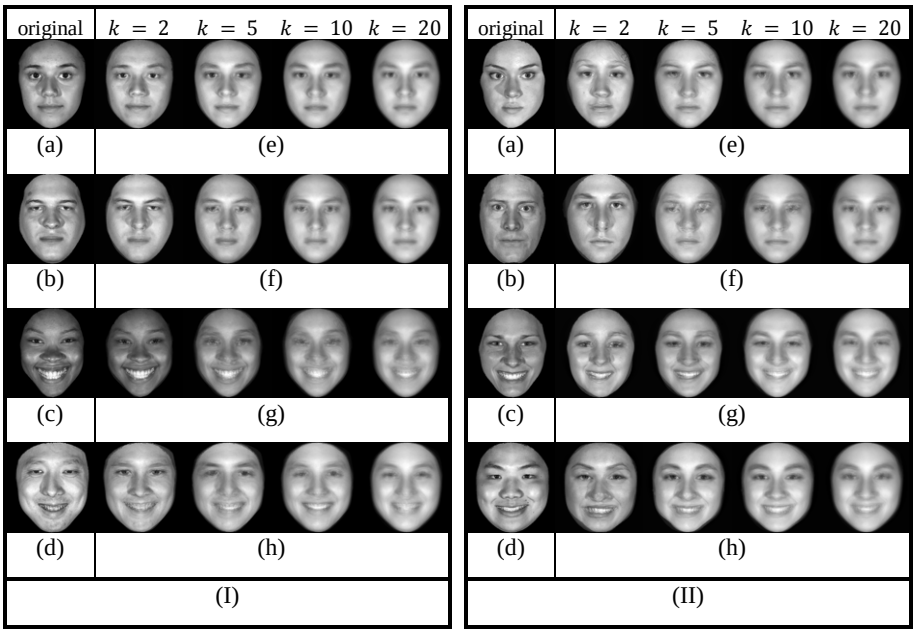
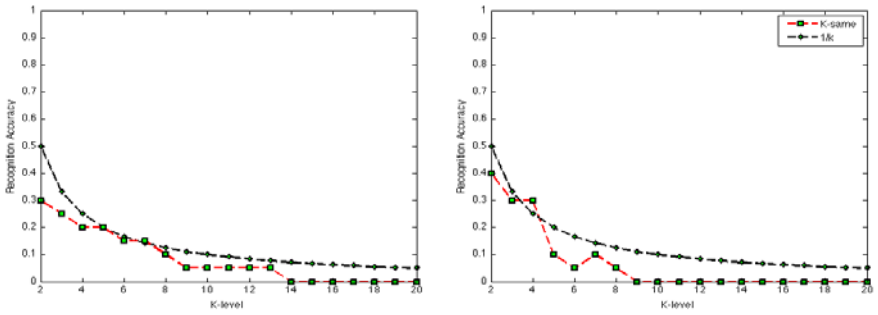


Fig. 7. Examples of (I) original seen and (II) original unseen face images as well as their de-identified faces generated by the method of *k*-SameClass-Eigen for $k = 2$, 5, 10, and 20 from left to right: (a) original female neutral, (b) original male neutral, (c) original female smiley, (d) original male smiley, and (e) - (h) de-identified results for (a) - (d).



(a) (b)

Fig. 8. PCA recognition risk of the de-identified probe images against the original gallery images. (a) Probe images are selected from the gallery. (b) Probe images are from outside the gallery.

4.3 Evaluation of Data Utility

In order to evaluate the algorithm's power of retaining data utility, this work measures the rate at which the same expression is measured from a de-identified face image as from its original image. Again, both seen and unseen probe sets are used in the experiments with each set containing 20 randomly selected face images from within and outside the gallery, respectively. The results are presented in Fig. 9. The proposed face de-identification algorithm *k*-SameClass-Eigen can always retain the expression on a seen (or known) probe face and therefore delivers a perfect expression accuracy. For the unkown probe faces, the accurary is between 80% and 95% with an average of 86%. The lower expression accuracy with the unknown faces is due to the fact that the LDA has been trained with the known faces in the gallery and the LDA projection obtained in this way may fail to correctly classify an unknown face. When the expression on a unknown face has been incorrectly classified, the same incorrect expression with the measured expression intensity will be imposed onto the de-identified face by the *k*-SameClass-Eigen method using (2).

4.4 Evaluation of the Ability to Change Expression Intensity

In order to evaluate the algorithm's ability to adjust expression intensity and the visual quality of the result images, experiments are conducted in this work where the expression intensity on the de-identified neutral faces is continuously varied between 5 (completely neutral) and -5 (very happy). The range of the expression intensity is defined following the results of LDA training (refer to Fig. 5). Fig. 10 displays the visual results for (I) an example male face and (II) an example female face, respectively. As displayed in Fig. 10, the proposed algorithm has the ability to switch between facial expressions and continuously adjust the intensity of the expression. Furthermore, the visual quality of the result images remains good across the various expression intensity values.

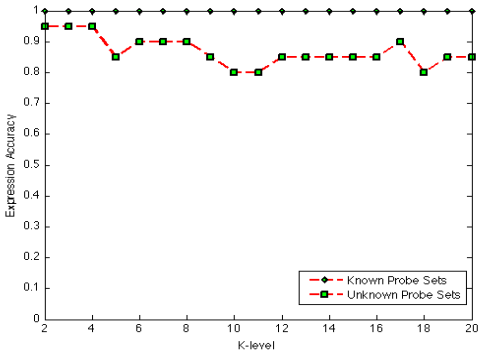


Fig. 9. Expression accuracy for both of the probe images.

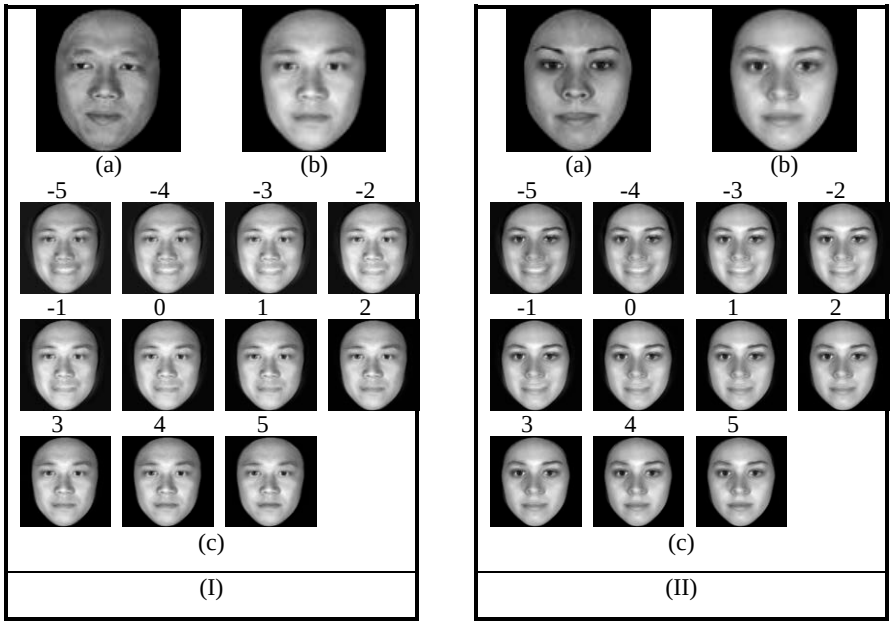


Fig. 10. Results of (I) an example male face and (II) an example female face. (a) original, (b) de-identified, and (c) mutations of (b) for various expression intensities (intensity values are given above the corresponding face image).

5 Discussion and Conclusions

A new face de-identification algorithm *k*-SameClass-Eigen has been proposed in this paper with a goal to preserve privacy as well as retain both the facial expression class and the expression intensity. Experimental results show that it is able to limit the recog-

nition risk to below 1/*k*. Furthermore, it can always retains the expression on a face image as long as the expression has been measured correctly by the LDA classifier. In practice, the accuracy of the LDA classifier can be enhanced by the use of a larger training set. Finally, *k*-SameClass-Eigen is capable of changing the expression intensity on a face to any value within the valid range. As facial expression is naturally a continuous motion presenting various degrees of intensity, the proposed algorithm has a great potential with the de-identification of real-world images and videos.

References

1. L. Sweeney, Surveillance of Surveillances camera watch project, http://dataprivacylab.org/dataprivacy/projects/camwatch, 2005.
2. J. Crowley, J. Coutaz, F. Berard, "Perceptual user interfaces: things that see," Communications of the ACM, vol. 43, pp. 54–64, 2000.
3. C. Neustaedter, and S. Greenberg, "Balancing privacy and awareness in home media spaces," Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, in conjunction with the 5th Int'l Conf. Ubiquitous Computing (UBICOMP), Seattle, WA , 2003.
4. M. Boyle, C. Edwards, and S. Greenberg, "The effects of filtered video on awareness and privacy," Proc. ACM Conf. Computer Supported Cooperative Work, 2000.
5. E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," IEEE Trans. Knowledge and Data Eng., vol. 12, no. 2, pp. 232 – 243, February 2005.
6. C. Neustaedter, S. Greenberg, and M. Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," ACM Trans. Computer Human Interactions (TOCHI), vol. 13, issue 1, March 2006.
7. L. Sweeney, "k-Anonymity: a model for protecting privacy," Int'l J. Uncertainty, Fuzziness, and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.
8. I. T. Jolliffe, Principal Component Analysis, 2nd ed., New York: Springer-Verlag New York, Inc, 2002.
9. M. Turk, and A. P. Pentland, "Eigenfaces for recognition," J. Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.
10. P. N. Belhumeur, J. P. Hespanha, D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, pp. 711–720, 1997.
11. R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics 7 (2): 179–188, 1936.
12. R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," Workshop on Privacy-Enhanced Technologies, 2005.
13. L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3D facial expression database for facial behavior research," Int'l Conf Automatic Face and Gesture Recognition, 2006.