Integrating Utility into Face De-Identification

Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney

Data Privacy Laboratory, School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213
{rgross, eairoldi, malin, latanya}@cs.cmu.edu

Abstract. With the proliferation of inexpensive video surveillance and face recognition technologies, it is increasingly possible to track and match people as they move through public spaces. To protect the privacy of subjects visible in video sequences, prior research suggests using *ad hoc* obfuscation methods, such as blurring or pixelation of the face. However, there has been little investigation into how obfuscation influences the usability of images, such as for classification tasks. In this paper, we demonstrate that at high obfuscation levels, *ad hoc* methods fail to preserve utility for various tasks, whereas at low obfuscation levels, they fail to prevent recognition. To overcome the implied tradeoff between privacy and utility, we introduce a new algorithm, k-Same-Select, which is a formal privacy protection schema based on k-anonymity that provably protects privacy *and* preserves data utility. We empirically validate our findings through evaluations on the FERET database, a large real world dataset of facial images.

1 Introduction

Walk through the streets of any metropolitan area and your image is captured on an ever increasing number of closed-circuit television (CCTV) surveillance cameras and webcams accessible via the Internet. Consider some recent statistics. As of 2002, a survey of New York City's Times Square revealed there exist over 250 different CCTV's in operation [1]. In addition, researchers with the Camera Watch project of Carnegie Mellon University estimate that there are over 10,000 webcams focused on public spaces around the United States [2]. The dramatically decreasing costs of surveillance equipment and data storage technologies guarantees that these numbers will continue to increase.

Surveillance systems, including automated face recognition, must be held accountable to the social environments in which they are implemented [3]. Given the ubiquity with which surveillance systems are creeping into society, protection of images already captured and stored, must be developed. Currently, most attempts at enabling privacy in video surveillance and automated face recognition systems have been approached from an *ad hoc* perspective. For instance, researchers studying telecommuting have investigated the degree to which simple de-identification methods, such as "blurring" or "pixelating" an image, prevents the recipient of the image from determining the identity of the individual [4, 5].

In these studies, it has been repeatedly demonstrated that identity is sufficiently protected. As a result, the European Union's directive 95/46/EC of the Data Protection Act explicitly states that once an image is pixelated, it can be shared for research or law enforcement purposes. A simple Internet search produces several companies specializing in the development and commercialization of privacy protecting surveillance equipment - the justification being that their system can pixelate images.[1]

While simple filtering methods might prevent a human from recognizing subjects in an image, there is no guarantee that recognition will always be thwarted. Moveover, a computer that applies face identification and recognition technology is much more adept at seeing through the obfuscation. Thus, when *ad hoc* de-identification methods are tested against automated recognition systems, there is evidence that claimed protections are nothing more than superficial [6]. As a result, it is a naïve and dangerous assumption that simple de-identification methods provide privacy protection. Now that such techniques have been incorporated into various legal statutes, we suspect that it is only a matter of time before lawsuits challenging the Data Protection Directive and similar specifications will begin.

In prior research, it was demonstrated that face recognition can be sufficiently thwarted using models built on formal privacy methods [6]. The k-Same algorithm, provides the guarantee that a face recognition system can not do better than $1/k$ in recognizing who a particular image corresponds to. Moreover, this level of protection will hold up against any recognition system, human or computer, such that it is unnecessary to experimentally validate if images subject to the k-Same algorithm will be sufficiently protected. However, though privacy can be guaranteed, there is no accompanying guarantee on the utility of such data. In previous research, it was demonstrated that k-Samed images look like faces, but it is unknown if the image communicates information necessary for surveillance or classification purposes.

The goal of this paper is two-fold:

- Provide experimental evidence regarding how *ad hoc* methods can not simultaneously protect privacy and provide data utility in face images, and
- Develop and demonstrate an algorithm which provides formal privacy protection that maintains data utility on classification challenges, such as gender and expression characterization.

The remainder of this paper is organized as follows. In Section 2 we survey related work. Section gives an overview of the face recognition algorithms used in the experiments. Section 4 defines face de-identification and introduces the k-Same-Select algorithm. In Section 5 we evaluate privacy protection and data utility for both *ad hoc* and formal protection algorithms. We conclude by discussing the findings of the paper in 6.

[1] For example see IES Digital's "SiteScape[2]" - `http://www.iesdigital.com/pdfs/ssbrochure.pdf`.

2 Related Work

In this paper we investigate technical issues regarding anonymity in face recognition systems. This issue is a specific topic within the more general area of how to protect privacy in communicated images. In this section, we briefly review several proposed methods and the difference between formal models of privacy and *ad hoc* strategies. It should be noted that this research concerns the degree to which anonymity can be guaranteed from a technological perspective and, as such, we do not consider issues regarding security or policy concerns, such as who is permitted to view the images, time limits on image storage, or encryption strategies for secure storage and communication.

There exist a number of methods by which an image can be obfuscated for privacy protection. In particular, proposed methods include pixelation and various distortion filters, such as replacement of an image with a shadow, thresholding for edge presentation only, and Gaussian blurs [4,5,7–11]. This set of obfuscating techniques prevent the rendering of the original image, but they do not provide intuition as to whether or not privacy is maintained. To test if these techniques do protect privacy, researchers investigate the degree to which these methods fool human observers of the obfuscated video [7,10,11]. Researchers ask individuals if they can tell what they are observing, such as "Can you determine the gender of the subject in the video?" or "Can you determine who specifically is in this video?". If the human can not answer such questions correctly, then it is claimed that privacy is being protected.

The previous is a feasible model of privacy, provided the entity receiving the video is a human without access to accompanying information, such as a database of images to compare the incoming video to. Senior et al. [12] note that if the human or computer on the receiving side of the obfuscated video has access to references images (e.g. non-obfuscated images of the video subjects), then care must be taken to ensure that transmitted images do not reveal information which can be used to match to a reference image. In this space, we must take into account how the obfuscation affects an automated systems' ability to systematically analyze the video and perform identity determination through such techniques as face, gait, or scene (e.g. location) recognition. Through a prototype called PrivacyCam, Senior et al. demonstrated how incoming video feeds can be segmented, stored in a database, and managed using multi-level security techniques. The type of segmentation process, and the semantics of the segments, are ultimately controlled by the operators of the camera, but they can correspond to a wide range of features, from a general background vs. foreground to a more specific characterization of the individuals observed in a scene. Once the video is segmented, it can be generalized or granulated, depending on the access level of the individual requesting to view the video. For instance, a public transportation manager needs to know the number of people per subway stops, while a public health investigator may need more detail in terms of facial expression to see if the people are coughing or sneezing. In order to facilitate levels of access, the authors recommend that each segment be revealed at a different levels of granularity per access level. For very basic scene analysis, each segment

could be tracked using a different color (e.g. background displayed as the color black and individuals in the foreground displayed as blue)

Though Senior et al. present a privacy protection schema, there are no proofs or empirical analysis of whether or not their methods defeat recognition systems. Rather, this task was first addressed in 2003 by Alexander and Kenny, who investigated how individuals can fool automated face recognition systems. [13]. In their study, they consider *ad hoc* methods, such as when the subject wears sunglasses, paints their face with various designs, and shines a laser pointer at the lens of the camera. Their findings reported recognition rates at varying levels of such protection (e.g. the level of tint in the sunglasses) and provided levels at which the computer could not complete correct recognition. Yet, this evaluation of privacy is performed with the belief that the computer is not adaptive. For example, the authors submit an image of a subject wearing sunglasses and probe a database of faces in which no one is wearing sunglasses. This is a naive strategy, since the computer (or the human provided with obfuscated image) can detect how an image has been augmented.

The issue of adaptive recognition is a serious and realistic concern, which is addressed in recent research by Newton et al. [6]. In their analyses, they assume a computer can mimick the obfuscation technique within the recognition system's gallery of images. They demonstrated that recognition rates of obfuscated images against nonobfuscated images are low as observed by Alexander and Kenny [13]. However, the recognition rates soar when the computer can augment the gallery of searched images. In certain cases it was discovered that obfuscation, such as pixelation, actually increases recognition rates above the baseline tests. Newton et al. [6] concluded that the main reason why *ad hoc* methods of privacy protection in video and face recognition systems are fallible is because there is no formal model of how privacy can be compromised or protected. To rectify this concern, Newton et al. [6] proposed a new obfuscating model called k-Same. The details of the k-Same method are discussed below, but from a model perspective, it is a formal protection strategy based on the k-anonymity framework of Sweeney [14]. In general, k-anonymity stipulates that for every piece of protected data, there must exist k pieces of data in the original dataset to which the piece of data could be representative of. Translated to the k-Same model, each face image presented to a face recognition system could be representative of k faces in the gallery.

The version of k-Same introduced in [6] protects privacy and preserve detail in facial structure, however, there are no guarantees of the utility of the data (e.g. "What is the gender of the face?" or "What is the face's expression?"). In this paper, we investigate how utility can be incorporated into the k-Same model of privacy protection.

3 Face Recognition Algorithms

Automatic recognition of human faces has been an active area of research in computer vision and related fields since the 1970s [15,16]. In this section we describe

in detail the two face recognition algorithms used in experiments: 1) Principal Component Analysis, a standard academic benchmark and 2) the commercial face recognizer FaceIt, which was among the best performing systems in the Face Recognition Vendor Test of 2000 [17] and 2002 [18].

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a method for the unsupervised reduction of dimensionality [19]. An image is represented as an n-dimensional vector, where each dimension corresponds to a pixel. Given a set of N sample images $\{x_1, x_2, \ldots, x_N\}$, we define the *total scatter* matrix $S_T = \sum_{k=1}^{N}(x_k - \mu)(x_k - \mu)^T$ where μ is the mean of the data. PCA determines the orthogonal projection Φ in

$$y_k = \Phi^T x_k, k = 1, \ldots, N$$

that maximizes the determinant of the total scatter matrix of the projected samples $y_1, \ldots, y_N$:

$$\Phi_{opt} = \arg\max_{\Phi} \mid \Phi^T S_T \Phi \mid = [\phi_1 \phi_2 \ldots \phi_m]$$

where $\{\phi_i | i = 1, 2, \ldots, m\}$ are the n-dimensional eigenvectors of S_T corresponding to the m largest eigenvalues (typically $m << n$). In order to compare two face images x and y in a PCA subspaces, we project the images into the subspace using the projection defined by the eigenvectors $\phi_1, \phi_2, \ldots, \phi_m$ and compute the Mahalanobis distance: $d_M(x, y) = (x - y)^T S_T^{-1}(x - y)$.

3.2 Identix: FaceIt

FaceIt's recognition module is based on Local Feature Analysis (LFA) [20]. This technique attempts to overcome two major problems of PCA. First, the application of PCA to a set of images yields a global representation of the image features that is not robust to variability due to localized changes in the input [21]. Second, the PCA representation is non-topographic. In other words, nearby values in the feature representation do not necessarily correspond to nearby values in the input image. To address these problems, LFA uses localized image features. The feature images are then encoded using PCA to obtain a compact description. According to Identix, FaceIt is robust against variations in pose variation of up to 35° in all directions, lighting, skin tone, eye glasses, facial expression, and hair style.

4 Face De-Identification and Data Utility

4.1 Definitions

For the following we consider all face images i to be vectors of fixed size n with values between 0 and 255. We will use sets of face images, where we assume that

no two images within the same set come from the same person. In particular we will refer to the *gallery* set of face images $\mathcal{G}$, which contains images of *known* individuals and the *probe* set of face images $\mathcal{P}$, which contains images of *unknown* subjects. We can then define face recognition as follows:

Definition 1 (Face Recognition). *Given a set of probe images $\mathcal{P}$ and a set of gallery images $\mathcal{G}$, the face recognition function $f_R : \mathcal{P} \rightarrow \mathcal{G}$ associates one face image of the gallery to every face image in the probe set.*

Let $\mathcal{F}_{\mathcal{P},\mathcal{G}} = \{f_R^1, f_R^2, \ldots, f_R^n | f_R^j : \mathcal{P} \rightarrow \mathcal{G}, j = 1, \ldots, n\}$ be a set of face recognition functions.

Definition 2 (Face Recognition Performance). *Let $Eval_f : [\mathcal{F}_{\mathcal{P},\mathcal{G}}, \mathcal{P}, \mathcal{G}] \rightarrow [0 \ldots 1]$ be an evaluation function which, for given probe and gallery sets $\mathcal{P}, \mathcal{G}$ associates a face recognition function $f_R \in \mathcal{F}_{\mathcal{P},\mathcal{G}}$ with the fraction of correctly recognized face images in $\mathcal{P}$.*

In this paper we will discuss different face de-identification methods.

Definition 3 (Face De-Identification). *Let $\mathcal{M}_o = \{m_1, m_2, \ldots, m_i\}$ and $\mathcal{M}_d = \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_i\}$ be face images sets. We then define face de-identification as image transformation $f_D : \mathcal{M}_o \rightarrow \mathcal{M}_d$ such that $f_D(m_j) = \hat{m}_j, m_j \neq \hat{m}_j, j = 1, \ldots, i$.*

The implicit goal of a face de-identification method f_D is to remove identifying information from face images, so that $Eval_f(f_R, f_D(\mathcal{P}), \mathcal{G}) < Eval_f(f_R, \mathcal{P}, \mathcal{G})$. In Section 5 we will compare $Eval_f(f_R, f_D(\mathcal{P}), \mathcal{G})$ for different face recognition functions f_R and face de-identification functions f_D. k-anonymity provides a formal model of privacy protection [14]. Newton et al. extended k-anonymity to face image sets as follows [6]:

Definition 4 (k-Anonymized Face Set). *We call a probe set of face images k-anonymized, if for every probe image there exist at least k images in the gallery to which the probe image correctly corresponds.*

The second focus of this paper is on data utility functions, which assign face images to one of multiple, mutually exclusive classes.

Definition 5 (Data Utility Function). *Let $\mathcal{M}$ be a set of face images. We then define $u : \mathcal{M} \rightarrow \{c_1, c_2, \ldots, c_k\}$ as data utility function, which associates each face image in $\mathcal{M}$ with exactly one class $c_j, j = 1, \ldots, k$.*

We assume that for each face image i the *correct* class c_j is known. Examples of image classes include facial expressions $\{neutral, smile\}$, gender $\{male, female\}$, and eye status $\{open, closed\}$.

Definition 6 (Data Utility Performance). *Let $\mathcal{U} = \{u^1, u^2, \ldots, u^m\}$ be a set of data utility functions. We define $Eval_u : [\mathcal{U}, \mathcal{M}] \rightarrow [0 \ldots 1]$ as evaluation function which computes the fraction of correct class associations for a given data utility function u and a face image set $\mathcal{M}$.*

In Section 5 we will compare $Eval_u(u^j, f_D(\mathcal{P}))$ for different data utility functions u^j and face de-identification functions f_D.

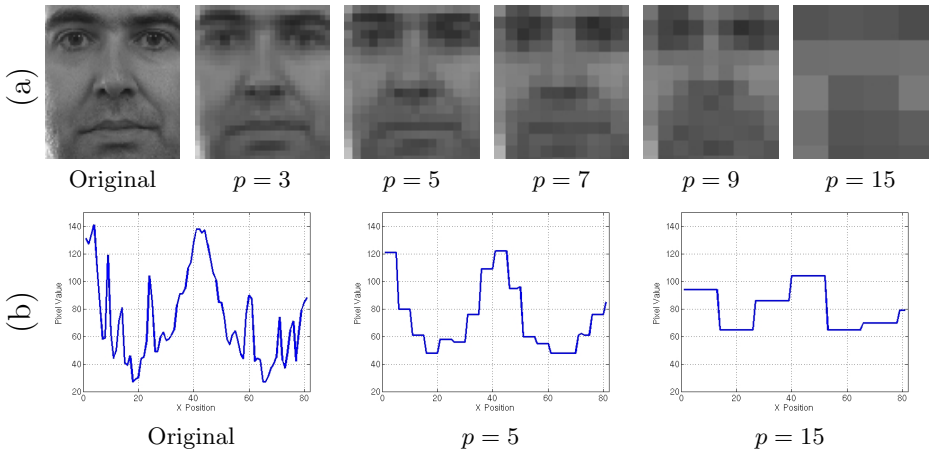


Fig. 1. Pixelation applied to face images. (a) shows example images of applying pixelation with different p factors. In (b) we plot pixel intensity value across the whole image at roughly eye height of the original image. The plots for $p = 3$ and $p = 7$ illustrate the subsampling effect of the pixelation operation.

4.2 Naïve De-identification Methods

In this section we describe two *ad-hoc* de-identification methods typically used in both previous studies [4, 5] and the popular press.

Pixelation The process of pixelation reduces the information contained in an image through subsampling. For a given pixelation factor p, image sub-blocks of size $p \times p$ are extracted and replaced by the average pixel value over the sub-block. As the value of p increases more information is removed from the image. See Figure 1(a) for examples of applying pixelation with different factors p to face images. The subsampling effect of pixelation is further illustrated in Figure 1(b) where we plot the distribution of pixel intensity values across a face for original and pixelated images.

Blurring To blur an image, each pixel in the image is replaced by a weighted average of the pixel's neighborhood. A popular choice for the weighting function is a Gaussian kernel, which weights pixels near the center of the neighborhood more heavily. In two-dimensions, for coordinates x and y, the Gaussian blurring operator is defined as $G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$ [22]. The standard deviation σ controls the size of the neighborhood. The blurred image is then computed as convolution of the original image with the Gaussian blurring operator. Figure 2 shows example images of applying blurring with different σ values to a face image. The averaging effect of image blurring is illustrated in Figure 2(b) where

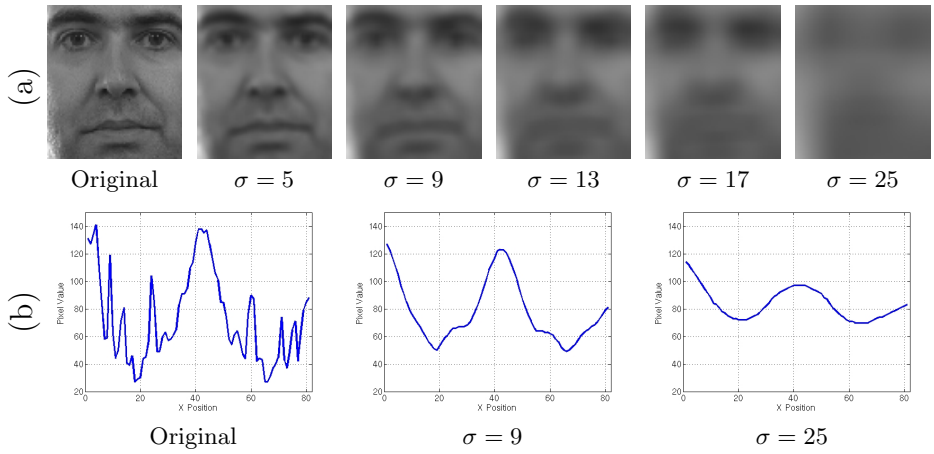


Fig. 2. Blurring applied to face images. (a) shows example images for multiple levels of blurring. Similar to Figure 1 we plot pixel intensity values across the face image in (b). The plot for $\sigma = 9$ and $\sigma = 25$ illustrates how blurring removes information from the image through averaging.

input : Face set $\mathcal{M}_o$, privacy constant k, with $|\mathcal{M}_o| \geq k$
output: De-identified face set $\mathcal{M}_d$

$\mathcal{M}_d \leftarrow \emptyset$
for $i \in \mathcal{M}_o$ **do**
  if $|\mathcal{M}_o| < k$ **then**
    $k = |\mathcal{M}_o|$
  end
  Select the k images $\{j_1, \ldots, j_k\} \in \mathcal{M}_o$ that are closest to i according to L_2 norm.
  $avg \leftarrow \frac{\sum_{m=1}^{k} j_m}{k}$
  Add k copies of avg to $\mathcal{M}_d$
  Remove $j_1, \ldots, j_k$ from $\mathcal{M}_o$
end

Algorithm 4.1: k-Same algorithm.

we again plot the distribution of pixel intensity values across a face for both original and blurred images.

4.3 De-identification using *k*-Same

Here we introduce a new variant of the k-Same algorithm. The k-Same algorithm was first introduced by Newton et al. in [6]. Intuitively, k-Same works by taking the average of k face images in a face set and replacing these images with the average image. A version of the algorithm is described in Figure 10. Image sets

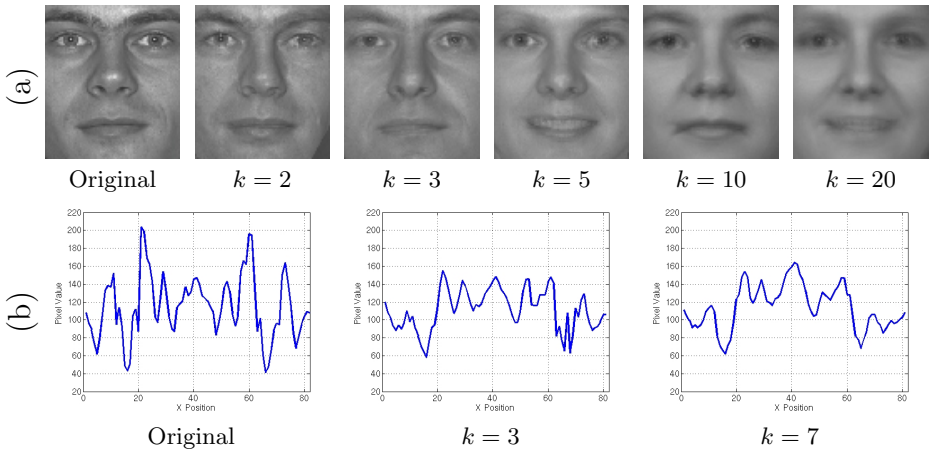


Fig. 3. Faces de-identified using the k-Same algorithm. (a) shows example images for different values of k. In contrast to pixelation and blurring the resulting image is still a face. This is also evident in the pixel distribution plots in (b). While the plots for $k = 3$ and $k = 7$ are different from the original plot they still show characteristics of the face (two local minima in the position where the pupils are).

input : Face set $\mathcal{M}_o$, privacy constant k, $|\mathcal{M}_o| \geq k$, data utility function u
output: De-identified face set $\mathcal{M}_d$

1 $\mathcal{M}_d \leftarrow \emptyset$
2 Let $\mathcal{M}_{o_1}, \ldots, \mathcal{M}_{o_l} \subset \mathcal{M}_o$ with $\mathcal{M}_{o_i} = \{x \in \mathcal{M}_o | u(x) = c_i\}$
3 $\mathcal{M}_{d_i} = ksame(\mathcal{M}_{o_i}, k), i = 1, \ldots, l$
4 $\mathcal{M}_d = \cup_{i=1}^{l} \mathcal{M}_{d_i}$

Algorithm 4.2: k-Same-Select algorithm.

de-identified using the k-Same algorithm are k-anonymized (see [6] for a proof). Figure 3(a) shows example images of a k-Samed face for different values of k. In Figure 3(b) we again show the distribution of pixel intensities across original and k-Samed face images. While the distributions change, overall characteristics such as the two local minima in the location of the pupils are maintained.

4.4 De-identification using *k*-Same-Select

One of the shortcomings of the k-Same algorithm is the inability to integrate data utility functions. Notice that the k-Samed images in Figure 3 appear to change facial expressions from neutral in the original image to smile in the image for $k = 5$. In order to address this problem we propose the k-Same-Select algorithm, summarized in Figure 4. Intuitively the algorithm partitions the input set of face images into mutually exclusive subsets using the data utility function and applies the k-Same algorithm independently to the different subsets. Due to the usage

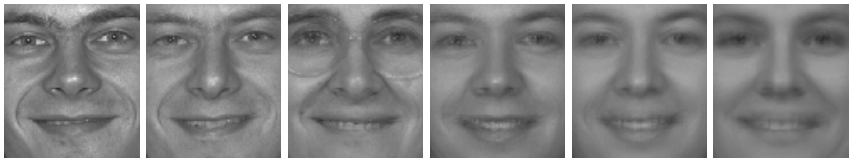


Fig. 4. Faces de-identified using the k-Same-Select algorithm with an expression data utility function. Note that facial expression stays constant across different levels of k-anonymization unlike in the examples shown in Figure 3.

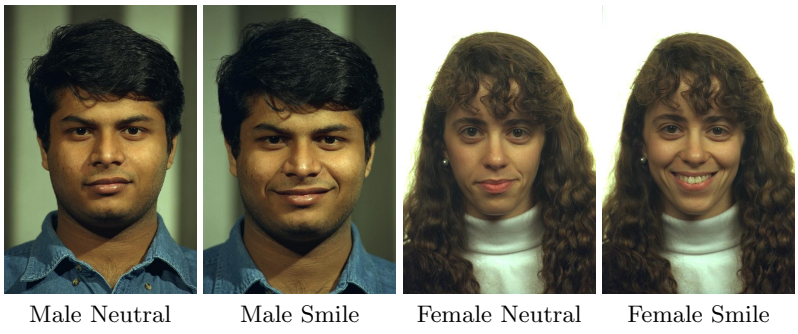


Fig. 5. Examples from the FERET database [23] showing a male and a female subject displaying a neutral and a smile expression.

of the k-Same algorithm, k-Same-Select guarantees that the resulting face set is k-anonymized.

5 Experiments

5.1 Image Database: FERET

Our experiments are based on images from the recently released color version of the FERET database [23]. We use images of 833 subjects (474 male, 359 female), ranging in age from 20 to 70 years old. Experiments involving a gender data utility function report results on a subset of 584 subjects for which images showing both neutral and smile expressions are available. See Figure 5 for example images.

5.2 Face Localization and Registration

For the PCA and gender/expression recognition experiments we use the manually determined locations of the eyes, the tip of the nose and the center of the mouth (which are distributed as part of the FERET database) to geometrically

normalize the images for translation and rotation. We furthermore scale the images to a fixed size of 81x92 pixels. In experiments involving FaceIt the original images of size 512x768 are employed, since FaceIt relies on built-in face detection routines for image normalization and alignment.

5.3 Evaluation of Privacy Protection

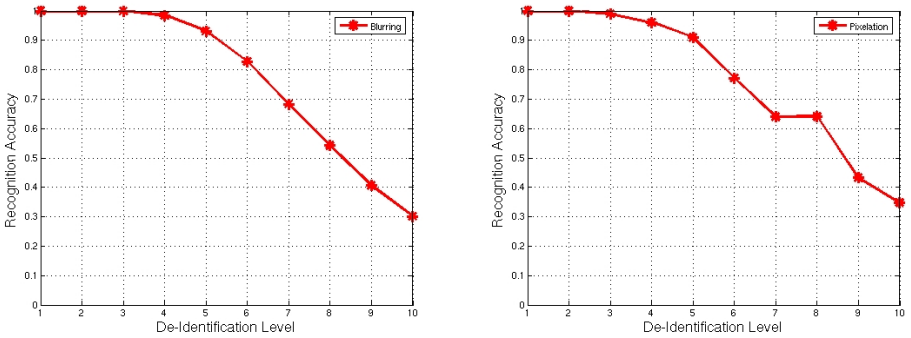
Evaluation Sets Following Phillips et al. [23] we distinguish between *gallery* and *probe* images. The gallery contains the images of known individuals against which the unknown images in the probe sets are matched. All results reported here are based on non-overlapping gallery and probe sets. We use the *closed universe* model for evaluating the performance, meaning that every individual in the probe set is also present in the gallery. For PCA recognition we randomly choose 20% of the subjects for computation/training of the eigenspace (see Section 3.1).

Results Simulating a potential real world scenario, we evaluate both PCA and FaceIt using the original, unaltered images in the gallery set and images de-identified by blurring, pixelation, and application of k-same and k-same-select in the probe set. In the experiments we vary the pixelation parameter p (sub-block size) between 1 and 21 and the blurring parameter σ (variance) between 1 and 37. PCA results are shown in Figure 6. In the case of both blurring and pixelation, recognition accuracies stay high ($> 90\%$) for up to 5 levels of de-identification. In contrast, recognition accuracies for all variants of the k-same and k-same-select algorithms are below the theoretical maximum of $\frac{1}{k}$. Recognition accuracies for FaceIt are shown in Figure 7. Unlike PCA, FaceIt has not been trained on the evaluation image set. Recognition accuracies are therefore generally lower for FaceIt than PCA. Nevertheless, the same observations hold. For both blurring and pixelation higher levels of de-identification have to be choosen in order to protect privacy.

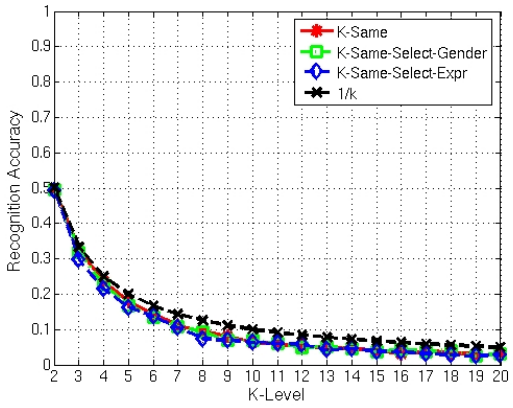
5.4 Evaluation of Data Utility

Experiment Setup In order to evaluate data utility we set out to perform two classification tasks: gender and expression classification. For each task we use a Support Vector Machine classifier with a linear kernel [24]. We partition the dataset in 5 equally sized subsets, training in turn on four subsets and testing on the remaining fifth, reporting the average accuracy over the five experiments (5-fold cross-validation). This classification accuracy is our measure of data utility. Although the de-identification methods measure the intensity of the de-identification on different scales (e.g. level of blur vs. level of pixelation), for the purposes of this discussion, we normalize all methods to a common intensity scale which ranges from 0% to 100% de-identification

Results We evaluate data utility for both gender and expression data on original, pixelated, blurred, k-Samed, and k-Same-Selected data, using the same



Blurring Pixelation



k-same, k-same-select

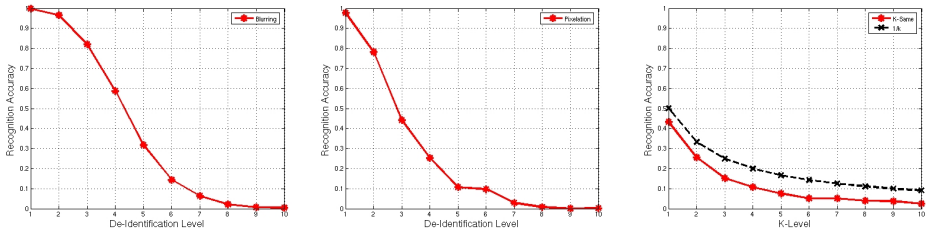
Fig. 6. PCA recognition rates for unaltered gallery images and de-identified probe images. Recognition accuracies stay above 90% for both blurring and pixelation for up to 5 levels of de-identification. In contrast, recognition accuracies for all variants of the k-same and k-same-select algorithms are below the theoretical maximum of $\frac{1}{k}$.

levels of de-identification as described above. Figure 8 shows the results of the experiments. Data utility decreases for blurring, pixelation, and k-Same, and *increases* for k-Same-Select.

6 Discussion

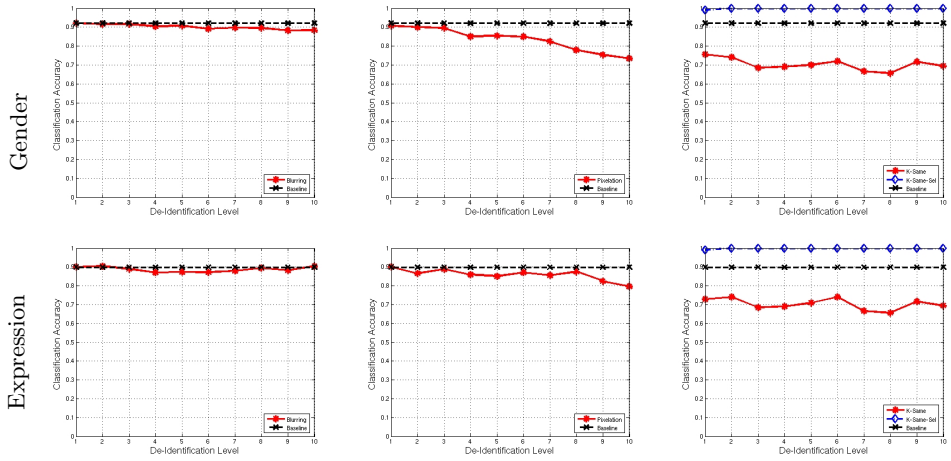
In this section, we discuss several notable findings that emerged from our experiments, as well as some of the limitations and possible extensions to this research.

Note, in the experiments of the previous section, the notion of utility was quantified in terms of classification accuracy. Although the de-identification



Blurring Pixelation k-same, k-same-select

Fig. 7. FaceIt recognition rates for unaltered gallery images and de-identified probe images. Similar to the PCA case shown in Figure 6 higher levels of de-identification have to be choosen for blurring and pixelation in order to protect privacy. Accuracies on images de-identified using the k-same algorithm again stay below the theoretical maximum of $\frac{1}{k}$.



Blurring Pixelation k-Same, k-Same-Select

Fig. 8. Data utility evaluation. We show accuracies for the gender and expression classification tasks for images de-identified using blurring, pixelation, k-Same, and k-Same-Select. Data utility decreases for blurring, pixelation, and k-Same, and *increases* for k-Same-Select.

methods measure the intensity of the de-identification on different scales (e.g. level of blur vs. level of pixelation), for the purposes of this discussion, we normalize all methods to a common intensity scale which ranges from 0% to 100% de-identification.

6.1 Formal Methods Are Better For Protecting Privacy

In previous studies, it has been claimed that *ad hoc* de-identification methods, such as pixelation and blurring prevent humans from reliably recognizing the identity of images de-identified [4, 5, 7–9, 11]. However, as our experiments demonstrate, these methods can not prevent a computer from reliably performing recognition; even at 40% de-identification the PCA algorithm achieves almost perfect recognition. In Figures 1 and 2, this roughly translates into pixelation level $p = 9$ and blur level $\sigma = 14$. Furthermore, at a de-identification level of 70% the computer recognition accuracy remains as high as 70%, or 7 correct recognitions out of every 10 probes! This corresponds to pixelation level $p = 15$ and blur level $\sigma = 25$. In contrast, the recognition rate for k-Same and k-Same-Select is controllable and inversely proportional to k, since both are derivative of the k-anonymity formal protection model. For example, when $k = 2$, a de-identification of 5%, recognition is approximately 50%. Moreover, it can be validated that both algorithms consistently permit lower recognition rates than the *ad hoc* methods (see Figure 6).

6.2 Flexibility of Formal Methods

As the level of de-identification increases, the *ad hoc* de-identification methods and k-Same incur a substantial loss in data utility in comparison to k-Same-Select. For the *ad hoc* methods, the main reason for this loss is that the specifications of pixelation and blur are inflexible. They can not be modified to explicitly account for a notion of data utility during the de-identification process. Similarly, while k-Same provides provable guarantees regarding privacy protection, for any protection level k, there is no criteria for utility preservation. However, the k-Same-Select algorithm provides an ability to preserve data utility. Unlike the previous de-identification methods, k-Same-Select is a *flexible* de-identification method; that is, it can translate any preconceived notion of data utility into de-identified images that encode that notion.

6.3 The Privacy/Utility Trade-Off

The pixelation, blur, and k-Same face de-identification methods exhibit a trade-off between privacy protection and data utility. This is because each of these algorithms de-identify images at the cost of loss of accuracy in the classification tasks. k-Same-Select overcomes this trade-off by integrating prior knowledge of the notion of utility. In doing so, k-Same-Select allows for the preservation of discriminative features during the de-identification process and provides a slight increase in classification accuracy over the prior methods.

The k-Same-Select algorithm basically functions as a mixture of a de-identification algorithm, k-Same, with a simple stratified learning techniques that dampens the sampling variability of the discriminating features. As a result, k-Same-Select provides an increased potential for data utility preservation in comparison to both k-Same and *ad hoc* methods. In a real-world implementation of

k-Same-Select, we can make use of semi-supervised learning techniques [25] and of co-training [26], to learn gender and expression of probe images that are not excluded from the original probe image set. However, it is unclear whether k-Same-Select will increase or decrease the data utility in a real-world setting since stratified sampling techniques tend to help classification, whereas semi-supervised learning techniques tend to harm it. In future research, we expect to investigate the degree to which semi-supervised learning can facilitate the classification process.

We conclude that k-Same-Select is the only known algorithm that poses a challenge to the trade-off between privacy and utility. In our experiments, it was able to increase utility for any given level of privacy protection in a controlled setting. We believe it will be able to maintain the data utility for any given desired level of privacy protection in the real-world.

6.4 Beyond Preconceived Notions of Utility

In this research, k-Same-Select uses known labels (i.e. gender and expression classes) during the de-identification process. However, we do not need to know the actual gender and expression of each image in the probe. This information is unnecessary because, as stated above, we can reliably estimate such information. Given an initial probe image set, we can run semi-supervised learning methods to build classifiers for the prediction of a new image's class. Thus, in order to apply k-Same-Select, only a *preconceived notion* of data utility is needed.

In order for k-Same-Select to be universally applicable, as an off the shelf tool, we need to make data utility and de-identification operationally independent. We believe this is possible, and this is the focus of on-going research.

6.5 Conclusions

In this paper, we studied the degree to which *ad hoc* and formal face de-identification methods preserve data utility. Two distinct classification problems, expression and gender prediction, were specified and a sophisticated machine learning method, in the form or support vector machines, was employed to measure the classification accuracy of de-identified images. Privacy protection was measured in the form recognition accuracy against standard academic and state-of-the-art commercial face recognition systems. Our experiments demonstrate that formal face de-identification algorithms always dominate *ad hoc* methods in terms of providing privacy protection (i.e. incorrect face recognition). Furthermore, we demonstrated that formal de-identification methods can be extended to explicitly model criteria for utility preservation. As a result, we introduced a new formal face de-identification method, k-Same-Select, which is superior to prior de-identification methods in both privacy protection and utility preservation. The main drawback of our research is a need to specify criteria prior to de-identification occurs and, as a result, in future research we hope to determine more a set of more generalized criteria.

References

1. Belsie, L.: The eyes have it - for now. Christian Science Monitor (November 7, 2002)
2. Sweeney, L.: Surveillance of surveillance camera watch project (2004)
3. Bowyer, K.: Face recognition technology and the security vs. privacy tradeoff. IEEE Technology and Society (2004) 9–20
4. Crowley, J., Coutaz, J., Berard, F.: Things that see. Communications of the ACM **43** (2000) 54–64
5. Neustaedter, C., Greenberg, S.: Balancing privacy and awareness in home media spaces. In: Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, in conjunction with the 5th International Conference on Ubiquitous Computing (UBICOMP), Seattle, WA (2003)
6. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying facial images. IEEE Transactions on Knowledge and Data Engineering **17** (2005) 232–243
7. Boyle, M., Edwards, C., Greenberg, S.: The effects of filtered video on awareness and privacy. In: ACM Conference on Computer Supported Cooperative Work, Philadelphia, PA (2000) 1–10
8. Greenberg, S., Kuzuoka, H.: Using digital but physical surrogates to mediate awareness, communication, and privacy in media spaces. Personal Technologies **4** (2000)
9. Hudson, S., Smith, I.: Techniques for addressing fundamental privacy and disruption tradeoffs in awareness support systems. In: ACM Conference on Computer Supported Cooperative Work, Boston, MA (1996) 1–10
10. Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. ACM Transactions on Computer Human Interactions (TOCHI) (2005) in press.
11. Zhao, Q., Stasko, J.: Evaluating image filtering based techniques in media space applications. In: ACM Conference on Computer Supported Cooperative Work, Seattle, WA (1998) 11–18
12. Senior, A., Pankati, S., Hampapur, A., Brown, L., Tian, Y.L., Ekin, A.: Blinkering surveillance: enabling video surveillance privacy through computer vision. IBM Research Report RC22886 (W0308-109), T. J. Warson Research Center, Yorktown Heights, NY (2003)
13. Alexander, J., Kenny, S.: Engineering privacy in public: confounding face recognition. In: 3rd International Workshop on Privacy Enhancing Technologies, Dresden, Germany (2003) 88–106
14. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems **10** (2002) 557–570
15. Kanade, T.: Computer Recognition of Human Faces. Birkhauser (1977)
16. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: a literature survey. ACM Computing Surveys **35** (2003) 399–458
17. Blackburn, D., Bone, M., Philips, P.: Facial recognition vendor test 2000: evaluation report (2000)
18. Phillips, P.J., Grother, P., Ross, J.M., Blackburn, D., Tabassi, E., Bone, M.: Face recognition vendor test 2002: evaluation report (2003)
19. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 711–720

20. Penev, P., Atick, J.: Local feature analysis: A general statistical theory for object representati on (1996)
21. Gross, R., Yang, J., Waibel, A.: Face recognition in a meeting room. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France (2000)
22. Sonka, M., Hlavac, V., Boyle, R.: Image processing, analysis, and machine vision. 2nd edn. Brooks/Cole (1999)
23. Phillips, P.J., Wechsler, H., Huang, J.S., Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing **16** (1998) 295–306
24. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines. Cambridge University Press, Cambridge, UK (2000)
25. Seeger, M.: Learning with labeled and unlabeled data. Technical report, University of Edinburgh (2002)
26. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers (1998)