

Engineering Privacy in Public: Confounding Face Recognition

James Alexander and Jonathan Smith

Department of Computer and Information Science, University of Pennsylvania
200 South 33rd Street
Philadelphia, PA 19104-6389
jalex@cis.upenn.edu, jms@cis.upenn.edu

Abstract. The objective of DARPA’s Human ID at a Distance (HID) program “is to develop automated biometric identification technologies to detect, recognize and identify humans at great distances.” While nominally intended for security applications, if deployed widely, such technologies could become an enormous privacy threat, making practical the automatic surveillance of individuals on a grand scale. Face recognition, as the HID technology most rapidly approaching maturity, deserves immediate research attention in order to understand its strengths and limitations, with an objective of reliably foiling it when it is used inappropriately. This paper is a status report for a research program designed to achieve this objective within a larger goal of similarly defeating all HID technologies.

1 Introduction

Surveillance cameras, as with most information technologies, continually improve in performance while simultaneously decreasing in cost. This falling cost-of-entry for quality imaging has led to widespread deployment of cameras in public and semi-public areas, often with no public oversight of how these cameras are used. The possibility of pooling many lenses together into a larger sensor network is disturbing enough, allowing the possibility of tracking an individual’s movements wherever these lenses can see. In the past, monitoring such a network to follow everyone in view would have entailed prohibitive personnel costs. However, pairing together such a network with an identification-at-a-distance technology such as automatic face recognition, could greatly reduce or eliminate this final cost barrier. Although not yet a mature technology, we anticipate that face recognition will continue to advance with the other information technologies. Given the recent great increase in interest in leading-edge security technologies, conditions are good for rapid performance improvements, so the emergence of face recognition as a low-cost commodity may not be far off. In this paper, we begin an investigation into how current face recognition techniques might be defeated, both to explore their limitations in legitimate security applications, as well as preparing for the possibility of future misuses.

The chief contribution of this paper is the empirical results of testing various face recognition countermeasures against the first of several systems, as well as a framework for interpreting those results. In the next section, we provide the details of our experimental conditions and methods. Section 3 presents our preliminary findings: what did and did not work, the patterns we see emerging, and the beginnings of a framework for modeling those patterns. Section 4 offers some thoughts on measuring privacy problems in general, which is the ultimate goal of the research program that the current experiment has initiated. In Sect. 5 we briefly discuss some related efforts, and finally we conclude with some remarks on how we intend to proceed in the future.

2 Description of Experiment

The face recognition system used in this experiment is an implementation of the eigenfaces method [36], incorporating the optimizations that Moon and Phillips [17] recommend in order to achieve the best performance among eigenfaces systems. Specifically, it is a system that treats training images as vectors, stacks them into a matrix, and estimates eigenvectors for that matrix using principal component analysis (PCA). It then projects probe¹ images into the “face space” for which the eigenvectors form a basis, and uses a distance measure to find the closest matches to the projections of the training images into the same space. The particular distance metric used is the “Mahalanobis angle,” a novel metric proposed in [17] that performs the best overall among the seven such similarity measures tested in Moon and Phillips’s empirical evaluation. The measure simply takes the venerable “cosine” similarity measure and scales each axis in face space according to the actual variance in the data along that axis. For reference, it is defined as shown in (1), where \mathbf{x} and \mathbf{y} are the projections of the images being compared in face space and $z_i = 1/\lambda_i^{1/2}$, and λ_i is the eigenvalue corresponding to the i th eigenvector.

$$d_{\text{MA}}(\mathbf{x}, \mathbf{y}) = - \frac{\sum_{i=1}^k x_i y_i z_i}{\left[\sum_{i=1}^k (x_i)^2 \sum_{i=1}^k (y_i)^2 \right]^{1/2}} \quad (1)$$

The system uses these distances to produce a sorted list of possible matches in the training set, ideally in order of decreasing likelihood. Among other minor optimizations, this system also incorporates some simple image filtering to lessen sensitivity to varying light levels that might exist across its input image sets.

This particular eigenfaces implementation was actually used as a baseline in the FERET [23, 19] evaluation, a government study of the state of face recognition technology that ran from 1993 to 1997. An extremely valuable contribution that FERET made to the field of face recognition was its corpus: 14051 256x384

¹ The face recognition literature standardly uses the term *probe* for images to be identified, i.e. members of testing data sets, and uses *gallery* for the images that the system will learn face data from, i.e. training data.

eight-bit grayscale facial images. We used the subset of these images, 3816 images in total, that had been annotated for eye, nose and mouth position. The eigenfaces implementation we tested requires the position of the eyes as one of its inputs since it lacks the ability to locate the eyes unaided. The system uses these coordinates in order to scale the images to a standard size and to crop them with an elliptical template, removing irrelevant background information. Note that supplying this information to the system only makes our countermeasure task harder: in developing our techniques to fool the system, we had to ignore the possibility of preventing it from locating the eyes in the first place, which could prove to be the Achilles heel of more sophisticated systems.

We randomly divided the 3816 images into two sets, 2/3 training and 1/3 testing, in order to establish a base performance, measured as the fraction of answers which had at least one correct match in the top ten. The result (about 75%) was consistent with that reported by Moon and Phillips. This initial run was only needed to confirm that when we added new images at the next stage, we would not unexpectedly degrade the aggregate performance of the system.

Since the performance of the system employed in this first experiment is not state-of-the-art, we endeavored to give the system the best chance possible to get the right answer. Accordingly, we only used frontal images in the testing set and tried to keep lighting conditions reasonably controlled. We also planned to give the system several training examples to work with since, while the system only achieved 75% accuracy when many subjects had only a single training images (and some none at all), it achieves a much more respectable 91% accuracy when considering only subjects with 2 or more training images.

Since we wanted to generate images of our facial recognition countermeasures in real-world use, we couldn't just retouch images in the FERET database, and none of the subjects already in the corpus were available to us for re-imaging. Accordingly, we did not directly use any of the FERET images to test our disguise techniques, but rather we employed them only as distractors. Instead, we took a set of images of two new subjects with a 3 megapixel digital camera, selected 12 images of each subject, modifying them only to match the size and color depth of the FERET images. The set of images selected had small variations in pose and facial expression comparable to those subjects with 12 or more images in the corpus. After adding these new images, a random split of the augmented corpus was made and the experiment was repeated. Aggregate performance decreased slightly (74%), but the system successfully identified *all* of the probe images taken of the new subjects. We expected this high level of performance on these images because of the much larger than average training set (9 images each), and also because it was impossible to duplicate the lighting and equipment used to take the FERET images, so the system could use any such variations to differentiate our images from the distractors.

Next, we randomly split the this augmented corpus once more, and kept this split fixed for all further experiments described here. More precisely, the training set was kept fixed, but many images were added to the probe set: these images were of one of the new subjects employing various candidate countermeasures.

In generating these images, we made an attempt to control for lighting, pose and facial expressions, with varying degrees of success, in order compare the performance of pairs of countermeasures as accurately as possible. That is, we were interested in measuring the effect of the noise we were deliberately injecting into the system, as opposed to unrelated and unintentional noise. Over 40 images, modified only for size and color depth as before, were tested. Additionally, over 300 more were generated by a morphing algorithm (more on this in Subsect. 3.2) plus a few with minor digital image edits (which are explained in Subsect. 3.4). Figure 1 depicts some examples of our new probes: one of the baseline images as well as a couple of the more successful countermeasures.



Fig. 1. Some sample probes: undisguised, masked with nylon hose, and employing a laser

Finally, a follow-up experiment was done making use of the AR database [16], which contains a sequence of 26 images of each of 126 people. These image sequences consist of two sessions of 13 images each, varying lighting conditions and facial expression, plus several images with facial occlusions employing sunglasses and a winter scarf. We trained on 10 of the 12 images in each sequence where the facial expression was essentially neutral, reserving 2 of the neutral images for baseline testing, as well as testing the occluded images. These results of these tests lend some statistical weight to our findings, which are presented in the next section.

3 Results and Analysis

3.1 Scoring disguises

To analyze the output of the system, we need to model how a face recognition system is likely to be used. We assume a powerful adversary, deploying a face recognition on a large scale using high-quality imaging equipment, and that they

might have more than one reference image of each individual in their training data. Based on descriptions of fielded commercial systems, we further assume the system will return as its output a reasonably small number of candidate matches, which a system operator is expected to verify, either live or offline. Let N be the maximum number of images we think an operator could quickly and accurately verify for a face recognition system deployed on some large scale, i.e. a scale in which a large volume of faces per unit time is expected to be processed. Our goal in designing countermeasures is to keep as many correct identifications out of the top N slots as we can, and we want any correct matches to rank as far down the list as possible. Our evaluation metric, then, should weight those matches that appear earlier over those that appear later. For simplicity, we chose the most obvious decreasing linear function to model this, given as (3.1).

$$w_x(i) = \begin{cases} N - i + 1 & \text{if the candidate in the } i\text{th position} \\ & \text{really is } x \text{ (i.e. a match)} \\ 0 & \text{otherwise} \end{cases}$$

We then sum w_x over N and, for ease of interpretation, we normalize the sum to lie in the interval $[0, 1]$, giving us our score function, shown in (3.1). For concreteness, we chose to use $N = 10$ for the numbers reported here, but this choice is arbitrary and is tunable in our analysis software.

A good countermeasure, then, will score 0 or close to it: 0 represents no correct guesses in the top ten slots. A particularly ineffective disguise might score, for example, 0.6545, which represents correct guesses in positions 1, 2, 3, 6 and 7.

$$\text{score}(x) = \frac{\sum_{i=1}^N w_x(i)}{\sum_{i=1}^N i}$$

All commercial face recognition systems we are aware of also contain some tunable cut-off for their similarity measure so that the system does not display matches that fall below some minimum likelihood in their model. This allows an arbitrary increase in match accuracy in return for a corresponding decrease in recall, and hence an almost certain increase in the rate of false negatives. However, since we cannot tell where an adversary might set this parameter, we assume, for now, that that this threshold is set so that all top matches are taken seriously. This assumption leads to some difficulties that we will revisit in Subsect. 3.3.

In order to give the reader a frame of reference in which to interpret the “goodness” of any raw scores mentioned later in the text, a few performance statistics on our probe set under this score function might be useful. The whole probe set, including the FERET-supplied distractors, scored an average of 0.2482 with a standard deviation of 0.2365. Note that these statistics indicate that, in the average case, a probe has more than one correct answer in the top ten, and even at one standard deviation below the average, the system still scores better than one correct answer in the tenth position. The median score, 0.1818, corresponds to a correct answer in the first position.

As a baseline for comparison, we took several images of the test subject with a completely undisguised face, and as expected, these images were easily identified, all but one scoring about 0.5 (which is near the one standard deviation point above the mean). The leftmost image in Fig. 1 is representative of these baselines.

For reasons of space, we cannot reproduce here all of the images that scored the optimal zero under our score function. The center and right images in Fig. 1, however, are good examples of such images: in one the head is covered with dark nylon hose, whereas the right image results from shining a laser pointer into the camera lens. It is surprisingly easy to do hit the camera lens reliably provided the position of the camera is known, which admittedly is not a particularly realistic assumption, and it is nothing approaching subtle, not to mention potentially hazardous to bystanders. A bright flashlight appears to work just as well without being nearly as dangerous, but otherwise suffers from the same practical drawbacks. Simpler disguises that also worked very well were a pair of mirrored sunglasses, a scarf wrapped around the lower face as if for cold weather, a bandage wrapped around one eye as if the subject had suffered an eye injury, and wide, dark stripes painted onto the face with stage make-up, a technique we will explore more thoroughly later. Other techniques that worked well, but did not quite score zero, included less opaque dark glasses and a lighter-colored pair of nylon hose.

Our tests employing the AR database were consistent with those of our test subject when employing similar countermeasures. These results are summarized in Tab. 1; the AR sequence numbers of the images that comprise each class are provided for reference. The system was able to identify the baseline images with a very high degree of accuracy, but stumbles quite badly when the same faces are disguised with dark glasses, and not quite so badly when the lower face is covered with a scarf. This is consistent with results that have been reported elsewhere [10], suggesting that the eye area contains more discriminatory information than the mouth area.

Table 1. AR test results

Image Group	Accuracy	Mean Score	StdDev of Score
baseline (5, 18)	254/255 = 99.6%	0.6947	0.2330
sunglasses (8, 9, 10, 21, 22, 23)	115/765 = 15.0%	0.0344	0.1125
scarf (11, 12, 13, 24, 25, 26)	449/765 = 58.7%	0.2323	0.2751
all AR probes	818/1785 = 45.8%	0.2136	0.3042

A very interesting fact arises from examining the disguised AR subjects that were recognized, i.e. those with a score greater than zero: a majority of them score better than one match in the top ten. Moreover, most of these subjects also had more than three on their disguised images identified. A few unlucky souls were identified in all or nearly all of their disguised pictures! Clearly it is

not acceptable to just inform such individuals that they are out of luck, so this certainly bears further investigation. Subsection 3.4 discusses some preliminary observations.

Most of the attempts at disguise that did not succeed were not at all surprising: we tried several normal eyeglass frames with clear lenses, and these all performed comparably to the baseline images. This result is consistent with results reported in trials of commercial face recognition systems [21]. One result that was surprising at first, until a pattern began to emerge in the successful trials, was that a latex nose appliance that had been painted to match the skin of the subject also performed similar to the baseline, even though the apparent shape of the nose was considerably altered. Another try involved use of reflective make-ups in order to use the flash of our camera against it, dazzling it much as we had successfully done with the laser and flashlight. Unfortunately, not enough light was reflected back into the lens to have an effect, which was just as well since it is not clear that a covert face recognition system could employ a flash in any case.

A final interesting failure is an attempt to build a camera-dazzling device using infrared light-emitting diodes. The idea would be to reproduce the effect of the flashlight and laser in a way that would not be visible to an unaided observer. Unfortunately, while our cameras are sensitive to infrared wavelengths, they are not sensitive enough for this purpose: the result was barely visible under ambient light, even with flashes disabled. Based on the apparent brightness of the LEDs in our test photographs, we estimate that the light source would need to be 20 to 100 times brighter in order to have the desired effect. While there are inexpensive infrared LED lasers available that could achieve this, such an illumination level is well into the hazardous range in terms of potential damage to human eyes, particularly since the eye cannot reflexively protect itself from wavelengths it cannot detect in the first place.

We should note that none of the countermeasures we employed to truly devastating effect on the performance of the system under study are undetectable when the system is being monitored by a live observer. This means that in a supervised security application, our disguises would likely prove ineffective: a guard could ask an individual to remove dark glasses, for instance. But in a situation where an individual is being surveilled without his or her knowledge, in a public place or perhaps inside a store, it might be totally reasonable to be wearing moderately tinted lenses or to be bundled up for a cold winter day, without attracting undue attention.

3.2 Squeezing More From the Score

Treating all zero-score disguises as equivalent is somewhat unsatisfactory: some actually have correct matches in the teens, while others do not have a correct match until well after the hundredth position. It is not clear, however, that extending the score function to an arbitrarily large N (greater than 10) would be a viable approach: at some point the distance measure must lose statistical significance. Identifying this point precisely is not easy, however; the next subsection

will have more to say about this. In this subsection we describe a different approach that has some nice advantages: we chose an arbitrary score threshold of 0.0909^2 . We also chose one of the baseline images³ to use as a reference image, and morphed each of the images that scored less than the threshold into this reference image. A fragment of such a morph is shown in Fig. 2, which should be enough to give a good idea of what results from this process; the full sequence contains more intermediate images.



Fig. 2. A sample morph

We can now feed these synthetic images back into the recognition system to see how they fare. The results should give us a better idea of how much of a disguise is needed to meet any specific threshold: by counting the number of frames it takes each zero-scoring disguise to reach this threshold, we can compare their relative effectiveness.

Again, we cannot display much of our data in the present paper, and neither is there a good way of aggregating it, but we can highlight a few interesting

² This is arbitrary in the sense that we could have chosen any fixed threshold to investigate. This particular threshold corresponds to a single correct identification in the sixth position - one or more matches in the first through fifth position would result in a higher score than this.

³ ... the one that happened to have the average score of all the baselines. This choice is also arbitrary.

results. The top right picture in Fig. 2 is the first in its sequence to reach the 0.0909 threshold, after eight frames. In terms of lens opacity, it looks remarkably like an image employing a different pair of sunglasses, which scores exactly the same. Somewhat less robust than the sunglasses were the nylons shown in Fig. 1, degrading to our threshold after seven frames. A pair of white nylons, which are visibly more transparent than darker nylons, achieved a similar score to the morph frame of the grey nylons with a similar visual opacity. The disguise the proved most robust of all under this analysis used a simple, bright-white party mask, about which we will have more to say in Subsect. 3.4.

3.3 The Distance Metric and Performance Trade-offs

In further research, we would like to see if we can get some leverage directly from the distances that the eigenfaces system returns for each candidate in the image gallery, rather than just analyzing the ordering that this similarity score induces. Unfortunately, a literature search [1, 2, 6, 7, 33] indicates that there is no statistical foundation on which we can base a direct interpretation of this “Mahalanobis angle” similarity measure. Indeed, it seems that most of the similarity measures commonly used in the classification literature appeal to some notion of topological distance in the vector space in question, rather than the kind of probabilistic basis that we would wish for in the current study.

Another option is to refine our scoring model to account for the fact we mentioned at the beginning of this section: commercial systems have a way of discarding less likely matches in order to increase the accuracy of the matches, almost always at the expense of an increased rate of false-negatives. Consider Fig. 3: if an operator were to drop results with similarity below 400 or so, he could achieve nearly perfect accuracy. Actually doing so would be foolhardy, however, as our false-negative rate is high enough to make the system essentially useless. More conservative trade-offs are possible, though: the operator can raise accuracy to 76.2% while keeping the false-negatives below 10% using a cut-off of 267. This is still a pretty undesirable false-negative rate, however, so it is not obvious that an operator would make that choice for this system, hence the reason we did not use this scoring system for our present work. We will reconsider this choice when we work with more robust systems, which may have more obvious trade-offs available.

One useful thing this trade-off model would give us, however, is an explanation for the only serious anomaly in our data set: as mentioned before, an image with a scarf wrapped around the lower face scored zero, like many of the AR images did, however another image with the same scarf wrapped in the same way, with the addition a stocking cap that covers the top of the head (i.e. *more* of the face is actually obscured) actually has a correct match in the first position! Examining the similarity scores of this anomaly, however, reveal that the similarity levels are actually among worst over the face images in the data set, faring even worse than an image of a cat’s face, and comparable to a set of probes depicting various inanimate objects, none of which even have correct matches available in the training set. We therefore speculate that this correct

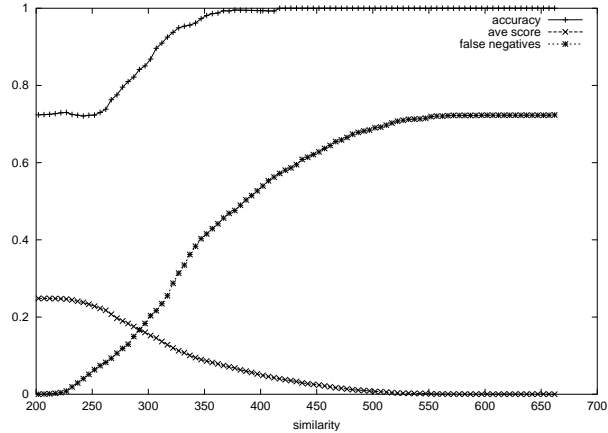


Fig. 3. Performance trade-offs

answer arose by chance, and this trade-off model would certainly rule it out as such, but we would like to be able to back that claim more rigorously.

3.4 Some Modeling Ideas

To conclude this section, we present the beginnings of an explanation that models the success (or lack thereof) of the disguise techniques we developed in the course of this experiment. The countermeasures fell into two broad classes: concealment behind an opaque object, or else overloading of the camera’s sensory apparatus with a bright light source. The degree to which each class worked can be summed up in a single word: *contrast*, i.e. large color difference between adjacent pixels. That is, the single most important factor seems to be to add as much contrast as possible, contrast that the system cannot predict based solely on the training photographs, or else to conceal or distract from a contrast that the system expects to find. The reason we think this is the case is that the system is likely to exploit contrast in order to identify candidate discriminating features that it uses to measure (dis)similarity. By adding new areas of contrast, or obscuring existing ones, we can distract the system from the real features, forcing it to map the features it expects to find to the closest coincidental match, ideally a wrong one. The natural world offers many examples that employ camouflage of this nature [9]. Some insects have abdomen coloration that looks like eyes, distracting predators from the real, more vulnerable organs. Also, as is well known, many animals have evolved take advantage of their coloration in order to confuse their features with the background environment.

Recall that some of the AR subjects seem especially hard to disguise. Our preliminary investigation into this fact has yielded some indications that it is helpful to ensure that any reference images of an individual that might end up in

a facial recognition database are as feature-deficient as possible. The individuals that had all, or almost all, of their images correctly identified, for instance, all wear glasses in all of their images, and prominent facial hair seems to be a factor as well. As is also the case when one is trying to fool a human observer, it seems likely that successful face recognition countermeasures could include *removing* extra identifying features as well as adding new distractions. Deeper investigation into this phenomenon is ongoing, and we are especially eager to see whether results like this occur in face recognition systems built on foundations other than eigenfaces.

Beyond employing some minimal level of contrast, we think the desired spurious matches are more likely to occur if we insert as many false features as possible, features not present in the training images. That is, it is better than not to have as many of these contrasts per unit area as we can fit in, provided the contrasting areas themselves are large enough to register as features.

Figure 4 gives some minimal pairs that should illustrate these ideas well. The leftmost image is one of our most robust disguises: a close-fitting party mask. Like our successes with make-up, this mask obscures very little structure of the face, but changes the color profile radically. The next image is nearly the same, only the color of the mask has been digitally modified to be a uniform flesh tone similar to the subject's own coloration. Even though the resulting simulated mask is still opaque, hiding the structure of the face the lies directly behind it, it performs quite poorly, with correct matches in the first, third, and ninth positions. The structure of the mask is the same, but we have all but lost the contrast that its bright color gave us before. The second image from the right is the face paint image that we mentioned before: it also scores a zero (the first correct match is in the sixteenth position). It is interesting to observe that this make-up striping technique resonates well with conventional wisdom on military facial camouflage [26]. Now consider the rightmost image: it employs precisely the same face paint, but covers almost all of the face rather than leaving square gaps that expose the true skin color. Even though we have obscured more of the face, this performs much worse, with a match in the very first position!

Our explanation for this is simply that we have eliminated almost all of the contrast that the striping, or the whiteness of the mask, gave to us. Examining the architecture of this eigenfaces system yields an obvious, partial explanation: the system starts by applying a histogram-equalization and scale-normalization filter to the raw images⁴, which allows it a degree of robustness over varying lighting conditions, and thus likely also global changes in skin tone. Essentially, if a color change is applied globally, the brightness and contrast of the image, as a whole, can be adjusted to give something resembling the same face with the expected color. However, if there are large contrasts in the new color pattern, as

⁴ We attempted an experiment that turned off this image processing phase to see if this accounted for the bulk of the performance difference between the striped image and the more fully painted image, however this triggered a bug in the linear algebra library that the system uses, causing it to crash. We have not yet had a chance to investigate fixing this problem.



Fig. 4. Minimal pairs

is the case with striping, there is no simple way to enhance dark portions of the image without washing out the lighter portions.

A simple way to model our attacks is to divide the face into grid zones, as illustrated in Fig. 5. The size of these zones, at the moment, is somewhat arbitrary, but we think that they should be on order of the size of the major facial features; an experiment is planned to find this critical size threshold. In order to get the best possible results under this model, we simply need to put as large a contrast as we can manage in each zone that doesn't already contain a large contrast, or else we need to hide an existing contrast by removing it entirely (e.g. by shaving off facial hair) or hiding it behind a larger, false contrast. The dark glasses and the white mask, then, hide an expected contrast behind an unexpected, large, distracting contrast, and do so across several zones. The striped face succeeds almost as well by establishing a medium contrast in the majority of facial zones (in terms of color difference, the dark glasses are more than twice as dark as the face paint). Indeed, our model would predict that a darker paint color would perform even better than the dark glasses, and indeed darkening the face paint color digitally achieves exactly the expected result, as did repeating the experiment with more contrastive make-up colors.

The model requires a further refinement in order to account for the greater importance of the eye area over the mouth area, as shown in our experiments with the AR images. We can do this by simply weighting the contribution of each score appropriately, which we should be able to do by estimating a probability distribution, based on our current result, for the importance of each zone's contribution, and testing it on novel disguises. Such work is currently in progress.

4 Measuring Privacy

This section offers a brief, likely somewhat controversial, foray into modeling privacy problems generally. We believe it to be possible to model all privacy problems, at a high level, in a common framework. In particular, we claim that

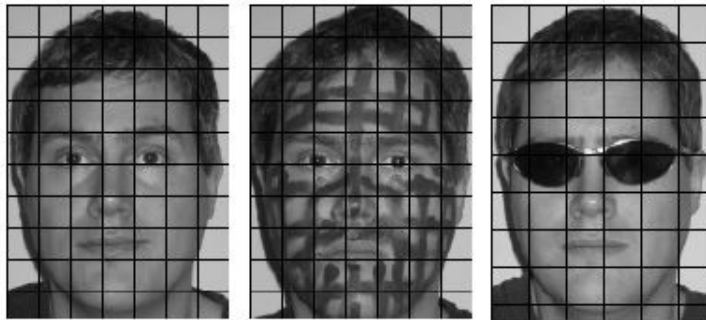


Fig. 5. Grid model

any privacy problem can be characterized as follows: an adversary knows for certain that a predicate of interest hold of some person in the world, but is uncertain of the identity of that person. For example, the adversary may know that x_1 crossed the view of a particular camera lens at time t_1 , or that x_2 bought corn flakes at his store at time t_2 , or that x_3 sent a packet through a mix network that emerged at time t_3 . The adversary, in attempting to identify the correct valuation of his predicate, builds a probability distribution on the set of all individuals in the world. The job of a privacy enhancing technology is simply to make sure that this probability distribution is not very informative: the correct individual(s) should not stand out.

Our goal in proposing this common framework is to develop a general privacy metric, suitable to serve as the “benefit” half of a cost/benefit ratio, to be used to evaluate any candidate solution meant to mitigate a given privacy problem. A suitable privacy metric should, in principle, be explainable to a mathematically-inclined lay person⁵, should be completely orthogonal to any cost metric, and, most importantly, needs to place reliable bounds on how effectively an adversary can unmask an individual trying not to be identified. That is, we want to be able to predict how flat we can make his probability distribution, or place a lower bound on the entropy of his distribution.

The value of a unified model for disparate privacy problems is most clear when one considers multimodal privacy attacks. For example, an adversary may try to correlate multiple biometric sensors, or might try to improve face recognition results using data mining on the purchase data from the store in which the target face image was captured. A common framework will allow us to model such sensor fusion. Also, if the benefit of countermeasures is measured on the same scale, an individual can more easily decide where best to spend his resources when combating more than one privacy problem at once.

⁵ ... at least as explainable, say, as computer performance metrics.

Our strategy for identifying a suitable metric is an empirical one: we will propose several candidate metrics and evaluate their effectiveness at predicting the performance of an adversary using several particular attacks. In the work currently in progress, we are concentrating on capturing the limits of biometric HID technologies [21, 4], however as our theoretical work continues to evolve, we continually keep in mind other areas of significant privacy concern, such as data mining [11, 35] and one of the most widely-studied privacy enhancing technologies, anonymous communication on public networks [25, 27, 28, 29, 34].

We conclude this section with a description of one of our most promising candidate metrics. As others have done for anonymity networks [5, 31], we draw our inspiration from information theory [32, 20]. Consider Fig. 6: we model identity as any rigid identifier, say a unique integer for each individual in the world. In going about her business, any individual broadcasts this identifier over a noisy channel, a channel which an adversary may be monitoring using one or more sensors. The individual, fortunately, controls some parameters of this noisy channel, while the adversary controls others. The individual wishes to exploit the parameters she controls in order to maximize the entropy in the adversary’s probability distribution, which he builds from his sensory information.

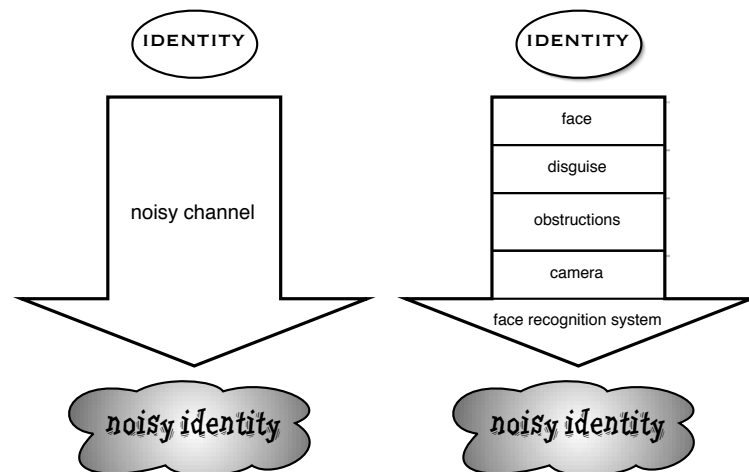


Fig. 6. A noisy channel model of privacy problems

Figure 6 also shows a possible instantiation of this model for the face recognition domain. For this problem, the individual’s face is a, perhaps imperfect, transmitter for her identity. A camera combined with a face recognition system is her adversary’s receiver for this signal. Fortunately, the individual has the options of using obstructions in her environment or actively disguising herself in order to inject more noise into the channel, hoping to limit what the adversary

can successfully receive. As is usual in a communication problem, the adversary would like to use his receiver to get the correct information through the channel despite noise. Contrary to the usual case, however, the individual does not wish to cooperate with this communication, and will do her best to disrupt it. Since the individual controls only part of the channel, she cannot direct all of the noise sources to her advantage, but by raising the noise floor, she should be able to directly affect the entropy in the adversary's model.

Figure 7 illustrates how we see our grid model fitting in as part of this noisy channel model. The grid coordinates together with the contents of that grid location make up the symbols in the message being transmitted over the noisy channel, which the adversary would like to decode into an identity. A successful disguise will ensure that not enough signal emerges from the transmitter to be successfully received.

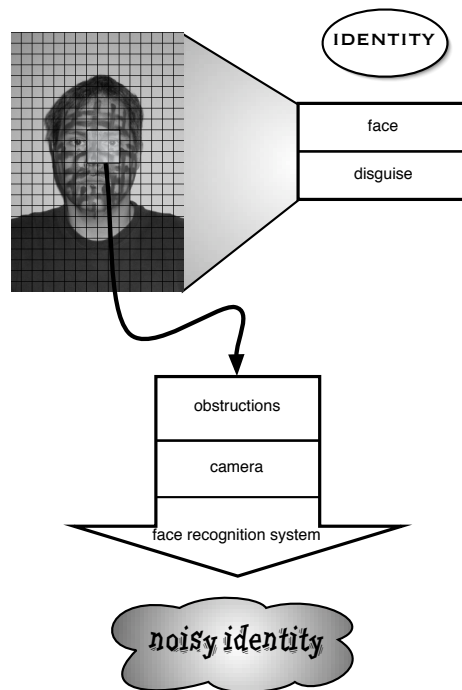


Fig. 7. Integration of grid and noisy channel models

A far as we are aware, this noncooperative communicative act is not an aspect of information theory that has been well explored. While it is well-understood how one might continue to successfully communicate over a jammed radio channel, there does not seem to be a theory of how to be an effective jammer, par-

ticularly when the jammer is also operating the transmitter. Also, prior models of noise in communication channels assume a randomly-distributed noise source, however our experimental work indicates that the most effective countermeasures require targeted, nonrandom noise. Overcoming these theoretical challenges is a top priority for further research.

5 Related Work

Beyond the development of new face recognition algorithms [12, 30, 36, 37, 38], the efforts most closely related to our facial recognition research are the FERET test itself [19, 24, 23] along with the work that followed directly from it [17, 3]. A very recent paper [10] made us aware of the existence of the AR database, and itself did a very quick evaluation of the occluded AR images against a different eigenfaces system as well as one of the best commercial systems, FaceIt⁶. Our work differs from these earlier evaluations primarily in perspective: while they are attempting to understand the boundaries of performance of current face recognition system in order to identify performance goals for future research systems, we are trying to find ways of reliably defeating those systems. Consequently, while we and the face recognition community are both interested in understanding why certain faces fail to get identified, our work is additionally interested in those faces that *are* identified when we would rather they were not. This different outlook calls for a significantly different evaluation methodology than pure statistical performance, as well as, for example, trying to find failure modes that those building new systems might decide to ignore, being outside the scope of their intended application.

Although our work is primarily motivated by wanting to prevent or discourage abusive uses of automatic face recognition technology, we certainly recognize that the technology also has perfectly legitimate applications in security and authentication. Indeed, our work should be seen as complementary rather than in opposition to the face recognition community: we expect it to be of value those developing new face recognition techniques and refining existing systems. While tests like those conducted during the FERET trials are useful for understanding how well the technology can perform, in any security application, it is at least as important to understand how and why the technology could fail.

Recent papers from one of the creators of the AR database [14, 15] actually focus explicitly on methods of improving performance in the presence of occlusions. Interestingly, his model for working around the occlusions has much in common with our grid model for evaluating occlusions: the face is divided up into zones that are modeled separately, and the output of the individual zone models are combined according to weights generated by a probabilistic likelihood model. We will certainly be interested in evaluating this system, or one like it, using our methodology.

⁶ We are working on obtaining a license for FaceIt, or a similarly robust system, for evaluation in our own framework.

We were recently made aware a preprint [39] of an interesting paper that makes use of the area of information theory we are exploring, channel capacity, in the privacy arena, but with a drastically different scope. It describes an interesting protocol where a consumer can reveal personal information to a market researcher (in exchange for something the consumer wants), but which limits the ability of the marketer to connect accurate information with specific individuals, while maintaining his ability to obtain aggregate information with a known margin of error. Attacks against this protocol are analyzing using Shannon's channel coding theorems.

6 Future Directions and Conclusions

The long-term goal of the research program that this paper initiates is to develop a generalized privacy metric. In order to evaluate competing solutions to any problem, engineers must first have a standard of measurement with which to evaluate the candidates. We would like privacy enhancing technology to emerge as a first-class engineering discipline, and a reliable metric is a prerequisite to that.

The present paper establishes some of the empirical foundations on which we intend to build this metric. In particular, it identifies a paradigm of successful countermeasures against one face recognition system, and develops a framework in which we can formalize the crucial properties of those countermeasures without unduly narrowing the scope of our investigation at this early stage.

In the near term, we will continue our investigation into defeating face recognition by expanding our dataset with more disguises applied to more subjects, as well as utilizing reliable synthetic imaging such as the morphing technique described in the present paper. The experiments will, of course, also be replicated using face recognition systems built upon significantly different principles than the one we have already studied. Similar experiments that attempt to counteract other developing HID technologies such as voice, gait [21], and iris recognition [4] will follow, and will be used to further refine and validate our metric.

Acknowledgments

This research is supported by ONR / DARPA F30602-99-1-0512. Portions of the research in this paper use the FERET database of facial images collected under the FERET program [23]. Thanks to Aleix Martínez for access to the AR database, and Ralph Gross and Jianbo Shi for access to their image annotations as well as valuable discussion.

References

- [1] Mark S. Aldenderfer and Roger K. Blashfield. *Cluster Analysis*. Sage Publications, 1984.
- [2] G. H. Ball. Data analysis in the social sciences: What about the details? In *Proc. AFIPS 1965 Fall Joint Computer Conference*, volume 27, pages 533–559, 1965.

- [3] J. Ross Beveridge, Kai She, Bruce Draper, and Geof H. Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535–542, 2001.
- [4] John Daugman. Iris recognition. *American Scientist*, 89(4):326–333, 2001.
- [5] Claudia Díaz, Stefaan Seys, Joris Claussens, and Bart Prenel. Towards measuring anonymity. In *The Second Workshop on Privacy Enhancing Technologies*, 2002.
- [6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, Inc., 2001.
- [7] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.
- [8] Simson Garfinkel. *Database Nation*. O’Reilly, 2000.
- [9] R. L. Gregory and E. H. Gombrich, editors. *Illusion in Nature and Art*. Gerald Duckworth and Co. Ltd., 1973.
- [10] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, 2001.
- [11] M. A. Hernández and S. J. Stolfo. A generalization of band joins and the merge/purge problem. <http://www.cs.columbia.edu/~sal/hpapers/mpjourn.ps>, 1996.
- [12] Steve Lawrence, C. Lee Giles, Ah Chung Tsoi, and Andrew D. Back. Face recognition: A convolutional neural network approach. *IEEE Transaction on Neural Networks, Special Issue on Neural Networks and Pattern Recognition*, 8(1):98–113, 1997.
- [13] Alberto Leon-Garcia. *Probability and Random Processes for Electrical Engineering*. Addison-Wesley, 1994.
- [14] A. M. Martínez. Recognition of partially occluded and/or imprecisely localized faces using a probabilistic approach. In *Proc. of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2000.
- [15] A. M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(6):748–763, 2002.
- [16] A. M. Martínez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center, Campus Universitat Autònoma de Barcelona, 1998. <http://rv11.ecn.purdue.edu/ARdatabase/ARdatabase.html>.
- [17] Hyeonjoon Moon and P. Jonathon Phillips. Computational and performance aspects of PCA-based face-recognition algorithms. *Perception*, pages 303–321, 2001.
- [18] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [19] NIST. FERET database. <http://www.itl.nist.gov/iad/humanid/feret/>, 2001.
- [20] Elements of Information Theory. *Thomas M. Cover and Joy A. Thomas*. Wiley-Interscience, 1991.
- [21] United States General Accounting Office. Technology assessment: Using biometrics for border security. <http://www.gao.gov/new.items/d03174.pdf>, 2002. Pub. number GAO-03-174.
- [22] George Orwell. *1984: a novel*. New American Library, 1961.
- [23] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing Journal*, 16(5):295–306, 1998.
- [24] P. Jonathon Phillips, Patrick J. Rauss, and Sandor Z. Der. FERET (face recognition technology) recognition algorithm development and test results. Technical Report ARL-TR-995, Army Research Laboratory, 1996.

- [25] Charles Rackoff and Daniel R. Simon. Cryptographic defense against traffic analysis. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, 1993.
- [26] JV Ramana Rao. *Introduction to Camouflage and Deception*. Defence Research and Development Organization, Ministry of Defense, New Delhi, 1999.
- [27] Jean-François Raymond. Traffic analysis: Protocols, attacks, design issues and open problems. In Hannes Federath, editor, *Designing Privacy Enhancing Technologies*, Lecture Notes in Computer Science (LNCS 2009), pages 10–29. Springer-Verlag, 2001.
- [28] Michael Reed, Paul Syverson, and David Goldschlag. Anonymous connections and onion routing. In *IEEE Journal on Selected Areas in Communication Special Issue on Copyright and Privacy Protection*, 1998.
- [29] Michael K. Reiter and Aviel D. Rubin. Crowds: Anonymity for web transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.
- [30] F. Samaria and F. Fallside. Automated face identification using hidden markov models. In *Proceedings of the International Conference on Advanced Mechatronics*, 1993.
- [31] Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In *The Second Workshop on Privacy Enhancing Technologies*, 2002.
- [32] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
- [33] Roberts R. Sokal and Peter H. Sneath. *Principles of Numerical Taxonomy*. W. H. Freeman and Company, 1963.
- [34] Paul Syverson, Gene Tsudik, Michael Reed, and Carl Landwehr. Towards an analysis of onion routing security. In *Workshop on Design Issues in Anonymity and Unobservability*, 2000.
- [35] J. F. Traub and Y. Yemini. The statistical security of a statistical database. *ACM Transactions on Database Systems*, 9(4):672–679, 1984.
- [36] Matthew A. Turk and Alex P. Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–591, 1991.
- [37] Dominique Valentin, Hervé Abdi, and Alice J. O’Toole. Categorization of identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. *Journal of Biological Systems*, 2(3):413–429, 1994.
- [38] Dominique Valentin, Hervé Abdi, Alice J. O’Toole, and Garrison W. Cottrell. Connectionist models of face processing: A survey. *Pattern Recognition*, 27:1209–1230, 1994.
- [39] Poorvi L. Vora. Towards a theory of variable privacy. In review, available at http://www.hpl.hp.com/personal/Poorvi_Vora/Pubs/plv_variable_privacy.pdf, 2002.
- [40] Matthew Wright, Micah Adler, Brian N. Levine, and Clay Shields. An analysis of the degradation of anonymous protocols. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS 2002)*, 2002.