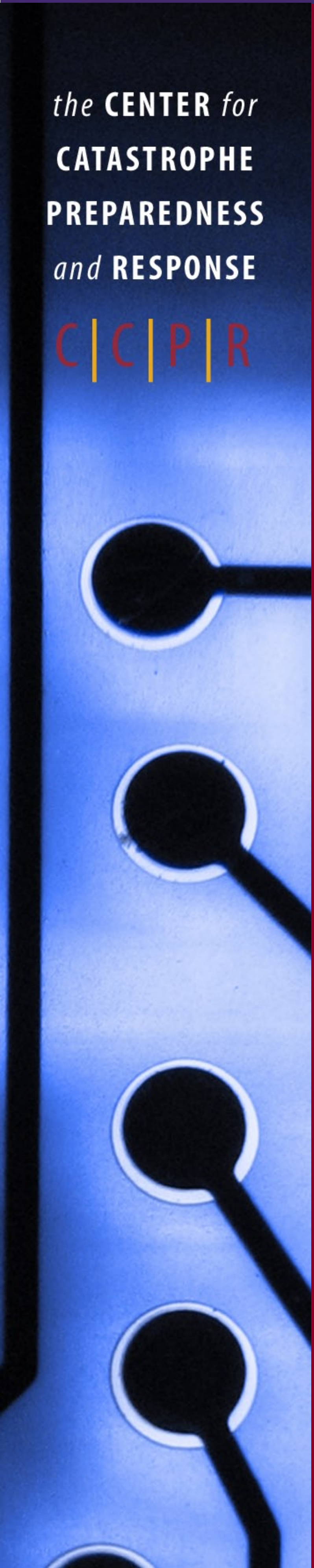
NEW YORK UNIVERSITY

the CENTER for
CATASTROPHE
PREPAREDNESS
and RESPONSE

C|C|P|R



Facial Recognition Technology

A Survey of Policy and Implementation Issues

Lucas D. Introna
Lancaster University, UK; Centre for the Study of Technology and Organization

and

Helen Nissenbaum
New York University; Department of Media, Culture, and Communication,
Computer Science, and the Information Law Institute

ACKNOWLEDGMENTS

Although responsibility for the final product is ours, we could not have produced this report without key contributions from several individuals to whom we are deeply indebted.

Jonathon Phillips and Alex Vasilescu generously shared their wisdom and expertise. Extended conversations with them, particularly in the early phases of the project, guided us to important sources and helped to focus our attention on crucial features of facial recognition and related technologies.

Solon Barocas and Travis Hall, who served as research assistants on the project, made invaluable contributions to all aspects of the report, locating sources, verifying factual claims, developing the executive summary, and carefully reading and making substantial revisions to the text. With a keen eye, Alice Marwick carefully read and edited final drafts. Ryan Hagen designed the report's cover and layout.

We are immensely grateful to Jim Wayman who, in the capacity of expert referee, carefully reviewed an earlier draft of the report. As a result of his many astute comments, contributions, and suggestions, the report was significantly revised and, we believe, enormously improved.

The authors gratefully acknowledge support for this work from the United States Department of Homeland Security. Helen Nissenbaum served as Principal Investigator for the grant, administered through the Center for Catastrophe Preparedness and Response at New York University.

EXECUTIVE SUMMARY

Facial recognition technology (FRT) has emerged as an attractive solution to address many contemporary needs for identification and the verification of identity claims. It brings together the promise of other biometric systems, which attempt to tie identity to individually distinctive features of the body, and the more familiar functionality of visual surveillance systems. This report develops a socio-political analysis that bridges the technical and social-scientific literatures on FRT and addresses the unique challenges and concerns that attend its development, evaluation, and specific operational uses, contexts, and goals. It highlights the potential and limitations of the technology, noting those tasks for which it seems ready for deployment, those areas where performance obstacles may be overcome by future technological developments or sound operating procedures, and still other issues which appear intractable. Its concern with efficacy extends to ethical considerations.

For the purposes of this summary, the main findings and recommendations of the report are broken down into five broad categories: performance, evaluation, operation, policy concerns, and moral and political considerations. These findings and recommendations employ certain technical concepts and language that are explained and explored in the body of the report and glossary, to which you should turn for further elaboration.

1. *Performance*: What types of tasks can current FRT successfully perform, and under what conditions? What are the known limitations on performance?
    a. FRT has proven effective, with relatively small populations in controlled environments, for the verification of identity claims, in which an image of an individual's face is matched to a pre-existing image "on-file" associated with the claimed identity (the verification task). FRT performs rather poorly in more complex attempts to identify individuals who do not voluntarily self-identify, in which the FRT seeks to match an individual's face with any possible image "on-file" (the identification task). Specifically, the "face in the crowd" scenario, in which a face is picked out from a crowd in an uncontrolled environment, is unlikely to become an operational reality for the foreseeable future.
    b. FRT can only recognize a face if a specific individual's face has already been added to (enrolled in) the system in advance. The conditions of enrollment—voluntary or otherwise—and the quality of the resulting image (the gallery image) have significant impact on the final efficacy of FRT. Image quality is more significant than any other single factor in the overall performance of FRT.
    c. If certain existing standards for images (ANSI INCITS 385-2004 and ISO/IEC 19794-5:2005) are met or exceeded, most of the current, top-performing FRT could well deliver a high level of accuracy for the verification task. Given that images at the site of verification or identification (the probe image) are often captured on low quality video, meeting these standards is no small feat, and has yet to be achieved in practice.
    d. Performance is also contingent on a number of other known factors, the most significant of which are:

        - *Environment*: The more similar the environments of the images to be compared (background, lighting conditions, camera distance, and thus the size and orientation of the head), the better the FRT will perform.
        - *Image Age*: The less time that has elapsed between the images to be compared, the better the FRT will perform.
        - *Consistent Camera Use*: The more similar the optical characteristics of the camera used for the enrollment process and for obtaining the on-site image (light intensity, focal length, color balance, etc.), the better the FRT will perform.
        - *Gallery Size*: Given that the number of possible images that enter the gallery as near-identical mathematical representations (biometric doubles) increases as the size of the gallery increases, restricting the size of the gallery in "open set" identification applications (such as watch list applications) may help maintain the integrity of the system and increase overall performance.

   e. The selection and composition of images that are used to develop FRT algorithms are crucial in shaping the eventual performance of the system.

2. *Evaluations:* How are evaluations reported? How should results be interpreted? How might evaluation procedures be revised to produce more useful and transparent results?
   a. Many of the existing evaluation results do not lend themselves to clear comparisons or definitive conclusions. The results of "close set" performance evaluations, for instance, which are based on the assumption that all possible individuals who might be encountered by the FRT are known in advance (i.e., there are no outside imposters), cannot be compared across different tests or with "open set" (i.e., where there could be imposters) performance figures, and do not reflect or predict performance of an FRT in operational conditions (which are always "open set"). "Close set" evaluation results are contingent on the size of the gallery and rank number (see below) in the specific evaluation; they are thus fundamentally incommensurate with one another. "Open set" evaluation results are equally difficult to compare, as there is no way to predict in advance the number of imposters an FRT might encounter and therefore produce a standard performance baseline.
   b. The current lack of publicly available data on operational (i.e., *in situ*)—as compared to laboratory—evaluations of FRT is a major concern for organizations that may want to consider the use of FRT. Without such evaluations, organizations are dependent on claims made by the FRT vendors themselves.
   c. Evaluations should always include tests under full operational conditions, as these are the only tests that offer a real-world measure of the practical capabilities of FRT. These results, however, should not be casually generalized to other operational conditions.
   d. More informative and rigorous tests would make use of gallery and evaluation images compiled by an independent third party, under a variety of conditions with a variety of cameras, as in the case of the current round of government-sponsored testing known as the Multibiometric Grand Challenge (MBGC).
   e. Evaluation results must be read with careful attention to pre-existing correlations between the images used to develop and train the FRT algorithm and the images that are then used to evaluate the FRT algorithm and system. Tightly correlated training (or gallery) and evaluation data could artificially inflate the results of performance evaluations.

3. *Operation*: What decisions must be made when deciding to adopt, install, operate, and maintain FRT?
   a. It is up to a system's developers and operators to determine at what threshold of similarity between a probe and gallery image (the similarity score threshold) they wish the system to recognize an individual. Threshold decisions will always be a matter of policy and should be context and use-specific.
   b. For instance, a system with a high threshold, which demands a high similarity score to establish credible recognition in the verification task, would decrease the number of individuals who slip past the system (false accept mistakes), but would also increase the number of individuals who would be incorrectly rejected (false reject mistakes). These trade-offs must be determined, with a clear sense of how to deal with the inevitable false rejections and acceptances.
   c. The rank number, which is the number of rank-ordered candidates on a list of the percent most likely matches for any given probe image, is a matter of policy determination. At rank 10, for example, successful recognition would be said to have occurred if the specific individual appeared as any of the top 10 candidates.
   d. The images that are used to develop and train the FRT algorithm and system should reflect, as much as possible, the operational conditions under which the system will perform, both in terms of the characteristics of the individuals in the images (ethnicity, race, gender, age, etc.) and the conditions under which the images are captured (illumination, pose, the orientation of the face, etc.). This will facilitate a high level of performance.
   e. There is an inherent trade-off in the identification task between the size of the gallery and performance; who, then, should be included in the gallery, and why?

4. *Policy concerns*: What policies should guide the implementation, operation, and maintenance of FRT?
    a. Given that a system performs best when developed for its specific context of use, FRT should be treated as purpose-built, one-off systems.
    b. Those who consider the use of FRT should have a very clear articulation of the implementation purpose and a very clear understanding of the environment in which the technology will be implemented when they engage with application providers or vendors.
    c. Integration with broader identity management and security infrastructure needs to be clearly thought through and articulated.
    d. The decision to install a covert, rather than overt, FRT will entail a number of important operational and ethical considerations, not least the related decision to make enrollment in the system mandatory or voluntary. In either case, special attention should be paid to the way in which enrollment is undertaken.
    e. FRT in operational settings requires highly trained and professional staff. It is important that they understand the operating tolerances and are able to interpret and act appropriately given the exceptions generated by the system.
    f. All positive matches in the identification task should be treated, in the first instance, as potential false positives until verified by other overlapping and/or independent sources.
    g. The burden placed on (falsely) identified subjects, for a given threshold, should be proportionate to the threat or risks involved.

5. *Moral and political considerations*: What are the major moral and political issues that should be considered in the decision to adopt, implement, and operate FRT?
    a. FRT needs to be designed so that it does not disrupt proper information flows (i.e., does not allow “private” information to be accessed or shared improperly). What defines “private” information and what is improper access or transmission is context-specific and should be treated as such.
    b. There are a number of questions that should be asked of any FRT or biometric identification system:
        - Are subjects aware that their images have been obtained for and included in the gallery database? Have they consented? In what form?
        - Have policies on access to the gallery been thoughtfully determined and explicitly stated?
        - Are people aware that their images are being captured for identification purposes? Have and how have they consented?
        - Have policies on access to all information captured and generated by the system been thoughtfully determined and explicitly stated?
        - Does the deployment of FRT in a particular context violate reasonable expectations of subjects?
        - Have policies on the use of information captured via FRT been thoughtfully determined and explicitly stated?
        - Is information gleaned from FRT made available to external actors and under what terms?
        - Is the information generated through FRT used precisely in the ways for which it was set up and approved?
    c. The implementation of FRT must also ensure that its risks are not disproportionately borne by, or the benefits disproportionately flow to, any particular group.
    d. The benefits of FRT must be weighed against the possible adverse effects it may have on subjects’ freedom and autonomy. The degree to which FRT may discourage the freedom to do legal and/or morally correct actions for fear of reprisal must be taken into account.
    e. FRT may create new security risks if not deployed and managed carefully. Any use of these technologies must, at a minimum, answer these questions:
        - Does the implementation of the system include both policy and technology enforced protection of data (gallery images, probe images, and any data associated with these images)?
        - If any of this information is made available across networks, have necessary steps been taken to secure transmission as well as access policies?

CONTENTS

1. Purpose and scope of this report

This report is primarily addressed to three audiences: decision-makers in law enforcement and security considering the purchase, investment in, or implementation of facial recognition technology (FRT); policy makers considering how to regulate the development and uses of facial recognition and other biometric systems; and researchers who perform social or political analysis of technology.

The main objective of the report is to bridge the divide between a purely technical and a purely socio-political analysis of FRT. On the one side, there is a huge technical literature on algorithm development, grand challenges, vendor tests, etc., that talks in detail about the technical capabilities and features of FRT but does not really connect well with the challenges of real world installations, actual user requirements, or the background considerations that are relevant to situations in which these systems are embedded (social expectations, conventions, goals, etc.). On the other side, there is what one might describe as the "soft" social science literature of policy makers, media scholars, ethicists, privacy advocates, etc., which talks quite generally about biometrics and FRT, outlining the potential socio-political dangers of the technology. This literature often fails to get into relevant technical details and often takes for granted that the goals of biometrics and FRT are both achievable and largely Orwellian. Bridging these two literatures—indeed, points of view—is very important as FRT increasingly moves from the research laboratory into the world of socio-political concerns and practices.

We intend this report to be a general and accessible account of FRT for informed readers. It is not a "state of the art" report on FRT. Although we have sought to provide sufficient detail in the account of the underlying technologies to serve as a foundation for our functional, moral, and political assessments, the technical description is not intended to be comprehensive.[1] Nor is it a comprehensive socio-political analysis. Indeed, for a proper, informed debate on the socio-political implications of FRT, more detailed and publicly accessible *in-situ* studies are needed. The report should provide a sound basis from which to develop such *in-situ* studies. The report instead attempts to straddle the technical and the socio-political points of view without oversimplifying either.

Accordingly, we have structured the report in nine sections. The first section, which you are currently reading, introduces the report and lays out its goals. In the second section, we introduce FRT within the more general context of biometric technology. We suggest that in our increasingly globalized world, where mobility has almost become a fact of social life, identity management emerges as a key socio-political and technical issue. Tying identity to the body through biometric indicators is seen as central to the governance of people (as populations) in the existing and emerging socio-political order, nationally and internationally, in all spheres of life, including governmental, economic, and personal. In the third section, we introduce FRT. We explain, in general terms, how the recognition technology functions, as well as the key tasks it is normally deployed to perform: verification, identification, and watch-list monitoring. We then proceed to describe the development of FRT in terms of the different approaches to the problem of automated facial recognition, measures of accuracy and success, and the nature and use of face image data in the development of facial recognition algorithms. We establish a basic technical vocabulary which should allow the reader to imagine the potential function of FRT in a variety of application scenarios. In section four, we discuss some of these application scenarios in terms of both existing applications and future possibilities. Such a discussion naturally leads to questions regarding the actual capabilities and efficacy of FRT in specific scenarios. In section five, we consider the various types of evaluation to which FRT is commonly subjected: technical, scenario, and operational. In technical evaluations, certain features and capabilities of the technology are examined in a controlled (i.e., reproducible) laboratory environment. At the other extreme, operational evaluations of the technology examine systems *in situ* within actual operational contexts and against a wide range of metrics. Somewhere in the middle, scenario evaluations, equivalent to prototype testing, assess the performance of a system in a staged setup similar to ones anticipated in future *in situ* applications. These different evaluations provide a multiplicity of answers that can inform stakeholders' decision-making in a variety of ways. In the final sections of the report, we focus on three of these aspects of concern: efficacy, policy, and ethical implications. In section six, we consider some of the conditions that may limit the efficacy of the technology as it moves from the laboratory to the operational context. In section seven, we consider some of the policy implications that flow from the evaluations that we considered in section five, and in section eight we consider some of the ethical implications that emerge from our understanding and evaluation of the technology. We conclude the report in the ninth section with some open questions and speculations.

2. Biometrics and identification in a global, mobile world ("why is it important?")

Although there has always been a need to identify individuals, the requirements of identification have changed in radical ways as populations have expanded and grown increasingly mobile. This is particularly true for the relationships between institutions and individuals, which are crucial to the well-being of societies, and necessarily and increasingly conducted impersonally—that is, without persistent direct and personal interaction. Importantly, these impersonal interactions include relationships between government and citizens for purposes of fair allocation of entitlements, mediated transactions with e-government, and security and law enforcement. Increasingly, these developments also encompass relationships between actors and clients or consumers based on financial transactions, commercial transactions, provision of services, and sales conducted among strangers, often mediated through the telephone, Internet, and the World Wide Web. Biometric technologies have emerged as promising tools to meet these challenges of identification, based not only on the faith that "the body doesn't lie," but also on dramatic progress in a range of relevant technologies. These developments, according to some, herald the possibility of automated systems of identification that are accurate, reliable, and efficient.

Many identification systems comprise three elements: *attributed identifiers* (such as name, Social Security number, bank account number, and drivers' license number), *biographical identifiers* (such as address, profession, and education), and *biometric identifiers* (such as photographs and fingerprint). Traditionally, the management of identity was satisfactorily and principally achieved by connecting attributed identifiers with biographical identifiers that were anchored in existing and ongoing local social relations.[2] As populations have grown, communities have become more transient, and individuals have become more mobile, the governance of people (as populations) required a system of identity management that was considered more robust and flexible. The acceleration of globalization imposes even greater pressure on such systems as individuals move not only among towns and cities but across countries. This progressive disembedding from local contexts requires systems and practices of identification that are not based on geographically specific institutions and social networks in order to manage economic and social opportunities as well as risks.

In this context, according to its proponents, the promise of contemporary biometric identification technology is to strengthen the links between attributed and biographical identity and create a stable, accurate, and reliable identity triad. Although it is relatively easy for individuals to falsify—that is, tear asunder—attributed and biographical identifiers, biometric identifiers—an individual's fingerprints, handprints, irises, face—are conceivably more secure because it is assumed that "the body never lies" or differently stated, that it is very difficult or impossible to falsify biometric characteristics. Having subscribed to this principle, many important challenges of a practical nature nonetheless remain: deciding on which bodily features to use, how to convert these features into usable representations, and, beyond these, how to store, retrieve, process, and govern the distribution of these representations.

Prior to recent advances in the information sciences and technologies, the practical challenges of biometric identification had been difficult to meet. For example, passport photographs are amenable to tampering and hence not reliable; fingerprints, though more reliable than photographs, were not amenable, as they are today, to automated processing and efficient dissemination. Security as well as other concerns has turned attention and resources toward the development of automatic biometric systems. An automated biometric system is essentially a pattern recognition system that operates by acquiring biometric data (a face image) from an individual, extracting certain features (defined as mathematical artifacts) from the acquired data, and comparing this feature set against the biometric template (or representation) of features already acquired in a database. Scientific and engineering developments—such as increased processing power, improved input devices, and algorithms for compressing data, by overcoming major technical obstacles, facilitates the proliferation of biometric recognition systems for both verification and identification and an accompanying optimism over their utility. The variety of biometrics upon which these systems anchor identity has burgeoned, including the familiar fingerprint as well as palm print, hand geometry, iris geometry, voice, gait, and, the subject of this report, the face. Before proceeding with our analysis and evaluation of facial recognition systems (FRS), we will briefly comment on how FRS compares with some other leading biometric technologies.

In our view, the question of which biometric technology is "best" only makes sense in relation to a rich set of background assumptions. While it may be true that one system is better than another in certain performance criteria such as accuracy or difficulty of circumvention, a decision to choose or use one system over another must take into consideration the constraints, requirements, and purposes of the use-context, which may include not only technical, but also social, moral and political factors. It is unlikely that a single biometric technology will be universally applicable, or ideal, for all application scenarios. Iris scanning, for example, is very accurate but requires expensive equipment and usually the active participation of subjects willing to submit to a degree of discomfort, physical proximity, and intrusiveness—especially when first enrolled—in exchange for later convenience (such as the Schiphol Privium system[3]). In contrast, fingerprinting, which also requires the active participation of subjects, might be preferred because it is relatively inexpensive and has a substantial historical legacy.[4]

Facial recognition has begun to move to the forefront because of its purported advantages along numerous key dimensions. Unlike iris scanning which has only been operationally demonstrated for relatively short distances, it holds the promise of identification at a distance of many meters, requiring neither the knowledge nor the cooperation of the subject.[5] These features have made it a favorite for a range of security and law enforcement functions, as the targets of interest in these areas are likely to be highly uncooperative, actively seeking to subvert successful identification, and few—if any—other biometric systems offer similar functionality, with the future potential exception of gait recognition. Because facial recognition promises what we might call "the grand prize" of identification, namely, the reliable capacity to pick out or identify the "face in the crowd," it holds the potential of spotting a known assassin among a crowd of well-wishers or a known terrorist reconnoitering areas of vulnerability such as airports or public utilities.[6] At the same time, rapid advancements in contributing areas of science and engineering suggest that facial recognition is capable of meeting the needs of identification for these critical social challenges, and being realistically achievable within the relatively near future.

The purpose of this report is to review and assess the current state of FRT in order to inform policy debates and decision-making. Our intention is to provide sufficient detail in our description and evaluation of FRT to support decision-makers, public policy regulators, and academic researchers in assessing how to direct enormous investment of money, effort, brainpower, and hope—and to what extent it is warranted.

3. Introduction to FRT ("how does it work?")

Facial recognition research and FRT is a subfield in a larger field of pattern recognition research and technology. Pattern recognition technology uses statistical techniques to detect and extract patterns from data in order to match it with patterns stored in a database. The data upon which the recognition system works (such as a photo of a face) is no more than a set of discernable pixel-level patterns for the system, that is, the pattern recognition system does not perceive meaningful "faces" as a human would understand them. Nevertheless, it is very important for these systems to be able to locate or detect a face in a field of vision so that it is only the image pattern of the face (and not the background "noise") that is processed and analyzed. This problem, as well as other issues, will be discussed as the report proceeds. In these discussions we will attempt to develop the reader's understanding of the technology without going into too much technical detail. This obviously means that our attempts to simplify some of the technical detail might also come at the cost of some rigor. Thus, readers need to be careful to bear this in mind when they draw conclusions about the technology. Nevertheless, we do believe that our discussion will empower the policymaker to ask the right questions and make sense of the pronouncements that come from academic and commercial sources. In order to keep the discussion relatively simple, we will first discuss a FRT in its normal operation and then provide a more detailed analysis of the technical issues implied in the development of these systems.

3.1. FRT in operation

3.1.1. Overview

Figure 1 below depicts the typical way that a FRS can be used for identification purposes. The first step in the facial recognition process is the capturing of a face image, also known as the *probe image*. This would normally be done using a still or video camera. In principle, the capturing of the face image can be done with or without the knowledge (or cooperation) of the subject. This is indeed one of the most attractive features of FRT. As such, it could, in principle, be incorporated into existing good quality "passive" CCTV systems. However, as we will show below, locating a face in a stream of video data is not a trivial matter. The effectiveness of the whole system is highly dependent on the *quality*[7] and characteristics of the captured face image. The process begins with face detection and extraction from the larger image, which generally contains a background and often more complex patterns and even other faces. The system will, to the extent possible, "normalize" (or standardize) the probe image so that it is in the same format (size, rotation, etc.) as the images in the database. The normalized face image is then passed to the recognition software. This normally involves a number of steps such as *extracting* the features to create a biometric "template" or mathematical representation to be compared to those in the reference database (often referred to as the *gallery*). In an identification application, if there is a "match," an alarm solicits an operator's attention to verify the match and initiate the appropriate actions. The match may either be true, calling for whatever action is deemed appropriate for the context, or it may be false (a "false positive"), meaning the recognition algorithm made a mistake. The process we describe here is a typical identification task. FRS can be used for a variety of tasks. Let us consider these in more detail.

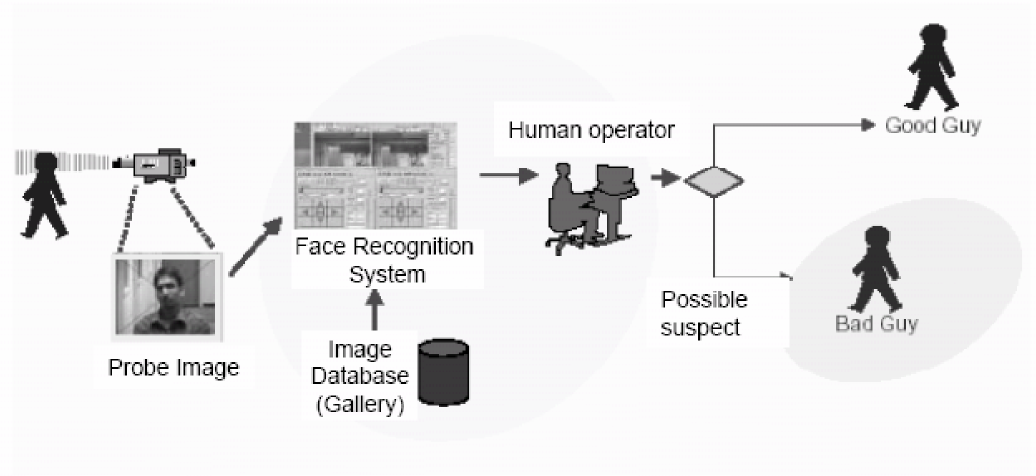
3.1.2. FRS tasks

FRS can typically be used for three different tasks, or combinations of tasks: verification, identification, and watch list.[9] Each of these represents distinctive challenges to the implementation and use of FRT as well as other biometric technologies.

Verification ("Am I the identity I claim to be?")

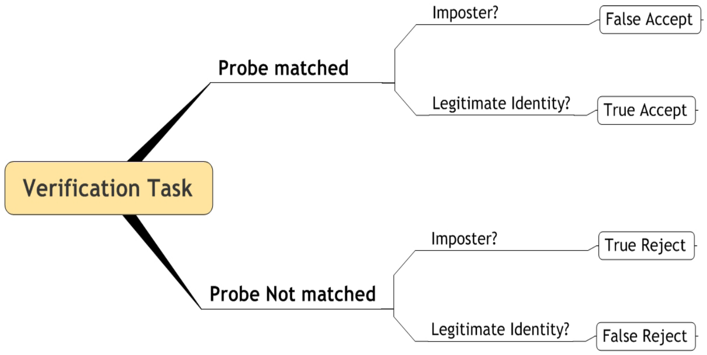
Verification or authentication is the simplest task for a FRS. An individual with a pre-existing relationship with an institution (and therefore already enrolled in the reference database or gallery) presents his or her biometric characteristics (face or probe image) to the system, claiming to be in the reference database or gallery (i.e. claiming to be a legitimate identity). The system must then attempt to match the probe image with the particular, claimed template in the reference database. This is a *one-to-one* matching task since the system does not need to check every record in the database but only that which corresponds to the claimed identity (using some form of identifier such as an employee number to access the record in the reference database). There are two possible outcomes: (1) the person is not recognized or (2) the person is recognized. If the person is not recognized (i.e., the identity is not verified) it might be because the person is an imposter (i.e., is making an illegitimate identity claim) or because the system made a mistake (this mistake is referred to as a *false reject*). The system may also make a mistake in accepting a claim when it is in fact false (this is referred to as a *false accept*). The relationship

Figure 1: Overview of FRS[8]



between these different outcomes in the verification task is indicated in Figure 2 . It will also be discussed further in section 3.1.3 below.

Figure 2: Possible outcomes in the verification task

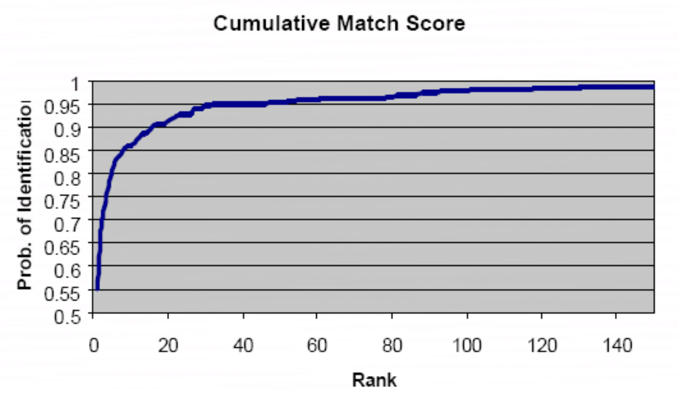


Identification ("Who am I or What is my identity?")

Identification is a more complex task than verification. In this case, the FRS is provided a probe image to attempt to match it with a biometric reference in the gallery (or not). This represents a *one-to-many* problem. In addition, we need to further differentiate between closed-set identification problems and open-set identification problems. In a *closed-set* identification problem we want to identify a person that *we know* is in the reference database or gallery (in other words for any possible identification we want to make we know beforehand that the person to be identified is in the database). *Open-set* identification is more complex in that *we do not know in advance* whether the person to be identified is or is not in the reference database. The outcome of these two identification problems will be interpreted differently. If there is no match in the closed-set identification then we know the system has made a mistake (i.e., identification has failed (a false negative)). However in the open-set problem we do not know whether the system made a mistake or whether the identity is simply not in the reference database in the first instance. Real-world identification applications tend to be open-set identification problems rather than closed-set identification problems.

Let us assume a *closed-set identification problem* to start with. In this case the system must compare the probe image against a whole gallery of images in order to establish a match. In comparing the probe image with the images in the gallery, a similarity score is normally generated. These similarity scores are then sorted from the highest to the lowest (where the lowest is the similarity that is equal to the operating threshold). This means that a higher threshold would generate a shorter rank list and a lower threshold would generate a longer list. The operator is presented with a ranked list of possible matches in descending order. A probe image is correctly identified if the correct match has the highest similarity score (i.e., is placed as "rank 1" in the list of possible matches). The percentage of times that the highest similarity score is the correct match for all individuals submitted is referred to as the *top match score*. It is unlikely that the top match score will be 100% (i.e., that the match with the highest similarity score is indeed the correct match). Thus, one would more often look at the percentage of times that the correct match will be in the *n*th rank (i.e., in the top *n* matches). This percentage is referred to as the "closed-set" *identification rate*.

Figure 3: Cumulative Match Score[10]



The performance of a closed-set identification system will typically be described as having an identification rate at rank *n*. For example, a system that has a 99% identification rate at rank 3 would mean that the system will be 99% sure that the person in the probe image is in either position 1, 2, or 3 in the ranked list presented to the operator. Note that the final determination of which one the person actually happens to be is still left to the human operator. Moreover, the three faces on the rank list might look very similar, making the final identification far from a trivial matter. In particular, it might be extremely difficult if these faces are of individuals that are of a different ethnic group to that of the human operator who must make the decision. Research has shown that humans have extreme difficulty in identifying individuals of ethnic groups other than their own.[11] A graph that plots the size of the rank order list against the identification rate is called a *Cumulative Match Score* (also known as the *Cumulative Match Characteristic)* graph, as shown in Figure 3.

As indicated in Figure 3, the identification problem in open-set evaluations is typically described in a different manner since a non-match might be a mistake (the identity was in the reference database but was not matched) or it might be that the person was not in the database at all. Thus, open-set identification provides an additional problem, namely how to separate these two possible outcomes. This is important for a variety of reasons. If it is a mistake (i.e., a false negative) then the recognition can be improved by using a better quality probe image or lowering the recognition threshold (i.e., the threshold used for similarity score between the probe and the gallery image). If, however, it is a true negative then such actions may not be beneficial at all. In the case of resetting the threshold it might lead to overall performance degradation (as will discuss below). This underscores the importance of having contextual information to facilitate the decision process. More specifically, open-set identification ought to function as part of a broader intelligence infrastructure rather than a "just in case" technology (this will also be discussed further below). The relationship between these different outcomes in the identification task is indicated in Figure 4 below. It will also be discussed further in section 3.1.3 below.

Watch list ("is this one of the suspects we are looking for?")

The watch list task is a specific case of an *open-set* identification task. In the watch list task, the system determines if the probe image corresponds to a person on the watch list and then subsequently identifies the person through the match (assuming the identities of the watch list are known). It is therefore also a *one-to-many* problem but with an open-set assumption. When a probe is given to the system, the system compares it with the entire gallery (also known in this case as the watch list). If any match is above the operating threshold, an alarm will be triggered. If the top match is identified correctly, then the task was completed successfully. If however the person in the probe image is not someone in the gallery and the alarm was nonetheless triggered, then it would be a false alarm (i.e., a false alarm occurs when the top match score for someone not in the watch list is above the operating threshold). If there is not an alarm then it might be that the probe is not in the gallery (a true negative) or that the system failed to recognise a person on the watch list (a false negative). The relationship between these different outcomes in the watch list task is indicated in Figure 5 below. It will also be discussed further in section 3.1.3 below.

Figure 4: Possible outcomes in the identification task

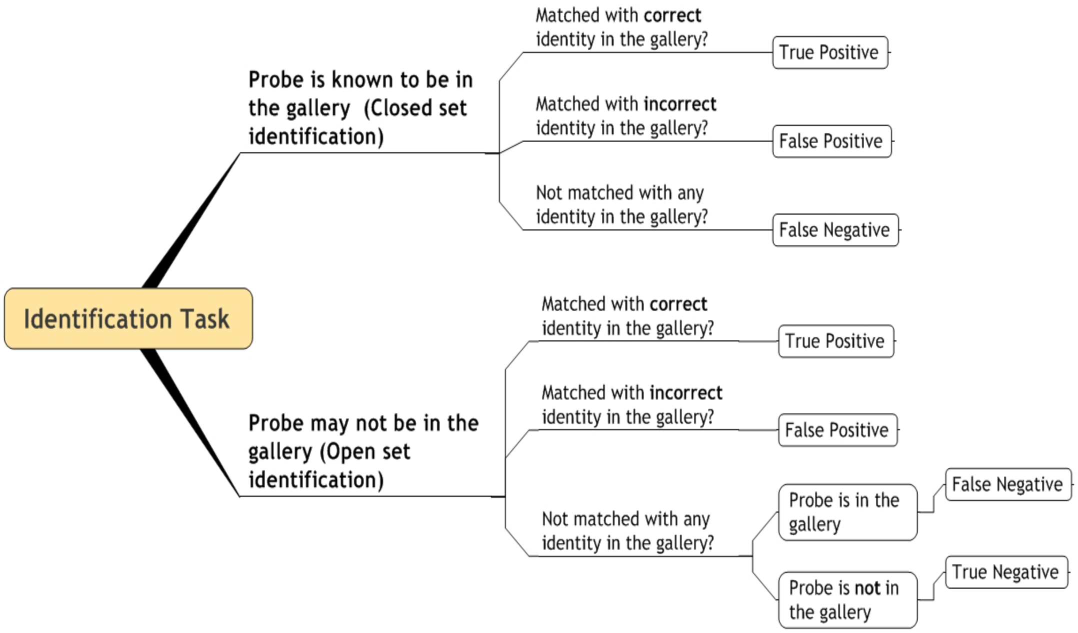
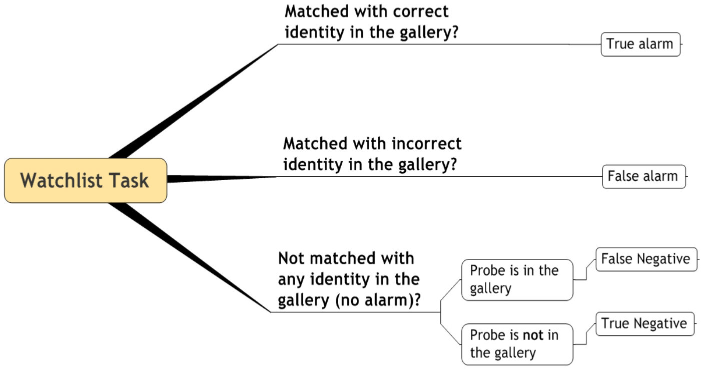

Identification Task
Probe is known to be in the gallery (Closed set identification)
Matched with correct identity in the gallery?
True Positive
Matched with incorrect identity in the gallery?
False Positive
Not matched with any identity in the gallery?
False Negative
Probe may not be in the gallery (Open set identification)
Matched with correct identity in the gallery?
True Positive
Matched with incorrect identity in the gallery?
False Positive
Not matched with any identity in the gallery?
Probe is in the gallery
False Negative
Probe is not in the gallery
True Negative

Figure 5: Possible outcomes in the watch list task


Watchlist Task
Matched with correct identity in the gallery?
True alarm
Matched with incorrect identity in the gallery?
False alarm
Not matched with any identity in the gallery (no alarm)?
Probe is in the gallery
False Negative
Probe is not in the gallery
True Negative

3.1.3. Interpreting FRS performance against tasks

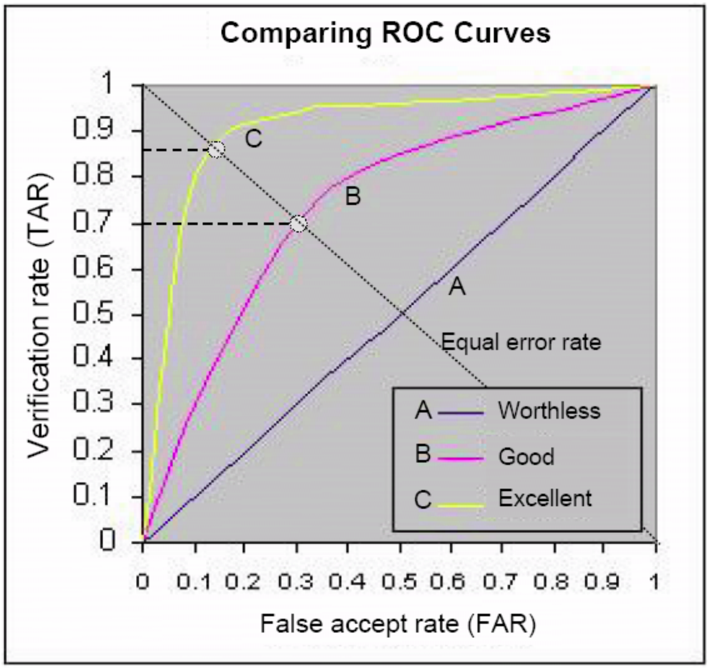
The *matching* of a probe against the gallery or reference database is never a simple binary decision (i.e., matched or not matched). The comparison between the probe and the template in the reference database produces a *similarity score*. The identity claim is accepted if the similarity score meets the threshold criteria and rejected if it does not meet it.[12] These thresholds are determined by implementation choices made with regard to specific operational conditions (in considering this threshold rate, one might want to refer to the discussion of the equal error rate below).

When setting the threshold there is always a tradeoff to be considered. For example, if the threshold for a similarity score is set too high in the verification task, then a legitimate identity claim may be rejected (i.e., it might increase the *false reject rate* (FRR)). If the threshold for a similarity score is set too low, a false claim may be accepted (i.e., the *false accept rate* (FAR) increases). Thus, within a given system, these two error measures are one another's counterparts.[13] The FAR can only be decreased at the cost of a higher FRR, and FRR can only be decreased at the cost of a higher FAR.

The Receiver Operating Characteristic (ROC) graph represents the probability of correctly accepting a legitimate identity claim against the probability of incorrectly accepting an illegitimate identity claim for a given threshold. Because the ROC allows for false positives from impostors, it is the metric used in open-set testing, both for verification and identification. To make this relationship more evident, let us consider the three ROC curves in the graph in Figure 6 for a verification task. In this graph, we see that the ROC curve for system A indicates that this system cannot discriminate at all. An increase in the verification rate leads to exactly the same level of increase in the FAR for any chosen operating threshold. This system cannot discriminate in either direction. This will be equal to a system of random decision-making in which there is an equal probability of being accepted or rejected irrespective of the operating threshold.

System B is better because one can obtain a large degree of improvement in the verification rate for a small increase in the FAR rate, up to a verification rate of approximately 70%. After this point there is an exponential increase in the FAR for small increases in the verification rate of the system. System C is the best system since there is a relatively small increase in the FAR for a large increase in verification rate up to a rate of approximately 86%.

Figure 6: Example ROC curves for three different systems in the verification task



Performance accuracy in the open-set case is therefore a two-dimensional measurement of both the verification (or true accept rate) and false accept rates *at a particular threshold*.[14] The perfect system will give 100% verification for a 0% FAR. Such a system does not exist and probably will never exist except under very constrained conditions in controlled environments, which will be of little, if any, practical use. An alternative approach is to use the Detection Error Trade-off (DET) Curve. The DET curves typically plots matching error rates (false non-match rate vs. false match rate) or decision error rates (false reject rate vs. false accept rate).

Some authors also use the *equal error rate* (EER) curve to describe the recognition performance of a FRS. The equal error rate is the rate at which the FAR is exactly equal to the FRR. This is represented by the straight line connecting the upper left corner (coordinates 0, 1) to the lower right corner (coordinates 1, 0). The equal error rate is the point at which the ROC curve intersects with the ERR curve—this is approximately 70% for System B, 86% for System C, and 50% for System A. This seems correct, as we would expect a system such as System A that randomly accepts or rejects identities (based on perfect chance) to have a 50% likelihood to either accept or reject an identity—given a large enough population and a large enough number of attempts. We must however note that one point on the curve is not adequate to fully explain the performance of biometric systems used for verification. This is especially true for real life applications where operators prefer to set system parameters to achieve either a low FAR or high probability of verification. Nevertheless, it might be a good starting point when thinking about an operating policy. It would then be a matter of providing a justification for why one might want to move away from it. For example, one might want to use the system as a filtering mechanism where one decreases the FAR (and simultaneously increase the FRR) but put in place a procedure to deal with these increased incidents of false rejects. Or one might want to determine and assign costs to each type of error, for instance the social cost of misidentifying someone in a particular context or the financial costs of granting access based on misidentification. Managers and policymakers might then settle on what is perceived to be a suitable trade-off. Obviously it is never as clear cut as this—especially if the particular implementation was not subject to adequate *in situ* operational evaluation.

Sometimes the ROC is presented in a slightly different way. For example, in the Face Recognition Vendor Test 2002, the FAR was represented with a logarithmic scale because analysts are often only concerned with FAR at verification rates between 90% and 99.99%. Remember that a FAR of 0.001 (verification rate of 99.9%) will still produce a 1/1000 false accept rate. It is possible to imagine an extreme security situation where the incidence of impostors is expected to be high, and the risk of loss very great, where this may still be unacceptable. It is important to understand the graphing conventions used when interpreting ROC graphs (or any statistical graph for that matter).

We now have a sense of how a FRS works, the sort of tasks it does, and how successes in these tasks are reported. Let us now describe the development of these systems in more detail by considering the following:

- The typical recognition steps performed by an facial recognition algorithm
- The different types of facial recognition algorithms
- The different types of image data used in the facial recognition process.

3.2. The development of FRS

In order to appreciate the complexity (and susceptibilities) of FRT, we need to get a sense of all the complex tasks that make up a system and how small variations in the system or environment can impact on these tasks. We will endeavor to keep the discussion on a conceptual level. However, from time to time, we will need to dig into some of the technical detail to highlight a relevant point. We will structure our discussion by starting with the key components (algorithms) of the system and then look at data and environment. The intention is to give the reader a general sense of the technology and some of the issues that emerge as a result of the technical design features and challenges, rather than providing a state of the art discussion.

3.2.1. Facial recognition algorithms

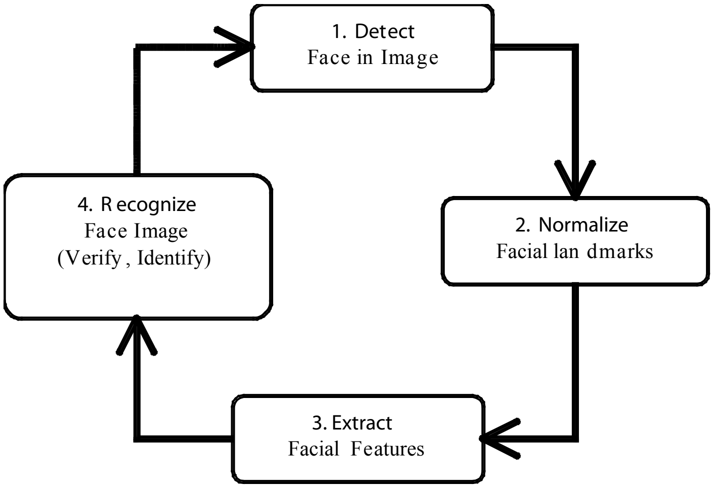
Steps in the facial recognition process

Let us for the moment assume that we have a probe image with which to work. The facial recognition process normally has four interrelated phases or steps. The first step is face detection, the second is normalization, the third is feature extraction, and the final cumulative step is face recognition. These steps depend on each other and often use similar techniques. They may also be described as separate components of a typical FRS. Nevertheless, it is useful to keep them conceptually separate for the purposes of clarity. Each of these steps poses very significant challenges to the successful operation of a FRS. Figure 7 indicates the logical sequence of the different steps.

Detecting a face: Detecting a face in a probe image may be a relatively simple task for humans, but it is not so for a computer. The computer has to decide which pixels in the image is part of the face and which are not. In a typical passport photo, where the background is clear, it is easy to do, but as soon as the background becomes cluttered with other objects, the problem becomes extremely complex. Traditionally, methods that focus on facial landmarks (such as eyes), that detect face-like colors in circular regions, or that use standard feature templates, were used to detect faces.

Normalization: Once the face has been detected (separated from its background), the face needs to be normalized. This means that the image must be standardized in terms of size, pose, illumination, etc., relative to the images in the gallery or reference database. To normalize a probe image, the key facial landmarks must be located accurately. Using these landmarks, the normalization algorithm can (to some degree) reorient the image for slight variations. Such corrections are, however, based on statistical inferences or approximations which may not be entirely accurate. Thus, it is essential that the probe is as close as possible to a standardized face.[15] Facial landmarks are the key to all systems, irrespective of the overall method of recognition. If the facial landmarks cannot be located, then the recognition process will fail. Recognition can only succeed if the probe image and the gallery images are the same in terms of pose orientation, rotation, scale, size, etc. Normalization ensures that this similarity is achieved—to a greater or lesser degree.

Figure 7: Steps in the facial recognition prcess



Feature extraction and recognition: Once the face image has been normalized, the feature extraction and recognition of the face can take place. In feature extraction, a mathematical representation called a *biometric template* or *biometric reference* is generated, which is stored in the database and will form the basis of any recognition task. Facial recognition algorithms differ in the way they translate or transform a face image (represented at this point as grayscale pixels) into a simplified mathematical representation (the "features") in order to perform the recognition task (algorithms will be discussed below). It is important for successful recognition that maximal information is retained in this transformation process so that the biometric template is sufficiently distinctive. If this cannot be achieved, the algorithm will not have the discriminating ability required for successful recognition. The problem of biometric templates from different individuals being insufficiently distinctive (or too close to each other) is often referred to as the generation of *biometric doubles* (to be discussed below). It is in this process of mathematical transformation (feature extraction) and matching (recognition) of a biometric template that particular algorithms differ significantly in their approach. It is beyond the scope of this report to deal with these approaches in detail. We will merely summarize some of the work and indicate some of the issues that relate to the different approaches.

Face recognition algorithms[16]

The early work in face recognition was based on the geometrical relationships between facial landmarks as a means to capture and extract facial features. This method is obviously highly dependent on the detection of these landmarks (which may be very difficult is variations in illumination, especially shadows) as well as the stability of these relationships across pose variation. These problems were and still remain significant stumbling blocks for face detection and recognition. This work was followed by a different approach in which the face was treated as a general pattern with the application of more general pattern recognition approaches, which are based on photometric characteristics of the image. These two starting points: *geometry* and the *photometric* approach are still the basic starting points for developers of facial recognition algorithms. To implement these approaches a huge variety of algorithms have been developed.[17] Here we will highlight three of the most significant streams of work: Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), and Elastic Bunch Graph Matching (EBGM).

Principal Components Analysis (PCA)

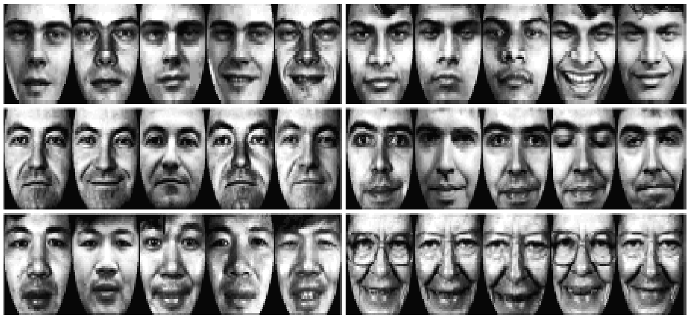
The PCA technique[18] converts each two dimensional image into a one dimensional vector. This vector is then decomposed into orthogonal (uncorrelated) principle components (known as eigenfaces)—in other words, the technique selects the features of the image (or face) which vary the most from the rest of the image. In the process of decomposition, a large amount of data is discarded as not containing significant information since 90% of the total variance in the face is contained in 5-10% of the components. This means that the data needed to identify an individual is a fraction of the data presented in the image. Each face image is represented as a weighted sum (feature vector) of the principle components (or eigenfaces), which are stored in a one dimensional array. Each component (eigenface) represents only a certain feature of the face, which may or may not be present in the original image. A probe image is compared against a gallery image by measuring the distance between their respective feature vectors. For PCA to work well the probe image must be similar to the gallery image in terms of size (or scale), pose, and illumination. It is generally true that PCA is reasonably sensitive to scale variation.

LDA: Linear Discriminant Analysis

LDA[19] is a statistical approach based on the same statistical principles as PCA. LDA classifies faces of unknown individuals based on a set of training images of known individuals. The technique finds the underlying vectors in the facial feature space (vectors) that would maximize the variance *between* individuals (or classes) and *minimize* the variance within a number of samples of the same person (i.e., within a class).

If this can be achieved, then the algorithm would be able to discriminate between individuals and yet still recognize individuals in some varying conditions (minor variations in expression, rotation, illumination, etc.). If we look at Figure 8 we can see that there is a relatively large amount of variation between the individuals and small variations between the varieties of poses of the same individual. To do this the algorithm must have an appropriate training set. The database should contain several examples of face images for each subject in the training set and at least one example in the test set. These examples should represent different frontal views of subjects with minor variations in view angle. They should also include different facial expressions, different lighting and background conditions, also examples with and without glasses if appropriate. Obviously, an increase in the number of varying samples of the same person will allow the algorithm to optimize the variance between classes and therefore become more accurate. This may be a serious limitation in some contexts (also known as the small sample size problem). As for PCA, LDA works well if the probe image is relatively similar to the gallery image in terms of size, pose, and illumination. With a good variety in sampling this can be somewhat varied, but only up to a point. For more significant variation other non-linear approaches are necessary.

Figure 8: Example of variation between and within classes[20]



Elastic Bunch Graph Matching (EBGM)

EBGM relies on the concept that real face images have many nonlinear characteristics that are not addressed by the linear analysis methods such as PCA and LDA—such as variations in illumination, pose, and expression. The EBGM method places small blocks of numbers (called "Gabor filters") over small areas of the image, multiplying and adding the blocks with the pixel values to produce numbers (referred to as "jets") at various locations on the image. These locations can then be adjusted to accommodate minor variations. The success of Gabor filters is in the fact that they remove most of the variability in images due to variation in lighting and contrast. At the same time they are robust against small shifts and deformations. The Gabor filter representation increases the dimensions of the feature space (especially in places around key landmarks on the face such as the eyes, nose, and mouth) such that salient features can effectively be discriminated. This new technique has greatly enhanced facial recognition performance under variations of pose, angle, and expression. New techniques for illumination normalization also enhance significantly the discriminating ability of the Gabor filters.

3.2.2. Developmental image data

An important part of the development of FRT is the training of the algorithms with a set of images usually referred to as the *developmental set*. This is done by exposing the algorithms to a set of images so that the algorithms can learn how to detect faces and extract features from these faces. The designers will study the results of exposing the algorithms to the training set and fine-tune the performance by adjusting certain aspects of the algorithm. It should be clear that the selection and composition of the developmental set images will be very important in shaping the eventual performance of the algorithms. Indeed, the developmental set should at least reflect, as much as possible, the operational conditions under which the system will perform or function, both in terms of the characteristics of the individuals in the images (ethnicity, race, gender, etc.) and the conditions under which the images are captured (illumination, pose, size of image, etc.). There is also an issue when it comes to the developmental sets used in the evaluation of a FRS. If it is possible to improve the performance of algorithms by having a close correlation between the developmental set and the evaluation set, then one needs to look very critically at the degree to which both the developmental and evaluation sets actually reflect the potential operational conditions under which that the technology will perform. All results of evaluations should be read, understood, and interpreted *relative to* sets against which they were developed and evaluated. Thus, it is important to note that it is not very helpful to evaluate a system against a set that is not representative of the data it will need to deal with in actual operational conditions *in situ* when making decision about actual implementation scenarios. This is especially true if the operational conditions or subject populations are likely to change from the initial point of evaluation. Determining appropriate thresholds for a FRS based on evaluations conducted under different conditions or with a radically different subject population would be problematic indeed.

3.2.3. The gallery (or enrolled) image data

The gallery is the set of biometric templates against which any verification or identification task is done. In order to create the gallery, images of each individual's face needs to be *enrolled* by the FRS. Enrollment into the system means that images have to go through the first three steps of the recognition process outlined above (i.e., face detection, normalization and feature extraction). This will then create a biometric template—stored in the gallery—against which probe images will be compared. It is self-evident that (from the discussion of algorithms above) that success of the verification and identification tasks will be significantly impacted by the close relationship between the images of the developmental database, the enrolled database and the probes.

The gallery can be populated in a variety of ways. In a typical verification scenario, the individual willingly surrenders his or her face image in a controlled environment so as to ensure a high quality image for the gallery. However, in some cases, especially in the case of identification or watch lists, the gallery image may not have been collected under controlled conditions.

3.2.4. The probe image data

It is true that the similarity of collection conditions of the probe image to the gallery and developmental images can make a significant difference in the performance of all FRS. Images collected under expected conditions will be called "good quality." Without a good quality probe image, the face and necessary landmarks, such as the eyes, cannot be located. Without the accurate location of landmarks, normalization will be unsuccessful, which will affect the performance of all algorithms. Without the accurate computation of facial features, the robustness of the approaches will also be lost. Thus, even the best of the recognition algorithms deteriorate as the quality of the probe image declines. Image quality is more significant than any other single factor in the overall performance of FRS.[21] According to the American National Standards Institute International Committee for Information Technology Standards (ANSI/INCITS) 385-2004 Face Recognition Format for Data Interchange, a good quality face image for use on passports:

- Is no more than 6 months old
- Is 35-40mm in width
- Is one in which the face takes up 70%-80% of the photograph
- Is in sharp focus and clear
- In one in which the subject is looking directly at the camera
- Shows skin tones naturally
- Has appropriate brightness and contrast
- Is color neutral
- Shows eyes open and clearly visible
- Shows subject facing square onto camera
- Has a plain light-colored background
- Has uniform lighting showing no shadows
- Shows subject without head cover (except for religious purposes)
- Where eye glasses do not obscure the eyes and are not tinted
- Has a minimum of 90 pixels between the eye centers.

More recently the International Standard Organization/International Electrotechnical Commission (ISO/IEC) released a very similar standard for "Best Practices for Face Images" (ISO/IEC 19794-5). If these ANSI and ISO/IEC standards are met (for both the gallery and the probe image) most of the top FRS will deliver a very high level of performance. It should be noted that this standard was created for images to be held with JPEG compression on e-passports, and thus in limited storage space. Recent NIST testing (FRVT 2006) has shown that higher resolution images than specified in the standard can lead to better performance. Thus, 19794-5 is only a reasonable standard under the limitation of storage space to the 2kB generally afforded on e-passports.

It seems obvious that such an image will not be easy to capture without the active participation of the subject. The surreptitious capture of face images is unlikely to meet these ideals and therefore liable to severely limit the potential performance of an FRS based on such images. The two most significant factors that affect the performance of FRS are *pose variation* (rotation of head in the X, Y, and Z axes) and *illumination* (the existence of shadows). It has been claimed that "variations between the images of the same face due to illumination and viewing direction are almost always larger than image variations due to change in face identity."[22] This will have important consequences when we discuss facial images obtained from elevated CCTV cameras. Pose variation and illumination problems make it extremely difficult to accurately locate facial landmarks. We will discuss these in more detail below. Different types of inputs have been proposed in order to deal with these problems. Let us consider these briefly.

Video stream input

Face recognition from image sequences captured by video cameras would seemingly be able to overcome some of the difficulties of pose and illumination variation since multiple poses will be available as the subject moves through a space. The information from all of these different angles could be collated to form a composite image that ought to be reasonably accurate. Additional temporal information can also augment spatial information. This form of input however also poses significant challenges:

- The quality of video is often low with a lot of clutter in the scene that makes face detection very difficult.
- Face images in video sequences are often very small (15 by 15 pixels), so obviously the ISO/IEC 19794-5 requirement for a minimum of 90 interoccular pixels cannot be met. Accurate detection and normalization is challenging with such small images.

There is a significant amount of research being done in this area, including the US government funded Multiple Biometric Grand Challenge (MBGC), but considerable challenges remain. It might be possible that this approach could be combined with other biometric data that can be collected "at a distance," such as gait recognition. Such combined (multi-modal) approaches seem to be a very promising avenue of research. Nevertheless, it seems clear that systems based on video tracking, detection, and recognition are in the early stages of development.

Three-dimensional (3D) input

Three-dimensional inputs seem like a logical way to overcome the problems of pose and illumination variation. A 3D profile of a face ought to provide much more information than a 2D image. Although this may be true, it is quite difficult to obtain an accurate 3D image in practice. 3D images are collected using 3D sensing technologies. There are currently three approaches: passive stereo, structured lighting and laser. In the first two of these approaches, it is very important that there is a known (and fixed) geometric relationship between the subject and the sensing devices. This means that it is necessary for the subject to participate in the capturing of the probe image or that the environment be controlled to such a degree that the geometric relationships can be determined with a certain degree of accuracy. This requirement will constrain the sort of applications that can use 3D input. In addition, it has been shown that the sensing approach is in fact sensitive for illumination variation: according to Bowyer, et al., "changes in the illumination of a 3D shape can greatly affect the shape description that is acquired by a 3D sensor."[23] A further complication is that FRS based on 3D shape images alone seem to be less accurate than systems that combine 2D and 3D data.[24]

Infra-red (IR) input

Another area of research concerns infrared thermal patterns as an input source. The thermal patterns of faces are derived primarily from the pattern of superficial blood vessels under the skin. The skin directly above a blood vessel is on average 0.1 degree centigrade warmer than the adjacent skin. The vein tissue structure of the face is unique to each person (even identical twins); the IR image is therefore also distinctive. The advantage of IR is that face detection is relatively easy. It is less sensitive to variation in illumination (and even works in total darkness) and it is useful for detecting disguises. However, it is sensitive to changes in the ambient environment, the images it produces are low resolution, and the necessary sensors and cameras are expensive. It is possible that there are very specific applications for which IR would be appropriate. It is also possible that IR can be used with other image technologies to produce visual and thermal fusion. Nevertheless, all multi-modal systems are computationally intensive and involve complex implementation issues.

Now that we have considered all the elements of a typical FRS, one might ask about the actual capabilities of the technology and potential implementation scenarios. In the next section, we will consider these issues. Of particular import for policy makers is the actual capability of the technology as evaluated by independent actors or agencies and in realistic operational situations.

4. Application scenarios for facial recognition systems (FRS)

Armed with this description of the core technical components of facial recognition and how they function together to form a system, we consider a few typical applications scenarios envisioned in the academic literature and promoted by systems developers and vendors. The examples we have selected are intended to reflect the wide-ranging needs FRS might serve, as well as diverse scenarios in which it might function.

In the scenario that we have called "the grand prize," an FRS would pick out targeted individuals in a crowd. Such are the hopes for FRS serving purposes of law enforcement, national security, and counterterrorism. Potentially connected to video surveillance systems (CCTV) already monitoring outdoor public spaces like town centers, the systems would alert authorities to the presence of known or suspected terrorists or criminals whose images are already enrolled in a system's gallery, or could also be used for tracking down lost children or other missing persons. This is among the most ambitious application scenarios given the current state of technology. Poor quality probe images due to unpredictable light and shadows in outdoor scenes, unpredictable facial orientation, and "noise" from cluttered backgrounds make it difficult for an FRS in the first place to even pick out faces in the images. Challenges posed by the lack of control inherent in most scenarios of this kind are exacerbated by the likelihood of uncooperative subjects. Additionally CCTV cameras are generally mounted high (for protection of the camera itself), looking down into the viewing space, thus imposing a pose angle from above which has been shown to have a strong negative impact on recognition[25] and operate at a distance for which obtaining adequate (90 pixel) interoccular resolution is difficult. In a future section we will see how the BKA "Fotofandung" test overcame these usual limitations.

In other typical application scenarios, one or more of the complicating factors may be controlled. Still in watch list mode with uncooperative targets such as terrorist or criminal suspects, an FRS setup might obtain higher quality probe images by taking advantage of the control inherent in certain places, such as portals. For example, in airports or sports arenas, foot traffic may be fairly stringently controlled in queues, turnstiles, passport inspection stations, or at security checkpoints where officers may, even indirectly, compel eye contact with passengers. (An application of this kind occurred at the Tampa Super Bowl XXXV, where spectators underwent surreptitious facial scans as they passed through stadium turnstiles.) A similar configuration of incentives and conditions exist in casinos, where proprietors on the lookout for undesirable patrons, such as successful card-counters, have the advantage of being able to control lighting, seating arrangements and movement patterns, but mount the cameras in the ceiling, making fully automated recognition impossible. An extension envisioned for such systems would follow targeted individuals though space, for example by tracking a suspected shoplifter moving up and down store aisles or a suspected terrorist making his or her way through an airport, but basing the recognition on clothing color or the top of the head, not the face, so that cameras could be ceiling mounted.

Scenarios in which FRS may be used for authentication or verification purposes include entry and egress to secured high-risk spaces, for example military bases, border crossings, and nuclear power plants, as well as access to restricted resources, such as personal devices, computers, networks, banking transactions, trading terminals, and medical records. In these environments, not only is movement controlled, cooperation is structured by the way incentives are organized, for example, subjects benefiting in some way from successful operation of a FRS.

There may also be scenarios that mix and match different elements. For example in controlled settings, such as an airplane-boarding gate, an FRS may be used in place of random checks merely to screen passengers for further investigation. In casinos, strategic design of betting floors that incorporates cameras at face height with good lighting, could be used not only to scan faces for identification purposes, but possibly to afford the capture of images to build a comprehensive gallery for future watch list, identification, and authentication tasks.[26] For the moment, however, such applications are not in place.

FRS might be used as back-end verification systems to uncover duplicate applications for benefits in scenarios that require other forms of identification. The United States, United Kingdom, New Zealand, Pakistan, and other countries have explored this approach for passport and visa applications and the states of Colorado and Illinois for issuance of drivers' licenses.[27] Another typical scenario, in the context of law-enforcement, is for police to apply FRS to the task of verifying the identity of lawfully detained and previously known individuals and to check whether the individual matches the images in a gallery of "mug shots."

Application scenarios likely to be less politically charged are those in which FRS does not require or call on a centralized gallery database. It is possible to construct a smartcard system in which facial image data is embedded directly in ID cards, for instance in drivers' licenses, passports, etc. In these instances, such as the Australian Customs SmartGate system to be discussed, probe images are simply compared against embedded images. Other applications not commonly discussed are for entertainment contexts, for example, to facilitate human-robot or human computer interaction, or virtual-reality training programs.[28]

When considering scenarios in which FRS may be applied, the question is not usually FR or nothing, but FR or another biometric, such as fingerprint, hand geometry, or iris scan. FRT has several advantages, as it imposes fewer demands on subjects and may be conducted at a distance without their knowledge or consent. Whether or when to select FRT also critically depends on how well it performs, not only how well each of the technical components described in previous sections functions, and not only how well they all work together, but how well they function in the scenarios and for the purposes to which they are applied. As far as we know, few application scenarios have been rigorously tested, or, at least, few results of such tests have been made public. There have, however, been several notable evaluations of FRT. In the sections that follow, we report on them, including the rare *in situ* applications we could find. A preview of these findings, in a nutshell, is that FRS functions best for verification tasks with galleries of high quality images.

5. The evaluation of FRT and FRS (does it actually work?)

One could divide the evaluations of FRT/FRS into three categories or types: *technological, scenario,* and *operational*.[29] Each of these evaluation types focuses on different aspects, uses a different approach, and serves a different purpose. Nevertheless, they should all feed into each other to form a coherent and integrated approach to evaluation. Ideally, the evaluation of a system that will serve a particular purpose starts with a technology evaluation, followed by a scenario evaluation, and finally an operational evaluation.

The purpose of a *technology* evaluation is to determine the underlying technical capabilities of a particular FRT against a database of images collected under previously determined conditions. Technology in this context is understood to be the different types of facial recognition algorithms. The evaluation is normally performed under laboratory conditions using a standardized data set that was compiled in controlled conditions (ensuring control over pose, illumination, background, resolution, etc.). Consequently, the evaluation determines the performance of the algorithms only under these specific conditions. The standardized setup and controlled conditions mean that such evaluations are always to a large degree repeatable, but do not extend to other collection conditions and other populations. The results of technology evaluations could be used by developers to refine their technology, but only under the tested conditions. However, the evaluation can also be used by potential customers to select the most appropriate technology for their particular application-requirements, provided that those requirements are the same as the test-image collection conditions. The most prominent example of technology evaluation in FRT is the Face Recognition Vendor Tests (FRVT) and the Facial Recognition Grand Challenge (FRGC) conducted by

National Institute of Standards and Technology (NIST). These will be discussed below.

The purpose of *scenario* evaluations is to evaluate the overall capabilities of the entire system for a specific application scenario, designed to model a real-world environment and population. This would include the image-capturing component (cameras, video, etc.), the facial recognition algorithms, and the application to which they would be put to use. Scenario evaluations are not always completely repeatable, but the approach used in conducting the evaluation can always be repeated. Scenario evaluations are more complex to set up and may take several months or even years to complete. They are often designed for multiple trials under varying conditions. Results from a scenario evaluation typically show areas that require future system integration work, as well as providing performance data on systems as they are used for a specific application. An example of a scenario evaluation is the Identix (FaceIT) scenario evaluation reported by Bone and Blackburn[30] and the *BioFace* evaluations which were performed in Germany.[31] The results of these will be discussed below.

The purpose of *operational* evaluation is to evaluate a system *in situ* (i.e., in actual operational conditions). Operational evaluations aim to study the impact of specific systems on the organization of workflow and the achievement of operational objectives. Operational evaluations tend not to be repeatable. These evaluations typically last from several months to a year or more since operational performance must be measured prior to the technology being embedded and again after implementation so that operational conditions and objectives can be compared. At present, there are limited publicly reported data available on operational evaluation of facial recognition systems. We will discuss the data that we could access below.

The lack of publicly available data on scenario and operational evaluations of FRT is a major concern for organizations that want to consider the use of FRT. Without such evaluations, organizations are often dependent on claims made by vendors of FRT. Moreover, it should be noted that evaluations do have a limited "shelf life." Evaluations such as those done at the National Physical Laboratory or the National Institute of Standards and Technology may require 2 or more years to design, execute and document, but if an evaluation is older than 18 months, the performance results may be outdated. Nevertheless, by reviewing older evaluations one can learn a lot about the sort of issues that are relevant to the actual performance of the technology. This is often helpful in interpreting more recent evaluations. Let us consider the technical, scenario, and operational evaluations that are publicly available in a bit more detail.

5.1. FRT technology evaluations

5.1.1. The Face Recognition Vendor Tests of 2002 (FRVT 2002)

The FRVT 2002[32] evaluation was a significant technology evaluation that followed in the footsteps of the earlier FRVT 2000 and the FERET evaluations of 1994, 1995, and 1996. In the FRVT 2002, ten FRS vendors participated in the evaluations. These were independent evaluations sponsored by a host of organizations such as Defense Advanced Research Projects Agency (DARPA), the Department of State, and the Federal Bureau of Investigation. The FRVT of 2002 were more significant than any of the previous evaluations due to:

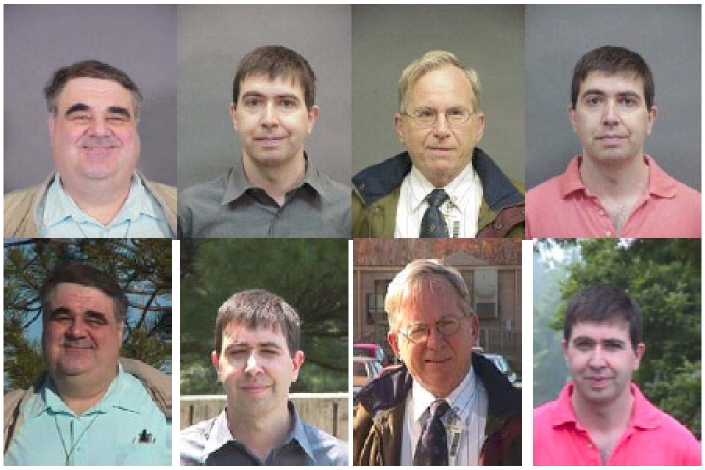
- The use of a large gallery (37,437 individuals)
- The use of a medium size database of outdoor and video images
- Some attention given to demographics

<u>*The data set of FRVT 2002*</u>

The large database (referred to as the HCInt data set) is a subset of a much larger database provided by the Visa Services Directorate, Bureau of Consular Affairs of the Department of State. The HCInt data set consisted of 121,589 images of 37,437 individuals with at least three images of each person. All individuals were from the Mexican non-immigrant visa archive. The images were typical visa application-type photographs with a universally uniform background, all gathered in a consistent manner.

The medium sized database consisted of a number of outdoor and video images from various sources. Figure 9 gives an indication of the images in the database. The top row contains images taken indoors and the bottom contains outdoor images taken on the same day. Notice the quality of the outdoor images. The face is consistently located in the same position in the frame and similar in orientation to the indoor images.

Figure 9: Indoor and outdoor images from the medium database[33]



The results of FRVT 2002

In order to interpret the results of FRVT2002 appropriately we should take note of the fact that FRVT2002 was a *closed-set evaluation*. For the identification task, the system received an image of an unknown person (assumed to be in the database). The system then compared the probe image to the database of known people. The identification performance of the top systems is indicated in Figure 10 below.

With the very good images (i.e., passport-type images) from the large database (37,437 images), the identification performance of the best system at rank one was 73% at a FMR of 1%—note that this performance is relative to database size and applies only to a database of exactly 37,437 images.

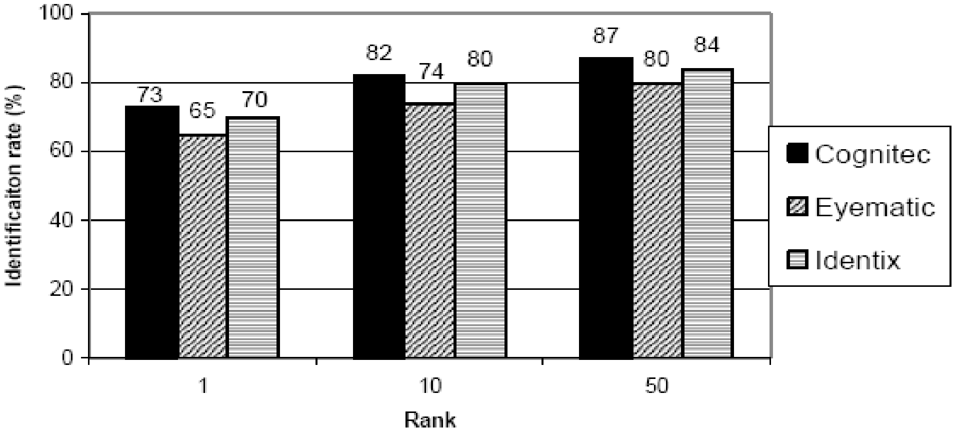
What are the factors that can detract from "ideal" performance? There might be many. The FRVT 2002 considered three:

- Indoor versus outdoor images
- The time delay between the acquisition of the gallery image and the probe image
- The size of the database.

The identification performance drops dramatically when outdoor images are used, in spite of the fact that they can be judged as relatively good (as can be seen in Figure 10). For the best systems, the recognition rate for faces captured *outdoors* (i.e., less than ideal circumstances) was only 50% at a FMR of 1%. Thus, as the evaluation report concluded, "face recognition from outdoor imagery remains a research challenge area."[34] The main reason for this problem is that the algorithm cannot distinguish between the changes in tone, at the pixel level, caused by a relatively dark shadow versus such a change caused by a facial landmark. The impact of shadows identification may be severe if it happens to be in certain key areas of the face.

As one would expect, the identification performance also decreases as time increases between the acquisition of the database image and the newly captured probe image presented to a system. FRVT 2002 found that for the top systems, recognition performance degraded at approximately 5% per year. It is not unusual for the security establishment to have a relatively old photograph of a suspect. Thus, a two-year-old photograph will take 10% off the identification performance. A study by NIST found that two sets of mugshots taken 18 months apart produced a recognition rate of only 57%, although this performance cannot be

Figure 10: Performance at rank 1, 10, and 50 for the three top performers in the evaluation with gallery of 37,437 individuals[35]



compared directly to FRVT 2002 because of the different database size.[36] Gross, et al. found an even more dramatic deterioration.[37] In their evaluation, the performance dropped by 20% in recognition rate for images just two weeks apart. Obviously, these evaluations are not directly comparable because of the "closed-set" methodology. Nevertheless, there is a clear indication that there may be a significant deterioration when there is a time gap between the database image and the probe image.

What about the size of, or number of subjects in, the database? For the best system, "the top-rank identification rate was 85% on a database of 800 people, 83% on a database of 1,600, and 73% on a database of 37,437. For every doubling of database size, performance decreases by two to three overall percentage points."[38] What would this mean for extremely large databases in a "closed-set test?" One might argue that from a practical point of view, it is immaterial because no real world applications are closed-set. Consequently, the government-funded facial recognition evaluation community has switched to "open-set" tests and to ROC/DET reporting metrics, for which approximate scaling equations are known.

To conclude this discussion, we can imagine a very plausible scenario where we have a large database, less than ideal images due to factors such as variable illumination, outdoor conditions, poor camera angle, etc., and relatively old gallery images. Under these conditions, performance would be very low, unless one were to set the FMR to a much higher level, which would increase the risk that a high number of individuals would be unnecessarily subjected to scrutiny. Obviously, we do not know how these factors would act together and they are not necessarily cumulative. Nevertheless, it seems reasonable to believe that there will be *some* interaction that might lead to a compound affect.

The FRVT 2002 report concluded that face recognition in uncontrolled environments still represents a major hurdle. The FRVT 2006 report, which was released in 2007, partially responds to this issue and shows that that there was a significant increase in the performance of the technology. This was partly due to the introduction of the Facial Recognition Grand Challenge.

5.1.2. The facial recognition grand challenge (FRGC)

The FRGC was designed to create a standardized research environment within which it would be possible for FRT to achieve an order of magnitude increase in performance over the best results in the FRVT 2002. The open-set performance selected as the reference point for the FRGC is a verification rate of 80% at a false accept rate of 0.1%. This is equal to the performance level of the top three FRVT 2002 participants. In this context, an order of magnitude increase in performance was therefore defined as a verification rate of 98% at the same fixed FAR/FMRof 0.1%. FRGC moved to open-set metrics, discarding as immaterial any closed-set results because actual implementations are always open-set.

The FRGC data set

The data for the FRGC experiments was collected at the University of Notre Dame in the 2002-2003 and 2003-2004 academic years. Students were asked to sit for a session in which a number of different images were collected. In total, a session consisted of four controlled still images (studio conditions with full illumination), two uncontrolled still images (in varying illumination conditions such as hallways and outdoors), and one three-dimensional image (under controlled illumination conditions). Each set of uncontrolled images contained two expressions, smiling and neutral. See
Figure 11 for a sample of the data collected in a typical session. The data collected in these sessions was divided into two sets or partitions, a *gallery set* and a *validation* or *evaluation set*. The data in the training set was collected in the 2002-2003 academic year. The gallery set was split into two gallery partitions. The first is the large still image gallery set, which consists of 12,776 images from 222 subjects, with 6,388 controlled still images and 6,388 uncontrolled still images. The second is a smaller set of 6,601 images which consists of 943 3D images, 3,773 controlled still images and 1,886 uncontrolled still images. Images for the validation set were collected during the 2003-2004 academic year.

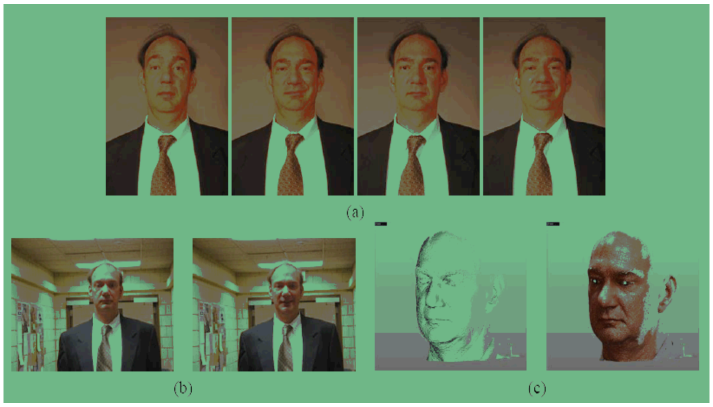


Figure 11: Images for the FRGC[39]

The validation set contains 32,056 images from 466 subjects collected over the course of 4,007 subject sessions and resulting in 4,007 3D images, 16,028 controlled still images, and 814 uncontrolled still images. The data set of FRGC is summarised in Table 1.

There are a number of things to note about the FRGC data set:

- The data set covers a relatively small set of subjects (i.e., less than 500). The data consists of very high quality images. For example, the ISO/IEC 19794-5 standard requires at least 90 pixels between the eyes, with 120 considered normative (already considered a very good quality image). Most of the images in the data set exceeded this requirement.
- The time delay between multiple images of one subject is relatively small, just one academic year. Aging was one of the problems identified by FRVT 2002.
- There seems to be many elements in the "uncontrolled" images that are in fact controlled. For example, the size and location of face in the image frame.

It should be clear from the above that it would not make sense to extrapolate too much from the FRVT 2002, the FRVT 2006, or the FRGC results. The results should rather be seen as a fixed baseline against which developers of the technology can measure themselves rather than as a basis for predicting how the technology might perform under operational conditions.

Experiments and results of FRGC

The FRGC designed a number of experiments to focus the development of FRT on a number of key areas: high resolution still 2D images, high resolution multiple still 2D images, and 3D still images. These experiments are summarised in Table 2 below.

Table 1: FRGC Data (approximately 50,000 images)

	Gallery set		Validation set (Probe)
Data collected	Students at University of Notre Dame 2002/2003 academic year		Students at University of Notre Dame 2003/2004 academic year
Subsets	Gallery set 1	Gallery set 2	
Type of images	Still images (2D)	3D images Controlled 2D Uncontrolled 2D	3D images Controlled 2D Uncontrolled 2D
Number of images Total	12,776 images	6,601 images	32,056 images
2D Controlled	6,388	3,772	16,028
2D Uncontrolled	6,388	1,886	8,014
3D	0	943	4,007
Number of subjects	222 subjects	222	466
Subject sessions	9-16 per subject (mode = 16)	943	4,007 1 – 22 per subject
Pixel size between eyes (average)	Controlled - 261 Uncontr. - 144 3D – 160	Controlled - 261 Uncontr.- 144 3D - 160	Controlled - 261 Uncontr. - 144 3D - 160

Table 2: FRGC experiments

Type of Exp	Gallery Image	Probe Image	Purpose	Number of results
Experiment 1	High resolution controlled still 2D image	High resolution controlled still 2D image	Standard facial recognition problem	17
Experiment 2	High resolution controlled multiple still 2D images	High resolution controlled multiple still 2D image	Evaluate the effect of multiple images	11
Experiment 3	3D facial images (both the shape and texture)	3D facial images (both the shape and texture)	Evaluate recognition with 3D images	10
Experiment 3s	3D facial images (shape only)	3D facial images (shape only)	Evaluate recognition with 3D images - shape	4
Experiment 3t	3D facial images (texture only)	3D facial images (texture only)	Evaluate recognition with 3D images - texture	5
Experiment 4	High resolution controlled multiple still 2D images	High resolution single *uncontrolled* still 2D image	Standard facial recognition problem - the difficult problem as identified by FRVT 2002	12
Experiment 5	3D facial images (both the shape and texture)	High resolution controlled still 2D image	Evaluate recognition with 3D and 2D images (standard problem)	1
Experiment 6	3D facial images (both the shape and texture)	High resolution single *uncontrolled* still 2D image	Evaluate recognition with 3D and 2D images (difficult problem)	

There were 19 organizations (technology developers) that took part in the FRGC. Not all participated in every experiment. In total, 63 experiments were conducted. This means that the participants completed an average of 3 to 4 experiments each although there was an uneven distribution with only one submission of results for experiments 5 and 6 (as seen in the last column of Table 2).

The interim results of the FRGC are shown in Figure 12. At first glance, these results suggest very significant improvements as compared to the performances in the FRVT 2002. They are certainly impressive. However, one needs to evaluate them relative to the data set upon which they are based. As indicated above the data set consisted of high quality images of a relatively small set of subjects.

The most significant conclusions one might draw from the interim results of the FRGC are:

- The performance of FRT seems to be steadily improving.
- Uncontrolled environments are still a significant problem. The mean performances were still lower than the top performer in the FRVT 2002.
- 3D recognition using both shape and texture do not necessarily provide better results than high quality 2D images.

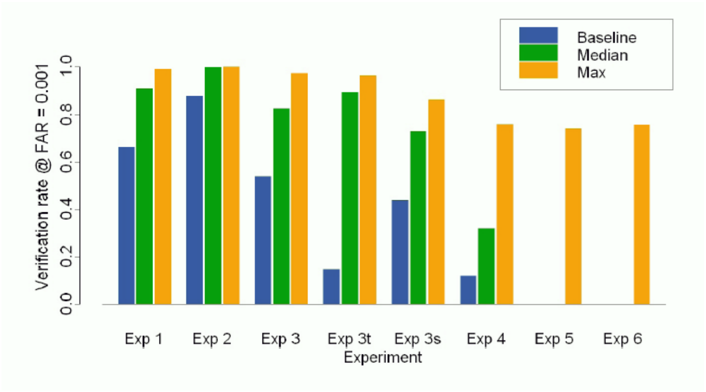


Figure 12: The interim results of the FRGC[40]

5.1.3. The Face Recognition Vendor Tests of 2006 (FRVT 2006)

The widely reported FRVT of 2002 was followed by the FRVT 2006 evaluation. As was the case for FRVT 2002, this evaluation was an independent assessment performed by NIST and sponsored by organizations such as the Department of Homeland Security, the Director of National Intelligence, the Federal Bureau of Investigation, the Technical Support Working Group, and the National Institute of Justice. Some of the key features of this evaluation were:

- The use of high resolution 2D still images
- The use of 3D images (both a shape and texture channel)
- The evaluation of algorithm performance as compared to human performance
- Simultaneous evaluation of iris recognition technology (which will not be discussed here).

The evaluation took place in 2006-2007 and the report was released in March 2007.[41]

The data sets of FRVT 2006

Three different data sets were used in FRVT 2006. The first was a multi-biometric data set consisting of very high-resolution still frontal facial images and 3D facial scans as well as iris images. The very high-resolution images were taken with a 6 megapixel Nikon D70 camera and the 3D images with a Minolta Vivid 900/910 sensor. The second data set is the high-resolution data set, which consisted of high-resolution frontal facial images taken under both controlled and uncontrolled illumination. The high-resolution images were taken with a 4 megapixel Canon PowerShot G2. The average face size for the controlled images was 350 pixels between the centers of the eyes and 110 pixels for the uncontrolled images. The data for the very high-resolution as well as the high-resolution data sets were collected during the fall 2004 and spring 2005 semesters at the University of Notre Dame. The subjects were invited to participate in acquisition sessions at roughly weekly intervals throughout the academic year. Two controlled still images, two uncontrolled still images, and one three-dimensional image were captured at each session. Figure 13 shows a set of images for one subject session. The controlled images were taken in a studio setting and are full frontal facial images taken with two facial expressions (neutral and smiling). The uncontrolled images were taken in varying illumination conditions (e.g., hallways, atria, or outdoors). Each set of uncontrolled images contains two expressions (neutral and smiling).

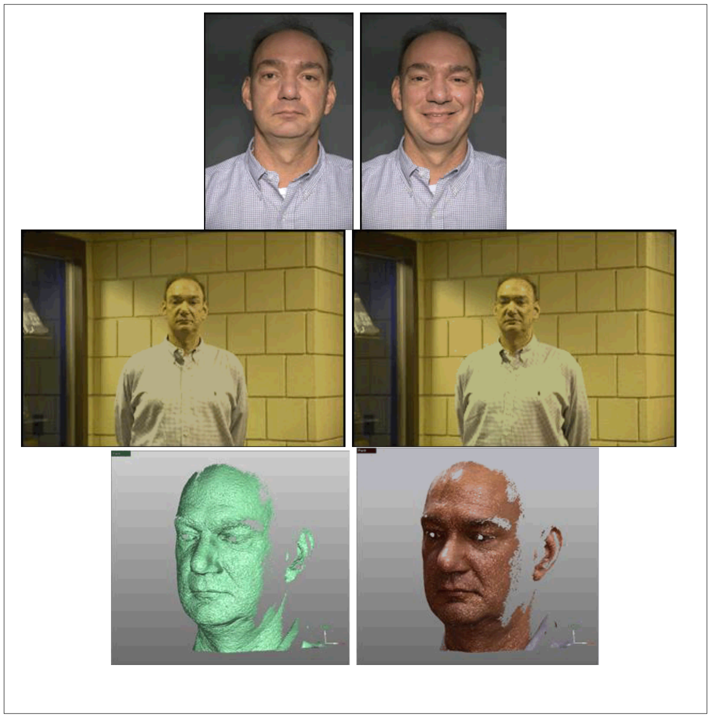


Figure 13: Examples of the facial images used in the FRVT 2006 evaluation[4]

The third data set was a low-resolution data set, consisting of low-resolution images taken under controlled illumination conditions. In fact, the low-resolution data set was the same data set used in the HCInt portion of the FRVT 2002 evaluation. The low-resolution images were JPEG compressed images with an average face size of 75 pixels between the centers of the eyes. The difference in image size between the FRVT 2002 and 2006 evaluation is quite significant, which raises some questions about the comparability of these evaluations (as represented in

Figure 14 below). It must be noted that none of the data sets were at the ISO/IEC 19794-5 required resolution. We will return to this issue. Another important aspect of the Notre Dame data set is the fact that it only included a small number of subjects (less than 350) and was not racially balanced.

The results of FRVT 2006

The FRVT 2006 evaluation reports an order-of-magnitude improvement in recognition performance over the FRVT 2002 results as indicated in Figure 14 below. This indicates that in FRVT 2002 the best algorithms were 80% accurate (at a false accept rate of 0.1%); in FRVT 2006, the best algorithms were 99% accurate (at a false accept rate of 0.1%). This indicates a massive improvement in the technology. The other important conclusion of the evaluation is that the best algorithms outperform humans in the identification task. For this experiment, 26 undergraduate students were shown 80 (40 male, 40 female) faces that were determined by the automated algorithms to be moderately difficult to identify. A face pair is moderately difficult if approximately only half of the algorithms performed correctly in matching a face to the right person. (This protocol in selecting the face pairs to present to the human examiners has been strongly criticized by some groups.) Both of these conclusions are significant and have very important policy implications. As such, they need to be submitted to further scrutiny.

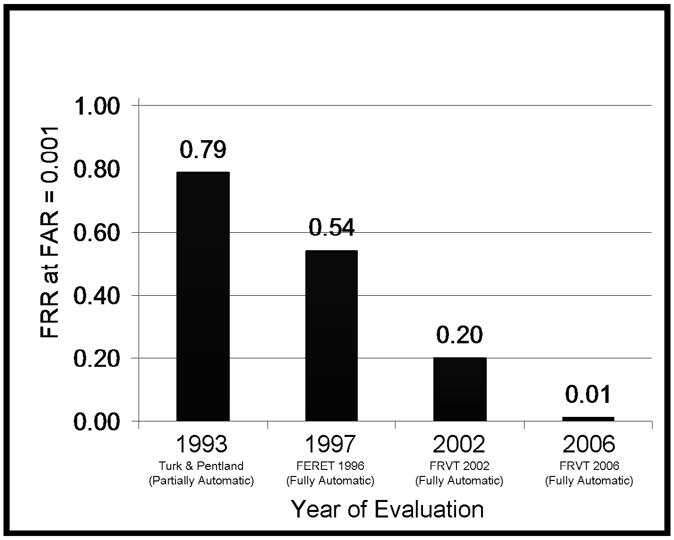


Figure 14: Comparative results of evaluations from 1993-2006[43]

We should not use Figure 14 and FRVT 2006 to conclude that the performance of the technology is being compared under comparable circumstances or similar conditions. Authors of the NIST study would suggest that this is not the case at all. We would all argue that it would be more accurate to suggest that these are the relative performances given the conditions of the evaluations at the time. Unfortunately, however, it is exactly this figure that is often used by the press and vendors to make inappropriate claims about FRT. Let us consider some of these evaluation conditions in more detail.

One of the significant differences between FRVT 2002 and 2006 is the high *quality of the images* used in FRVT 2006. As indicated above, the controlled still images had an average of 400 pixels between the centers of the eyes; the uncontrolled still images had an average of 190 pixels between the centers of the eyes. In contrast, the images in the large data set for FRVT 2002 have an average face size of only 75 pixels between the centers of the eyes. In other words, the information available (at the pixel level) to the algorithms in 2006 was potentially twenty-five times greater than that of 2002. NIST considered this increase in resolution owing to more advanced cameras to be part of the improvement in "technology." What might this mean in potential performance terms? According to the FRVT 2006 report, the results between the very high-resolution data set and the low-resolution data set of 2002 indicates a difference of 4% in recognition rates for the best performing algorithms. This is important if we take into account that the typical passport photo is 960 pixels high and 840 pixels wide (i.e., required to have at least 90 pixels between the centers of the eyes).

One can also question whether this is a realistic comparison of the two different evaluations (except as a relative measure or ranking of the algorithms against each other within a particular evaluation). In the report, it was suggested that "[s]ince performance was measured on the low-resolution data set in both the FRVT 2002 and the FRVT 2006, it is possible to estimate the improvement in performance due to algorithm design."[44] This seems to be a reasonable claim. However, we would suggest that it is not a fair comparison since the FRTV 2002 data set had been in the public domain between 2002 and the evaluation in 2006. It is well known that many developers of FRT used the FRVT 2002 data set as a developmental set to support their ongoing development efforts. It would have been a more appropriate comparative evaluation to re-run the 2002 algorithms against the 2006 data set. However, even this might have not have been an appropriate evaluation since the 2006 algorithms were also developed against the FRGC data set *which were collected at the same time and under the same conditions as the FRVT 2006 data set.* The FRGC and the FRVT 2006 data sets were both collected at the University of Notre Dame using the same equipment and set-up. This means that many of the potential factors that may influence recognition rates were kept constant or controlled for. Thus, one might have reasonable grounds to question the improvements in performance between the FRGC and the FRVT 2006 if the only difference was the difference in subjects. Our discussion below of scenario evaluations and operational evaluations will show that recognition performance is extremely sensitive to any variation from the developmental data set. A true evaluation of the technology will need to take into account such variations. *We would argue that it would be a more rigorous and fair evaluation if the evaluation data set was compiled using a variety of cameras and settings.*

Unfortunately, some of the major weaknesses identified in FRVT 2002 were not evaluated in FRVT 2006. The number of subjects in the very high- and high-resolution data set was not more than 350. The low-resolution data set included 37,437 individuals (however, as we indicated, this is not really a true evaluation as some of this data set was already in the public domain and most likely used in the developmental process). Our discussion of the BioFace II scenario evaluation below will indicate that the issue of "biometric doubles," in which the "nearest neighbor" becomes nearer than the expected within-class variation of a typical individual, is very significant in even moderately large data sets. Another significant issue identified in FRVT 2002 was the time delay between the capture of the gallery image and the probe image. However, we do not want to argue that there have not been significant improvements in the technology, when "technology" is taken to include both the algorithms and the imaging systems. We simply want to caution against taking the results of the evaluation out of context. Unfortunately, vendors of these technologies only report the headline results without providing the context within which these claims need to be evaluated. For policymakers, it is important to assess these results in context.

One of the novel aspects of FRVT 2006 was to compare algorithm performance against human identification performance. This is important as one can claim that even if algorithms are not perfect they ought to be considered as a viable option if they perform better than the alternative (i.e., human operators). The report shows that the best algorithms outperform humans (given the conditions of the experiment). Again we would caution against taking these claims out of context. A number of aspects of the experimental design are worth some further scrutiny, in particular the following:

- The use of undergraduate students as the human identifiers. Why use undergraduate students rather than security personnel that are experienced in the identification task? If one is going to compare the best algorithms against humans, one should use the most experienced humans (indeed, humans that would in the normal circumstance do the identification task). We are not suggesting that they would necessarily perform better. It just seems more appropriate to do the evaluation on this basis. Unfortunately, trained security personnel are not as readily available as college students for a number of reasons, including professional sensitivity and national security.
- Why are the algorithms used to identify face images that are moderately difficult to identify? Would it not be better to get humans to identify the images that are moderately difficult to identify? Or perhaps construct a data set that comprises an equal number of images that are "moderately difficult" as defined by the algorithms and the humans respectively?

In summary, it seems to us that one might have arrived at a different result if one set up the evaluation differently. Thus, it is important to evaluate the results of FRVT 2006 in the context of the conditions of the evaluation. Indeed, we would argue that it would be more appropriate to do a comparative evaluation of human and algorithm performance under realistic operational conditions if the result is to feed into policy debates and decisions, as we will discuss below.

Finally, it is also worth mentioning that technology evaluations are just one element of an overall evaluation. The really significant results, with regard to the feasibility of the technology, are the performance of these algorithms as part of specific scenarios in operational conditions. As the FRT expert Jim Wayman notes: "As with all of the FRVT reports, results need to be interpreted with caution as this is a "technology," not a "scenario" or "operational" evaluation […T]he test gives us *little predictive information* about the performance of current facial recognition algorithms in real-world immigration environments."[45] This will be discussed in the next section.

5.2. FRT scenario evaluations

Scenario evaluations are important as they represent the first steps out of the laboratory environment. These evaluate the overall capabilities of the *entire system for a specific application scenario*. This would include the image-capturing component (cameras, video, etc.), the facial recognition algorithms, and the application within which they will be embedded. In this section, we will report on two such evaluations.

5.2.1. BioFace II scenario evaluations

The BioFace II evaluation was conducted in 2003 and followed the BioFace I project.[46] BioFace I consisted of the creation of an image database that would form the baseline for subsequent BioFace evaluations. The BioFace evaluations are joint projects of the Federal Office for Information Security (FOIS) in Bonn, Germany, and the Federal Office of Criminal Investigation (BKA) in Wiesbaden, Germany, with additional assistance provided by the Fraunhofer Institute for Computer Graphics Research (IGD). Vendors were invited to submit their algorithms and systems for evaluation. The following vendors submitted their technology for evaluation

- ZN Vision Technologies
- Controlware GmbH
- Cognitec Systems GmbH (one of the top performers in the FRVT 2002, 2006)
- Astro Datensysteme AG

The evaluation of the technology in BioFace II was based on a very large set of 50,000 images which were assembled in the BioFace I project. The evaluation consisted of two phases. Phase 1 evaluated the algorithms in both the verification task and the identification task. Phase 2 evaluated whole vendor systems in an identification scenario. Let us consider the results of the evaluation.

<u>Phase 1 evaluation (facial recognition algorithms)</u>

Facial recognition algorithms and the verification task

The evaluation of the verification scenario was conducted by attempting to match an image of a person (identified with a unique identification number) from a probe image with at least one image of that same person (identified with the same unique identification number) in the gallery. The gallery contained an image for every person in the probe database as identified by the unique identification number. In every verification attempt, the two images were compared with each other and the degree of agreement between the two facial images (the matching score) was recorded. Databases containing images of 5,000, 10,000, 20,000, and 50,000 persons were investigated in order to establish the impact of the size of the database on the verification performance.

As this is a verification evaluation, the size of the database against which the verification task was performed did not have an effect on the matching scores produced. Furthermore, it seems that age differences between images of the same person did not pose any serious problems to the algorithms. Tests were conducted using images of the same person captured up to ten years apart. Although the matching scores declined as the images were further apart, the difference was not so significant as to bring into question the overall verification results. However, this is not the full picture. The true discriminating capability of the algorithm is to generate a sufficiently distinct biometric template so that the image of a person cannot only be verified against images of that person (as determined by the unique identification number) but also fail to verify against images (or biometric templates) of all other persons in the gallery database. When compared to the entire gallery, the probe image would produce a high level of false accepts. In other words, many templates would overlap. Such overlapping obviously increased as the size of the database increased. This suggested that although the images were distinct, some the biometric templates were almost identical. *Thus, as the database increases in size in an identification application, the probability of the occurrence of biometric doubles will also increase.* In practice, this will mean that the matching score threshold needs to be set at a relatively high level to prevent the false acceptance of a biometric double, but only in identification applications in which the entire database is searched. In verification applications, size of the database is immaterial. Such high thresholds will then also lead to a relatively high level of false rejects and thus significantly bring down the overall identification rate. This could cause problems in unmanned identification, but not verification scenarios with large databases.

It should also be noted that the relative success of the algorithms in the verification task was also due to the high quality of images in the probe and gallery databases. The report concluded that: "*the suitability of facial recognition systems as (supporting) verification systems is neither proved nor disproved by BioFace II. The stability of the scoring of matches proves that the systems possess the reliability that is necessary. However, the systems do not provide the reliable differentiation between "biometric twins" that is necessary for their use in practice.*"[47]

Facial recognition algorithms and the identification task (closed-set)

In the evaluation of the identification scenario, the images of 116 persons were presented to the algorithms. The gallery against which these were compared contained 305 images—and therefore often more than one image, in different sizes—of the 116 persons. But as each of the 116 probe images had at least one mate in the database, this was a closed-set evaluation. In addition to these images, the overall gallery contained 1,000, 5,000, or 50,000 filler images of persons other than these 166 persons. The closed-set identification outcome was a rank 10 list of the best matches per identification run against the total database. A match was considered to have been made if the person appeared in the list of ten best matches. In the identification scenario, the size of the database turned out, as expected, to have a significant influence on the recognition performance of the systems. From the tests, it emerged that as the gallery increased, more non-matches (or false matches) displaced matches (or true matches) in the rank 10 list.[48] In other words, the systems made more and more mistakes.

The report concluded that "the suitability of facial recognition systems as (supporting) identification systems is neither proved nor disproved by BioFace II. However, in the identification scenario there is less room for compensating for the weaknesses of the systems as regards separating matches from non-matches than in the verification scenario, so that in this case further improvements to the algorithms are imperative before the systems are suitable for use."[49] In other words, the report suggests that some significant development would be necessary before FRT could be considered suitable for identification purposes over large databases—this general finding holding for real-world open-set applications as well. We are a bit perplexed as to why closed-set results were reported at all, given that all real applications are indeed open-set. The recent results of the FRVT 2006 might suggest such improvements have taken place. However, this will need to be proven in a scenario and operational evaluations before it can really inform policy debates.

Phase 2 evaluation (FRS)

In phase 2 of the system test, an actual implementation of an FRS in an operational situation was evaluated. The FRS to be evaluated was integrated into the normal access control process for employees. Normally, the employees, upon arrival, would pass through several turnstiles and then enter themselves into a time registration system that records their time of entry (this data was also used as a comparative basis in the evaluation).

During the evaluation, the route from the central turnstile to the time registration terminal was also monitored by the facial recognition systems (as indicated in Figure 15). Twenty employees volunteered to take part in the

Figure 15: The entrance where the system evaluation was conducted[50]

evaluation. As the volunteers passed through the area, they consciously looked at the cameras. The system being evaluated was expected to identify the person in a database which also contained 500 images of other persons in addition to one image of each volunteer. To be a reasonable operational test, we must assume that all persons are law abiding and that no unauthorized person would ever attempt to use this system. A person was deemed to have been successfully identified when that person's image was included in the rank 5 list of best matches.[51] Since impostors are assumed not to exist, no attempt was made to determine if someone not in the database would rank among the top 5.

During the evaluation, the volunteers were enrolled into the database using two different approaches. In the first test, the volunteers were enrolled by standing in front of the cameras that were used in the actual implementation of the recognition system. In the second test, the volunteers were enrolled by being photographed using a digital camera. All the systems performed better using the images captured in the first test. From this we might conclude the following:

- The more similar the environment of the images to be compared (background, lighting conditions, camera distance, and thus the size of the head), the better the facial recognition performance.
- The greater the difference in the optical characteristics of the camera used for the enrollment process and for photographing the probe image (light intensity, focal length, colour balance, etc.), the worse the facial recognition performance.

All in all, two out of the four systems tested had a false non-match rate[52] of 64% and 68% respectively in the first test and 75% and 73% in the second test. This means that the best system in the best circumstances (test 1) correctly identified the volunteers only 36% of the time and in the worst case (test 2) only 27% of the time. The two other systems had false non-match rates of 90% and 98% in test 1 and 99% and 99.7% in test 2. This means that they were in fact not able to recognise any of the subjects. The weaker of these two systems was so unreliable that it was only available for use for the first few days of the evaluation. In each case, the recognition performance was not nearly as good as claimed in the promotional material of the vendors. Since impostors were not tested, we can also conclude that no unauthorized person was permitted access.

The evaluation also questioned the quality of the support provided by the vendor or distributor of the system. Some of the vendors and distributors were unable or unwilling to provide the necessary support. Often they were not able to answer technical questions themselves. They were also often dependent on experts who had to be flown in to get the systems up and running and carry out any necessary troubleshooting on-site.

It is difficult to generalize from such a small-scale experiment in an impostor-free environment. Nevertheless, it indicates that in actual operational evaluations the performance of the systems were in fact significantly lower than in the technology evaluations and significantly lower than claimed by the vendors. It also indicated the importance of doing full system evaluations in realistic operational conditions which would include the possibility of an access attempt by someone not authorized. Finally, the difference between the first and second tests indicates the sensitivity of the systems to environmental factors. Some of the difference could also obviously be due to the difference in the metric used – i.e. the change from rank 10 to rank 5 as the definition of 'recognition'.

5.2.2. Chokepoint scenario evaluation using FaceIT (Identix)

Another of the few scenario evaluations publicly available was performed by the US Naval Sea Systems Command (NAVSEA), sponsored by the Department of Defense Counterdrug Technology Development Program Office in 2002.[53]

The purpose of this evaluation was to assess the overall capabilities of entire systems for two chokepoint scenarios: verification and watch list. A chokepoint is a supervised controlled entry point to a secure area. In this evaluation, individuals walking through the chokepoint look toward an overt FRS operating in either verification or watch list mode. In verification mode, an individual approaches the chokepoint and makes their identity known using a unique identifier such as a smart card, proximity card, magnetic stripe card, or PIN. The FRS compares probe images of the individual's face with face images stored in the database for that identity. If the probe image and gallery image do not match within certain threshold criteria, an operator is notified (i.e. the person is denied access and needs to be investigated). In watch list mode, probe images of the individual's face are compared with a watch list of face images in the database.

If a match has a score greater than a certain threshold, an operator is notified.

For the verification scenario, a custom system manufactured by Identix was tested. For the watch list scenario, two off-the-shelf systems from Identix were tested: the FaceIt Surveillance system and the Argus system, respectively. FaceIt Surveillance has been on the market for a number of years while Argus was first introduced in late 2001.

In order to do a variety of experiments on verification and watch list tasks the data for the experiments was collected in the form of video footage that was played back (for every experiment) as input (probe images) to the FRS while varying threshold parameters and database sizes. A variety of experiments were performed to also study the impact of eyeglasses and enrollment image quality.

<u>Data set for scenario evaluation</u>

Data set for verification

The enrollment of images into the gallery database was performed according to vendor instructions. To be enrolled into the gallery database, the volunteers stood in front of a camera connected directly to the system with uniform illumination and a clean white background. This was performed only once for each individual. Sample enrollment images are shown in Figure 16. The important thing to notice about these images is the even illumination provided by the controlled lighting. The probe images were obtained by recording video from the camera(s) attached to the system as volunteers stood in specific locations. Operators instructed the users to stand at a marked location 4 feet 2 inches in front of the camera, remove hats and tinted glasses, and slowly tilt their heads slightly forward and backward about one to two inches while looking at the camera with a neutral expression. The camera tilt was adjusted for the height of the user, as recommended by the vendor. Once the camera adjustments were made, the video recorder was placed in record mode for ten seconds then stopped. Once users were enrolled and had posed for recorded video segments, the rest of the evaluation for this scenario was performed without user participation.

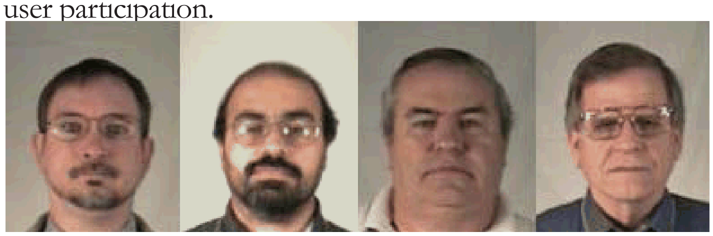


Figure 16: Enrollment images[54]

The advantage of using a video recording is that the evaluation can be rerun for a variety of different conditions as well as for future evaluation of the systems. There was a time difference of 0-38 days between enrollment image collection and the recording of the video.

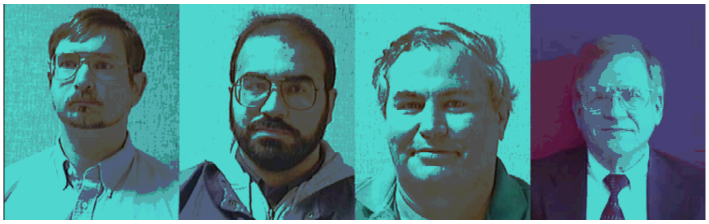
Data set for watch list

For the watch list scenario, the enrollment database was created using existing security badge images. A request was made to obtain the badge images of all company employees and contractors for use in this evaluation. Security personnel agreed to this request and provided a database containing 14,612 images. During the course of the data collection effort, the images of volunteers were identified and additional images were selected at random to create gallery databases of 100, 400, and 1,575 images to be used for the data analysis. An example of the badge images used is shown in Figure 17 below. Although the images do not always have uniform illumination, they mostly have a consistent frontal orientation with a clean and consistent background. There was a time difference of 505-1,580 days between the capture of the images in the gallery database and when the probe images were collected using the video recordings.

<u>Results from the scenario evaluation</u>

Evaluation results of the verification scenario

In this scenario, users passing through the chokepoint stand in front of the system camera, present their assigned identity, and wait for a matching decision based on a one-to-one comparison. If the system returns a matching score that meets the threshold criteria, the user is accepted by the system and allowed access. Otherwise the user is rejected by the system and denied access. During the verification imposters would try and gain access by using the identification number assigned to another user.



The results of the verification evaluation, as indicated in Figure 18, shows that FRT can be used successfully for *verification* if the gallery image (enrollment image) is of high quality and the probe image is of high quality (i.e., both are captured in a controlled environment). Because this was a verification application with users presenting an identity, database size did not impact the experiment. It also shows that there is a clear trade-off between the valid users rejected and impostors accepted (for the various threshold rates). It is conceivable that an error rate where 2.2% of valid users are rejected and 2.4% of impostors are accepted is manageable in a relatively small application.

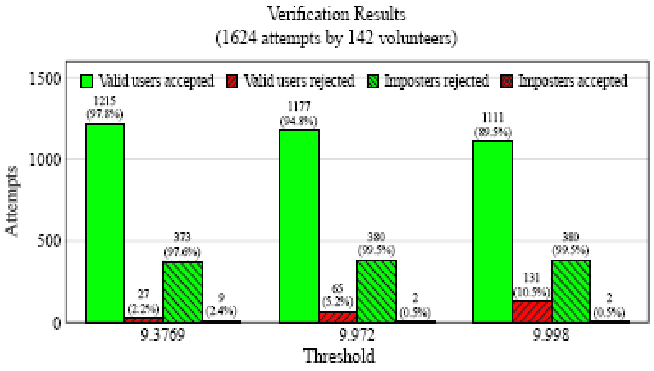


Figure 18: Verification results[56]

The closed-set identification performance against watch list size is summarized in Figure 19 below. This indicates that the best performance was achieved by the Identix Argus system with a recognition rate of 8% (with a watch list of 100), dropping down to 3% (for a watch list of 1,575). If the badge images in the watch list are replaced by images captured in the verification evaluation—i.e., in a controlled environment with the same camera equipment as used for the evaluation—then the identification performance increases to 37%.

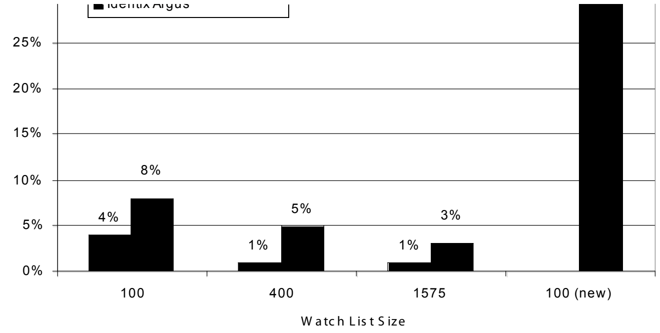


Figure 19: Watch list results[57]

This indicates the impact that good quality images (both gallery and probe) can have on system performance. However, in reality it is more likely that the images in the gallery of suspects being sought would be of a lower quality in uncontrolled environments.

Figure 20 shows the open-set ROC curve for three different sizes of the watch list database. It also clearly shows that the size of the watch list can have a relatively significant impact on the identification rate, as is well-known in the literature and for which an adequate predictive models exist—i.e. it is possible to estimate the impact of database size on recognition performance for open-set systems.

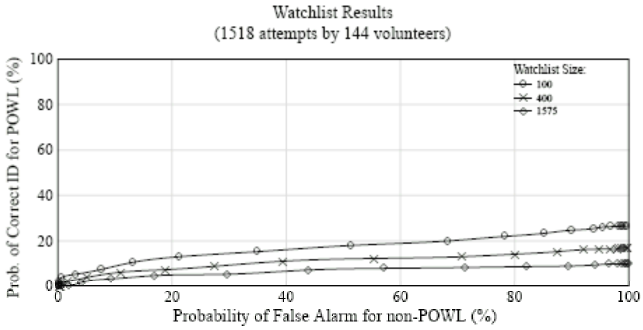


Figure 20: ROC curve for watch list results[58]

We have now looked at some scenario evaluations. Unfortunately, there are not many of these available for public consideration and scrutiny. This is significant in itself. Nevertheless, these evaluations do offer a consistent message. These evaluations suggest that FRT is somewhat proven for the verification task (under certain conditions) but performs very badly in the identification and watch list tasks, whether closed-set or open-set (as compared to the lab conditions). It might not be more informative if FRT were always evaluated in full operational conditions, as a "ground truth" is harder to assess and the factors that control errors cannot be evaluated. This is what we consider in the next section. We must also add that these scenario evaluations are now dated and that there has been a significant improvement in performances in technology evaluations (as seen in FRVT 2006). However, it is still an open question as to how these improvements will translate into operational improvements.

5.3. FRT operational evaluations

Operational evaluations are mostly performed by governments in response to specific operational requirements. As such, the results of such evaluations are mostly not publicly available. Since they are specific to the operational conditions of a particular application, these evaluation results may also not be generalizable. Nevertheless, such evaluations should be able to provide a more realistic sense of the actual capabilities of the technology. But because these are operational systems, with vulnerabilities potentially exploitable by those seeking to defeat the system, it is therefore not surprising that they are not made public by system owners. Unfortunately, most of the operational results that are publicly available are not the outcome of a systematic and controlled evaluation (as was discussed above). Nevertheless these results suggest performance well below the performances achieved in technology evaluations (at the time).

For example, the American Civil Liberties Union (ACLU) obtained data about the use of FRT by the Tampa Police Department as well as the Palm Beach International Airport.[59] In the Tampa case, the system was abandoned because of the large number of false positive alarms it generated. As far as could be ascertained, it did not make a single positive identification of anybody on the watch list. In the Palm Beach Airport case, the system achieved a mere 47% correct identifications of a group of 15 volunteers using a database of 250 images.[60] In Newham, UK, the police admitted that the FaceIt system had, in its two years of operation, not made a single positive identification, in spite of working with a relatively small database. One could argue that there might not have been the potential for a match to be made as none of the individual in the database actually appeared on the street. Nevertheless, the system could not "spot" a *Guardian* journalist, placed in the database, who intentionally presented himself in the two zones covered by the system.[61] These non-scientific, anecdotal cases indicate the complexity of real world application of the technology. As suggested, it may not be appropriate to generalize from these experiences (especially given the fact that they are now relatively dated). Nevertheless, they do raise questions that FRT providers need to answer if they want policymakers to become more confident about the capabilities of FRT in operational conditions. Most importantly, they indicate the importance of making implementation and procurement decisions on the operational *in situ* evaluation of the technology.

Beyond these more anecdotal case studies reported in the media, we do have access to at least two studies that were done in a relatively systematic manner and are publicly available. The first example is the SmartGate system where FRT has been deployed for the verification task. The second is an operational evaluation by the German Federal Criminal Police Office (BKA) of FRT in the identification task.

5.3.1. Australian SmartGate FRS for the verification task

SmartGate is an automated border processing system that allows a self-processing passport check normally performed by a Customs Officer. The system makes use of the Cognitec FRT to perform the face-to-passport check. Cognitec was one of the top performers in the FRVT 2002 and 2006.

There have been three versions of SmartGate deployed operationally since 2002. This paper will discuss the current version only. The current version of SmartGate is deployed at Brisbane, Cairns and Melbourne airports and may be used by arriving adult Australian and New Zealand citizens *carrying e-passports*. Australia Customs anticipates roll-out to 5 more airports in 2009. When performing the verification task, the system compares the face with the image on the passport, which requires no specialized enrollment beyond the passport issuance process.

SmartGate is currently a two-step process. Volunteering passengers first approach a kiosk and open and insert their e-passports to be read. The facial image on the passport chip is transferred into the system. The passenger must answer at the kiosk several 'health and character' questions, such as "Have you been to a country with pandemic yellow-fever within the last 5 days?" If the passenger is determined to be eligible to use SmartGate (over 18 years old, NZ or AU citizen, and on the expected arrival manifest submitted to Customs by the airlines), and the passport has been successfully read and all questions answered, a paper ticket is issued which the passenger takes to the SmartGate exit where the facial recognition occurs. If any issues develop at this point, such as a failure to read the passport, the passenger is referred to the standard immigration queue. Kiosks and exists from the Melbourne airport SmartGate implementation are shown in Figure 21.

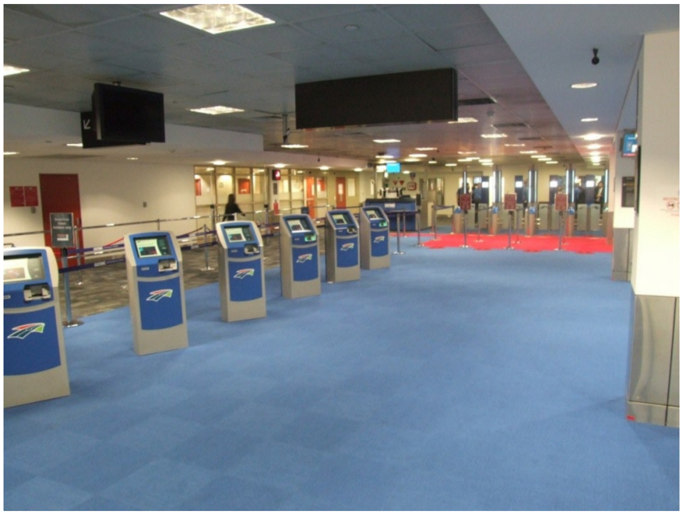


Figure 21: SmartGate implementation at Melbourne airport

At the exit, passengers insert the ticket obtained from the kiosk and look into a tower of 3 cameras, each at a different height. The cameras collect multiple images per second until a well-centered image is matched to the image retrieved from the e-passport. The system does not compare the exit or passport photos to any database. If no match can be made at the exit, the passenger is referred to the head of the standard primary queue, giving SmartGate volunteers who fail the face recognition activity priority processing at this point.

It was reported that by October, 2008[62] close to 100,000 transactions had been made with the current version of SmartGate (which has been in operation since August 2007). The FRR was reported as less than 9%. Several explanations were given as to why the FRR was this high:

a) Passengers did not understand that they needed to look directly into the tower of cameras, focusing their attention instead on the ticket insertion process, which was down and to the right.
b) Some passport images were not ICAO compliant (ISO/IEC 19794-5).
c) The ICAO standard failed to consider some sources of problems (such as photo printing on matte paper or use of gauze lens filters).

Australia Customs suggests that the first issue can be addressed through user habituation, better signage, and in-flight videos. Regarding the second issue, Australia Customs claims to be working with the passport issuance agency, the Department of Foreign Affairs and Trade (DFAT), to tighten the inspection process for self-submitted photos in the passport process. To address the third issue, the international standards organization responsible for ISO/IEC 19794-5 is reviewing the guidelines for acceptable passport photos as given in the standard. No attempt was made to quantify the various sources of the problems and it was stated that many FRR cases were caused by multiple apparent problems.

Australian Customs has released no figures on the FAR, only stating that several thousand attempts using Customs officers as impostors were made at both Melbourne and Brisbane airports and that the measured FAR was considerably below the design specification of 1%. A number of issues might nevertheless be highlighted:

- *Biometric Doubles*: During the roll-out of the original SmartGate system in 2002, two journalists with similar facial features, who had previously fooled facial recognition systems in other contexts, swapped passports during a press briefing and fooled the system. The impostor trials reported by Customs Australia involved only random exchanges of passports among customs officers. These tests do not give information on how well-organized groups could target the system by using persons previously established as "biometric doubles" on similar systems.
- *Aging*: By ICAO mandate, passports have a maximum lifespan of 10 years. How will recognition performance be impacted as the early e-passports reach 10 years of age? FRVT 2002 indicated that FRT is very sensitive to the aging effect.
- *Security*: How secure are e-passports? Can they be hacked? Is it possible to clone e-passports? There is some evidence that this can be done.[63] The security of biometric data is a major issue, not just for FRT, but for all biometric systems, especially in a context where it is generally assumed that biometrics cannot be falsified.
- *Function creep*: There has already been unconfirmed suggestions, reported in the press, that SmartGate might be used for the watch list task. We were not able to get confirmation if this claim is true or not. However, it seems clear that there is a general issue with regard to the way biometric data collected in one context (with the informed consent of the user) may serve purposes in another context that

a user has not necessarily consented to.

The SmartGate application of FRT indicates that the technology may now have matured to the level where it might be appropriate for the verification task in very specific controlled situations. However, the real challenge is in the identification task.

5.3.2. German Federal Criminal Police Office (BKA) evaluation of FRT in the identification task

Between October 2006 and January 2007, the German Federal Criminal Police Office (BKA) evaluated three FRS for purposes of identification at the rail terminal in the city of Mainz.[64] Two hundred commuters volunteered to be the "suspects" to be identified. The main aim was to identify suspects as they went through a chokepoint (in this case the escalators and the stairs as indicated in Figure 22). Four different scenarios were investigated:

- Recognition achievements on the escalator with daylight
- Recognition achievements on the escalator at night
- Recognition achievements on the stairs with daylight
- Recognition achievements on the stairs at night

An interesting aspect of this test is that the surveillance cameras were placed at the height of the faces being observed, not at the elevated, dysfunctional angle common for CCTV cameras.

On a daily basis, an average of 22,673 persons passed through the chokepoints. The false match rate was set at 1%, which would mean an average of 23 incorrect identifications per day that would need to be further investigated. Lighting was the most significant factor. In the daylight, recognition rates of 60% were achieved. However, at night time (when the area was lit by artificial light), the recognition rates dropped to as low as 10-20%, depending on the system being evaluated. The impact of participant movement on the stairs and escalator on the recognition rates was less than expected. On average, the recognition rates on the stairs (where the persons moved more rapidly) were 5-15% lower than on the escalators where persons would tend to move slower and more consistently. The evaluation also found that the technical setup of the system, in particular the camera technology being used, had a significant impact on the recognition rates.

The report concluded that indoor areas with constant lighting conditions could lend themselves to FRS with reasonable recognition rates but that variation in lighting conditions (darkness, back light, direct sun exposure, etc.) leads to significant deterioration in recognition rates. Moreover, since high-quality frontal images are required (for both the gallery image and the probe image), some cooperation of the subject would be necessary.

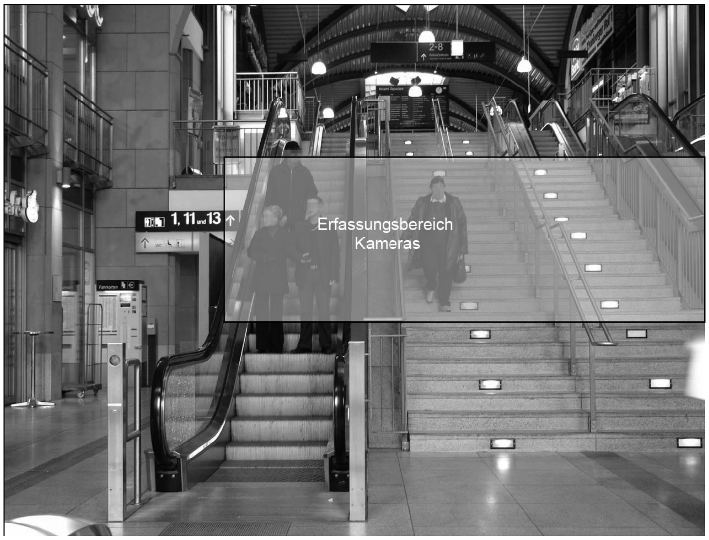


Figure 22: Escalators and stairs act as chokepoints. Box indicated area covered by FRS[65]

The report also emphasized that false alarms will require additional resources for follow-up and that further investigation would only be possible if the identified subject remained in the area long enough to be apprehended. Overall, the report concludes that FRT is *not yet suitable as a system for general surveillance in order to identify suspects on a watch list*. The German Federal Data Protection Commissioner Peter Schaar also expressed concern with the use of an immature technology. He suggested that it was especially problematic with regard to false positives "which, in the event of a genuine hunt, render innocent people suspects for a time, create a need for justification on their part and make further checks by the authorities unavoidable."[66]

The operational evaluation of FRT for identification purposes indicates that there are still some significant problems to be solved before the identification of the "face in the crowd" scenario, often seen as the ultimate aim of FRT, becomes a reality.

5.4. Some conclusions and recommendations on FRT and FRS evaluations

A number of general conclusions can be drawn from these evaluations that are relevant for the way others might interpret evaluations and how future evaluations might be conducted.

1. There is a general lack of publicly available scenario and operational evaluations of FRT. This means that policymakers and users often need to depend on technology evaluations such as the FRVT (which cannot be extrapolated to operational implementations) and the information provided by vendors (which are obviously not independent and are always the results of very small tests). *Recommendations*: Publicly funded scenario and operational evaluations are needed to support policy makers in making decisions about the appropriate use of FRT.
2. Vendors of FRT often use results from the technology evaluations (FRVT, FRGC, etc.) to make claims about their products more generally without providing the context of such evaluations. This leads to misleading conclusions about the efficacy of the technology. The evaluations above indicated that there is a significant deterioration in performance as one moves from the technology evaluations to operational conditions. *Recommendation*: Policy makers need to be informed of the context in which these results are being referenced. Hopefully this report will help to prevent the inappropriate use of evaluation data by vendors and the media.
3. Most of the evaluations available tend not to focus on some of the key problems that FRT ultimately will need to deal with such as, (1) large populations (the biometric double problem), (2) a significant age difference between gallery and probe image (the time delay or freshness/staleness problem) and (3) relatively uncontrolled environments (illumination, rotation, and background). *Recommendation:* It will be important for the development of FRT that technology evaluations incorporate more of these factors into the evaluation data set. The design of the evaluation image set is fundamental to understanding the results achieved in the evaluation.
4. There seems to be no publically available evaluation of falsification strategies. If the public is to trust the technology they need to be assured that it is secure and reasonably trust worthy. *Recommendation:* Publicly funded, systematic evaluations of falsification strategies, such as for example using known biometric doubles to gain access or to generate a false positive or false negative, are needed.
5. The current evaluation typology (technology, scenario and operational) does not necessarily include the evaluation of financial aspects as well as the evaluation of the ethical and political dimensions. *Recommendation:* It is suggested that more contextual and life cycle evaluations might be needed which might include financial evaluation as well as an ethical and political evaluation (to be discussed below)
6. It seems clear that no single biometric will be able to do all the work (especially with regard to identification), as such multi-biometric systems will probably be the future route of development. *Recommendation:* Evaluations should increasingly focus on multi-biometric systems as is happening in the NIST MBGC.

Taken together, the evaluations discussed above suggest that FRT has been proven effective for the verification task with relatively small populations in controlled environments. In the next section, the conditions that may limit the efficacy of FRS in operational conditions will be considered.

6. Conditions affecting the efficacy of FRS in operation ("what makes it not work?")

Given the discussion of the technical operation of FRT above, as well as the consideration of the various evaluations of the technology, it would be appropriate now to highlight the conditions that may limit the efficacy of an FRS in operational conditions (*in situ*). This is particularly important for decision makers and operational managers as it is often difficult to understand the technical jargon used by developers and vendors of FRT and what the results of the evaluations might mean in practice. What follows is not an exhaustive list but it will cover what we believe to be the most important elements given the current state of the technology.

6.1. Systems not just technologies

FRS are very sensitive to small variations in operational conditions. The scenario evaluations (BioFace and the chokepoint study) as well as the operational evaluations (SmartGate and the BKA study) reported above clearly suggest that the performance of FRT needs to be evaluated as *whole operational systems within operational* *conditions—i.e., in situ*. There are significant differences in performance when the technology is moved from the laboratory to the operational setting. Indeed, the research suggests that the technology is very sensitive to small variations in operational conditions.[67] This clearly also has important implications for the ongoing maintenance of these systems once implemented. It will be necessary to make sure that the implementation is sufficiently robust and sustainable in ongoing operational conditions. The operational conditions need to be carefully managed once implementation is complete. FRT is not "plug and play" technology. FRS need sustained and ongoing care if they are to perform at the levels that might make them feasible in the first place. This obviously raises questions regarding the cost to maintain the integrity of the system over the long term. What sort of infrastructure, practices, and staff need to be put in place to ensure this?

6.2. The gallery or reference database

The successful operation of a FRS in the identification mode is critically dependent on the key characteristics of the gallery database: image quality, size, and age. Image quality is one of the most important variables in the success of FRS. The performance of the recognition algorithms in locating the face and extracting features can only be as good as the images it is given. To be included in the database, the images need to be *enrolled*. This means the images need to go through a translation process (steps 1-3 of the recognition process as indicated in Figure 7 above) in order to create the biometric template. As the size of the identification database increases, the probability that two distinct images will "translate" into a very similar biometric template increases. This is referred to as the *biometric double* or *twin*. Obviously, biometric doubles lead to a deterioration of the identification system performance as they could result in false positives or false negatives. Thus, the decision whether to include an image in the gallery is a very important one. It might be better to exclude low quality images (even in important cases) rather than adding them "just in case." Restricting the database size in order to maintain the integrity of the system is an important priority. The temptation to increase the gallery will lead to a deterioration of the system performance, eventually at the cost of identifying those important high-risk cases in need of identification and apprehension. Very clear policies of prioritization are necessary.

Of course, in a verification scenario, the problems are different, but related. Enrollment image quality is still a major issue, but verification systems do not suffer from increasing false positives as the number of enrolled individuals increases. A major issue impacting verification systems, however, is to maintain image quality at the point of verification, including directions to the data subjects to maintain the proper pose angle with respect to the camera and to emulate the facial expression on the enrollment image (which might have long since been forgotten).

Another important consideration is the age of the image. The FRVT 2002 and BioFace evaluations have shown that the recognition performance deteriorates rapidly as the age difference between the gallery and the probe image increases. This is especially true for younger and older individuals. It is not clear that an image older than five years will achieve a good result. FRVT 2002 found that for the top systems, performance degraded at approximately 5% points per year in a closed-set test. Other studies have found significantly higher levels of deterioration.[68] Because we cannot freely translate between closed-set results and the real-world of open-set applications, we cannot make any quantitative predictions as to the performance degradation expected in practice. What is clear, however, is that use of old images (as much as 10 years old in the passport case) will cause problems for FRS.

The problem of biometric doubles can to some degree be managed by including multiple images, especially high quality images, of a person in the gallery.[69] Research has also indicated that the combination of 2D and 3D images can improve the performance of the system.[70] It has also shown that 3D images are susceptible to many of the problems of 2D images, especially the problem of illumination.[71] Ideally, the face images in the gallery should conform to the ANSI and the ISO/IEC good practice guidance and standard for face biometric images mentioned above.

6.3. Probe image and capture

The BioFace evaluation has shown that a FRS performs at its best if the conditions under which the probe images are captured most closely resembles that of the gallery image. In a verification scenario, this can be to some extent controlled since the active participation of the subject is guaranteed (for example, in the case of driver's license or passport photographs). In the identification task, one might not have the active cooperation of the individuals or might not be able to replicate the conditions of the gallery image. In this scenario, the difference (or likely difference) between the gallery image and the probe image is a very important consideration. For example, in the BioFace evaluation, the individuals were enrolled based on their badge images, which were very clear frontal images that covered the majority of the frame but which did not have uniform illumination. In the evaluation, the participants were asked to look directly at the camera as they approached the chokepoint. To the ordinary human eye, the gallery and probe images looked very similar. Yet, in spite of this level of control, the performance of the system was still very low. This underscores the complexity of the identification task. The task is obviously further complicated if the identification gallery (or watch list) database is itself large. It seems clear that in the scenario where one has an uncontrolled environment, and a probe is compared to a poor quality image in the gallery, performance is going to be poor. This suggests that the "face in the crowd" scenario, where a face is picked out from a crowd and matched with a face in the gallery, is still a long way off. Some researchers in the field suggested, in conversations with the authors, that it might take another decade to get there—if at all.

A number of other factors can confuse the recognition systems. Interestingly, however, the impact of facial hair and clear eyeglasses on the performance of the system is strongly debated.[72]

6.4. Recognition algorithms

Not all algorithms are equally good at all tasks. Algorithms differ in the way they define "facial features" and whether or not those features are located with respect to facial "landmarks," such as eyes, mouth, nose, etc. All algorithms need good quality images to function well. However, some are more susceptible to certain types of disturbances. As was discussed above, decomposition algorithms treat the recognition problem as a general pattern recognition problem, but chose the basis vectors for the decomposition based on a developmental database of faces. This approach is often sensitive to variations in rotation and position of the face in the image. Performance also degrades rapidly with pose changes, non-uniform illumination, and background clutter. In contrast, these systems are quite robust in dealing with very small images. This approach is most appropriate for applications where the image conditions are relatively controlled. In contrast, EBGM-based algorithms are much more robust against variations in lighting, eyeglasses, facial expression, hairstyle, and individual's pose up to 25 degrees. However, they are obviously still heavily dependent on the extracted facial features in the first instance and may be dependent upon consistent estimation of landmark points. It is important that an appropriate implementation algorithm be used.[73] As developers start to combine algorithms these considerations may become less important.

The same algorithm can function in very different ways depending on the developmental data set that was used to develop the system. Generally, one can say that the range and diversity of the developmental set will set the boundaries for the diversity of probe images that the algorithm will be able to deal with. However, it is also true that the closer the match between the conditions of the probe image and the gallery image the higher the likelihood that the system will perform well.

6.5. Operational FRR/FAR thresholds

The discussion above has shown that there are clear tradeoffs to be made when it comes to the operation of FRS. The selection of the system performance threshold determines these tradeoffs. The performance threshold can be understood as the level of certainty to which the system must perform. For example, in the verification scenario, one might decide that it is important that no imposters be accepted. This will require that a very high threshold for the matching or similarity score be set (say a FAR of 0.002). However, this will mean that valid identities are rejected (i.e., the FRR will increase). In the BioFace evaluation, a 0.5% FAR equated to a 10.5% FRR (i.e., the percent of valid identities that were rejected). It is unlikely that such a threshold rate could be used in installations with high throughput levels (such as airports) as it would generate a large amount of false rejects that would need to be dealt with by human operators.

For example, the SmartGate system discussed above reported a false reject rate (FRR) of approximately 1 in 11 (9%). What does this mean for the false accept rate? According to a report of the Australian Customs and Immigration service, the system works on a false accept rate of well below 1%. This would require exceptionally high quality image data to achieve. Unfortunately, there is no independent publicly available evaluation to confirm these reported figures. Nevertheless, it is clear that there are important choices to be made in deciding on the error rates one is prepared to accept for a given level of image data quality. If the gallery or probe image quality is low, then a high threshold will generate significant levels of false accepts or false rejects. These will then need to be dealt with in an appropriate manner.

6.6. Recognition rates and covariates of facial features: system biases?

One of the questions that often come up is whether different facial features related to race, gender, etc., make it easier or harder to recognize an individual. For example, the FRVT 2002 closed-set evaluation suggested that recognition rates for males were higher than females. For the top systems, closed-set identification rates for males were 6% to 9% points higher than that of females. Likewise, recognition rates for older people were higher than younger people. For 18 to 22 year-olds, the average identification rate for the top systems was 62%, and for 38 to 42 year-olds, 74%. For every ten-year increase in age, performance increases on average 5% through age 63. Unfortunately, the FRVT could not evaluate the effects of race as the large data set consisted of mostly Mexican non-immigrant visa applicants. However, subsequent research, using Principal Component Analysis (PCA) algorithms, has indeed confirmed some of the biases found in the FRVT 2002 evaluation, noting a significant racial bias but no gender bias.[74] These biases were confirmed using balanced databases and controlling for other factors. This study concluded that: "Asians are easier [to recognize] than whites, African-Americans are easier than whites, other race members are easier than whites, old people are easier than young people, other skin people are easier to recognize than clear skin people."[75]

Differences in algorithm design and systemic features can create problematic interactions between variables (i.e., there can be troublesome covariates). It is very difficult to separate these covariates. The interaction between these covariates has lead to some conflicting results in a variety of experiments.[76]

One might ask why these biases are important. If algorithms operate on high threshold tolerances, then it is more likely that individuals within certain categories might receive disproportionately greater scrutiny.[77] Moreover, these facial feature covariates may interact with other factors outlined in this section to create a situation where system recognition risks (mistakes) are disproportionately experienced by a specific group based on gender, race, age, etc. If this is the case, then it could be very problematic for the actual operation of the system—especially when the assumption is made, as it is often the case, that technology is neutral in its decision making process.

6.7. Situating and staffing

FRT is not so robust that it could or should be "black boxed" (i.e., sealed off from human intervention). FRS would need ongoing human intervention (often of high-level expertise) to ensure its ongoing operation. Moreover, the system will depend on human operators to make decisions on cases of either false rejection or false identification. It is entirely likely that a false identification can occur since there is likely to be a significant similarity between the targeted person and the probe image. How will the staff deal with this? They may assume that it is a true positive and that the other two elements in the identity triad have been falsified. The operators may even override their own judgments as they may think that the system "sees something" that they do not. This is likely as humans are not generally very good at facial recognition in high pressure situations.[78] This becomes increasingly significant if taken together with the other factors discussed above. Indeed, it might be that under these conditions, the bias group (African-Americans, Asians, dark skinned persons, and older persons) may be subjected to disproportionate scrutiny.

We would suggest that FRS in operational settings require highly trained and professional staff to operate. It is important that they understand the operating tolerances[79] and are able to interpret and act appropriately given the exceptions generated by the system. They should also be supported with the necessary identity management infrastructure to deal with situations of ambiguity—such as systems to investigate the other two elements in the identity triad. This is vital if public confidence in the technology is to be ensured.

7. Some policy and implementation guidelines ("what important decisions need to be considered?")

Having considered technical and operational issues, we now put these insights to use to inform key policy decisions regarding FRT in particular contexts. In this section, we outline broad policy considerations to help decide whether FRT is appropriate for a particular context, and spell out policy considerations to guide operational protocol if the decision to implement FRT has been rendered.

7.1. Some application scenario policy considerations

A decision whether to invest in or use FRT depends on a number of factors, the most salient of which are the specific purpose and function FRT will perform within a broader identity management and security infrastructure. In all cases, it should be noted that FRT is not a general technology (in the way that a dishwasher is a general technology that is relatively context independent). Every application of FRT is highly specific to the particular environment in which it will function. We would suggest that each application is so specific to its context that one should consider each implementation as being purpose-built—i.e., FRS should be seen as one-off systems.

7.1.1. FRS, humans, or both?

There is no doubt that FRT is developing very rapidly. FRVT 2006 indicated that FRT could, under certain conditions, outperform humans. This is particularly true in the following scenario:

- In the verification task
- Where high quality images exist (both in the gallery and in the probe image)
- Where a large amount of data needs to be processed

An example of such an application is the use of FRT to check if a driver attempts to procure multiple drivers licenses (or passports) under different names. The human operator can then be used to deal with exceptions. Humans are good in more difficult or nuanced situations (especially if they are dealing with their own ethnic group).[80] In such a scenario, careful consideration needs to be given to how the various parts of the task are distributed between humans and computers. If humans are used to deal with the exceptions, then these humans should be trained and have a high level of expertise in the additional verification and identification tasks that may be required to establish identity.

7.1.2. Verification and identification in controlled settings

It is clear from the research that FRT has matured to the point where it is possible to consider its use for verification task in highly controlled environments. SmartGate is an example of a relatively successful application. Such applications (as is the case in iris scanning) will require the active participation of subjects. Key questions, therefore, include: How will the participation of the subject be secured (where, when, and under what conditions)? Will the service be rolled out as a replacement or in addition to existing identification procedures, and could this create a form of tiered service? Who will be enrolled first? How will civil liberty issues (discussed below) be addressed?

7.1.3. Identification in semi-controlled settings

It might be possible to consider the use of FRT as a filtering mechanism to aid identification when managing high levels of throughput such as in airports, subway stations, etc. Such applications could be seen as high-risk applications that may need considerable upfront investment to develop and for ongoing tuning of the system to changing environmental conditions (as discussed in the BKA study and below). They should also function as part of a larger security infrastructure in which they fulfill a very specific purpose. They should never be used as a "just in case" technology. For example, it might be possible to create semi-controlled conditions that would allow one to get relatively high quality images of passengers as they disembark from an aircraft. One might further have intelligence that suggests that certain individuals might try to enter the country using specific airports. If one had relatively good quality images of these suspects (to place in the gallery) then one could use this to filter potential suspects from specific destinations as they disembark. In such a scenario, the FRT functions in a well defined way as part of a broader intelligence and security infrastructure. In our view, this is possible but must still be seen as a high-risk (or high cost) application scenario in the sense that there may be many false positives requiring further investigation.

7.1.4. Uncontrolled identification at a distance ("grand prize")

The current state of FRT does not support identification in uncontrolled environments, especially in crowd situations. The BKA evaluation indicated that even moderately controlled environments (such as well-lit escalators) only produce a 60% recognition rate, even with high quality gallery images. Should the technology advance in ways to overcome the problems discussed above, these types of applications are likely to be the most politically sensitive for liberal democracies (as discussed below).

7.2. Some implementation guidelines

Once the decision is made that FRS is an appropriate technology and it is clear how it will function within a broader intelligence and security strategy, a number of operational policies need to be specified.

7.2.1. Clear articulation of the specific application scenario

It is vital that potential customers of FRT have a very clear articulation of the implementation purpose and environment when they engage with application providers or vendors. Integration with the broader identity management and security infrastructure needs to be clearly thought through and articulated. What will be the recognition tasks? How will these tasks interact with other identity management and security operations? What will be the specific environment in which the system will function? What are the constraints that the specific environment imposes?

7.2.2. Compilation of gallery and watch list

When FRT is used to perform identification or watch list tasks, users should be able to answer a number of key questions:

- Who do we include in the gallery/watch list and why? (As we have seen, restricting the size of the gallery is a significant performance question.)
- What is the quality of the gallery and probe images?
- What is the likely age difference between the gallery and the probe images?
- What are the demographics of the anticipated subjects? (It is important that the system has been trained with images that at least reflect as broad a range possible of the demographics of the use-context.)
- What other resources are available, linking together all components of the identity triad? Final identification by means of triangulation with other identity elements is essential, as FRS must always function within the context of a larger security infrastructure.
- Have we adopted satisfactory policies governing the sharing of gallery images with others?

It is also important to try to ensure that there is as much similarity as possible between the enrollment conditions (when creating the gallery) and the probe images (captured for verification of identification) in terms of lighting, background, orientation, etc. (FRVT 2006 showed that this could lead to very significant improvements in performance). It is recommended that one should always at least use the ANSI 385 2004 good practice guidance for face biometric images and the ISO/IEC 19794-5 standard for minimum gallery image quality. Of course, FRVT 2006 also showed that improvements in performance can be gained by going beyond these standards.

7.2.3. From technology to integrated systems

It is important for users to make sure that the facial recognition supplier or vendor has the capability and track-record to deliver fully integrated operational systems. The BioFace evaluation showed that implementation expertise is not widespread and could represent a significant risk. This is especially important in light of the fact that FRS are one-off customized implementations. It is important that a system be extensively tested and piloted before its use is approved. It is also likely that the facial recognition implementation will require ongoing fine-tuning to the local conditions for it to perform at its full potential. As discussed, evaluations have shown that small variations can have dramatic effects on performance.

7.2.4. Overt or covert use?

Another important policy issue is whether the system is to be used overtly or covertly. Obviously these two options call for very different sorts of implementations. Covert use, specifically, may also raise civil liberty implications that need to be carefully considered (see discussion below).

7.2.5. Operational conditions and performance parameters

As discussed above, setting suitable performance thresholds is crucial. The rate at which individuals move through the relevant chokepoints in the system is an important consideration as the BKA study showed. High volumes of traffic with low Bayesian priors and low thresholds (i.e., high number of false positives) will require a lot of resources and careful design of the space to deal with all the potential false positives in an appropriate manner. This is best done in the context of an overall security strategy as discussed above. It is essential that system operators understand the relationship between these system performance parameters and the actual performance of the system *in situ*. This means that the systems should only be operated by fully trained and professional staff. Standardized policy and practices need to be developed for establishing the relevant thresholds and for dealing with alarms. These policies should also be subject to continual review to ensure the ongoing performance of the system. Setting appropriate performance parameters is often a matter of trial and error that needs ongoing tuning as the system embeds itself within the operational context.

7.2.6. Dealing with matches and alarms

There is the risk with FRT that individuals are treated as "guilty until proven innocent." In an identification scenario, we recommend that all matches be treated, in the first instance, as potential false positives until verified by other independent sources (such as attributed and biographical identifiers). This underscores the fact that the FRS must form part of an overall identity management program within a security and intelligence infrastructure. Identity management and security cannot be delegated to FRT. It can only act *in support of* specific targeted security and intelligence activities. Further, how one deals with matches and alarms must be suitable for the context. For example, one might have a very different set of practices in an airport, casino, or a prison. This means that one needs to consider carefully the timeframe, physical space, and control over the subject as they flow through the system.

8. Moral and political considerations of FRT

This report has considered technical merits of FRT and FRS, particularly as they function in real-world settings in relation to specific goals. Although certain barriers to performance might be overcome by technical breakthroughs or mitigated by policies and guidelines, there remains a class of issues deserving attention not centered on functional efficiency but on moral and political concerns. These concerns may be grouped under general headings of privacy, fairness, freedom and autonomy, and security. While some of these are characteristically connected to facial recognition and other biometric and surveillance systems, generally, others are exacerbated, or mitigated, by details of the context, installation, and deployment policies. Therefore, the brief discussion that follows not only draws these general connections, it suggests questions that need addressing in order to anticipate and minimize impacts that are morally and politically problematic.

8.1. Privacy

Privacy is one of the most prominent concerns raised by critics of FRS. This is not surprising because, at root, FRS disrupts the flow of information by connecting facial images with identity, in turn connecting this with whatever other information is held in a system's database.[81] Although this need not in itself be morally problematic, it is important to ascertain, for any given installation, whether these new connections constitute morally unacceptable disruptions of entrenched flows (often regarded as violations of privacy) or whether they can be justified by the needs of the surrounding context. We recommend that an investigation into potential threats to privacy be guided by the following questions:

- Are subjects aware that their images have been obtained for and included in the gallery database? Have they consented? In what form?
- Have policies on access to the gallery been thoughtfully determined and explicitly stated?
- Are people aware that their images are being captured for identification purposes? Have and how have they consented?
- Have policies on access to all information captured and generated by the system been thoughtfully determined and explicitly stated?
- Does the deployment of an FRS in a particular context violate reasonable expectations of subjects?
- Have policies on the use of information captured via the FRS been thoughtfully determined and explicitly stated?
- Is information gleaned from a FRS made available to external actors and under what terms?

- Is the information generated through the FRS used precisely in the ways for which it was set up and approved?

Although notice and consent are not necessary for all types of installations, it is essential that the question be asked, particularly in the context of answers to all the other privacy-related questions. If, for example, policies governing the creation of the gallery, the capture of live images, and access by third parties to information generated by the systems are carefully deliberated and appropriately determined, notice and consent might be less critical, particularly if other important values are at stake. It is also clear that requirements will vary across settings, for example in systems used to verify the identity of bank customers versus one used to identify suspected terrorists crossing national borders.

Whatever policies are adopted, they should be consistent with broader political principles, which in turn must be explicit and public. Generally, any changes in a system's technology or governing policies from the original setting for which it was approved requires reappraisal in light of impacts on privacy. For example, subjects might willingly enroll in a FRS for secure entry into a worksite or a bank but justifiably object if, subsequently, their images are sold to information service providers and marketing companies like ChoicePoint or DoubleClick.[82] The general problem of expanding the use and functionality of a given FRS beyond the one originally envisioned and explicitly vetted is commonly known as the problem of "function creep."

8.2. Fairness

The question of fairness is whether the risks of FRS are borne disproportionately by, or the benefits flow disproportionately to, any individual subjects, or groups of subjects. For example, in the evaluations discussed above, noting that certain systems achieve systematically higher recognition rates for certain groups over others—older people over youth and Asians, African-Americans, and other racial minorities over whites—raises the politically charged suggestion that such systems do not belong in societies with aspirations of egalitarianism. If, as a result of performance biases, historically affected racial groups are subjected to disproportionate scrutiny, particularly if thresholds are set so as to generate high rates of false positives, we are confronted with racial bias similar to problematic practices such as racial profiling. Beyond thorny political questions raised by the unfair distribution of false positives, there is the philosophically intriguing question of a system that manages disproportionately to apprehend (and punish) guilty parties from one race, ethnicity, gender, or age bracket over others. This question deserves more attention than we are able to offer here but worth marking for future discussion.[83]

8.3. Freedom and Autonomy

In asking how facial recognition technology affects freedom and autonomy, the concern is constraints it may impose on people's capacity to act and make decisions ("agency"), as well as to determine their actions and decisions according to their own values and beliefs. It is important to stress that the question is posed against a backdrop of existing expectations and standards of freedom and autonomy, which recognize that freedom and autonomy of any person is legitimately circumscribed by the rights of others, including their freedom, autonomy, and security.

Let us consider an incident reported in *Discover* about a facial recognition installation at the Fresno Yosemite International Airport:[84]

> "[The system] generates about one false positive for every 750 passengers scanned," says Pelco vice president Ron Cadle. Shortly after the system was installed, a man who looked as if he might be from the Middle East set the system off. "The gentleman was detained by the FBI, and he ended up spending the night," says Cadle. "We put him up in a hotel, and he caught his flight the next day."[85]

It seems from this quote that an individual was detained and questioned by the FBI because he triggered the alarm and "looked as if he might be from the Middle East." It is of course possible that the FBI had other reasons to detain the individual (not reported in the quote). We have not been able to corroborate the facts surrounding this incident but would still like to pose it as an interesting, and potentially informative, anecdote to flesh out some of the issues surrounding freedom and autonomy.

This anecdote illustrates several of the moral and political pitfalls not of FRT per se but how it is installed and implemented, as well as the policies governing its operation. To begin, it raises questions of fairness (discussed above) as it might suggest people of certain ethnicities might be burdened disproportionately. It may also suggest a challenge to the "presumption of innocence" enjoyed by citizens in liberal democracies, meaning that interference with freedom and autonomy requires a clear showing of "probable cause." (This criticism applies to many surveillance installations in public places.)

In the Fresno-Yosemite incident, it seems as if the FBI placed the burden of proof on the individual to produce additional identification, impeding his movement, and, by derailing his travel plans, curtailing his autonomy. This transforms the question of determining acceptable levels of false positives from a merely operational (technical) one into an ethical one. Moreover, one must weigh the burden placed on falsely identified subjects, for a given threshold, against the threat or risks involved. For example, when travelers passing through metal detectors cause an alarm, the burden of a manual body search, which takes a minute or two, is likely to be seen by most people as proportionate to the risk involved. However, if a falsely identified person is detained, and as a result misses a flight, many people might consider this a disproportionate burden, particularly if the identification process was done covertly without the individual's awareness or meaningful consent. There is also the less tangible but nevertheless serious risk of humiliation in being pulled out of line for further investigation.

In societies that value freedom and autonomy, it is worth questioning whether the burden of requiring individuals to follow routes optimal for system performance rather than routes most efficacious for achieving their own goals is acceptable. Related puzzles are raised by the question of whether taking advantage of a central virtue of FRT, the capacity to identify covertly and at a distance, is acceptable for free societies whose political bedrock includes presumption of innocence and meaningful consent. Meaningful consent recognizes subjects as decision makers by providing them information and the capacity to accept or reject conditions of the system (for example, allowing people to opt out of a particular service or place if it requires enrollment in a system and identification). Autonomy is also at stake when a nation-state or an organization, upon endorsing the deployment of FRT, must take steps to enroll citizens (employees, customers, members, etc.) into the gallery. Under what conditions, if ever, is it appropriate to coerce participation?[86] Even when FRT functions in a filtering role, certain assumptions are made about the subjects that "trigger" alarm. Subjecting citizens to the scrutiny of FRS can be conceived as investigating them in the absence of probable cause and a violation of civil liberties.

A more general issue raised by biometric identification technologies is how they affect the distribution of power and control by shifting what we might call the landscape of identifiability. Where identification is achieved through the administration of FRT, subjects may be identified by operators, systems managers, and owners, who themselves remain anonymous and, often, even unseen. This imbalance may feel and amount to a power imbalance, which needs to be questioned and justified. Even the mundane, relatively trivial experience of a sales call in which one is personally targeted by an unknown caller, elicits a sense of this power imbalance, even if fleeting.

Not being able to act as one pleases for fear of reprisal is not necessarily problematic if one happens to want to act in ways that are harmful to others, and clearly, there may be times and circumstances in which other considerations might trump freedom and autonomy, for example, in dealing with dire security threats. Our view is that in order to achieve balanced ends, FRT must function as part of a intelligence and security infrastructure in which authorities have a clear and realistic vision of its capacities and role, as well as its political costs.

8.4. Security

Acceptance of facial recognition and other biometric identification systems has generally been driven by security concerns and the belief that these technologies offer solutions. Yet, less salient are the security threats posed by these very systems, particularly threats of harm posed by lax practices dealing with system databases. Recent incidents in the UK and US suggest that institutions still do not deserve full public trust in how they safeguard personal information. In the case of biometric data, this fear is magnified many times over since it is generally assumed to be a non-falsifiable anchor of identity. If the biometric template of my face or fingerprint is used to gain access to a location, it will be difficult for me to argue that it was not me, given general, if problematic, faith in the claim that "the body never lies." Once my face or fingerprint has been digitally encoded, however, it can potentially be used to act "as if" it were me and, thus, the security of biometric data is a pressing matter, usefully considered on a par with DNA data and evidence. A similar level of caution and security needs to be established. In our view, minimally, the following questions ought to be raised:

- Does the implementation of the system include both policy and technology enforced protection of data (gallery images, probe images, and any data associate with these images)?
- If any of this information is made available across networks, have necessary steps been taken to secure transmission as well as access policies?

Two other issues, seldom identified as security issues, bear mentioning. One is the indirect harm to people who opt out, that is, refuse to enroll. About any system whose implementation is justified on grounds that subjects have consented to enroll or participate, it is essential to ask what the cost is to those who choose not to. Consent cannot be considered meaningful if the harm of not enrolling is too great. Finally, a system whose threshold allows too many false negatives, that is, offenders to be systematically overlooked, poses an almost greater threat than no system at all as it imbues us with a false sense of security.

8.5. Concluding comments on the moral and political considerations

As with functional performance, moral and political implications of FRS are best understood in a context of use and against a material, historical, and cultural backdrop. Most importantly, however, this report recommends that moral and political considerations be seen as *on a par* with functional performance considerations, influencing the design of technology and installation as well as operational policies throughout the process of development and deployment and not merely tacked on at the end. Finally, it is important to assess moral and political implications of a FRS not only on its own merits but in comparison with alternative identification and authentication systems, including the status-quo.

9. Open questions and speculations ("what about the future?")

There are good reasons to believe that it will still be some time before FRT will be able to identify "a face in the crowd" (in uncontrolled environments) with any reasonable level of accuracy and consistency. It might be that this is ultimately an unattainable goal, especially for larger populations. Not because the technology is not good enough but because there is not enough information (or variation) in faces to discriminate over large populations—i.e. with large populations it will create many biometric doubles that then need to be sorted out using another biometric. This is why many researchers are arguing for multi-modal biometric systems. Thus, in the future we would expect an increased emphasis on the merging of various biometric technologies. For example, one can imagine the merging of face recognition with gait recognition (or even voice recognition) to do identification at a distance. It seems self-evident that these multi-modal systems are even more complex to develop and embed in operational context than single mode systems. It is our view that the increasing reliance on biometric and pattern recognition technologies do represent a significant shift in the way investigation and security is conducted. There is an ongoing need to evaluate and scrutinize biometric identification systems given the powerful nature of these technologies—due to the assumption that falsification is either impossible or extremely difficult to do.

End of report

Appendix 1: Glossary of terms, acronyms and abbreviations

Glossary of terms

Attributed identifier — An attributed piece of personal information (e.g., a (unique) name, Social Security number, bank account number, or driver's license number)

Biographical identifier — An assumed piece of personal information (e.g., an address, professional title, or educational credential)

Biometric — See *biometric characteristic*

Biometric characteristic — A biological and/or behavioral characteristic of an individual that can be detected and from which distinguishing *biometric feature*s can be repeatedly extracted for the purpose of automated recognition of individuals

Biometric data subject — An individual from whom *biometric features* have been obtained and to whom they are subsequently attributed

Biometric double — A face image which enters the *gallery* as a sufficiently similar *biometric template* to a preexisting image that belongs to a different individual

Biometric feature — A *biometric characteristic* that has been processed so as to extract numbers or labels which can be used for *comparison*

Biometric feature extraction — The process through which *biometric characteristic*s are converted into *biometric template*s

Biometric identification — Search against a *gallery* to find and return a sufficiently similar *biometric template*

Biometric identification system — A face recognition system that aims to perform *biometric identification*

Biometric identifier — See *Biometric reference*

Biometric probe — *Biometric characteristics* obtained at the site of verification or identification (e.g., an image of an individual's face) that are passed through an algorithm which convert the characteristic into b*iometric features* for comparison with *biometric templates*

Biometric reference — one or more stored biometric samples, biometric templates or biometric models attributed to a biometric data subject and used for comparison. For example a face image on a passport

Biometric reference database — A *gallery* of stored *biometric templates* obtained through *enrollment*

Biometric sample — Information or computer data obtained from a biometric sensor device. Examples are images of a face or fingerprint

Biometric template — Set of stored biometric features comparable directly to biometric features of a probe biometric sample (see also *biometric reference*)

Biometric twin — See *Biometric double*

Biometric verification — The process by which an identification claim is confirmed through *biometric comparison*

Candidate — A *biometric template* determined to be sufficiently similar to the *biometric probe*, based on a *comparison score* and/or rank

Closed-set Identification — A biometric task where an unidentified individual is known to be in the database and the system attempts to determine his/her identity. Performance is measured by the frequency with which the individual appears in the system's top rank (or top 5, 10, etc.), often reported using the *cumulative match score or characteristic.*

Comparison — A process of comparing a biometric template with a previously stored template in the reference database in order to make an identification or verification decision.

Comparison score — Numerical value (or set of values) resulting from the comparison of a *biometric probe* and *biometric template*

Cumulative Match Characteristic (CMC) — A method of showing measured accuracy performance of a biometric system operating in the closed-set identification task by comparing the rank (1, 5, 10, 100, etc.) against the identification rate

Database — See *Gallery*

Database image — See *Biometric template*

Decision boundary — A limit, based on *similarity scores*, at which a face recognition algorithm, technology, or system is set to operate

Developmental set — A set of face images that the developers use to train the algorithm to detect and extract features from a face

Dissimilarity score — See *Distance score*

Distance score — *Comparison score* that decreases with similarity

Enrollment — The process through which a *biometric characteristic* is captured and must pass in order to enter into the image *gallery* as a *biometric template*

Equal error rate (EER) — The rate at which the *false accept rate* is exactly equal to the *false reject rate*

Evaluation set — A set of *biometric template*s, generally separated out from the *training set*, which are exposed to a facial recognition algorithm in order to evaluate its performance

Face template — See *Biometric template*

Facial features – The essential distinctive characteristics of a face, which algorithms attempt to express or translate into mathematical terms so as to make recognition possible.

Facial landmarks — Important locations in the face-geometry such as position of eyes, nose, mouth, etc.

False accept — An incorrect acceptance of a false claim to existence or non-existence of a candidate in the reference database during the verification task

False accept rate (FAR) — A statistic used to measure biometric performance when performing the verification task. The percentage of times a face recognition algorithm, technology, or system falsely accepts an incorrect claim to existence or non-existence of a candidate in the database over all comparisons between a *probe* and *gallery image*

False alarm — A metric used in open-set identification (such as watch list applications). A false alarm is when an alarm is incorrectly sounded on an individual who is not in the biometric system's database, or an alarm is sounded but the wrong person is identified

False Alarm Rate (FAR)— A statistic used to measure biometric performance when operating in the open-set identification (sometimes referred to as watch list) task. This is the percentage of times an alarm is incorrectly sounded on an individual who is not in the biometric system's database, or an alarm is sounded but the wrong person is identified.

False match rate (FMR) — See *false accept rate*

False negative — An incorrect non-match between a *probe* and a candidate in the *gallery* returned by a face recognition algorithm, technology, or system

False non-match rate (FNMR) — See *false reject rate*

False positive — An incorrect match between a *biometric probe* and *biometric template* returned by a face recognition algorithm, technology, or system

False reject — An incorrect non-match between a *biometric probe* and *biometric template* returned by a face recognition during the verification task

False reject rate — A statistic used to measure biometric performance when performing the verification task. The percentage of times a face recognition algorithm, technology, or system incorrectly rejects a true claim to existence or non-existence of a match in the *gallery*, based on the comparison of a *biometric probe* and *biometric template*

Gallery — A database in which stored *biometric templates* reside

Gallery image — See *Biometric template*

Grand prize — The surreptitious identification of an individual's face at a distance in uncontrolled settings, commonly described as the "face in the crowd" scenario

Identification— A task where the biometric system searches a database for a biometric template that matches a submitted biometric sample (probe), and if found, returns a corresponding identity

Identification rate — A metric used in reporting results of "closed-set" tests to indicate the probability that a probe and a candidate in the gallery be matched at Rank k when a probe is searched against the entire reference database.

Identification task — See *Biometric identification*

Identity triad — Identity resolution by way of *attributed identifier*s, *biographical identifier*s, and *biometric characteristic*s

Impostor — A person who submits a biometric sample in either an intentional or inadvertent attempt to claim the identity of another person to a biometric system

Match — A match is where the similarity score (of the probe compared to a biometric template in the reference database) is within a predetermined threshold

Matching score (deprecated) — See *Comparison score*

Normalization — The adjustment of the size, scale, illumination, and orientation of the face in *biometric probe* and *biometric template*s to ensure commensurability

Open-set Identification — A biometric identification task where an unidentified individual is not known to be in the reference database when the system attempts to determine his/her identity. Performance is normally reported in terms of recognitions rates against false alarm rates

Print — See *Biometric template*

Probe biometric sample — See *Biometric probe*

Probe image — See *Biometric probe*

Rank list — A rank ordered candidate list of the percent most likely matches for any given probe image

Receiver Operating Characteristics (ROC) — A method of reporting the accuracy performance of a facial recognition system. In a verification task the ROC compares false accept rate vs. verification rate. In an open-set identification task the ROC compares false alarm rates vs. detection and identification rate

Recognition — A generic term used in the description of biometric systems (e.g. face recognition or iris recognition) relating to their fundamental function. The term "recognition" does not inherently imply the verification, closed-set identification or open-set identification (watch list)

Recognition at a distance — The explicit or surreptitious identification or verification of an individual based on an image acquire from afar and without the use of an intrusive interface

Recognition rate — A generic metric used to describe the results of the repeated performance of a biometric system to indicate the probability that a probe and a candidate in the gallery be matched

Reference biometric feature set — See *Biometric template*

Similarity score — A value returned by a biometric algorithm that indicates the degree of similarity or correlation between a biometric template (probe) and a previously stored template in the reference database

Three-dimensional (3D) algorithm — A recognition algorithm that makes use of images from multiple perspectives, whether *feature-based* or *holistic*

Threshold — Numerical value (or set of values) at which a *decision boundary* exists

Top match score — The likelihood that the top match in the *rank list* for the probe image of an individual is indeed the same individual in the database image

Training set — A set of face images to which a facial recognition algorithm is initially exposed in order to train the algorithm to detect and extract features from a face

True accept rate (TAR) — 1-*false reject rate*

True reject rate (TRR) — 1-*false accept rate*

Validation set — See *Evaluation set*

Verification task — See *Biometric verification*

Watch list — The surreptitious attempt to identify a non-self-identifying individual by comparing his or her *probe image* to a limited set of database image

Acronyms and Abbreviations

ANSI INCITS — American National Standards Institute International Committee for Information Technology Standards

BKA — Federal Office of Criminal Investigation, Wiesbaden, Germany

DARPA — Defense Advanced Research Projects Agency

EBGM — Elastic Bunch Graph Matching

EER — Equal Error Rate

FAR — False Accept Rate

FERET — The Face Recognition Technology program

FOIS — Federal Office for Information Security, Bonn, Germany

FRGC —Face Recognition Grand Challenge

FRR — False Reject Rate

FRS — Face/Facial Recognition System

FRT — Face/Facial Recognition Technology

FRVT —Face Recognition Vendor Test

ICA — Independent component analysis

ICAO — The International Civil Aviation Organization

IGD — Fraunhofer Institute for Computer Graphics Research

ISO/IEC — International Standard Organization/ International Electro technical Commission

JPEG — Joint Photographic Experts Group

LFA — Local Feature Analysis

NAVSEA — US Naval Sea Systems Command

NIST — National Institute of Standards and Technology

PCA — Principal Component Analysis

ROC — Receiver Operating Characteristic

SVM — Support Vector Machines

TAR — True Accept Rate

TRR — True Reject Rate

Appendix 2: Works cited

Agre, Philip E. "Your Face is Not a Bar Code: Arguments Against Automatic Face Recognition in Public Places." *Whole Earth*. 106 (2001): 74-77.

Beveridge, J.Ross, Givens, Geof H., Phillips, P. Jonathan, Draper, Bruce A., and Lui, Yui Man. "Focus on Quality: Predicting FRVT 2006 Performance." *8th IEEE International Conference on Automatic Face and Gesture Recognition,* Amsterdam, The Netherlands, 17-19 September 2008. Available: http://www.cs.colostate.edu/~ross/research/papers/yr2008/focusFRVT2006.pdf

"BioFace: Comparative Study of Facial Recognition Systems." Bundesamt für Sicherheit in der Informationstechnik, 2003. Available: http://www.igd.fhg.de/igd-a8/en/projects/biometrie/bioface/BioFaceIIReport.pdf.

Bliss, Rhonda. Correspondence via e-mail. 23 January 2009.

Blackburn, Duane M. *Face Recognition 101: A Brief Primer*. Department of Defense Counterdrug Technology Development Program Office. 07 April 2003. Available: http://www.frvt.org/DLs/FR101.pdf.

Boggan, Steve. "'Fakeproof' E-Passport is Cloned in Minutes." *The Times (UK)*. 6 August 2008. Available: *http://www.timesonline.co.uk/tol/news/uk/crime/article4467106.ece*.

Bone, J. M., and Duane M. Blackburn. *Face Recognition at a Chokepoint: Scenario Evaluation Results*. Arlington: Department of Defense Counterdrug Technology Development Program Office, 2002.

Borchers, Detlef and Robert W. Smith. "Mixed reactions to the facial-features recognition technology project of the BKA." *Heise Online*. 16 July 2007. Available: http://www.heise.de/english/newsticker/news/92722.

Bowyer, Kevin W., Kyong Chang, and Patrick Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition." *Computer Vision and Image Understanding*. 101.1 (2006): 1-15.

Brey, Philip. "Ethical Aspects of Face Recognition Systems in Public Places." *Journal of Information, Communication & Ethics in Society*. 2.2 (2004): 97-109.

Brooks, Michael. "Face-off." *New Scientist*. 175.2399 (2002).

"The Constitutional Right to Anonymity: Free Speech, Disclosure and the Devil." *The Yale Law Journal,* 70.7 (1961): 1084-1128.

Davis, Natalie. *The Return of Martin Guerre*. Cambridge: Harvard University Press, 1983.

Daum, Henning. "Influences of Image Disturbances on 2D Face Recognition." *Audio- and Video-Based Biometric Person Authentication*. Ed. Takeo Kanade. Berlin: Springer, 2005, 900-908.

Etemad, K. and R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *Journal of the Optical Society of America A*, 14. 8 (1997): 1724-1733.

"Face Recognition: Part 2." *Biometric Technology Today*. 15.10 (2007): 10-11.

Furl, N., J. P. Phillips, and A. J. O'Toole. "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis." *Cognitive Science*. 26 (2002): 797-815.

Garfinkle, Simson. "Future Tech." *Discover*. 23.9 (2002): 17-20.

"German casinos secured with facial recognition." *Biometric Technology Today*. 14.11/12 (2006): 12.

Gesichtserkennung als Fahndungshilfsmittel. Deutsches Bundeskriminalamt. *Abschlussbericht* (2007) [in German]. Available: http://www.bka.de/kriminalwissenschaften/fotofahndung/pdf/fotofahndung_abschlussbericht.pdf.

Givens, G., J. R. Beveridge, B. A. Draper, and D. Bolme. "A Statistical Assessment of Subject Factors in the PCA Recognition of Human Faces." *Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop*, 8 (2003): 1-9.

Givens, G., J. R. Beveridge, B. A. Draper, P. Grother, and P. Phillips. "How Features of the Human Face Affect Recognition: a Statistical Comparison of Three Face Recognition Algorithms." *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2 (2004): 381–388.

Gross, R., J. Shi, and J. F. Cohn, J.F. "Quo vadis Face Recognition?" (2001). Available: http://dagwood.vsam.ri.cmu.edu/ralph/Publications/QuoVadisFR.pdf.

Harcourt, Bernard E. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: University of Chicago Press, 2007

Harcourt, Bernard E. "A Reader's Companion to Against Prediction: A Reply to Ariela Gross, Yoram Margalioth, and Yoav Sapir on Economic Modeling, Selective Incapacitation, Governmentality, and Race." *Law & Social Inquiry*. 33.1 (2008): 265-283.

Heisele, B., P. Ho, J. Wu, T. Poggio. "Face recognition: component-based versus global approaches." *Computer Vision and Image Understanding*. 91.1 (2003): 6-21.

Husken, M., M. Brauckmann, S. Gehlen, and C. von der Malsburg. "Strategies and benefits of fusion of 2D and 3D face recognition." *Proc. IEEE Workshop on Face Recognition Grand Challenge Experiments*, San Diego, CA, 20-25 June 2005.

Introna, L., and D. Wood. "Picturing algorithmic surveillance: the politics of facial recognition systems." *Surveillance and Society*, 2.2/3 (2004): 177-198.

Jenkins, R. and A. M. Burton. "100% Accuracy in Automatic Face Recognition." *Science*. 319.5862 (2008): 435-435.

Juels, A., Molnar, D., and Wagner, D. "Security and Privacy Issues in E-passports." IEEE/CreateNet SecureComm, 2005. Available: *http://eprint.iacr.org/2005/095.pdf*

Kemp, R., N. Towell, and G. Pike. "When seeing should not be believing: photographs, credit cards and fraud." *Applied Cognitive Psychology*. 11.3 (1997): 211-222.

Lease, David R. "Factors Influencing the Adoption of Biometric Security Technologies by Decision Making Information Technology and Security Managers," Diss. Capella University, 2005.

Lu, Xiaoguang. "Image Analysis for Face Recognition." Personal notes, Dept. of Computer Science &. Engineering, Michigan State University,, May 2003. Available: *http://www.face-rec.org/interesting-papers/General/ImAna4FacRcg_lu.pdf*

McLindin, Brett. "Improving the performance of Two-Dimensional Facial Recognition Systems." Diss. University of South Australia, 2005. Available: http://www.library.unisa.edu.au/adt-root/public/adt-SUSA-31102005-074958/index.html.

Meek, J. "Robo cop: Some of Britain's 2.5 million CCTV cameras are being hooked up to a facial recognition system designed to identify known criminals. But does it work?" *Guardian*, 13 June 2002.

Moses, Y.. Y. Adini, and S. Ullman. "Face Recognition: The Problem of Compensating for Changes in Illumination Direction." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 19.7 (1997): 721-732.

National Science and Technology Council (NCST) Subcommittee on Biometrics. *Face Recognition*. 7 August 2006. Available: http://www.biometrics.gov/Documents/FaceRec.pdf

Nissenbaum, Helen. "Privacy as Contextual Integrity." *Washington Law Review*. 79.1 (2004): 101-139.

Norris, Clive. "From Personal to Digital: CCTV, the panopticon, and the technological mediation of suspicion and social control." In *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*. Ed. David Lyon. London: Routledge, 2003: 249-279.

Phillips, P. J., A. Martin, C. L. Wilson, and M. Przybocki. "An Introduction to Evaluating Biometric Systems." *Computer*, 33.2 (2000): 56-63.

Phillips, P. Jonathon, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone. Face Recognition Vendor Test 2002. Arlington: DARPA, 2003.

Phillips, P. J., P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. "Overview of the Face Recognition Grand Challenge." *Proc. IEEE Computer Vision and Pattern Recognition*. June 2005: 947-954.

Phillips, P. J., P. Rauss, and S. Der. *FERET Recognition Algorithm Development and Test Report*. Arlington: US Army Research Laboratory, 1996. Available: http://www.frvt.org/DLs/FERET3.pdf

Phillips, P. Jonathon, W. Todd Scruggs, Alice J. O'Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, and Matthew Sharpe. "FRVT 2006 and ICE 2006 Large-Scale Results." Arlington: National Institute of Standards and Technology, 29 March 2007. Available: http://www.frvt.org/FRVT2006/docs/FRVT2006andICE2006LargeScaleReport.pdf

Schauer, Frederick F. *Profiles, Probabilities, and Stereotypes*. Cambridge: Harvard University Press, 2003

Stanley, J. and B. Steinhardt. "Drawing a Blank: the Failure of Facial Recognition in Tampa, Florida." Washington DC: American Civil Liberties Union, 2002.

Turk, M.A. and A.P. Pentland, "Face Recognition Using Eigenfaces." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Maui, HI, 3-6 June 1991: 586-591.

Turk, MA & AP Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neurosicence*, 3.1 (1991): 71-86

van der Ploeg, Irma. "The illegal body: 'Eurodac' and the politics of biometric identification." Ethics and Information Technology. 1.4 (2004): 295-302.

Wayman, Jim. "NIST test results unveiled." *Biometric Technology Today*. 15.4, (2007): 10-11.

Wayman, Jim. "Facial recognition from e-passports: Australian Customs SmartGate." *ROBUST 2008 Conference on Biometrics*, Honolulu, HI, Nov. 2-5, 2008.

Zhao, W., R. Chellappa, A. Rosenfeld, and J. Phillips. "Face recognition: A literature survey." *ACM Computing Surveys*, 35.4 (2003): 399-458.

Appendix 3: Companies that supply FRT products

Company	Location	Website	FRVT2002	FRVT2006
Acsys Biometrics Corp.	Burlington, Canada	http://www.acsysbiometrics.com	√	
Animetrics, Inc.	Conway, NH	http://www.animetrics.com		√
Asia Software Ltd.	Almaty, Kazakhstan	http://www.asia-soft.com/frs/en/main		
Aurora Computer Services Ltd.	Northampton, United Kingdom	http://www.facerec.com		
Bioscrypt [Acquired by L-1 Identity Solutions in 2008; Bioscrypt acquired A4 Vision in 2007]	Toronto, Canada	http://www.bioscrypt.com		
C-VIS Computer Vision und Automation GmbH [Acquired by Cross Match Technologies]	Palm Beach Gardens, FL	http://www.crossmatch.com	√	
Carnegie Mellon University	Pittsburgh, PA	http://www.ri.cmu.edu/labs/lab_51.html		√
Cognitec Systems GmbH	Dresden, Germany	http://www.cognitec-systems.de	√	√
Cybula Ltd.	York, United Kingdom	http://www.cybula.com		
Diamond Information Systems (DIS)	Beijing, China	http://www.disllc.net		√
FaceKey Corp.	San Antonio, TX	http://www.facekey.com		
FacePrint Global Solutions Inc.	Fresno, CA			
Genex Technologies [Acquired by Technest Holdings, Inc. in 2008]	Bethesda, MD	http://www.genextech.com		
Geometrix, Inc. [Acquired by ALIVE Tech in 2006]	Cumming, GA	http://www.alivesecurity.com		√
Guardia	Gilleleje, Denmark	http://www.guardia.com		√
IconQuest Technology	Atlanta, GA	http://www.iconquesttech.com	√	

L-1 Identity Solutions [Formed in a merger between Identix, Inc. and Viisage in 2006]	Stamford, CT	http://www.identix.com	√	√
NEC	Tokyo, Japan	http://www.nec.com/global/solutions/biometrics/technologies_b02.html		
Neurotechnology [Formerly Neurotechnologija]	Vilnius, Lithuania	http://www.neurotechnology.com		
Neven Vision [Acquired by Google in 2006; Formerly Eyematic Interfaces, Inc.]	Mountain View, CA	http://www.google.com/corporate	√	√
New Jersey Institute of Technology (NJIT)	Newark, NJ	http://www.cs.njit.edu/liu/facial recognition VPlab/index.html		√
Nivis, LLC	Atlanta, GA	http://www.nivis.com		√
Old Dominion University	Norfolk, VA	http://www.lions.odu.edu/org/vlsi/demo/vips.htm		√
OmniPerception Ltd.	Surrey, United Kingdom	http://www.omniperception.com		
Omron	Kyoto, Japan	http://www.omron.com/r_d/coretech/vision		
Panvista Limited	Sunderland, United Kingdom	http://www.panvista.co.uk		√
Peking University, Center for Information Science	Peking, China	http://www.cis.pku.edu.cn/vision/english/vision_1.htm		√
PeopleSpot Inc.	Beijing, China	http://www.peoplespotinc.com/en/index.htm		√
Rafael Armament Development Authority Ltd.	Haifa, Israel	http://www.rafael.co.il		√
RCG	Selangor, Malaysia	http://www.rcg.tv		
SAGEM SA	Paris, France	http://www.sagem-securite.com/eng		√
Samsung Advanced Institute of Technology (SAIT)	Seoul, South Korea	http://www.sait.samsung.com/eng/main.jsp		√
Speed Identity AB	Mediavägen, Sweden	http://www.speed-identity.com		
Tili Technology Limited				√
Toshiba Corporation	Tokyo, Japan	http://www.toshiba.co.jp/worldwide/about/index.html		√
Tsinghua University	Beijing, China	http://www.ee.tsinghua.edu.cn/English2006/index.htm		√

University of Houston	Houston, TX	http://www.cbl.uh.edu/URxD		√
VisionSphere Technologies Inc.	Ottawa, Canada	http://www.visionspheretech.com	√	
Visiphor Corp. [Formerly Imagis Technologies Inc.]	Burnaby, Canada	http://www.visiphor.com	√	
XID Technologies Pte Ltd	Singapore	http://www.xidtech.com		

Endnotes

[1] For those interested in a deeper grasp of technical issues, the works cited in the report should serve as a good point of departure for investigating this active area of science and engineering.

[2] Even those fail, however, as we learn in poignant historical tales like *The Return of Martin Guerre*. See Natalie Davis, *The Return of Martin Guerre*, Cambridge: Harvard University Press, 1983.

[3] The Schiphol Privium system allows passengers priority processing at passport control by using iris scanning. Other benefits are also linked to the system such as priority parking, etc.

[4] See, for instance, Irma van der Ploeg on a discussion of the use of finger printing and some of its stigmatizing consequences: "The illegal body: `Eurodac' and the politics of biometric identification," *Ethics and Information Technology*, 1.4 (2004): 295-302.

[5] Advances in "on the move" systems seem likely to extend the range of iris scanning but unlikely to match the recognition-at-a-distance potential of facial recognition.

[6] It is worth emphasizing that FRS can only recognize a probe image only if the same individual's image is already enrolled in the system's gallery.

[7] In this report we will often refer to the 'quality' of images. When referring to the 'quality' of the image, we mean the degree to which the image conforms to the ISO/IEC 19794-5 standard of best practice and the ANSI/INCITS 385-2004 standard—to be discussed below.

[8] Source is P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002,* Arlington: DARPA, 2003.

[9] Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002,* Arlington: DARPA, 2003.

[10] Source is Duane M. Blackburn, "Face Recognition 101: A Brief Primer," Department of Defense Counterdrug Technology Development Program Office. 07 April 2003. Available: *http://www.frvt.org/DLs/FR101.pdf.* Note that this graph does not comply with ISO/IEC 19795-1 because the size of the database is omitted. This is very important in closed-set evaluations otherwise it makes direct comparison of performance impossible.

[11] N. Furl, J. P. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science* 26 (2002): 797-815.

[12] One might think of these thresholds as level of confidence (as in statistical terms) or tolerance levels (as in the level of risk one is prepared to accept).

[13] This relationship can be expressed in the following equations: True Accept Rate + False Reject Rate = 1 or False Accept Rate + True Reject Rate = 1.

[14] Watch list performance can also be reported in a ROC graph where the ROC plots the trade-off between the *recognition rate* (true positive rate) and the *false alarm rate* (false positive rate).

[15] By 'standard' we mean conforming to some prior standard such as the ISO/ANSI standards to be discussed below.

[16] This section is based on a very useful introduction to face recognition prepared by the National Science and Technology Council, available at http://*www.biometrics.gov/Documents/FaceRec.pdf,*and James Wayman, Nicholas Orlans, Qian Hu, Fred Goodman, Azar Ulrich, and Valorie Valencia, "Technology Assessment for the State of the Art Biometrics Excellence Roadmap," Vol. 2, The MITRE Corporation, 2008, available at *http://www.biometriccoe.gov/SABER/index.htm.*

[17] For a very comprehensive survey of the variety of approaches by leading researchers in the field, see W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, 35.4 (2003): 399-458. Phillips is the National Institute of Standards and Technology FRT evaluator.

[18] Turk, MA & Pentland AP, Face Recognition Using Eigenfaces, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3-6 June 1991, Maui, Hawaii, USA, pp.586-591 and - Turk, MA & AP Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, 3.1 (1991): 71-86

[19] K. Etemad, R. Chellappa, "Discriminant Analysis for Recognition of Human Face Images," *Journal of the Optical Society of America A*, 14. 8 (1997): 1724-1733.

[20] Source is National Science and Technology Council (NCST) Subcommittee on Biometrics. Face Recognition. 7 August 2006. Available: http://www.biometrics.gov/Documents/FaceRec.pdf.

[21] In fact, quality is defined as the factor impacting performance.

[22] Y. Moses, Y. Adini, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19.7 (1997): 721-732.

[23] Kevin W. Bowyer, Kyong Chang, and Patrick Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Computer Vision and Image Understanding*, 101.1 (2006): 1-15.

[24] Ibid. See also P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the Face Recognition Grand Challenge," *Proc. IEEE Computer Vision and Pattern Recognition*, (June 2005): 947-954.

[25] Brett McLindin, "Improving the performance of Two-Dimensional Facial Recognition Systems," Diss. University of South Australia, 2005, Available: *http://www.library.unisa.edu.au/adt-root/public/adt-SUSA-31102005-074958/index.html.*

[26] The Casino Esplanade in Hamburg, Germany, implemented such a system in 2007. See "German casinos secured with facial recognition," *Biometric Technology Today*, 14.11/12 (2006): 12.

[27] Rhonda Bliss, Program Assistant and Victims Advocate for the Colorado DMV Investigations Unit confirmed via e-mail that the Colorado DMV actively utilizes Digimark Facial Recognition Technology.

[28] See Philip E. Agre, "Your Face is Not a Bar Code: Arguments Against Automatic Face Recognition in Public Places" *Whole Earth*, 106 (2001): 74-77; Philip Brey, "Ethical Aspects of Face Recognition Systems in Public Places," *Journal of Information, Communication & Ethics in Society*, 2.:2 (2004): 97-109; Clive Norris, "From Personal to Digital: CCTV, the panopticon, and the technological mediation of suspicion and social control" In *Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, Ed. David Lyon, London: Routledge, 2003: 249-279; W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, 35.4 (2003): 399-458.

[29] P. J. Phillips, A. Martin, C. L. Wilson, and M. Przybocki “An Introduction to Evaluating Biometric Systems,” *Computer*, 33.2 (2000): 56-63.

[30] J. M. Bone and D. M. Blackborn, “Face Recognition at a Chokepoint: Scenario Evaluation Results.” Department of Defense Counterdrug Technology Development Program Office, 14 November 2002.

[31] “BioFace: Comparative Study of Facial Recognition Systems” Bundesamt für Sicherheit in der Informationstechnik, 2003, Available: http://www.igd.fhg.de/igd-a8/en/projects/biometrie/bioface/ BioFaceIIReport.pdf

[32] P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002*, Arlington: DARPA, 2003.

[33] Source is P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002*, Arlington: DARPA, 2003.

[34] Ibid., 2.

[35] Source is P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002*, Arlington: DARPA, 2003.

[36] Michael Brooks, “Face-off,” *New Scientist*, 175.2399 (2002).

[37] R. Gross, J. Shi, and J. F. Cohn, J.F. “Quo vadis Face Recognition?,” (2001), Available: http://dagwood.vsam.ri.cmu.edu/ralph/Publications/QuoVadisFR.pdf.

[38] P. Jonathon Phillips, Patrick Grother, Ross Micheals, Duane M. Blackburn, Elham Tabassi, and Mike Bone, *Face Recognition Vendor Test 2002* Arlington: DARPA, 2003, 21

[39] Source is P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, “Overview of the face recognition grand challenge,” Computer Vision and Pattern Recognition (CVPR), 1 (2005): 947-954.

[40] Ibid.

[41] Report available at http://www.frvt.org/FRVT2006/default.aspx

[42] Source is P. Jonathon Phillips, W. Todd Scruggs, Alice J. O’Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, Matthew Sharpe, “FRVT 2006 and ICE 2006 Large-Scale Results,” Arlington: National Institute of Standards and Technology, 29 March 2007, Available: http://www.frvt.org/FRVT2006/docs/FRVT2006andICE2006LargeScaleReport.pdf

[43] Ibid.

[44] Ibid., 24.

[45] Emphasis added; Jim Wayman, “NIST test results unveiled,” *Biometric Technology Today*, 15.4, (2007): 10-11.

[46] Report available: *http://www.igd.fhg.de/igd-a8/en/projects/biometrie/bioface/BioFaceIIReport.pdf*. There also were subsequent BioFace III and GioFace IV evaluations that were not released. The authors repeatedly tried to secure access to these reports without success.

[47] Bundesamt für Sicherheit in der Informationstechnik, “BioFace: Comparative Study of Facial Recognition Systems” (2003), 7, Available: http://www.igd.fhg.de/igd-a8/en/projects/biometrie/bioface/ BioFaceIIReport.pdf

[48] Note that rank 5 is used in Phase II, making these two phases incommensurate.

[49] “BioFace: Comparative Study of Facial Recognition Systems” Bundesamt für Sicherheit in der Informationstechnik, 2003, 8, Available: http://www.igd.fhg.de/igd-a8/en/projects/biometrie/bioface/ BioFaceIIReport.pdf

[50] Ibid.

[51] Note that rank 10 was used in phase I so results are not strictly commensurate.

[52] Note that the false non-match rate is defined here as not appearing in the top 10 ranked images. This is not a standard definition of false non-match rate.

[53] J. M. Bone and D. M. Blackborn, “Face Recognition at a Chokepoint: Scenario Evaluation Results” Department of Defense Counterdrug Technology Development Program Office, 14 November 2002.

[54] Ibid.

[55] Ibid.

[56] Ibid.

[57] Ibid.

[58] Ibid.

[59] J. Stanley and B. Steinhardt, “Drawing a Blank: the Failure of Facial Recognition in Tampa, Florida,” Washington DC: American Civil Liberties Union, 2002.

[60] Michael Brooks “Face-off,” *New Scientist*, 175.2399 (2002).

[61] J. Meek, (2002) “Robo cop: Some of Britain's 2.5 million CCTV cameras are being hooked up to a facial recognition system designed to identify known criminals. But does it work,” *Guardian*, 13 June 2002.

[62] Wayman, Jim. “Facial recognition from e-passports: Australian Customs SmartGate.” ROBUST 2008 Conference on Biometrics, Honolulu, HI, Nov. 2-5, 2008.

[63] Boggan, Steve. “’Fakeproof’ E-Passport is Cloned in Minutes.” The Times (UK). 6 August 2008. Available: http://www.timesonline.co.uk/tol/news/uk/crime/article4467106.ece. , See also Juels, A., Molnar, D., and Wagner, D. “Security and Privacy Issues in E-passports.” IEEE/CreateNet SecureComm, 2005. Available: *http://eprint.iacr.org/2005/095.pdf.*

[64] Deutsches Bundeskriminalamt. *Gesichtserkennung als Fahndungshilfsmittel. Abschlussbericht*, 2007 (in German), Available: http://www.bka.de/kriminalwissenschaften/fotofahndung/pdf/fotofahndung_abschlussbericht.pdf.

[65] Source is Deutsches Bundeskriminalamt, Gesichtserkennung als Fahndungshilfsmittel, Abschlussbericht, 2007 (in German), Available: http://www.bka.de/kriminalwissenschaften/fotofahndung/pdf/fotofahndung_abschlussbericht.pdf.

[66] Detlef Borchers and Robert W. Smith, “Mixed reactions to the facial-features recognition technology project of the BKA,” *Heise Online*, 16 July 2007, Available: http://www.heise.de/english/newsticker/news/92722.

[67] For a comprehensive list of all the variables that can influence the performance of a FRS in operational settings, see Brett McLindin, “Improving the performance of Two-Dimensional Facial Recognition Systems,” Diss. University of South Australia, 2005, Available: *http://www.library.unisa.edu.au/adt-root/public/adt-SUSA-31102005-074958/index.html.*

[68] See Michael Brooks “Face-off,” *New Scientist*, 175.2399 (2002) and R. Gross, J. Shi, and J. F. Cohn, J.F. “Quo vadis Face Recognition?,” (2001), Available: http://dagwood.vsam.ri.cmu.edu/ralph/Publications/QuoVadisFR.pdf.

[69] A study reported in *Science* also suggests that the simple process of image averaging, where multiple images of a person are merged into an ‘average’ face, can dramatically boost automatic face recognition. See R. Jenkins and A. M. Burton, “100% Accuracy in

Automatic Face Recognition." *Science*, 319.5862, (2008): 435-435.

[70] See M. Husken, M. Brauckmann, S. Gehlen, and C. von der Malsburg, "Strategies and benefits of fusion of 2D and 3D face recognition*," Proc. IEEE Workshop on Face Recognition Grand Challenge Experiments*, San Diego, CA, 20-25 June 2005 and Kevin W. Bowyer, Kyong Chang, and Patrick J. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *Computer Vision and Image Understanding,* 101.1 (2006): 1-15.

[71] Ibid.

[72] Henning Daum, "Influences of Image Disturbances on 2D Face Recognition." *Audio- and Video-Based Biometric Person Authentication*, Ed. Takeo Kanade, Berlin: Springer, 2005, 900-908 and J.R. Beveridge, et al,
Beveridge, J.Ross, Givens, Geof H., Phillips, P. Jonathan, Draper, Bruce A., and Lui, Yui Man. "Focus on Quality: Predicting FRVT 2006 Performance." 8th IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, 17-19 September 2008. Available:
http://www.cs.colostate.edu/~ross/research/papers/yr2008/focusFRVT2006.pdf

[73] For more information, see Zhao et al., "Face Recognition: A Literature Survey" and B. Heisele, P. Ho. J. Wu, T. Poggio, "Face recognition: component-based versus global approaches," *Computer Vision and Image Understanding*, 91.1 (2003): 6-21 and Lu, Xiaoguang. "Image Analysis for Face Recognition." Personal notes, Dept. of Computer Science &. Engineering,. Michigan State University,. Personal notes, May 2003. Available: http://www.face-rec.org/interesting-papers/General/ImAna4FacRcg_lu.pdf

[74] G. Givens, J. R. Beveridge, B. A. Draper, and D. Bolme, "A Statistical Assessment of Subject Factors in the PCA Recognition of Human Faces," *Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop*, 8 (2003): 1-9.

[75] Ibid., 8. See also N. Furl, J. P. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science* 26 (2002): 797-815.

[76] G. Givens, J. Beveridge, B. Draper, P. Grother, and P. Phillips, "How Features of the Human Face Affect Recognition: a Statistical Comparison of Three Face Recognition Algorithms," *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2 (2004): 381–388.

[77] For further information, see L. Introna and D. Wood, "Picturing algorithmic surveillance: the politics of facial recognition systems," *Surveillance and Society*, 2.2/3 (2004): 177-198.

[78] See R. Kemp, N. Towell, and G. Pike, "When seeing should not be believing: photographs, credit cards and fraud," *Applied Cognitive Psychology*, 11.3 (1997): 211-222.

[79] Jim Wayman has suggested that the biggest problem with managing thresholds is the lack of Bayesian priors, which would allow us to go from "What is the probability of such a score given that this person is an impostor?" to "What is the probability that this person is an impostor, given this score?"

[80] Furl, N., J. P. Phillips, and A. J. O'Toole. "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis." *Cognitive Science*. 26 (2002): 797-815.

[81] We are adopting the approach to privacy described in Helen Nissenbaum, "Privacy as Contextual Integrity," *Washington Law Review*, 79.1 (2004): 101-139.

[82] While it is possible that facial recognition data could be sold to companies such as ChoicePoint or DoubleClick, facial recognition data was already shared by several state DMVs with Image Data, LLC- a producer of FRT- to be used in testing their technology TrueID in 1997. They then shared this data with the secret service, who was commissioning TrueID. When it came into the open that this transferral of data was occurring, the practice was put to a halt, although attempts to sue Image Data for damages were unsuccessful. See *http://ca10.washburnlaw.edu/cases/2004/02/02-1007.htm* or *http://epic.org/privacy/imagedata/image_data.html* for more information.

[83] See here for a recent, related debate: Frederick F. Schauer, *Profiles, Probabilities, and Stereotypes*, Cambridge: Harvard University Press, 2003, Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age,* Chicago: University of Chicago Press, 2007, Bernard E. Harcourt, "A Reader's Companion to Against Prediction: A Reply to Ariela Gross, Yoram Margalioth, and Yoav Sapir on Economic Modeling, Selective Incapacitation, Governmentality, and Race," *Law & Social Inquiry*, 33.1 (2008): 265-283.

[84] Simon Garfinkle, "Future Tech," *Discover*, 23.9 (2002): 17-20.

[85] This anecdote has not been independently confirmed, although it has been confirmed that Cadle recounted this story to Simson Garfinkle. As such it should be treated as a hypothetical situation that is plausible enough to be recounted by an industry spokesman as an instance of a false positive identification which was nonetheless quickly rectified. We would suggest that this plausibility enables the report to discuss its implications despite the possibility of its fabrication.

[86] For an excellent review of the debates over autonomy in Constitutional Law, please see: "The Constitutional Right to Anonymity: Free Speech, Disclosure and the Devil," *The Yale Law Journal 70*.7 (1961): 1084-1128.