NISTIR 6965

# FACE RECOGNITION VENDOR TEST 2002

## Evaluation Report

**March 2003**

**P. Jonathon Phillips[1,2], Patrick Grother[2], Ross J. Micheals[2], Duane M. Blackburn[3], Elham Tabassi[2], Mike Bone[4]**

[1]DARPA
3701 North Fairfax Dr.
Arlington, VA 22203

[2]National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899

[3]DoD Counterdrug Technology Development Program Office
17320 Dahlgren Rd, Code B07
Dahlgren, VA 22448

[4]NAVSEA Crane Division
300 Highway 361, Code 4041
Crane, IN 47522

***Sponsors and Supporters:***

Defense Advanced Research Projects Agency
Department of State
Federal Bureau of Investigation
National Institute of Justice
National Institute of Standards and Technology
Transportation Security Administration
ONDCP Counterdrug Technology Assessment Center
United States Customs Service
Department of Energy
Drug Enforcement Administration
Immigration and Naturalization Service
United States Secret Service
Technical Support Working Group
Australian Customs
Canadian Passport Office
United Kingdom Biometric Working Group

**This page intentionally blank.**

# Face Recognition Vendor Test 2002: Evaluation Report

**P. Jonathon Phillips[1,2], Patrick Grother[2], Ross J. Micheals[2], Duane M. Blackburn[3], Elham Tabassi[2], Mike Bone[4]**

[1]DARPA
3701 North Fairfax Dr.
Arlington, VA 22203

[2]National Institute of Standards and Technology
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899

[3]DoD Counterdrug Technology Development Program Office
17320 Dahlgren Rd, Code B07
Dahlgren, VA 22448

[4]NAVSEA Crane Division
300 Highway 361, Code 4041
Crane, IN 47522

## Abstract

*The Face Recognition Vendor Test (FRVT) 2002 is an independently administered technology evaluation of mature face recognition systems. FRVT 2002 provides performance measures for assessing the capability of face recognition systems to meet requirements for large-scale, real-world applications. Ten commercial firms participated in FRVT 2002. FRVT 2002 computed performance statistics on an extremely large data set—121,589 operational facial images of 37,437 individuals. FRVT 2002 1) characterized identification and watch list performance as a function of database size, 2) estimated the variability in performance for different groups of people, 3) characterized performance as a function of elapsed time between enrolled and new images of a person, and 4) investigated the effect of demographics on performance. FRVT 2002 shows that recognition from indoor images has made substantial progress since FRVT 2000. Demographic results show that males are easier to recognize than females and that older people are easier to recognize than younger people. FRVT 2002 also assessed the impact of three new techniques for improving face recognition: three-dimensional morphable models, normalization of similarity scores, and face recognition from video sequences. Results show that three-dimensional morphable models and normalization increase performance and that face recognition from video sequences offers only a limited increase in performance over still images. A new XML-based evaluation protocol was developed for FRVT 2002. This protocol is flexible and supports evaluations of biometrics in general.*

1. Please direct correspondence to Jonathon Phillips at jphillips@darpa.mil or jonathon@nist.gov

## Table of Contents

Abstract ..... 1
1. Introduction ..... 4
2. Background ..... 6
3. FRVT 2002 Evaluation Protocol ..... 9
4. Performance Statistics ..... 11
5. FRVT 2002 Test Design ..... 14
6. Image data sets ..... 15
    6.1 HCInt data set ..... 15
    6.2 MCInt data set ..... 16
7. High Computational Intensity Test Results ..... 17
8. Medium Computational Intensity Test Results ..... 30
9. Analysis and Discussion ..... 33
10. Conclusion ..... 40
Acknowledgements ..... 42
Appendix A ..... 43
    A.1 FRVT 2002 Evaluation Protocol Details ..... 43
    A.2 Statistics Details ..... 45
    A.3 HCInt Addendum ..... 48
    A.4 MCInt Addendum ..... 48
    A.5 Performance on Very Large Populations ..... 50
References ..... 53

## Tables and Figures


Table 1. Measurements of the size of the FERET Sep96 and FRVT evaluations ..... 6

Table 2. Summary of experiments performed in the FERET and FRVT evaluations ..... 7

Table 3. List of FRVT 2002 participants and tests completed ..... 14

Fig. 1. HCInt example images ..... 15

Fig. 2. Indoor and outdoor images from the NIST-NSWC-USF data set ..... 16

Fig. 3. Example of morphed images ..... 17

Fig. 4. Example of UT Dallas videos ..... 17

Fig. 5. Histogram of elapsed time between gallery and probe images of a person ..... 18

Fig. 6. Histogram of age distribution of people in the HCInt large gallery ..... 18

Fig. 7. ROC for HCInt large gallery and probe set ..... 19

Fig. 8. Standard error ellipses for verification performance ..... 20

Fig. 9. Plots showing the difference between normalized and non-normalized verification performance ..... 21

Fig. 10. Identification performance on HCInt large gallery and probe set ..... 22

Fig. 11. Rank 1 identification performance as a function of gallery size ..... 22

Fig. 12. Watch list performance ROC ..... 23

Fig. 13. Watch list performance for Cognitec as a function of rank for eight false alarm rates ..... 24

Fig. 14. Watch list performance as a function of gallery size for a false alarm rate of 0.01 ..... 25

Fig. 15. Rank 1 identification rate for temporal variation study ..... 26

Fig. 16. Verification performance for temporal variation study ..... 26

Fig. 17. CMC showing the difference in performance between male and female probes ..... 27

Fig. 18. ROC showing the difference in performance between male and female probe sets ..... 28

Fig. 19. Difference in verification performance for male and female probe sets ..... 28

Fig. 20. Rank 1 identification performance broken out by age ..... 29

Fig. 21. Interaction between age and sex for rank 1 identification on HCInt large gallery ..... 29

Fig. 22. Comparison of system performance on different categories of frontalprobes ..... 31

Fig. 23. The effect of still versus three-dimensional morphable models ..... 32

Fig. 24. Plot showing still versus video recognition for frontal imagery ..... 33

Fig. 25. Verification performances of face and fingerprint ..... 34

Fig. 26. Rank 1 identification rate as a function of gallery size for face and fingerprint ..... 35

Fig. 27. Interaction between age and sex identification on HCInt large gallery ..... 48

Fig. 28. ID performance of different categories of frontal probes ..... 48

Fig. 29. ID performance of still versus three-dimensional morphable models ..... 49

Fig. 30. ID performance of plot showing still versus video recognition for frontal imagery ..... 49

Fig. 31. Study of identification performance as a function of gallery size ..... 52

## 1. Introduction

The primary objective of the Face Recognition Vendor Test (FRVT) 2002 is to provide performance measures for assessing the capability of automatic face recognition systems to meet real world applications. Achieving this objective required an evaluation that was much larger and broader scale than any previous biometric evaluations. The increase in scale included the number of individuals in the evaluation as well as the detail and depth of analysis performed. This required designing a new biometric evaluation protocol and establishing a new standard for evaluations. This applies not only to face recognition, but biometric evaluations in general.

FRVT 2002 was an independently administered technology evaluation. Ten participants were evaluated under direct supervision of the FRVT 2002 organizers at the U.S. Government facility in Dahlgren, Virginia, in July and August 2002. The participants were tested on sequestered data sets. Sequestering the data sets guaranteed that all the participants were tested on a level field and the systems were tested on their recognition ability on new facial imagery. Digital data sets ensured that all the systems were tested on exactly the same data and performance among the systems could be directly compared.

The heart of FRVT 2002 was the high computational intensity test (HCInt) that measured system performance on 121,589 operational images of 37,437 people. The images were provided from the U.S. Department of State's Mexican non-immigrant visa archive. From this data, real-world performance figures on a very large data set were computed. Performance statistics were computed for verification, identification, and watch list tasks. The performance results on the HCInt show an improvement in the capabilities of the face recognition systems over the last two years. On comparable experiments in FRVT 2000, there has been a 50 percent reduction in error rates (Blackburn et al. 2001; Phillips et al. 2003).

In previous evaluations, performance was broken out by coarse imaging properties: facial images taken on the same day, images taken on different days, effects of pose changes, and effects of lighting changes. As face recognition technology matured from infancy to the first generation of commercial systems, this level of detail spurred research, identified the most promising approaches, and identified interesting research directions. Now, as face recognition technology is being considered for large-scale implementations and new research avenues are being developed, more detailed evaluations are required. FRVT 2002 is the first generation of large-scale face recognition evaluations. The broad scale of FRVT 2002 consists of an experimental design that, for the first time, investigates key imaging and demographic factors affecting performance.

One possible face recognition application is to identify an unknown person from a large database. Effectiveness of face recognition technology for this application depends upon how database size affects performance. Previously, the largest performance evaluation was conducted on a data set of 1,196 people (Phillips, Moon, et al. 2000). The FRVT 2002 measures recognition performance on databases of up to 37,437 people. This provides the first characterization of how database size affects performance.

In real-world applications, systems are required to recognize faces from images taken on different days. From earlier evaluations, we know that as the elapsed time between the enrolled image and a new image of person increases, recognition rates decrease. However, the rate at which performance declines had not been thoroughly investigated. In FRVT 2002, this temporal effect is reported in intervals of 60 days. This provides the first characterization of how elapsed time between image acquisitions affects performance.

As face recognition technology advances, new techniques are being proposed to improve performance. Evaluations provide an opportunity to assess if new techniques improve performance. FRVT 2002 assesses the impact of three new techniques: three-dimensional morphable models, normalization of similarity scores, and face recognition from video sequences. Three-dimensional morphable models correct for pose variations by fitting a three-dimensional model to a non-frontal facial image. The model then transforms the non-frontal facial image to a frontal facial image. The transformed frontal image

is then matched with other frontal images. Normalization is a post-matching, statistical technique for correcting differences in templates among people. Compensating for the statistical variation among people in a database potentially reduces the effect of differences among faces. The majority of face recognition systems perform recognition from still imagery. New techniques are being developed that explicitly incorporate temporal information in video sequences into face recognition algorithms. The theory is that the temporal information in video sequences contains more information than a still image and exploiting this information will improve performance. FRVT 2002 performed experiments to see if these three techniques improved performance. Experiments showed that three-dimensional morphable models and normalization could significantly improve performance. Experiments showed that for the video sequences in FRVT 2002, there was negligible change in performance.

An operational face recognition system will process and recognize different groups of people at different locations. Performance will vary for the different groups of people. To measure the potential variations among groups of people, FRVT 2002 calculates performance for different subsets of the 37,437 people in the Mexican Visa database. This provides an estimate of the variability of performance over different groups of people using a face recognition system.

One application for face recognition technology is to detect people on a watch list. The watch list task has elements of both the verification and identification tasks. In the watch list task, a photo of an unknown person is provided to a system and the system must determine if the unknown person is on a watch list. If a system determines that a person is on the watch list, the system then provides the identity of the unknown person. FRVT 2002 provides the first in-depth examination and analysis of the watch list task for multiple face recognition systems. Experiments measure the effect of watch list size on performance and performance as a function of the number of matches displayed. These results can provide a basis for establishing conditions under which watch lists can be effective in real-world applications.

Face recognition system performance is a function of the imaging conditions and the demographics of the population that will use the system. To date, evaluations of face recognition systems has not considered the effects of demographics on performance. FRVT 2002 examines the effects of sex and age of a person on performance. Experiments show that males are easier to recognize than females and older people are easier to recognize than younger people. Results also show that the difference in performance between males and females decreases with age. Thus the demographics of a data set do affect reported performance. Demographic information needs to be considered when setting specifications for operational systems.

As face recognition technology has matured and is being considered for more applications, the demand for evaluations is increasing. At the same time, the complexity and sophistication of the evaluations is increasing. This results in rising evaluation costs and increases the chance that mistakes will occur. To overcome these challenges, the FRVT 2002 introduced an evaluation protocol that is XML-based and contains a package of standard scoring routines. The XML specifications include formats for test data sets, raw output formats from systems, and performance scores. The scoring routines are designed to work with XML file formats. The FRVT 2002 evaluation protocol, XML specification, and scoring package are general and can be used to evaluate other types of biometrics. The FRVT 2002 evaluation protocol is a solid basis for establishing a standard evaluation protocol.

The complete FRVT 2002 report comes in the volumes: *Evaluation Report, Summary and Overview,* and *Technical Appendices* (all three are available at http://www.frvt.org). The *Evaluation Report* is the document you are reading. It contains a detailed description of the FRVT 2002 procedures, experiments, and results. The *Summary and Overview* briefly presents the key results from the FRVT 2002. The *Technical Appendices* provide supplementary material, detailed documentation of the FRVT 2002 evaluation protocol, and participant description of their systems and response to the FRVT 2002.

## 2. Background

FRVT 2002 follows four previous face recognition technology evaluations – three FERET evaluations (1994, 1995,1996) and FRVT 2000 (Phillips et al. 1996, 1998; Phillips, Moon, et al. 2000; Blackburn et al. 2001; Phillips et al. 2003). The FERET program introduced evaluations to the face recognition community and helped to advance face recognition from its infancy to the prototype system stage. By 2000, face recognition technology had matured from prototype systems to commercial systems. The Face Recognition Vendor Test 2000 (FRVT 2000) measured the capabilities of these systems and their technical progress since the last FERET evaluation. Public interest in face recognition technology had risen significantly by 2002. FRVT 2002 was designed to measure technical progress since 2000, to evaluate performance on real-life, large-scale databases, and to introduce new experiments to better understand face recognition performance.

In the literature, results are reported on three tasks: identification, verification, and watch list. Not every paper or report presents results for all three tasks. Computing performance requires two board sets. The first is the *gallery* which contains the database on individuals known to the system. The second is the *probe set* that contains the images (biometric signature) of an unknown individuals presented to the system for recognition. A *probe* is one signature in a probe set.

Each successive evaluation increased in size, difficulty and complexity, reflecting the maturing of face recognition technology as well as evaluation theory. Table 1 shows the increase in difficulty of the evaluations. Table 1 lists the number of signatures in each evaluation, the number of comparisons between signatures required to complete the evaluation, the evaluation time limit, and the minimum number of comparisons per second required to complete the evaluations with in the time limit. The FERET Aug94 and Mar95 evaluations are not included in the table because their designed was based on a different evaluation protocol. Table 2 summarizes the experiments conducted in the FERET and FRVT evaluations. The experiments are broken into different categories. The first category lists the different types of experiments conducted broken out by task and the maximum number of individuals in an experiment. The basic experiment category lists standard experiments included in face recognition evaluations. For the basic experiments, the gallery consisted of frontal facial images taken indoors. The indoor same day—expression change experiment consisted of probes taken on the same day as the gallery image, but with a change in expression. The indoor same day—illumination change experiments investigated the effects of changes of illumination indoors. In the indoor different day experiments, the probe image was taken on a different day than the gallery image. For the indoor different day—greater than 18 months, the probe image was taken at least 18 months after the gallery image of a person. In the outdoor same day experiments, the probe was taken outdoors on the same day as the gallery image. For the outdoor different day, the probe was taken outdoors on a different than the gallery image. In the pose experiments the probes were non-frontal facial images. In the left or right experiments the face is looking either to the left or right at varying angles. In the up or down experiments, the face is looking either up or down. The detailed analysis experiments look at results in greater depth. The resolution experiments examine the effects of changing the size of the face. The size of the face is changed by digital manipulation. The compression experiments examined the effect on performance of compressing the probe image. The media experiments examined the effect on performance of switching between

**Table 1. Measurements of the size of the FERET Sep96 and FRVT evaluations**

| Evaluation | No. of Signatures | No. of Comparisons | Evaluation Time Limt | Minimum No. of Comparisons per Second |
|---|---|---|---|---|
| **FERET Sep96** | 3,813 | ~14.5 million | 72 hours | 56 |
| **FRVT 2000** | 13,872 | ~192 million | 72 hours | 742 |
| **FRVT 2002 – MCINT†** | 9,612 | ~63 million | 264 hours | 66 |
| **FRVT 2002 - HCINT** | 12,1589 | ~15 billion | 264 hours | 15,555 |

†*Note:* The MCINT portion of FRVT 2002 is the only test in this chart that included "video" signatures. Signatures in all other tests were a single still image.

**Table 2. Summary of experiments performed in the FERET and FRVT evaluations**

| Measurable | FERET Aug94 | FERET Mar95 | FERET Sep96 | FRVT 2000 | FRVT 2002 |
|---|---|---|---|---|---|
| Largest number of individuals in: | | | | | |
| A verification experiment | | | 1,196 | 1,196 | 37,437 |
| An identification experiment | 498 | 831 | 1,196 | 1,196 | 37,437 |
| A watch list experiment | 25 | | | | 3,000 |
| Basic experiment categories | | | | | |
| Indoor same day –expression change | * | * | * | * | * |
| Indoor same day—illumination change | * | | * | * | * |
| Indoor different day | * | * | * | * | * |
| Indoor different day—greater than 18 months | | | * | * | * |
| Outdoor same day | | | | * | * |
| Outdoor different day | | | | | * |
| Pose—left or right | * | | | * | * |
| Pose—up or down | | | | | * |
| Detailed analysis | | | | | |
| Resolution of face | * | | | * | |
| Image compression | | | | * | |
| Media | | | | * | |
| Distance of face from camera | | | | * | |
| Standard error ellipses | | | | | * |
| Id. Performance as a function of gallery size | | | | | * |
| Watch list performance as a function of gallery size | | | | | * |
| Watch list performance as a function of rank | | | | | * |
| Technologies evaluated | | | | | |
| 3D morphable models | | | | | * |
| Normalization | | | | | * |
| Video | | | | | * |
| Demographic factors | | | | | |
| Sex | | | | | * |
| Age | | | | | * |
| Interaction between sex and age | | | | | * |

film and digital facial images. The distance experiment looked at the effects of distance from the camera on performance. Unlike the resolution experiment, these images were not digital manipulated. The error ellipse experiments looked the effects of different galleries on performance. The gallery size experiments looked the effects of increasing the gallery size on identification and watch list performance. The watch list rank experiment looked at the effect of increasing the number matches examined on performance. The technology experiments investigated three new techniques: three-dimensional morphable models, normalization, and video-based recognition algorithms. The demographic experiments examined the effect of sex and age on performance.

Biometrics testing is an active area. Unfortunately, most results are not publicly released. Reasons for not releasing results include sensitivity of performance figures for operational systems and proprietary rights associated with the performance data. Aside from security concerns, it is advantageous to the biometrics community if results are released. Releasing the results allows for peer assessment of the evaluation procedure and results, makes the biometrics community aware of the latest results, and informs the biometric community of the latest advances in testing

methodology. If a biometric evaluation is designed correctly, the results should be reproducible. This means that the results reported in one evaluation should be observed in evaluations conducted by different groups. For example, in FRVT 2002, we found that males are easier to identify than females. If this was a general property of face recognition algorithms, then other face recognition evaluations should also observe this phenomenon. If other groups do not observe this phenomenon, there could a problem in one of the groups' test methodologies or the phenomenon could just be a property of the data set in the evaluation or systems tested. If subsequent evaluations replicate the phenomenon, this provides further evidence that the phenomenon is a general property—in this example, that male faces are easier to identify than female faces. Public release of results helps the community to differentiate between general properties of a biometric and properties of a particular evaluation. As the testing community matures, this will evolve into a set of standard evaluation procedures.

Other publicly available evaluations have also significantly contributed to our understanding of biometric system performance. The NIST speaker recognition evaluations are annual technology evaluations that measure verification performance (Martin and Przybocki 2000). The overall goals of the evaluations have always been to drive the technology forward, to measure the state-of-the-art, and to find the most promising algorithmic approaches. As in the FERET and FRVT programs, the NIST speaker recognition evaluations have increased in size and scope since the first evaluation in 1996.

The Fingerprint Verification Competition (FVC) 2000 and 2002 are technology evaluations that measure verification performance of single-finger fingerprint devices (Maio et al. 2002a, 2002b). These evaluations were the first common benchmarks for fingerprint developers, thus allowing them to unambiguously compare their algorithms.

The *UK Biometrics Working Group's Biometric Test Programme Report* is a scenario evaluation report that compared six different biometrics (Mansfield et al. 2001). Representative face, fingerprint, hand geometry, iris, vein and voice recognition systems were tested for verification performance in a normal office environment with cooperative, non-habituated users. The objectives of the test program were to show the level of performance attainable by a selection of biometric systems, to determine the feasibility of demonstrating satisfactory performance through testing, to encourage more testing to be sponsored, and to promote methodologies contributing to the improvement of biometric testing. The report is groundbreaking in that it is the first open-source evaluation that directly compares performance of different biometrics for the same application. "Best practices in testing and reporting performance of biometric devices" was developed in part based on the lessons learned from performing the Biometrics Working Group's biometric test program (Mansfield and Wayman 2002).

San Jose State University has been evaluating biometric systems since 1994, and was the Biometric Consortium's National Biometric Test Center from 1997-2000. The Test Center explored a number of essential questions relating to the science underpinning biometric technologies. The Collected Works document from these endeavors is publicly available and contains evaluation results from the INSPASS Hand Geometry System, the Philippine AFIS System and numerous small-scale evaluations (Wayman 2000).

Bone and Blackburn (2002) performed a scenario on the watch list task and the effect of gallery size on verification performance. This study is the first publicly available evaluation to report in-depth on the watch list task.

## 3. FRVT 2002 Evaluation Protocol

A design principle and testing protocol describes how evaluations are designed and conducted. Design principles outline the core philosophy and guiding beliefs in designing an evaluation; the evaluation protocol provides the implementation details.

The FRVT 2002 evaluation protocol is based on the September 1996 evaluation protocol (Phillips, Moon, et al. 2000). The protocol added general biometric signatures, normalization of similarity scores, and an XML-based specification. The XML-based specification is extensible to other biometrics and is being used for fingerprint recognition evaluation.

The design of FRVT 2002, along with the FERET evaluations and FRVT 2000, followed the precepts for biometrics evaluations articulated in Phillips, Martin, et al. (2000). Succinctly stated, the precepts are:

1. Evaluations are designed and administered by groups that are independent of algorithm developers and vendors being tested.
2. Test data is sequestered and not seen by the participants prior to an evaluation.
3. The evaluation test design, protocol, and methodology are published.
4. Performance results are spread in a manner that allows for meaningful differences among the participants.

Points 1 and 2 ensure fairness in an evaluation. Point 1 provides assurance that the test is not designed to favor one participant over another. Independent evaluations help enforce points 2 and 4. In addition, point 2 ensures that systems are evaluated on their ability to generalize performance to new sets of faces, not the ability of the system to be tuned to a particular set of faces. When judging and interpreting results, it is necessary to understand the conditions under which algorithms and systems are tested. These conditions are described in the evaluation test design, protocol and methodology. Tests are administered using an evaluation protocol that identifies the mechanics of the tests and the manner in which the tests will be scored. In face recognition, the protocol states the number of images of each person in the test, how the output from the algorithm is recorded, and how the performance results are reported. Publishing the evaluation protocol, as recommended in point 3, lets the readers of published results understand how the results were computed.

Point 4 addresses the *'three bears'* problem. Phillips, Moon, et al. (2000) first articulated the 'three bears' problem in designing face recognition evaluations. The 'three bears' problem sets guiding principles for designing an evaluation of the right level of difficulty. If all the scores for all algorithms are too high and within the same error margin, then one cannot distinguish among the algorithms tested. In addition, if the scores are too high in an evaluation, then that is an indication that the evaluation was in reality an exercise in 'tuning' algorithm parameters. If the scores are too low, then it is not possible to determine what problems have been solved. The goal in designing an evaluation is to have variation among the scores. There are two sorts of variation. The first type is variation among the experiments in an evaluation. Most evaluations consist of a set of experiments, where each experiment reports performance on different problems in face recognition. For example, experiments might look at changes in lighting or subject pose of a face. The second type of variation is among algorithms for each experiment. The variation in performance among the experiments lets one know which problems are currently sufficiently solved for consideration in field testing, which problems are research problems, and which problems are beyond the capabilities of the field. The variation among algorithm performance lets one know which techniques are best for a particular experiment. If all the scores for all algorithms across all experiments are virtually the same, then one cannot distinguish among the algorithms.

The key elements that ease adoption of points three and four can be incorporated into the evaluation protocol. For FRVT 2002, this was the FRVT 2002 evaluation protocol. This evaluation

protocol was designed to assess the state of the art, advance the state of the art, and point to future directions of research. The ability to accomplish these three goals simultaneously was through a protocol whose framework allows for the computation of performance statistics for multiple galleries and probe sets. This allows for the FRVT 2002 evaluation protocol to solve the 'three bears' problem by including galleries and probe sets of different difficulties into the evaluation. This produces a comprehensive set of performance statistics that assess the state of the art, progress in face recognition, and point to future directions of research. The use of an XML-based specification allows for this evaluation protocol to become a formal standard for biometric evaluation.

The solution to the 'three bears' problem lies in the selection of images used in the evaluation. The characteristics and quality of the images are major factors in determining the difficulty of the problem being evaluated. For example, the location of the face in an image can affect problem difficulty. The problem is much easier if a face must be in the center of image compared to the case where a face can be located anywhere within the image. In FRVT 2002 data sets, variability was introduced by the size of the database, inclusion of images taken at different dates and both outdoor and indoor locations. This resulted in changes in lighting, scale, and background.

The testing protocol is based upon a set of design principles. The design principles directly relate the evaluation to the face recognition problem being evaluated. In particular, for FERET and FRVT 2000, the driving applications were searching large databases and access control. Stating the design principles allows one to assess how appropriate the FERET tests and FRVT 2000 are for a particular face recognition algorithm. Also, design principles assist in determining if an evaluation methodology for testing algorithm(s) for a particular application is appropriate.

The FRVT 2002 evaluation protocol consists of two parts. The first is the rules for conducting an evaluation, and the second is the format of the results that allow for scoring. The specific file format specifications are XML-based. Formal details can be found in *"Face Recognition Vendor Test 2002: The Technical Appendices."*

The input to an algorithm or system being evaluated is two sets of signatures, the *target* and *query sets*. Galleries and probe sets are constructed from the target and query sets respectively. The output from an algorithm is a *similarity measure* between all pairs of images from the target and query sets. A similarity measure is a numerical measure of how similar two faces are. Performance statistics are computed from the similarity measures. A complete set of similarity scores between all pairs of signatures from the target and query set is referred to as a *similarity matrix*. The first rule in the FRVT 2002 evaluation protocol is that a complete similarity matrix must be computed. This rule guarantees that performance statistics can be computed for all algorithms.

To be able to compute performance for multiple galleries and probe sets requires that multiple signatures of a person are placed in both the target and query sets. This leads to the second rule: Each signature in the target and query sets is considered to contain a unique face. In practice, this rule is enforced by giving every signature in the target and query sets a unique random identifier.

The third rule is that training is completed prior to the start of an evaluation. This forces each algorithm to have a general representation for faces, not a representation tuned to a specific gallery. Also, if training were specific to a gallery, it would not be possible to construct multiple galleries and probe sets from a single run. An algorithm would have to be retrained and the evaluation rerun for each gallery.

Using target and query sets allows us to compute performance for different categories of images. Possible probe categories include (1) gallery and probe images taken on the same day, (2) duplicates taken within a week of the gallery image, and (3) duplicates where the time between the images is at least one year. This is illustrated in the following example. A target and query set consists of the same set of facial images. Eight images of each face are taken. Each face is taken

both indoors and outdoors, with two different facial expressions on two different days. From these target and query sets, one can measure the effects of indoor versus outdoor illumination by constructing a gallery of indoor images and a probe set of outdoor images, both consisting of neutral expressions taken on the first day. Construction of similar galleries and probe sets would allow one to test the effects of temporal or expression changes. The effect of covariates such as age and sex of a person can also be measured. It is the ability to construct galleries from the target set and probe sets from the query set that allows the FRVT 2002 protocol to perform a detailed analysis.

The FRVT 2002 evaluation protocol allows for the computation of performance statistics for identification, verification, and watch list tasks. The protocol is sufficiently flexible that one can estimate performance using sub-sampling and re-sampling techniques. For example, galleries of varying sizes are created to measure the effects of gallery size on performance. To estimate the variability of performance, multiple galleries are created.

## 4. Performance Statistics

In FRVT 2002, performance results are reported for identification, verification, and watch list tasks. In addition to computing classical performance statistics, new statistical methods were developed to estimate variation in performance over multiple galleries and to explore the effect of covariates on performance. This section presents an overview of the performance statistics that describe the FRVT 2002 experiments. Mathematical details are provided in Appendix A.2.

The *identification* task allows one to ask how good an algorithm is at identifying a probe image. The question is not always "is the top match correct?," but rather "is the correct answer in the top $k$ matches?" The identification task is modeled after real-world law enforcement applications in which a large electronic mug-book is searched to identify suspects. In identification scoring, the similarity scores between the probe and all the gallery images are sorted. A higher similarity score means there is a greater resemblance between two faces. In identification, every probe has a match with one of the gallery images—this is known as the correct match (identification is a closed universe evaluation). A probe is correctly identified if the correct match has the highest similarity score. More generally, a probe has rank $k$ if the correct match is the $k^{th}$ largest similarity score. For example, a probe has rank 5 if the correct match is the fifth largest similarity score. A probe is correctly identified if it has rank 1. Rank can vary between 1 and the size of the gallery. The identification rate at rank $k$ is the fraction of probes that have rank $k$ or higher. Identification performance is plotted on a cumulative match characteristic (CMC). Figure 10 from Section 7 shows a CMC. The horizontal axis of the graph is rank on a logarithmic scale. A logarithmic scale was chosen to emphasize the lower ranks. Most applications are based on performance at lower ranks. The vertical axis is the identification rate. Identification is also known as 'one to many' matching because a probe is compared to an entire gallery.

In FRVT 2002, we assume the following operational model for biometric *verification* or *authentication* systems. In a typical verification task using, a subject presents his biometric signature to the system and claims to be a person in the system's gallery. The presented biometric signature is a probe. The system then compares the probe with the stored signature of the subject in the gallery. The comparison produces a similarity score. The system accepts the identity claim if the similarity score is greater than the system's operating threshold. The operational threshold is determined by the applications, and different applications will have different operational thresholds. Otherwise, the system rejects the claim. In analyzing verification performance, two cases need to be considered. The first case concerns when a person makes a legitimate identity claim—the person is who he claims to be. The performance measure for this case is the correct verification rate. The second case is when the claimed identity is incorrect—the person is not who he claims to be. The performance measure for the second case is the false accept rate. The ideal system would have a verification rate of 1.0 and a false accept rate of 0.0. All legitimate claims are

accepted and all false claims are rejected. However, there is a trade-off between the verification and false accept rates in real-world systems. By changing the operating threshold, the verification and false accept rate change. A higher operating threshold results in lowering the false accept rate and the verification rate.

Verification and false accept rates characterize *verification* performance. Both these performance rates cannot be maximized simultaneously; there is a trade-off between them. Verification performance is reported by showing the trade-off between the verification and false accept rates on a receiver operator characteristic (ROC). An example of a ROC is shown at figure 7, Section 7. The horizontal axis is the false accept rate (scaled logarithmically). A logarithmic scale is used because operating points of interest have low false accept rates and a logarithmic scale emphasizes low false accept rates. The vertical axis is the verification rate. Verification is also known as 'one to one' matching because a decision is made by comparing one probe signature with one gallery signature. Because probes presented to a system are not necessarily in the gallery, verification is considered an open universe task.

In FRVT 2002, we examine how *verification* performance varies under two conditions. The first is how performance varies with different galleries. This models the performance of a system that might be installed at different locations. The second is how performance varies for different classes of probes. For example, what is the difference in performance for male and female probes? Each combination of the gallery and probe sets generates a different ROC. To study the variation, it is necessary to combine results over a set of ROCs. One method of combining results is to measure the variation of the verification rate for each false alarm rate. This models the situation where one can readjust the operating threshold for each gallery or probe set. For many applications, this is not feasible or desirable. However, this is an appropriate technique for combining ROCs from multiple systems because it is not possible to set uniform operating thresholds across different systems. For the same system, it is possible to set one operating threshold across all galleries and probe sets. Using this 'base-operating threshold,' one computes the verification and false accept rate for each gallery and probe set. The resulting verification and false alarm rates will vary across the different galleries and probe sets. This method for computing variance in performance models the situation in which the operating threshold is set once for an application. Setting the base-operating threshold can be based upon an overall desired performance level for the population that will use the system. In FRVT 2002, the base-operating threshold is set based upon the system performance on an aggregate population. The base-operating threshold corresponds to a specific false accept rate on the aggregate population—this is referred to as the *nominal false accept rate*.

The last task examined was the *watch list*. The watch list task is a generalization of *verification* and *identification*. Like verification, watch list is an open universe task. In fact, verification and identification are special cases of the watch list task. In the watch list task, a system determines if a probe corresponds to a person on the watch list, and then identifies the person in the probe. When a probe is provided to the system, it is compared to the entire gallery. The gallery is also known as the watch list. As in identification, the similarity score between the probe and the gallery are sorted. If the top match is above an operating threshold, an alarm is declared and the estimated identity is reported. This is the verification portion of the watch list task. As in verification, there are two performance statistics: detection and identification rate, and false alarm rate. We will first look at the case where the identity of a probe is someone in the gallery (in the watch list). A probe is detected and identified if the probe is correctly identified and the correct match score is above the operating threshold. The detection and identification rate is the fraction of probes of people in the gallery who are detected and identified. In the second case, a probe is not of someone in the gallery. This type of probe is also referred to as an imposter. A false alarm occurs when the top match score for an imposter is above the operating threshold. Watch list performance is reported on a ROC, where the $x$-axis is the false alarm rate and the $y$-axis is the detection and identification rate (see figure 12 in Section 7). A watch list ROC plots the trade-off between the detection and identification rate, and false alarm rate.

In the general case, the system examines the top $k$ matches with the gallery. A probe of a person in the gallery is detected and identified at rank $k$ if the probe is of rank $k$ or less and the correct match is above the operating threshold. For imposter probes, a false alarm at rank $k$ occurs if any of the top $k$ matches are above the operating threshold. The full watch list performance is plotted along three axes: detection and identification rate, false alarm rate, and rank. The performance of a system is represented as a surface in the three-dimensional parameter space. When the size of the gallery is one, the watch list reduces to verification. Identification is the special case when the false alarm rate is 1.0. Formally, the identification CMC is the curve that results from intersecting the watch list performance surface with the false alarm rate equals 1.0 plane.

FRVT 2002 allowed participants to submit normalization procedures. Normalization is a post processing function that adjusts similarity scores based on a specific gallery. FRVT 2002 allowed a different normalization function for each task: identification, verification, and watch list. The input to a normalization routine is the set of all similarity scores between a probe and a gallery. The output is a new set of normalized similarity scores between a probe and a gallery. The normalization function attempts to adjust for variations among probes and to emphasize differences among the gallery signatures. If the gallery changes, then similarity scores need to be normalized again. This has implications for scoring techniques that require performance on multiple galleries. Traditionally, verification has been referred to as 'one to one' matching. This is because, in verification, one probe is matched with one gallery signature. Normalization requires that a probe be compared with a gallery. When normalization is applied, is verification still 'one to one' matching?

All the results in this report are performance statistics estimated from participant report similarity scores. As with all estimation procedures there is an uncertainty associated with the measurement. In face recognition and biometrics, it is an active area of research to develop techniques to measure the uncertainty of an estimator (Bolle et al. 2000; Beveridge et al. 2001; Micheals and Boult 2001; Moon and Phillips 2001). Unfortunately, there are no accepted methods for computing confidence intervals in the biometric community. Therefore, we do not place confidence intervals on performance statistics reported. Another measure of uncertainty is the variance of an estimator. For verification performance on the HCInt we report the observed variance of the verification performance estimators. In this case we were able to report the observed variance because of the large size of the HCInt data set.

Mansfield and Wayman (2002) describe four performance statistics that are not reported in FRVT 2002. These statistics provide an additional level of detail in performance and are mainly designed for scenario and fingerprint evaluations. FRVT 2002 views the systems being tested as black boxes and reports performance for the complete system. Two of the statistics are *failure to enroll* and *failure to acquire*. The analogy of *failure to enroll* in FRVT 2002 is that a template of sufficient quality could not be extracted for a gallery signature. In FRVT 2002, if a system could not extract a template for a gallery signature, then most likely, all matches with that signature would produce low similarity scores. This would be reflected in the performance statistics. The FRVT 2002 design is transparent to these types of internal errors that may be handled differently by each participant system. The analogy for *failure to acquire* is that a template of sufficient quality could not be extracted from a probe signature. This is handled in the same manner as a *failure to enroll*. *Binning error rate* and *penetration rate* are system-level statistics of the success of limiting the expense of one to many searches by suitably partitioning the gallery. FRVT 2002 is transparent to partitioning; any binning undertaken by a system is internal and not explicitly reported.

## 5. FRVT 2002 Test Design

FRVT 2002 was designed to allow participation by as many face recognition research groups and companies as possible. FRVT 2002 consisted of two sub-tests. These were the high computational intensity (HCInt) and medium computational intensity (MCInt) tests. The two sub-tests were designed to encourage broad participation in the evaluation. The HCInt was designed to evaluate the performance of state-of-the-art systems on extremely challenging real-world problems. The MCInt was designed to provide an understanding of a participant's capability to perform face recognition tasks with several different formats of imagery (still and video) under varying conditions. The MCInt was also designed to help identify promising new face recognition technologies not identified in the HCInt.

FRVT 2002 was announced on 25 April 2002 in Federal Business Opportunities, the Biometric Consortium list server (listserv), and in e-mails to members of the face recognition and biometrics communities. FRVT 2002 was open to all developers and providers of core face recognition technology. This included academia, research laboratories, and commercial companies. There were neither fees nor compensation for participating. Participants were required to provide all hardware, software, and personnel necessary to complete the test. Participants could take the HCInt, MCInt, or both. The participants and the tests they took are provided in table 3[2]. FRVT 2002 was administered at the U.S. Naval base at Dahlgren, Virginia between 10 July and 9 August 2002.

**Table 3. List of FRVT 2002 participants and tests completed**

| Participant | MCInt | HCInt |
|---|---|---|
| AcSys Biometrics Corp | X | |
| Cognitec Systems GmbH | X | X |
| C-VIS Computer Vision und Automation GmbH | X | X |
| Dream Mirh Co., Ltd | X | X |
| Eyematic Interfaces Inc. | X | X |
| Iconquest | X | |
| Identix | X | X |
| Imagis Technologies Inc. | X | X |
| Viisage Technology | X | X |
| VisionSphere Technologies Inc. | X | X |

All images and video sequences in FRVT 2002 were sequestered prior to the test and had not been seen by any participant. Testing on sequestered data has a number of advantages. First it provides a level playing field. Second, it ensures that systems are evaluated on the general face recognition task, not the ability to tune a system to a particular data set. FRVT 2002 was administered under strict U.S. Government supervision. All tuning and adjustment to participants' systems had to be complete prior to arrival at the test site.

The HCInt contained one target and query set. The target and query set were the same and contained 121,589 full frontal facial images (see Section 6.1 for details). The HCInt was fully automatic. In a fully automatic test, the input to the system is the facial image or video sequence. The system must automatically locate the face in the image and find any facial features that are required by the system. The HCInt had to be performed on the equivalent of three high-end workstations. Technical specifications for each participant's HCInt and MCInt systems are provided in the *Face Recognition Vendor Test 2002: Technical Appendices*. Participants were given 264 hours to complete the test and output the complete set of similarity files. The complete set of similarity files consisted of approximately 15 billion similarity scores.

The MCInt consisted of two sub-tests: *still* and *video*. The still and video sub-tests each had a target and query set. In both sub-tests the target and query set were the same. Both sub-tests were fully automatic. Participants had to complete both sub-tests in 264 hours on one high-end

2. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, or any other FRVT 2002 sponsor or supporter.

workstation. The output from the MCInt consisted of approximately 56 million similarity scores. The still target and query set consisted of 7,722 digital facial images. The types of facial images varied—details can be found in Section 6.2. The video target and query set consisted of 1,890 signatures. The signatures consisted of 882 still facial images and 1008 facial video sequences.

## 6. Image Data Sets

This section describes the data sets used in FRVT 2002. Some common aspects of all images used in FRVT 2002 are that they contained the face of exactly one individual, that the face (and neck and part of the chest) was the only foreground object, and that the image was acquired in the visible spectrum in color. All images provided to FRVT 2002 participants were in standard JPEG format.

### 6.1 HCINT DATA SET

The HCInt data set is a subset of a much larger collection provided by the Visa Services Directorate, Bureau of Consular Affairs of the U.S. Department of State. That collection consists of approximately 6.8 million images from about 6.1 million visa applicants collected from 1996 to 2002. The HCInt data set consisted of 121,589 images from the database. HCInt contained at least three images of each of the 37,437 subjects in the data set.

The images are of good quality and were gathered in a consistent manner. They were collected at U.S. consular offices using standard issue digital imaging apparatus whose specification remained fixed over the collection period. The result is a set of well-posed (i.e., frontal to within 10 degrees) images of cooperative subjects usually with a mouth-closed neutral facial expression. The subject usually occupies a quarter of the image area. The top of the subject's shoulders is almost always visible. The background is universally uniform, with a white background (in most cases). The names of the individuals were encoded to protect the privacy of the subjects. Year of birth was provided. Due to privacy considerations, representative images of the actual data set are shown in figure 1.

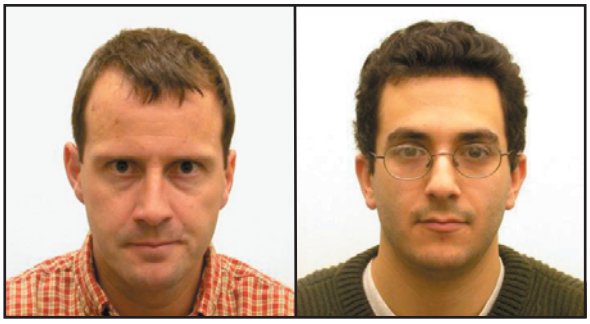

**Fig. 1. Images included here are reasonable representations of those used in the FRVT 2002 High Computational Intensity test.**

One concern with operational databases is the integrity of the identities associated with the images. There are two errors that can occur. First, two different persons can have the same identify. Since the identifier is based on a hash of the person's name and other identifying information, it is very unlikely that two people will have the same identifier. Second, a person has two identifiers. This usually occurs because of clerical errors. Since each individual in the data set has at least three images, the same clerical error would have had to occur three times. Thus, the inclusion of three images of each person in the data set makes it very unlikely that a person in the HCInt data set will have two identifiers.

### 6.2 MCINT DATA SET

In contrast to the HCInt image set, the MCInt data set is composed of a heterogeneous set of still images and video sequences of subjects in a variety of poses, activities and illumination conditions. The images originate from two sources. The first is the still facial image data set collected the National Institute of Standards and Technology (NIST), Naval Surface Warfare Center (NSWC, Dahlgren), and the University of South Florida (USF) between 1999 and 2002. The second, from The University of Texas at Dallas, consists of video and stills taken in 2001.

The NIST-NSWC-USF data set is comprised of images taken indoors and outdoors. Although the images were taken over more than three years at three sites, they were acquired in almost identical circumstances such that their geometric and photometric characteristics are approximately the same. While the imagery itself has consistent properties, the subject populations of the three sites differ. The USF segment consists of subject population of diverse ethnicity that is distinct from the older, predominantly Caucasian mix at NIST and NSWC. The images in figure 2 are examples of the gallery and probe set images used for testing changes in expression, overhead lighting and outdoor lighting, and elapsed time. The outdoor stills are characterized by changing background and directional sunlight illumination.

**Fig. 2. Indoor and outdoor images from the NIST-NSWC-USF data set. The top row contains images taken indoors and the bottom contains outdoor images taken on the same day as the indoor images.**

The MCInt image set included partitions to test the efficacy of using 3D morphable models as a preprocessing step in the recognition of non-frontal images (Blanz and Vetter 1999; Blanz et al. 2002). The concept is that a non-frontal image is fitted to a 3D model of human heads, and then the morphable model generates a frontal view based on the non-frontal image. In the MCInt, morphed images were generated by the algorithm of Blanz et al. (2002)(See Technical Appendice O).

The University of Texas at Dallas data set was created for use in human memory experiments. Close-range video clips and static images were taken of more than a hundred individuals on at least two different occasions to make duplicate sets separated by one week to six months. Included in the MCInt was a 63-person subset of this data set of subjects that appeared on two occasions.

*Still Images*. Nine high quality still images of each individual were taken from equally spaced viewpoints spanning angles from left profile to right profile.

*Exploration Video Sequences*. A video was taken of each individual's face moving through the nine facial poses used for the still images.

*Facial Speech Videos*. Two video clips were taken of individuals speaking, first in a neutral way, then in an animated way. Figure 4 shows two examples.

**Fig. 3. The top row shows original images. The subject, illuminated by one light source, is looking left and right at about 45 degrees, straight at the camera, and up and down at about 30 degrees. The second row shows the corresponding frontal reconstructions from the 3D morphable model. The center column shows the effect of fitting the morphable model to a frontal image.**

**Fig. 4. The rows show selected frames from examples of the UT Dallas "facial speech" videos lasting 150 frames. The two rows show the subject gathered on different occasions.**

## 7. High computational intensity test results

The HCInt target set contains 121,589 images from 37,437 individuals, with at least three images of each person. The query and target set are identical. The design of the HCInt experiments is based on the HCInt large gallery and HCInt large probe set. The large gallery consists of the earliest dated image for the 37,437 individuals in the target set. The large probe set consists of two images of each person in the gallery. The first image is the most recent image. This ensures that performance can be computed for the greatest possible elapsed time for each person in the HCInt data set. The second image of each person placed in the probe set is that with a median time between the gallery and all images of a person. For people with three images in the query set, the oldest image is placed in the gallery and the two most recent images are placed in the probe set. A histogram of the distribution of elapsed time between probe and gallery images is given in figure 5. The number of probes in each bin is broken out by sex.

The HCInt large gallery contains 18,468 males and 18,878 females (there were 91 individuals where the sex could not be determined). The numbers in figure 5 and figure 6 do not always add up. In figure 5 the histogram is terminated at 1,140 days. Some of the probes show the elapsed

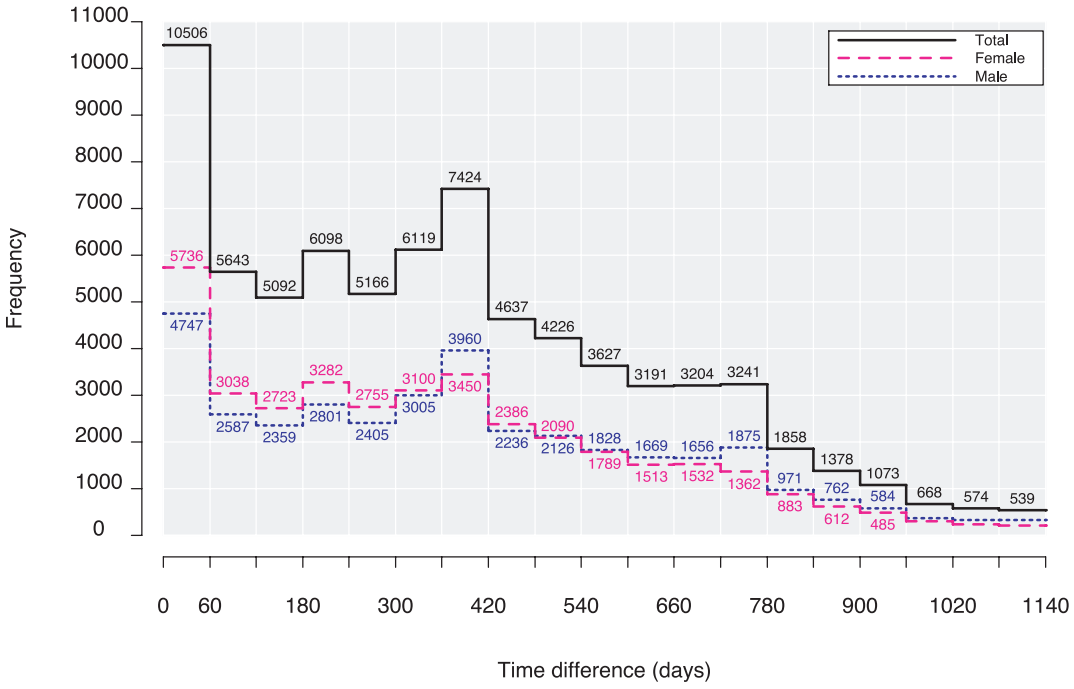


**Fig. 5. Histogram of elapsed time between gallery and probe images of a person. The time intervals are divided into 60-day bins. The number of probes in each bin is broken out by sex.**

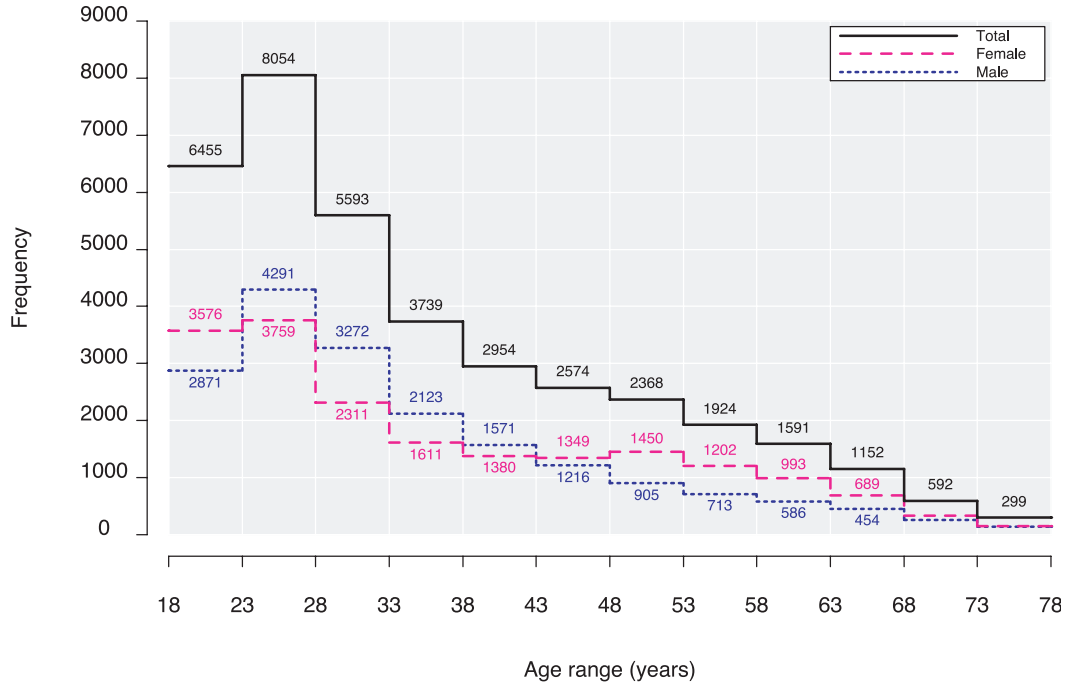


**Fig. 6. Histogram of age distribution of people in the HCInt large gallery. The number of people in each bin is broken out by sex.**

time is greater than 1,140 days, however, the number is sufficiently small so performance was not calculated. The people in the HCInt data set were divided into 12 age categories. Each age category is a five-year bin, with the first bin containing 18 to 22-year olds. The age of a person is set at the age when the gallery image was taken. A histogram documenting the distribution of elapsed time between probe and gallery images is shown in figure 6. The number of people in each bin is broken out by sex. The numbers do not add up to 37,437. People older than 78 were not included in the histogram. This is because there are too few of them to justify experimentation.

Additional galleries and probe sets are derived from the large gallery and probe set. The exact composition depends on the experiment run. Unless otherwise stated, if the image of a person is placed in the gallery, then the two images of that person in the large probe set are placed in the new probe set.

One of the new features introduced in FRVT 2002 is measuring the variance in the statistics computed to determine how performance changes if the people in the gallery are different. Computing the variance requires computing performance statistics on multiple galleries and probe sets. To perform the variance experiments, a set of *twelve HCInt small galleries* were generated from the large HCInt gallery. Each gallery consists of 3,000 individuals. The twelve galleries are disjoint. There are twelve corresponding small probe sets. A small probe sets consists of the two probe images in the large probe set for each person in the gallery. Thus, each small probe set consists of 6,000 images from 3,000 individuals; each probe is an image of a person in the gallery.

The first experiment is the *HCInt large gallery verification* experiment. Figure 7 shows verification performance on a ROC for the eight participants. The false accept rate axis is on a logarithmic scale. The verification performance was computed by the round-robin technique. The false alarm rate was computed from probes in the large probe set; i.e., the match and non-match distributions were both generated from the large probe set.

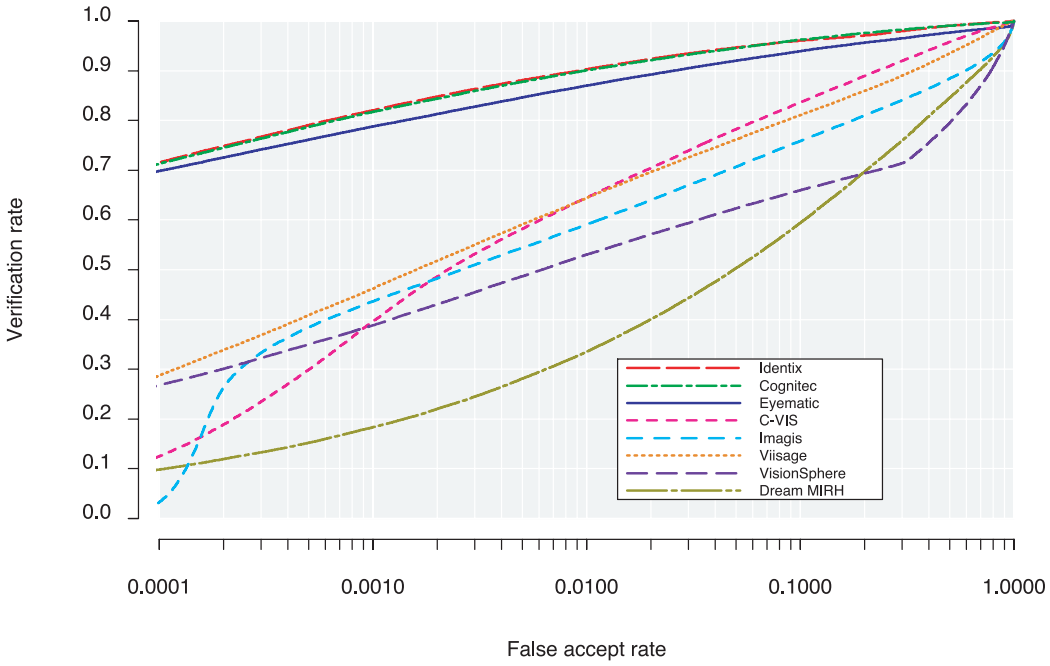


**Fig. 7. Verification performance ROC for eight participants on the HCInt large gallery and probe set. The gallery consisted of 37,437 individuals with one image per person, and the probe set consisted of 74,874 probes with two images per person.**

The results in figure 7 report verification performance for a single large gallery. The results do not address the important question of how performance varies if the people in the gallery are different. To measure this variation, verification performance was computed for the twelve HCInt small galleries. The false alarm rate was computed from 6,000 true imposters (two images of 3,000 individuals). The results for this experiment are presented in figure 8 for Cognitec, Eyematic, and Identix. The centerline is the aggregate performance for the twelve galleries. For selected operating points, performance was computed for the twelve small galleries and probe sets. For each of the twelve galleries, verification rates and false accept rates were computed. The false accept rate was computed with true imposters. Thus, at each operating threshold, there are twelve pairs of verification and false accept rates. A standard error ellipse was computed for each set of verification and false accept rates.

Error ellipses in figure 8 are two times the standard deviation of the verification and false accept rates along the appropriate axes. An ellipse gives an estimate of the range in performance that could result if the people in the gallery are changed. If the large gallery were larger, it would be possible to compute performance for more small galleries of size 3,000. The greater number of small galleries would increase the accuracy of the error ellipse. However, the size of the ellipses would not decrease as the number of small galleries increased. This is because the error ellipses are a function of the multiple small galleries, and composition of the small galleries reflects the natural variation in the population. The natural variation will always be present—more small galleries increase the accuracy of the estimated variation in the performance due to the natural composition of the population. In the HCInt the ellipses are estimated from disjoint galleries and probe sets. This avoids many of the issues associated with re-sampling techniques. Re-sampling techniques require making assumptions about the distributional properties of the similarity scores. Typical assumptions are that similarity scores are independent and identification distributed (iid). In interpreting the meaning of error ellipses, a number of subtle facts need to be noted. The error ellipses are not error bounds on the ROC. Rather, error ellipses are a measure of the variance in

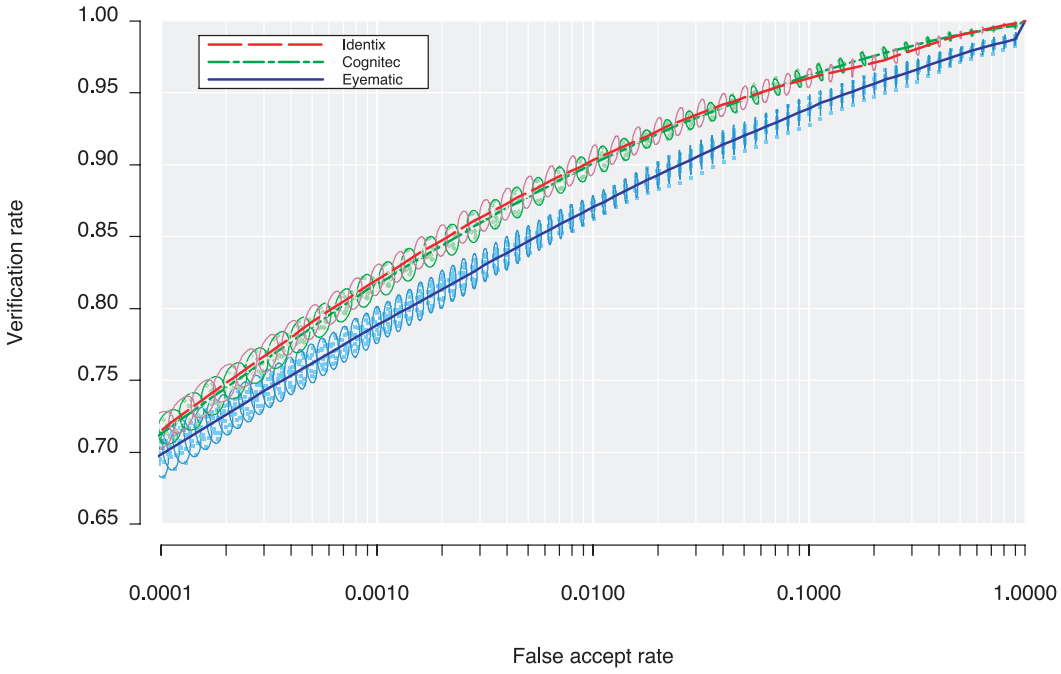


**Fig. 8. Standard error ellipses for verification performance for Cognitec, Eyematic, and Identix. The standard error was computed from twelve HCInt small galleries of size 3,000. The center line is the ROC performance for the aggregate of all twelve galleries. The ellipses are two times the standard deviation at selected performance points, and the points clustered around the ellipses represent the performance of one of the twelve galleries at the selected performance point.**

performance that occurs by changing the gallery. The standard error is an empirical estimate of the variation. They are not confidence intervals. Confidence intervals decrease in size as the number of samples increase. To estimate confidence intervals requires that one knows or can estimate the underlying distribution.

One of the recent innovations in face recognition has been the adoption of normalization. C-VIS, Eyematic, Identix, and Viisage submitted normalization routines in FRVT 2002. In two experiments, we investigate the contribution of normalization to verification and identification performance.

To measure the effect of normalization on verification, performance was computed for normalized and non-normalized versions for the twelve small galleries and probe sets. The difference in the verification rate was computed for four false accept rates: 0.0001, 0.001, 0.01, and 0.1. Figure 9 plots the normalized score minus the non-normalized score. Figure 9 contains four panels, one for each false alarm rate. In each panel, the mean change in verification rate over the twelve small galleries is plotted. For example, at a false accept rate of 0.001, the mean change in performance for Eyematic is 0.089. This means that normalization improved Eyematic's verification rate by 0.089. In this experiment, the improvement in verification rate over the twelve galleries was very stable. The variance in the change is sufficiently small that it would not show up in figure 9 if plotted. The results in figure 9 show that normalization improves performance. All verification performance results in FRVT 2002 are computed with normalized scores for those who submitted normalization functions.

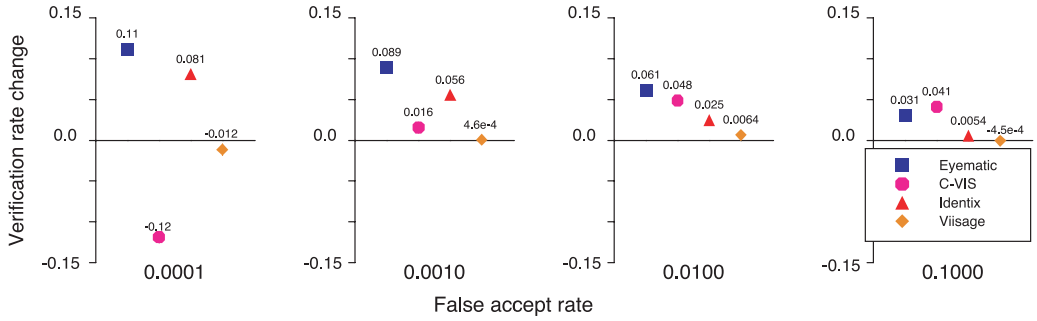


**Fig. 9. Plots showing the difference between normalized and non-normalized verification performance. Performance was computed for the twelve galleries in the experiment. Relative performance of the verification rate is shown for false accept rates of 0.0001, 0.001, 0.01, and 0.1.**

The same experiment was repeated with identification. The results showed that normalization did not improve performance for identification. Because normalization did not improve identification performance, all identification scores in FRVT 2002 are computed without normalization. For the watch list task, the improvement from normalized scores was comparable to verification, and therefore, watch list performance is computed using normalized scores.

The results of the *HCInt large gallery identification* experiment are shown in figure 10. Identification results are reported on a CMC with rank scaled logarithmically.

While the HCInt large gallery experiment measured performance on a large gallery of 37,437 individuals, the *HCInt gallery size* experiment examines how rank 1 identification performance decreases as gallery size increases. Figure 11 shows performance as a function of gallery size. The horizontal axis is gallery size on a logarithmic scale and the vertical axis is rank 1 identification performance. The performance was computed for thirteen difference gallery sizes: 25, 50, 100, 200, 400, 800, 1,600, 3,000, 3,200, 6,400, 12,800, 25,600, and 37,437. The spread of the galleries sizes is approximately logarithmic in gallery size.

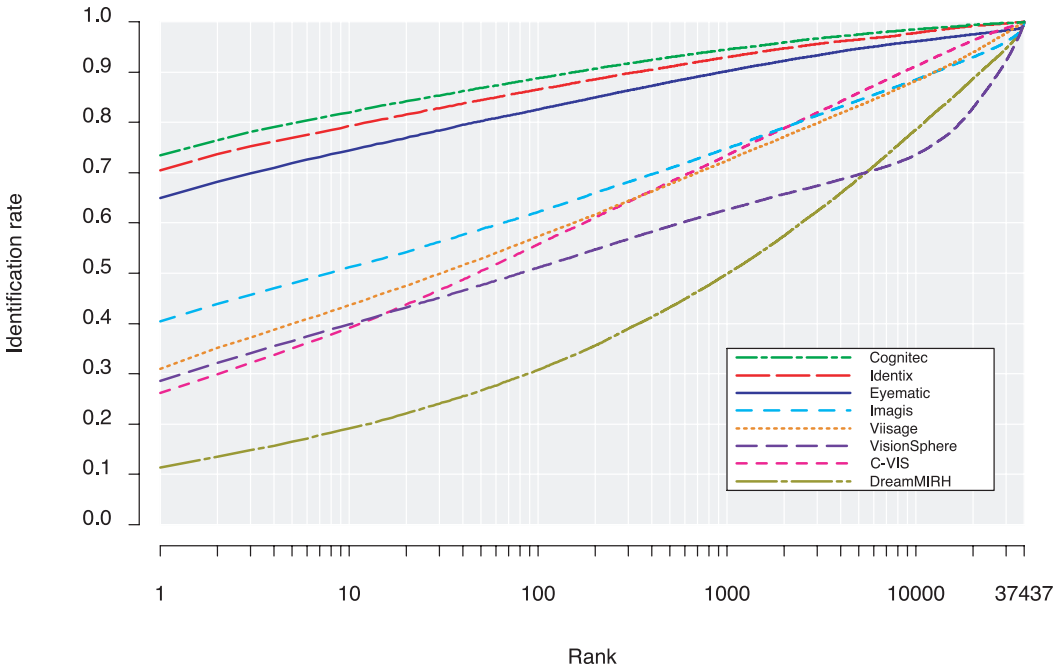


**Fig. 10. Identification performance reported on a CMC for the large gallery and probe set. The large gallery contained images of 37,437 individuals. The horizontal axis is on a logarithmic scale**

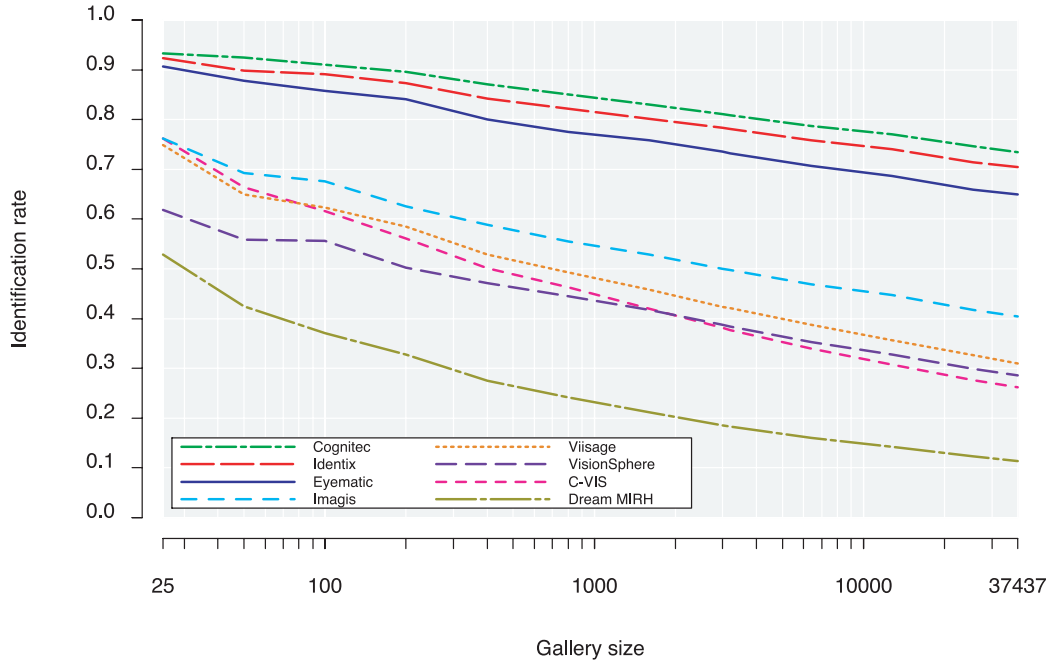


**Fig. 11. Rank 1 identification performance as a function of gallery size.**

The identification rates plotted in figure 11 are an average identification rate over a set of galleries. For each gallery size, disjoint galleries were generated and the mean identification was computed. Where possible, performance was computed from twelve galleries. However, the number of disjoint galleries generated was limited by the size of the HCInt data set. For example, it was not possible to generate more than one gallery of size 25,600 and 37,437. Where possible, performance is the average over multiple galleries. The averaging was done to provide a reliable estimate of performance.

The following experiments examined watch list performance: the watch list ROC experiment, the watch list rank experiment, and the watch list gallery size experiment. Performance of a system on a watch list task is characterized by four values: detect and identify rate, false alarm rate, gallery size, and rank. To understand the effect of each of the values on performance, we performed three watch list experiments. In each of the experiments, one or two of the values are held constant so that the effect of changing the remaining values can be examined.

The *watch list ROC* experiment computes a watch list ROC for all eight participants. In this experiment, performance was computed for twelve galleries. To provide a reliable estimate, average performance over the twelve galleries was plotted. All galleries consisted of 3,000 individuals. The probe set consisted of 6,000 probes of people in the gallery and 6,000 imposters (there were two images per person in the probe set). Rank 1 identification performance is computed. Specifically, is the top match with a given probe a person on the watch list? Figure 12 shows the watch list performance for all eight participants.

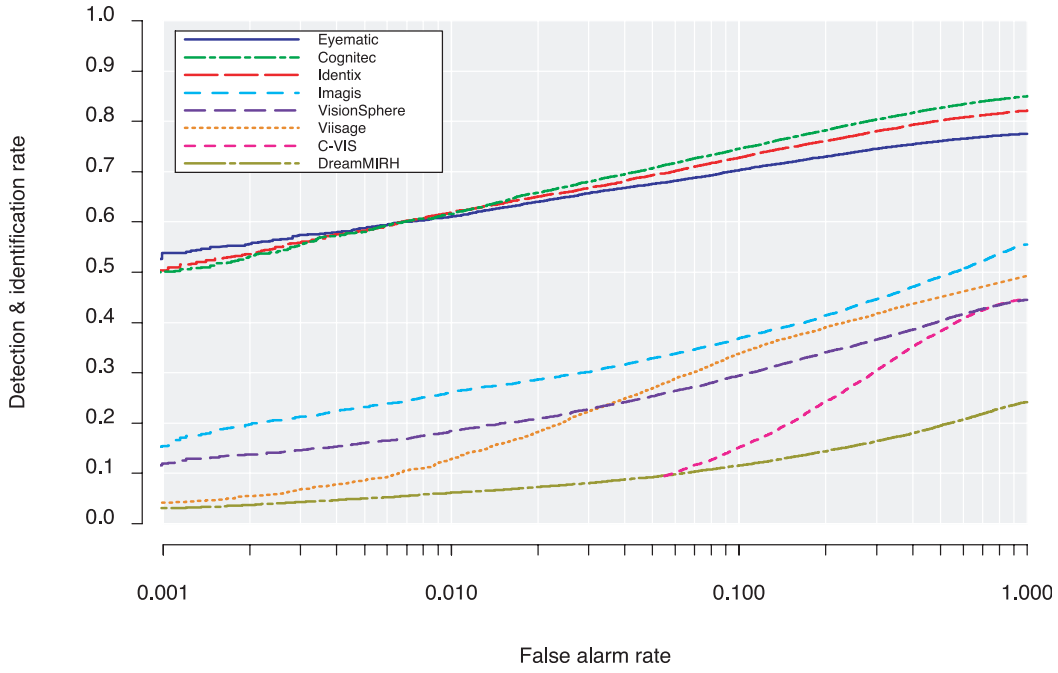


**Fig. 12. Watch list performance ROC. Gallery size is 3,000. Rank 1 performance is reported. The C-VIS system does not operate for the full range of false alarm rates plotted in this experiment.**

In a normal watch list task, the rank 1 match between a probe and a gallery is found. If the rank 1 match score is above a threshold, an alarm is issued. The general case finds the top $k$ matches between a probe and a gallery, and reports those matches above an operating threshold. The *watch list rank* experiment investigates the effect of examining the top $k$ matches on performance. In this experiment, results are only reported for Cognitec in order to keep the graph readable—the overall trend applies to all participants. In the watch list rank experiment, the gallery consists of 800 individuals. The probe sets consist of 1,600 probes of people in the gallery (two per person in the gallery) and 1,600 imposters of 800 individuals. Results for this experiment are shown in figure 13. Full CMCs are computed for eight false alarm rates. The $x$-axis is rank and the $y$-axis is the detection and identification rate. Each curve in the CMC is calculated for a given false alarm rate. The false alarm rate associated with each curve is adjacent to the curve on the right-hand side. For a false alarm rate of 0.01, performance is the same for all ranks. Even at a false alarm rate of 0.1, performance increases very slightly for $k = 2$. This suggests that in watch list applications, one only needs to examine the top match. It is worth noting that in figure 13 the curve with a false alarm rate of 1.0 is the CMC for identification against the gallery of 800 individuals.

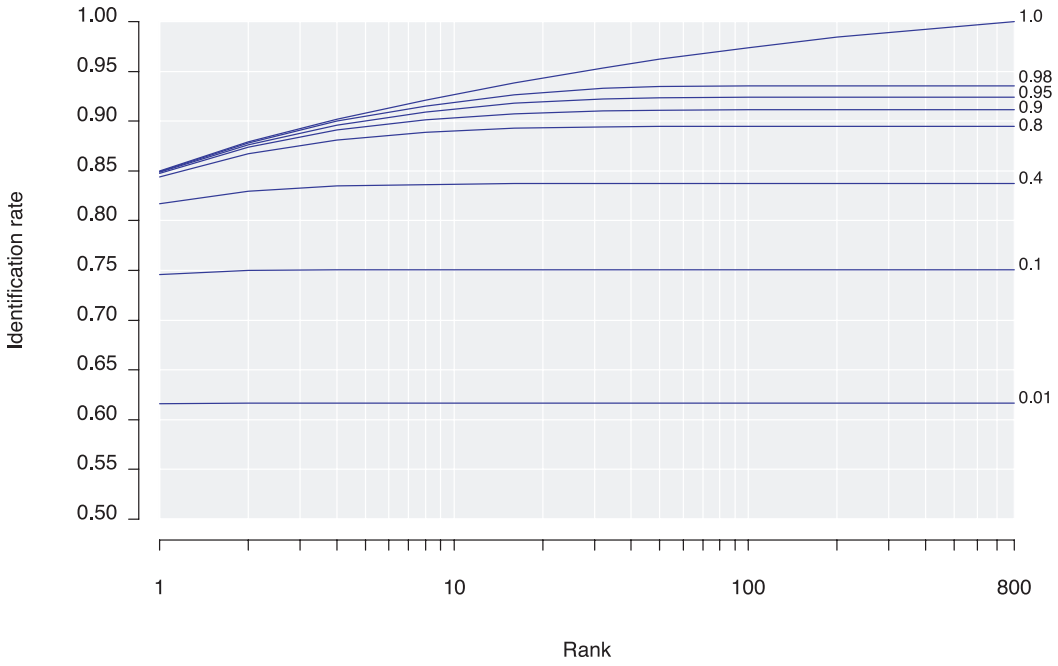


**Fig. 13. Watch list performance for Cognitec as a function of rank for eight false alarm rates. The false alarm rates for each curve are on the right side of the graph. The gallery size is 800. The top curve (false alarm rate of 1.0) is the CMC for the gallery in this experiment.**

The *watch list gallery size* experiment examines how watch list size (gallery size) affects performance. Performance was computed for eight galleries sizes: 25, 50, 100, 200, 400, 800, 1,600, and 3,000. The spacing of the gallery sizes is approximately logarithmic in size of the gallery. Each probe set contained two images of each person in the gallery and an equal number of imposters. To provide a reliable estimate of performance for each gallery size, watch list performance was computed on twelve disjoint galleries. The average performance is plotted in figure 14. The curves correspond to a false alarm rate of 0.01. The decrease in the detection and identification rate mirrors that of the decrease in identification. The decrease is approximately linear in the logarithm of the gallery size.

The performance analysis on the HCInt data set has, so far, concentrated on tasks: verification, identification, and watch list. The next step is to look at the effect of three covariates on performance. The covariates are: *elapsed time between acquisitions of gallery and probe images of a person, sex,* and *age of an individual*. Since there are multiple images of an individual in the data set, we fixed age by the age of an individual when the gallery image was taken.

Previous evaluations have examined the effect of temporal variations at a very coarse level. In past evaluations, the categories were: *Images Collected on the Same Day, Images Collected on Different Days,* and *Images Collection over a Year Apart*. FERET and FRVT 2000 showed that there was a major difference in performance between same-day images and images collected on different days. However, the rate at which performance declined as a function of elapsed time between gallery and probe could not be accurately estimated. This was due to the small size of the probe set.

In the HCInt, the effects of temporal variation are studied by computing identification and verification performance on a large gallery and probe set. The probe set is partitioned into 19 bins. Each bin corresponds to a 60-day interval. The first bin corresponds to 1 to 60 days. A probe is in the 1 to 60-day bin if the time between the acquisition of the gallery and probe image is between 1 and 60 days. The second bin corresponds to 61 to 120. The remaining 17 bins are constructed in a similar manner in 60-day intervals. Figure 5 shows the number of probes in each 60-day interval.

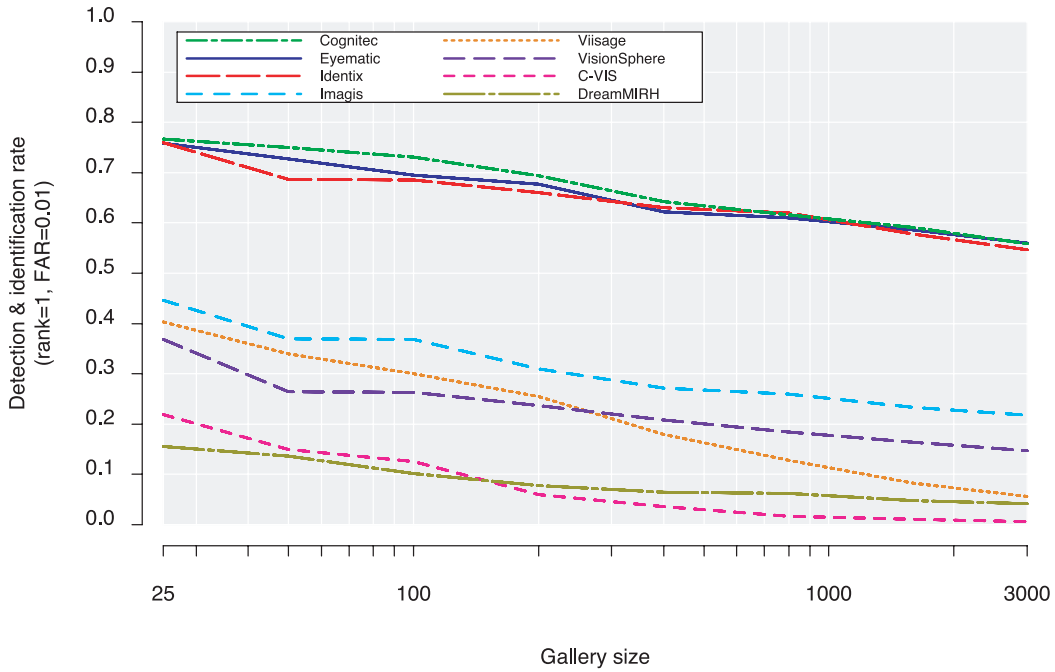


**Fig. 14. Watch list performance as a function of gallery size for a false alarm rate of 0.01. Rank 1 performance is measured.**

Both identification and verification performance were computed. For all 19 bins, performance was computed against the large gallery, and the size of the probe set varied according to the bin. Figure 15 shows the rank one identification rate. Figure 16 plots the verification rate at a nominal false accept rate of 0.01. The false accept rate is nominal because there is a change in the false accept rate for each bin. Details on how the verification and false alarm rate are computed for each bin are explained in more detail in Section 4. Briefly, from the large gallery verification experiment, the threshold that corresponds to a false accept rate of 0.01 is selected. Using this threshold, verification and false accept rates are computed for the different probe sets. The different probe sets correspond to the elapsed time bins. This results in different verification and false accept rates for each probe set. This method of computing effects for both verification and false accept rates is the basis for analysis in the HCInt covariate experiments.

Imaging conditions and elapsed time between images are one set of conditions that affect performance. These effects have been extensively studied in this and previous evaluations. In addition, the population in a database can significantly affect performance. Two examples are the sex and the age of people being recognized. Experiments on the HCInt show that the sex and age of a person affects performance.

To measure the effect of sex on performance, the HCInt large probe set was divided into male and female probe sets. Identification performance was computed for the male and female probe sets against the large gallery. Note: the large gallery was not divided into male and female galleries. For all systems, the identification rate for males was higher than the identification rate for females. Figure 17 shows the difference in identification performance for the male and female probe sets. For each rank, male minus female performance is plotted.

The calculation of the difference in performance for identification is straightforward; subtract the male and female identification rates. For verification, the procedure is a bit more complicated. Two statistics characterize the performance difference between males and females, the difference in the verification rate and the difference in the false accept rate. Figure 18 shows the difference in performance between the male and female probe sets in a ROC. The difference is shown by a

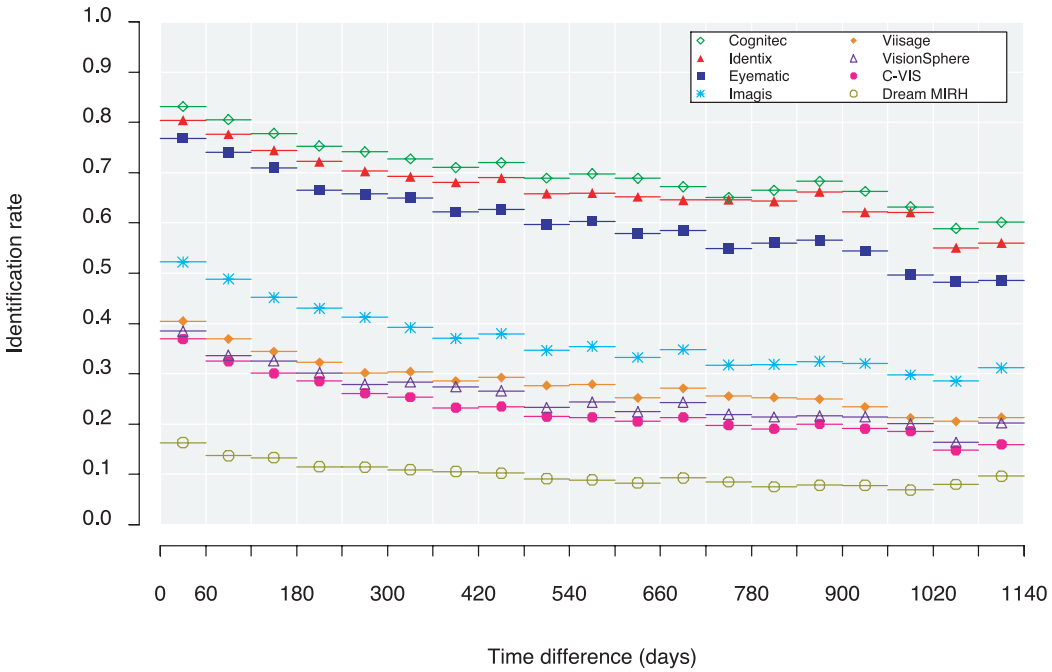


**Fig. 15. Rank 1 identification rate for temporal variation study. The identification performance is broken out by 60-day intervals. The identification rate is for probes against the HCInt large gallery.**

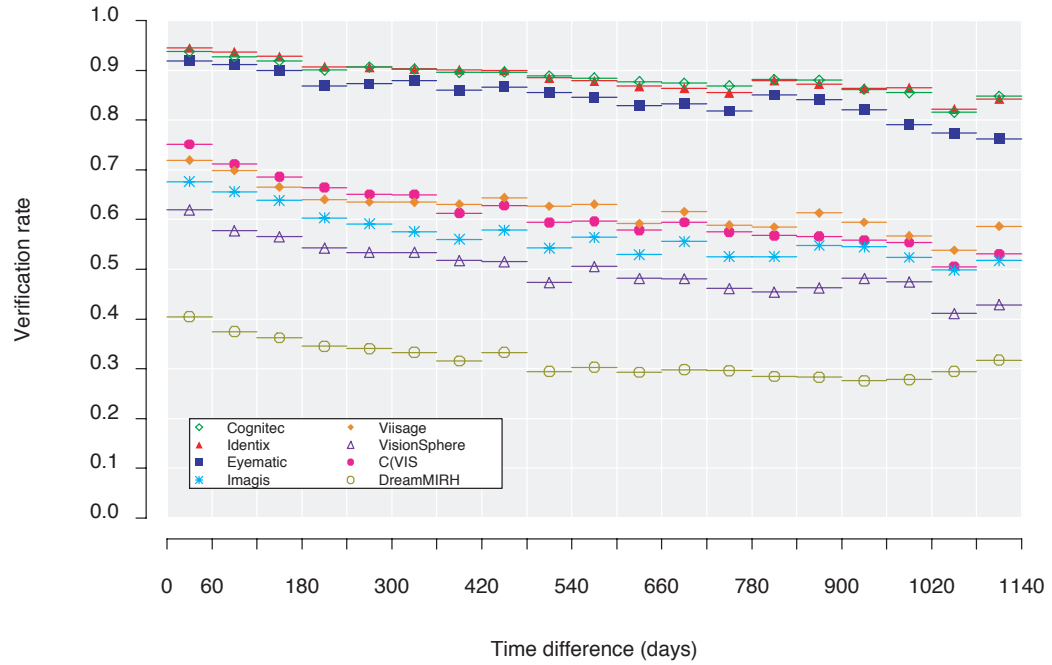


**Fig. 16. Verification performance for temporal variation study. The verification rate is given at a false accept rate of 0.01.**

series of line segments for each system. The dot at approximately the center of each line segment is the aggregate performance for both male and female probe sets. This point is found on the large gallery ROC in figure 7. The bottom end of each line segment is the corresponding performance point for the female probe set. The top end of each line segment is the corresponding performance point for the male probe set. For all systems, the verification rate for males is better than for females. However, the same is not true for the false accept rates. For Cognitec, Eyematic, Viisage,

VisionSphere, the false accept rate for males is better than for females (line segments go from the top left to bottom right). For these four systems, performance on males is better than females. For C-Vis, DreamMIRH, Identix, and Imagis, the false accept rate for females is better than males (line segments go from the top right to bottom left). For these systems, neither female nor male performance is better than the other. Rather, the decision depends on the scenario.

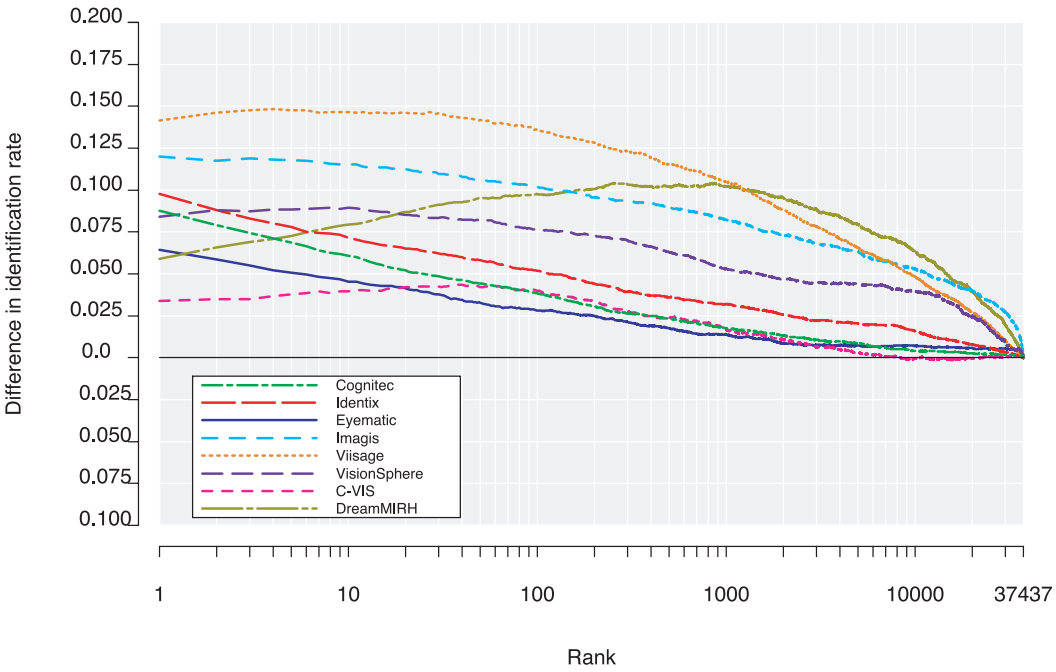


**Fig. 17. CMC showing the difference in performance between male and female probes. For each rank, male minus female performance is plotted. The CMC shows that the identification rate of males is better than females.**

Figure 17 and figure 18 break out performance by sex for the large gallery and large probe set, and show noticeable differences in performance between the sexes. This generates a follow-up question: Is the difference in performance consistent, or is it a property of the large gallery and large probe set? To address this question, verification and identification performance was broken out by sex and was computed on the twelve small gallery and probe sets. Verification and false accept rates were computed for both male and female probe sets. The performance figures were computed at a nominal false accept rate of 0.01. For each of the twelve galleries the male minus female performance figures were calculated for both verification and false accept rates. Figure 19 plots these differences for all twelve small galleries and participants. Each symbol is the performance of one of the participants in one of the twelve galleries. The *x*-axis is the difference in the false accept rate and the *y*-axis is the difference in the verification rate. For each participant, the results for all twelve galleries cluster. With the exception of three runs, the verification rate for males was better than females, and for the three exceptions, the rate is very close to zero. There is an order of magnitude difference between the two axes. This shows that there is a much larger change in the verification rate than in the false accept rate. The same computations were performed for identification, and the results are consistent with the results in figure 17. For all galleries and all participants, the identification rate for males was better than females. This shows that the bias towards males in the large gallery is consistent with the results of the twelve small galleries and provides strong evidence that the observed sex bias is a property of the HCInt data set.

The effect of age on performance has not been previously examined. To examine the effect of age on performance the HCInt large probe set was divided into 12 age categories. Each age category is a five-year bin, with the first bin containing 18 to 22 year olds. All probes of a person

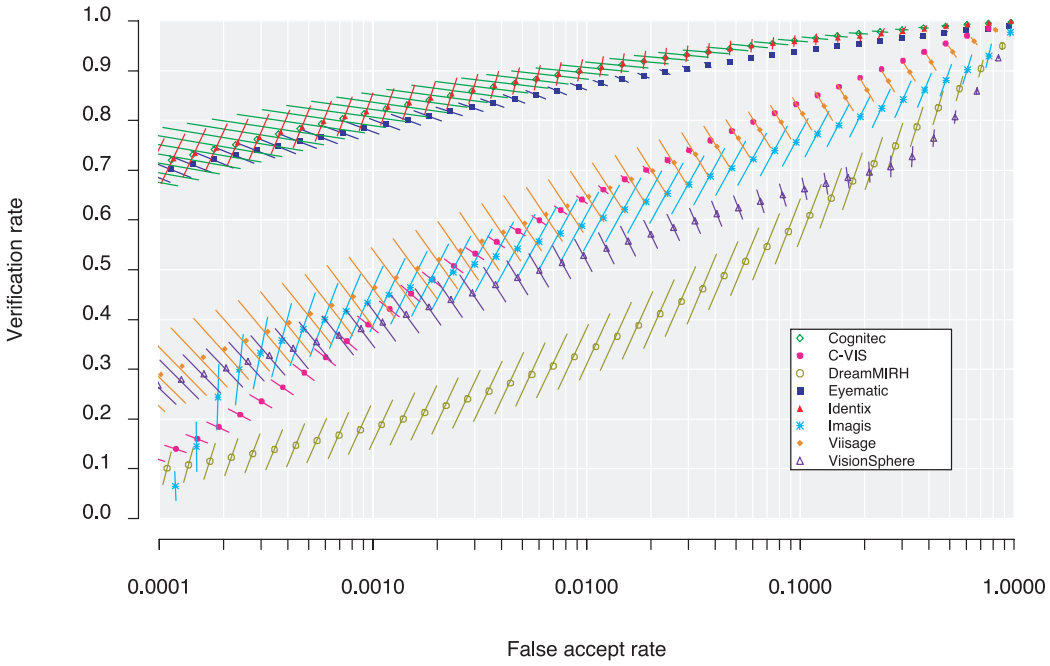


**Fig. 18. ROC showing the difference in performance between male and female probe sets. For each line segment, the center point is the aggregate performance for male and female probes. The end of the line segment below the center is the performance on the females, and the end above the center point is performance for the males. For each ROC, the difference in performance at multiple operating points is shown.**

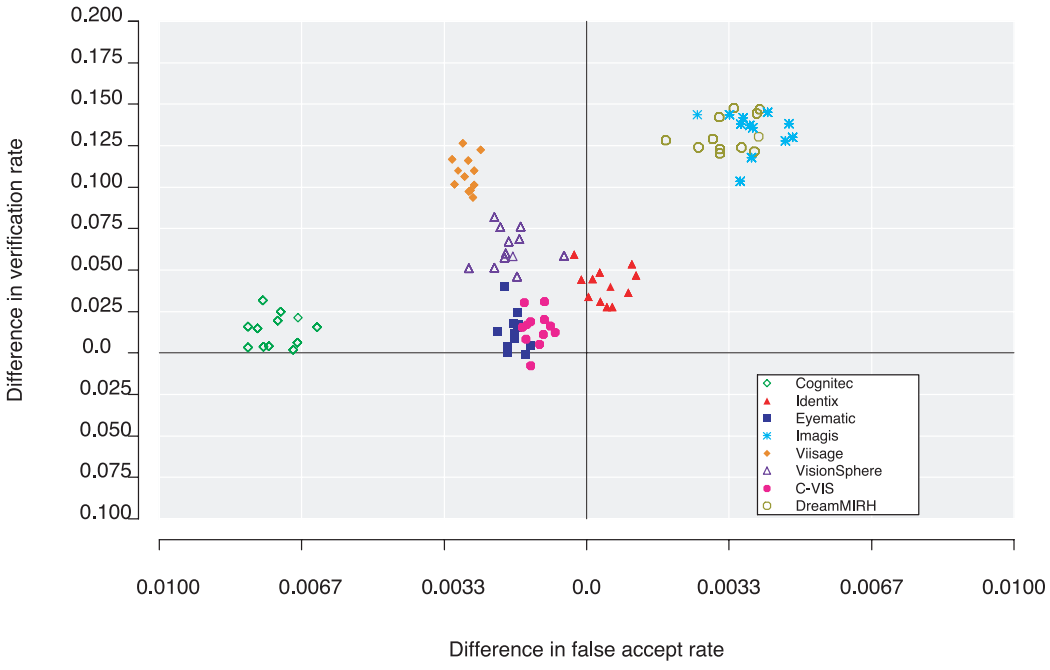


**Fig. 19. Difference in verification performance for male and female probe sets for the twelve small galleries and probe sets. The x-axis plots the change in the false accept rate and the y-axis plots the change in the verification rate. Note that there is an order of magnitude difference in the scales of the two axes.**

are placed in the same bin, and the bin is determined by the age of a person when the gallery image was taken. For example, if the gallery was taken when a person was 21 year old, and the probes where taken when the person was 24 and 25, then both probes were placed in the 18 to 22 year bin. The performance for all age bins was measured against the HCInt large gallery. The top rank identification performance was broken out by age and is plotted in figure 20.

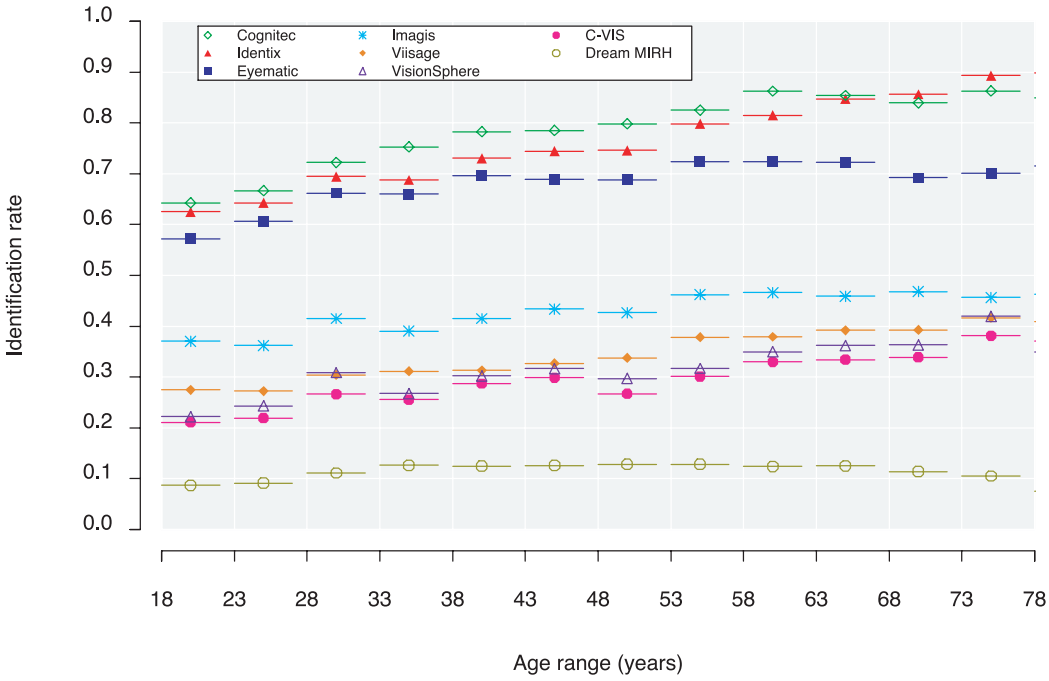


**Fig. 20. Rank 1 identification performance on the HCInt large gallery broken out by age.**

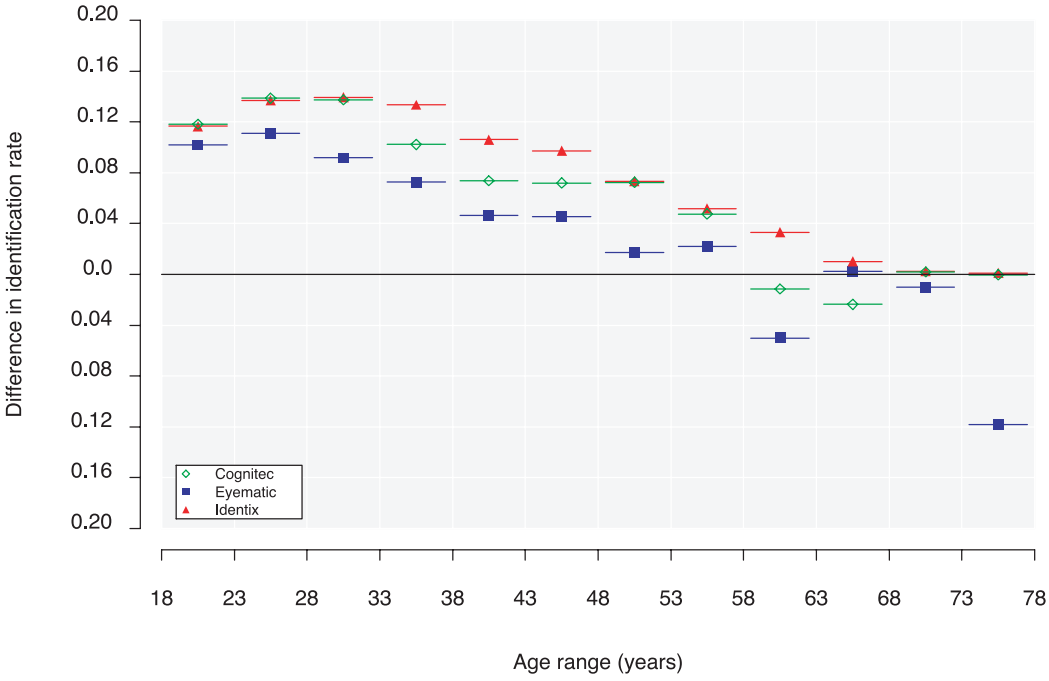


**Fig. 21. Interaction between age and sex for rank 1 identification on HCInt large gallery (for Cognitec, Eyematic, and Identix). Rank 1 performance for males minus females is plotted for each age bin.**

Figures 17-20 show that both the sex and age of a person affect performance. However, it is possible that the age and sex effects are the same. This could happen if the males are much older than the females. Thus, we need to check the interaction of the two factors. This is accomplished by dividing each of the 12 age probe sets into male and female probe sets. For each of the 12 age bins, we compute the difference in the male and female rank 1 identification rates (male minus female identification rate). The results for Cognitec, Eyematic, and Identix are shown in figure 21. Figure 21 shows that the difference in performance for the sexes decreases as age increases, and shows an interaction between sex and age. Figure 27 shows rank 1 identification performance for all participants.

The HCInt experiments have provided answers to a number of long-standing questions and raised new questions. The long-standing questions have been the effect of gallery size on performance, variation in performance for different galleries, effect of elapsed time on performance, and characterization of watch list performance. New questions emerge from studying the effect of covariates on performance. The new questions arise from the determination that sex and age, and their interaction, do affect performance. This will lead to new research and insights into how face recognition systems work as well as issues associated with deploying automatic face recognition systems.

## 8. Medium Computational Intensity Test Results

The MCInt was designed to measure performance under different imaging conditions. MCInt reports three sets of the experiments and their results. The first set investigates recognition from frontal images under different imaging conditions. The second set looks at how pose variations affect performance and the impact of using three-dimensional morphable models as a preprocessing stage. The third set compares recognition performance using still and video sequences. The experiments also provide another assessment—the effect of pose variations on performance.

The goal of the MCInt is to provide a broad assessment of progress and the state-of-the-art. To provide a clear and concise assessment, performance is characterized in the *Evaluation Report* by two statistics: *rank 1 identification rate* and *verification rate at a false accept rate of 0.01*. Full identification CMCs and verification ROCs for each experiment and participant can be found in the *Technical Appendices*. In figure 22 through figure 24, the verification rate is plotted for a false accept rate of 0.01. In figure 28 through figure 30 in Section A.4, the experiment performance is reported for rank 1 identification rate. Verification rate was selected for the main body of the report because verification performance is not a function of the gallery size. This allows for a degree of comparison among the results from the three sets of experiments. A strict comparison among the experiments is not possible because performance was computed from different galleries.

The first experiment is the *MCInt frontal face experiment,* which investigates how changes in lighting conditions affect face recognition from frontal mugshot style images. Performance is computed for five probe sets against the same gallery. The gallery consisted of 787 people taken under incandescent studio lighting. The subjects had a neutral expression. To provide an empirical upper bound on system performance, the first probe set consisted of images taken within five minutes of the gallery image under the same imaging conditions. This is referred to as the *indoor-same day probe* set. The indoor-same day probe set had a different expression than the gallery image and contained 786 images from 781 individuals. The gallery images the expression was usually neutral, and in the probe the expression was usually a smile.

The *indoor-overhead lighting* probe set consisted of images taken with five minutes of the gallery images but with overhead fluorescent illumination. This probe set contained 786 images from 786 individuals. The indoor-overhead lighting probe set tested the effects of indoor lighting changes on performance.

The *indoor-different day* probe set contained probes taken under the indoor studio setup on a different day than the gallery image of a person. This probe set contained 320 images from 320 individuals and measured the effects of temporal variation on indoor probes. The median elapsed time between gallery and probe image was 147 days and the maximum was 730 days.

The *outdoor-same day* probe set contained images taken outdoors on the same day as the gallery image of a person. This probe set contained 444 images from 435 individuals and measured the effects of outdoor illumination.

The *outdoor-different day* probe set contained images taken outdoors on a different day than the gallery image of a person. This probe set contained 145 images from 103 individuals. The outdoor-different day probe set measured the effects of temporal variation and outdoor illumination. The median elapsed time between gallery and probe image was 152 days and the maximum was 505 days.

The results from the MCInt frontal face experiment are presented in figure 22 and figure 28. Figure 22 shows verification performance at a false accept rate of 0.01 for each participant and each probe set. The *y*-axis is verification performance. The *x*-axis is categorical, with one category for each of the five probe sets. The marks on the vertical line above Indoor (same day) are verification performance for the indoor-same day probe set. The performance results for the other four probe sets are reported in the same manner. Lines are drawn between the results for each participant on different probe sets. This is done to make it easier to examine how the properties of different probe sets affect performance. Figure 28 plots identification performance in the same manner.

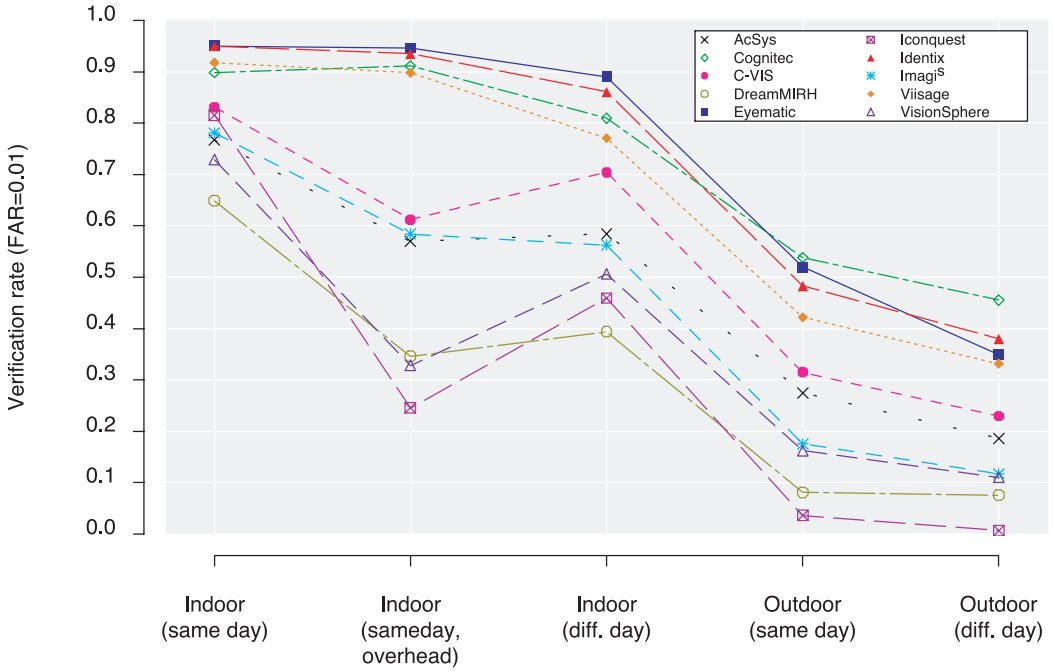


**Fig. 22. Comparison of system performance on different categories of probes. The verification rate at a false accept rate of 0.01 is plotted.**

The MCInt morphable model experiment examined the effects of pose variation and morphable models on performance. This experiment consisted of one gallery and nine probe sets. The gallery consisted of full frontal images of 87 individuals taken indoors under studio conditions with a single incandescent photo floodlight. Each probe set consisted of 87 images of 87 people. All probe images were taken within five minutes of the gallery image under the same conditions. The only difference is the pose.

Figure 3 shows examples of the nine probe sets. The *45 left* and *45 right* probe sets contained facial images facing 45 degrees to the left and right of center respectively. The 45 L and 45 R columns in figure 23 and figure 29 report verification and identification results for the 45 left and 45 right probe sets. Line segments are drawn between original and corresponding morphed probe sets to emphasize the effect of morphing. The *30 up* and *30 down* probe sets contain facial images facing 30 degrees up and down respectively. The performance results for these two probe sets are reported in the 30 U and 30 L columns. In the remaining five probe sets, a three-dimensional morphable model has been applied to the probes. See figure 3 for examples of applying a three-dimensional morphable model to a probe.

The *frontal morph* probe set provides a baseline for how morphing affects a system. In the frontal morph probe set, the morphable model is applied to the gallery images. Thus, the difference between a gallery and a frontal morph image is that the morphable model has transformed the probe. The results for the frontal morph probe set are in column frontal (morph). If a system were insensitive to the artifacts introduced by the morphable model, then the verification and identification rates would be 1.0. In Figure 23, sensitivity to morphable models range from 0.98 down to 0.45.

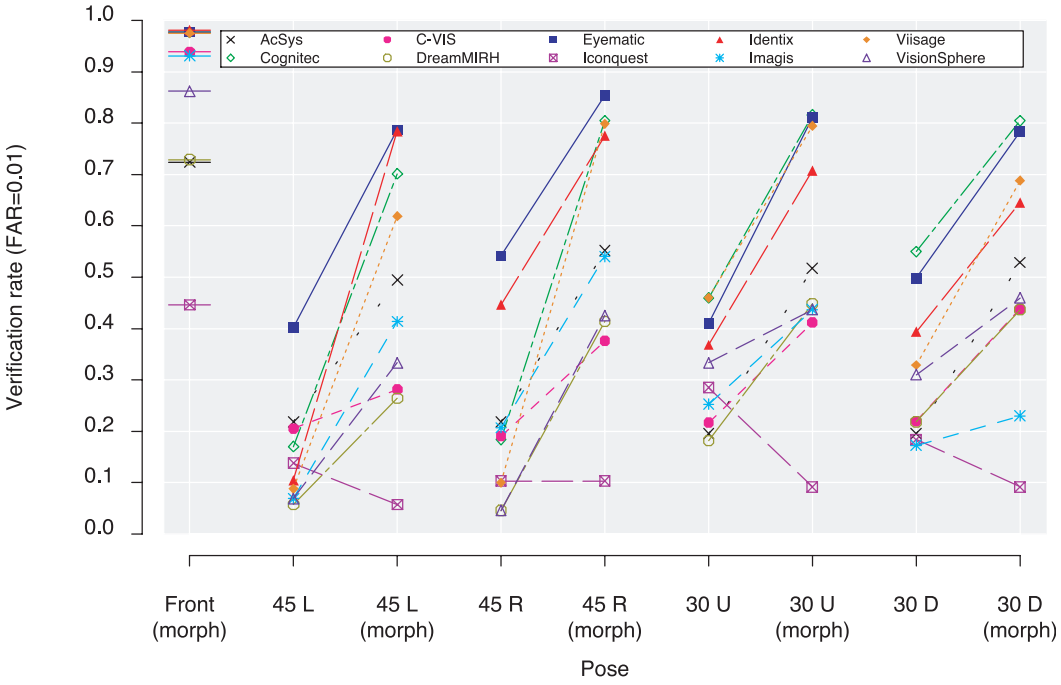


**Fig. 23. The effect of still versus three-dimensional morphable models. The verification rate at a false accept rate of 0.01 is plotted.**

To investigate the effects of morphable models, performance was computed for four probe sets: *45 left morphed, 45 right morphed, 30 up morphed,* and *30 down morphed*. These probe sets were produced by applying the morphable model to the 45 left, 45 right, 30 up, and 30 down probe sets respectively. The results for the morphed probe sets are in columns 45 L (morph), 45 R (morph), 30 U (morph), and 30 D (morph). The results show that with the exception of Iconquest, morphable models significantly improved performance.

The *MCInt still-video* experiment compares performance of still and video-style probe sets. This experiment consisted of one gallery of still full frontal digital images of 63 individuals taken indoors under studio conditions. In the still-video experiment, there are two probe sets: still and video. Both probe sets contained 63 signatures from 63 individuals. The probes in both probe

sets were taken indoors under studio conditions, consist of the same individuals, and are taken on different days than the gallery images. The still and video signature probes of a person were taken on the same day. The still probe set consisted of a single still full frontal image. A video probe was extracted from a 5 second video sequence of a person speaking (see figure 4). A video probe consisted of a sequence of 100 images (3.33 seconds), where each image was a frame in the original digital video sequence. We called this probe set video-style because the probe set consists of a sequence of frames.

The results of the MCInt still-video experiment are shown in figure 24 and figure 30. (Human inspection of the Viisage similarity files confirmed their performance on the still probe set.) With the exception of DreamMIRH and VisionSphere, performance was better on the still probe set than the video probe set. This was true for both verification and identification.

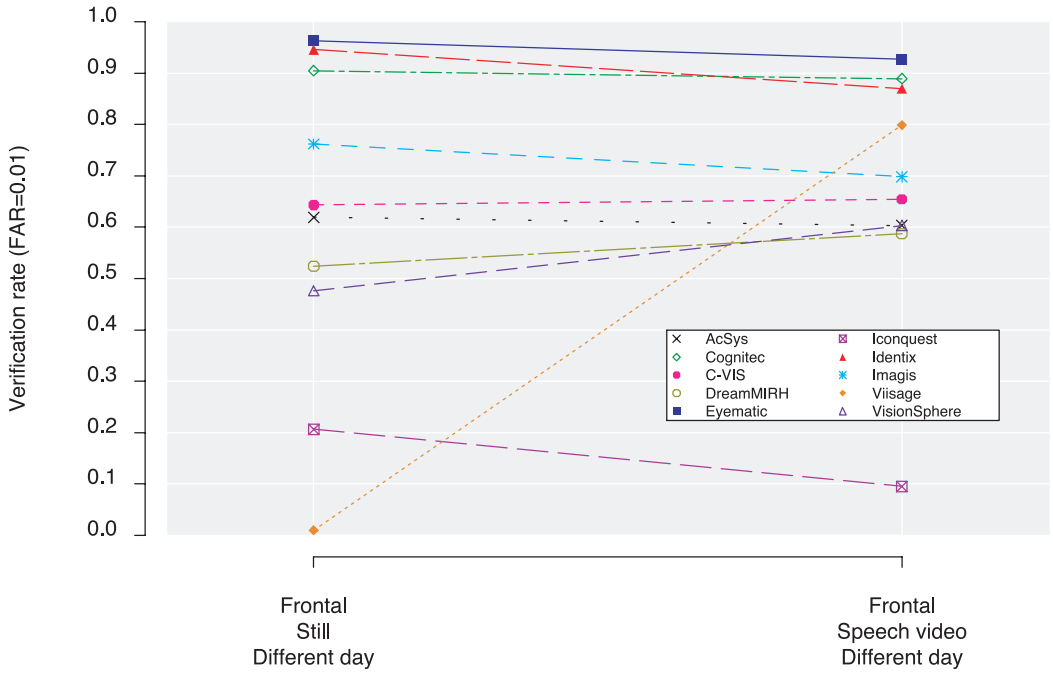


**Fig. 24. Plot showing still versus video recognition for frontal imagery. The verification rate at a false accept rate of 0.01 is plotted.**

## 9. Analysis and Discussion

The large number of images and people in the HCInt allow for a much more detailed analysis than in previous face recognition and biometric evaluations. The HCInt measures performance on indoor images taken on different days. Performance for this category was also reported in FRVT 2000 (Blackburn et al. 2001). Because the gallery sizes in FRVT 2000 and FRVT 2002 vary, comparing the results between the two must be restricted to verification. On comparable experiments, the error rate has dropped by approximately 50 percent between FRVT 2000 and 2002. This shows substantial improvement in performance since FRVT 2000.

A companion NIST study on fingerprints highlights progress in face recognition and points to directions of research to improve identification rates for large galleries (NIST 2002). Performance was computed for the verification test bed (VTB), an in-house NIST algorithm based on an automated fingerprint identification system (AFIS). The matching technology in the VTB is comparable to commercial systems in 1998. The recognition rates were computed from fingerprints provided to NIST by the U.S. Immigration and Naturalization Service's IDENT system.

Fingerprint verification performance was computed for twelve independent galleries and probe sets. The gallery consisted of the fingerprint of one index finger for 6,000 individuals. The probe set consisted of a single fingerprint for the 6,000 people in the gallery.

Figure 25 shows verification performance with error ellipses for the VTB, Cognitec, Eyematic, and Identix. At false accept rates around 0.0l, verification performance is comparable. In fact, this is the cross-over point between face recognition and fingerprint performance. At false accept rates below 0.01, fingerprint performance is better. At rates above 0.01, the best face recognition systems perform better. The good news is for false accepts around 0.01, that face recognition performance is now comparable to large-scale fingerprint systems available in 1998. This suggests that a dual face recognition and fingerprint system is viable.

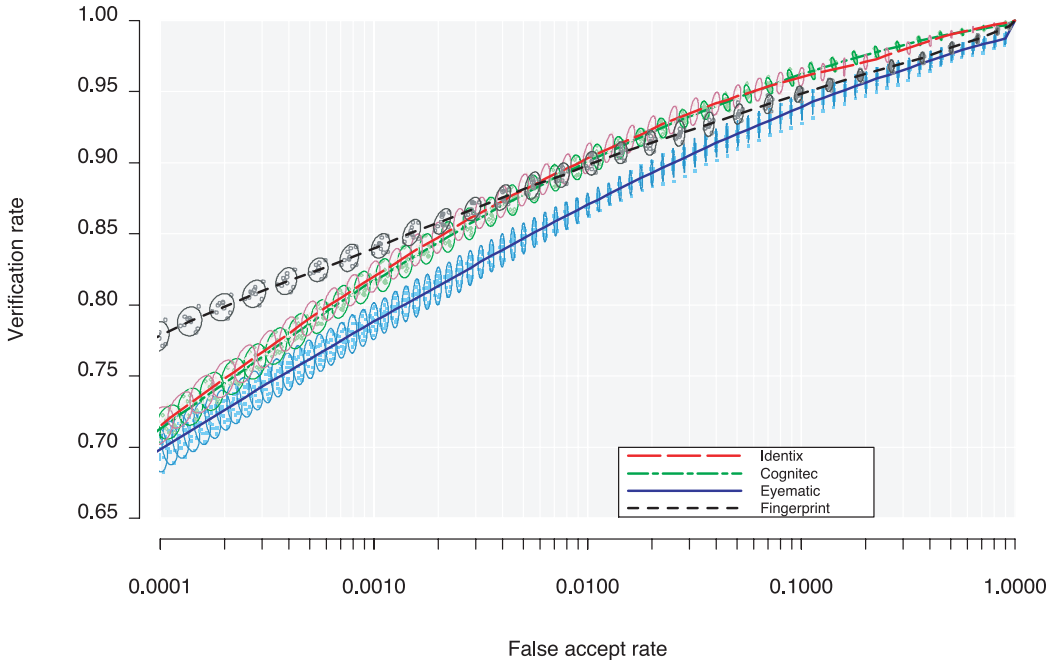


**Fig. 25. Verification performances of Cognitec, Identix, and Eyematic on Visa images and NIST VTB single fingerprint matcher on IDENT data.**

We now turn our attention to identification rates for large galleries. Figure 26 shows rank 1 identification rate as a function of gallery size for Cognitec, Identix, Eyematic, and NIST VTB (single index finger). The face and fingerprint curves do not completely overlap because identification rates were computed for different sized galleries for face and finger. The results from the VTB clearly show that fingerprint performance on identification is superior to face. In Appendix A.5, we present a *moment model* for predicting performance as a function of gallery size (see later in this section for a more detailed discussion). The moment model relates verification and identification performance. The model predicts that the verification rate for false accept rates at 0.001 and 0.0001 have the most significant impact on identification performance for large galleries. This could explain why face and finger performance are comparable at a false accept rate of 0.01, but do not have comparable identification performance. Verification performance at false accept rates of 0.001 and 0.0001 has a greater impact on identification rate.

The fingerprint verification ROC is flatter than the face ROC (figure 25). A ROC that is flat means that a system either recognizes a person or completely fails. This suggests that a fingerprint recognition algorithm is much more likely to recognize a majority with high confidence and the remainder with low confidence. Face recognition, on the other hand, is more "egalitarian;" it recognizes most people at about the same level.

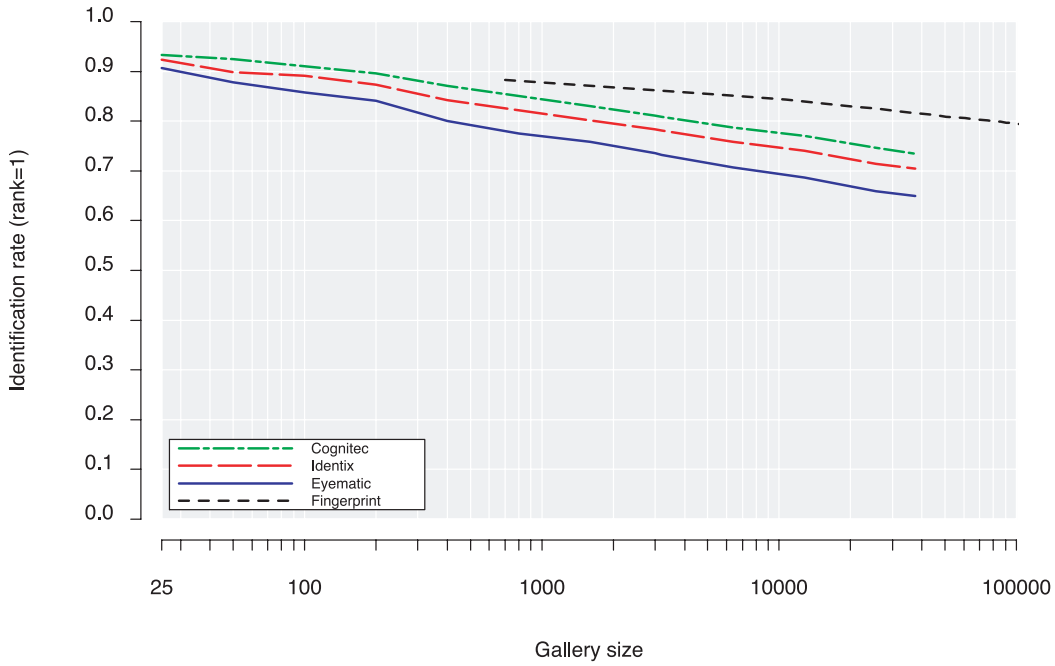


**Fig. 26. Rank 1 identification rate as a function of gallery size for Cognitec, Eyematic, Identix, and NIST VTB single fingerprint matcher on IDENT data.**

The results from FRVT 2002, the VTB, and the moment model indicate that to improve identification performance, effort should be made to increase verification performance for low false alarm rates. One possible approach is to concentrate on the vast majority of faces that are relatively easy to recognize and to not concentrate on the hardest faces to recognize. This would flatten out the resulting ROC, but at the same time, this could have a significant effect on identification performance for large galleries. To be able to increase the verification rate for very low false accept rates may require new features that are tailored to an individual or group of individuals. This differs from current approaches that select features that distinguish among a large population of people. Developing and evaluating performance for improvements at very low false accept rates will require experiments to be performed on large databases.

The large number of images and people in the HCInt data set made it possible to estimate the variance in performance statistics. For verification performance, error ellipses were computed. This provided an estimate of the change in verification when different people are in the gallery. Knowledge of the amount of variation is necessary for understanding the predictive value of an evaluation. If the error ellipses are large, the actual performance of a deployed system could be substantially different than performance predicted in an evaluation. If this is the case, the predictive value of the evaluation is small. Also, for applications in which a system will be deployed at multiple locations, the ellipses can provide an estimate of the range of performances that will be observed over the multiple locations. From a scientific point of view, error ellipses are one method of investigating the natural variation of faces in the population. This is because part of the variation that contributes to the error ellipses is generated by the natural variation in the population.

One of the primary purposes of the evaluation was to assess the state-of-the-art in face recognition. The HCInt provided a robust assessment of performance on digital visa images. The visa images were taken indoors under controlled lighting conditions. While the HCInt provides an assessment of performance for one category of images, the MCInt provides an assessment over a wider range of imaging conditions.

The results from the HCInt provide a baseline to calibrate the difficulty of image categories in the MCInt. The primary performance measure for comparing results from different experiments is the verification rate at a false accept rate of 0.01. Verification performance was selected because it is independent of gallery size. A specific false accept rate was selected because it makes the comparison manageable and is representative of a real-world operational setting.

To confirm that it is possible to compare results from the HCInt with the MCInt, a set of 1,024 visa images of 512 individuals was included in the MCInt. Verification results from the visa images in the MCInt and HCInt are comparable and consistent. Performance on the MCInt visa image experiment is within the error ellipses on the HCInt data.

We first examined the probe categories taken indoors. The frontal facial image experiment examined three categories of probes that were taken indoors: same-day different expression, same-day different illumination, and different day. Verification performance on the MCInt different-day probe set is comparable to the performance on the HCInt visa images. The top three performers do not have the same verification performance on the HCInt and MCInt indoor different day probe sets. However, the top three verification scores, without consideration of which participant produced the score, are comparable.

For the better systems, performance on the same-day expression and overhead lighting probe sets is comparable. The difference between these probe sets is primarily lighting. The indoor-different day and visa probe sets are comparable. Here again, the primary difference is lighting. This suggests that the best face recognition systems are not sensitive to normal indoor lighting changes.

Next we examined the effect of faces taken outdoors on performance. Compared to the indoor probe set, there was a large decrease in performance on the outdoor probe categories—same day and different day. For the better systems, the change in performance going from indoor to outdoor probes was much greater than the change going from same day to different-day probes. We also ran the experiment with a gallery consisting of outdoor images. We then measured performance on same-day and different-day outdoor probes. Performance was comparable to the results reported for the indoor gallery and different-day outdoor probe set (figure 22). This suggests that the main cause of the drop in performance from indoor to outdoor imagery is that the images were taken outdoors. MCInt recognition results on outdoor probe sets are consistent with the results of FRVT 2000. This indicates that recognition of faces in outdoor images needs to be a focus of research.

Figure 15 and figure 16 show the results for the HCInt time-lapse experiment. Identification performance dropped off roughly linearly in the time elapsed between acquisition of the gallery and probe images. Identification performance drops are approximately 0.05 points per year. For verification, performance dropped off more slowly than for identification. The next step was to make sure that the drop-off in performance was not confounded with the sex effect. To address this, performance was broken out by sex for each elapsed time bin. The results were consistent with previous results (figure 15 and figure 16). This shows that the time difference results and sex of the subject are not confounded.

FRVT 2002 looked at three new face recognition techniques: three dimensional morphable models, normalization of similarity scores, and face recognition from video sequences. The results of preprocessing non-frontal images with three-dimensional morphable models are shown in figure 23 and figure 29. The effects of pose changes in looking right and left, and up and down were examined. For all but one system, Iconquest, preprocessing with three-dimensional morphable models produced a substantial increase in performance.

Participants were not aware that the MCInt would include non-frontal images that had been preprocessed by three-dimensional morphable models. Therefore, the systems were not tuned to artifacts that the morphing may have introduced into the images, and the morphing technique was

not tuned to any of the systems. These results provide a baseline for improvements in recognition of non-frontal images after preprocessing by a morphable model.

Normalization is a post-processing procedure that adjusts for variations in the composition of galleries. Four of the evaluation participants submitted normalization routines. FRVT 2002 results show that for identification, normalization did not improve performance. For verification, three of the normalization routines increased performance; but the amount of increase depended on the false accept rate. Similar results were found for verification in the watch list experiments. Our results are consistent with those found by Mansfield et al. (2001) for fingerprints.

The third technique investigated was face recognition from video sequences. There are two theories supporting the hypothesis that face recognition from video is better than stills. The first is that there are more images of person acquired. The second is that video techniques can explicitly incorporate temporal information in a sequence the representation of a face. Video sequences can potentially improve the performance of a system at two stages. One is at the detection and location stage. A test of this stage would involve video where the face size and angle would vary through the video sequence. The other is at the recognition stage. FRVT 2002 examined performance at the recognition stage. To test recognition at this stage, the probe video sequences contained a face of a person talking—sometimes expressively. Each frame in the video sequence was the quality of a mugshot image. The difference between a mugshot image and the frames in the video was that the collection of frames contained a greater range of expressions and motions. By testing systems on high quality facial imagery, FRVT 2002 tested the recognition ability from video. If the sequences had contained people walking towards the camera where the face varied in size and orientation, the results would have confounded detection and recognition.

FRVT 2002 results show that with the exception of DreamMirh and VisionSphere, recognition performance did not improve using video. While DreamMirh and VisionSphere performance was better on video, their performance from video sequences was significantly worse than the recognition rate on stills for the best systems. Our results show that for the FRVT 2002 video sequences, and the system tested, pure recognition from video does not improve performance.

Previous evaluations have broken out performance by imaging conditions, but have not previously examined the effects of demographics on performance. The HCInt evaluation showed that demographics can have a significant effect on performance. The HCInt results showed that the sex and age of a person affect performance. The effect of race on performance was not examined because the overwhelming majority of faces in the HCInt data set were Mexican. It is known that the race of a person does affect algorithm performance (Furl et al. 2002). The second largest racial group in the HCInt data consisted of 629 people born in China. Because of the limited size of this population, we were not able to perform a detailed study. However, results based on the limited amount of data suggest that people of Chinese origin are easier to recognize than people born in Mexico. Thus, race is another demographic factor that needs to be considered when fielding a system.

It is important to differentiate between two types of effects of covariates on performance. The first examines the effects on a specific population or set of users of a biometric system. In the HCInt data set, FRVT 2002 measured the performance of a population that was roughly equally divided between male and female, contained significantly more young people than older people, and consisted mostly of Mexicans. The performance would have been different if the data set had consisted of primarily males in their twenties. Because of the observed effect of covariates on performance, it is recommended that prior to fielding a system, the performance for that system is estimated for the demographic make-up of the user population.

The second type of effect of covariates consists of fundamental properties of the systems or images themselves. The most prominent covariate result in the HCInt evaluations is that males are easier to recognize than females. Is the difference in performance a fundamental property of male

and female faces, or is it a function of ratio of male to females in the HCInt data set? To provide insight into this question, we ran a number of additional experiments on balanced data sets. A balanced data set consists of an equal number of gallery images from each covariate. By balancing on the covariate, we control for the effect of the covariate in the data set.

Two experiments were run to measure the effects of the sex of a person on performance. The first examined sex alone. The other looked at the interaction of sex and age. For the first experiment, two galleries were created of the same size: one all female and one all male. Identification performance for the male-only gallery was higher than the female-only gallery.

In the second experiment, one gallery was created that was balanced for both sex and age. Thus, there were the same number of male and female individuals in each age bin, and all age bins were the same size. The results are consistent with the unbalanced experiment. Both unbalanced and balanced experiments showed a sex effect, an age effect, and an interaction between the two.

One of the lessons of the covariate experiments is that covariates cannot be examined in isolation. The interactions among the covariates need to be examined. If the interactions are not examined, it is possible the observed effects of two covariates are from the same source. For example, it may have been the case the younger age bins had contained significantly more females and the older age bins may have contained significantly more males. If this were the case, the observed age and sex effects would have been caused by the same underling phenomena. But, as was observed in the HCInt data set, the performance gap between males and females declines with age. Without examining the interaction, this effect would not have been found.

There are a number of results in the literature that support the observation that base recognition rates are higher for males and females. Givens et al. (2002) found a small sex effect, and also found a statistically significant age effect. The effects are reported for a principal component-base face recognition algorithm using images from the FERET database. In the Givens et al. study, there were only two age groups: young and old. Our findings differ with those of Gross et al. (2001). Gross et al. found females easier to recognize than males from the AR database (Martinez and Benavente 1998). However, their findings were on a small data set of 130 people.

A related problem in face processing is automatically determining the sex of an unknown face. While there is an extensive literature on this subject, very few papers report classification rates for males and females. Moghaddam and Yang (2002) report male and female classification errors for eight algorithms. Their experiments were conducted on images from the FERET database. For all eight algorithms, the error rates for males were lower than females. Shaknarovich et al. (2002) find similar results on a database of images collected on the World Wide Web.

The results from FRVT 2002 and in the literature provide evidence that automatic recognition tasks are easier for males than for females. The underlying reason that males are easier to recognize is not known. Additional experiments are required to provide an explanation. Possible explanations range from facial hair on men to the general observation that women are more likely to have greater day-to-day variation in their appearance than men. However, follow-up experiments are required to determine the explanation for the bias.

It is known that the training set for a face recognition algorithm does affect its performance on different demographic groups. Systems were trained and tuned prior to starting the FRVT 2002. The FRVT 2002 evaluation protocol did not specify a training set or restrict the training set that a participant could use in training and tuning their system. Thus, the FRVT 2002 did not control for the composition of the training set, and this could have contributed to the covariate effects. However, the covariate effects existed for all eight systems. This suggests that the composition of

the training set is not the only possible source of the covariates effects. Additional experiments are required to determine the contribution of the composition of the training set to covariate effects.

The identification results in figure 11 show that identification performance decreases as gallery size increases. The rate of decrease in gallery size appears to be linear in the logarithm of gallery size. Performance decreases by approximately 0.06 for every order magnitude increase in gallery size (base 10). The one exception is DreamMirh. The log-linear performance in figure 11 is observed up to a gallery of 37,437. Does the log-linear performance continue indefinitely or change at some point? The same trend was observed for performance as a function of gallery in the watch list task (see figure 14). An empirical answer to this question requires experimental runs on even larger galleries. The other challenge is to develop a mathematical model to explain this phenomenon. One current model regards identification from a gallery of size $N$ as $N$ attempts at verification. Wayman (1999) and Daugman (2003) formalized this as a binomial model, where the binomial coefficient is constant for all gallery sizes. This model does not explain the observed data because the model predicts an exponential decrease in performance as gallery size increases. Appendix A.5 presents a more sophisticated model that is based on moments of the match distribution. This model predicts the log-linear behavior, but underestimates the identification rate. The probable cause of the underestimation is that the model assumes that the similarity scores are independent and identically distributed (iid). The iid assumption fails because there are complex interactions among the similarity scores.

The face recognition community, and biometrics in general, has developed a range of evaluations in terms of number of people and images. To provide a rough guide to evaluation size, we introduce the following nomenclature:

- Small: ~1,000 signatures and ~330 individuals
- Medium: ~10,000 signatures and ~3,300 individuals
- Large: ~100,000 signatures and ~33,000 individuals
- Very large: ~1,000,000 signatures and ~330,000 individuals
- Extremely large: ~10,000,000 signatures and ~3,300,000 individuals

Each size has its own role and place. A larger evaluation is not inherently better, especially when cost is considered. Most previous evaluations have been small, but they have had a positive impact on the development and assessment of biometrics.

The MCInt is a small to medium evaluation, and was able to differentiate between large and small effects on performance. For example, the MCInt results showed a large difference in performance between recognition of non-frontal images and non-frontal images that have been morphed. Thus, the MCInt results showed that morphable models improved performance for non-frontal images. In another example, the MCInt results showed a small change in performance for video versus still signatures, and we concluded that video signatures do not improve performance. An evaluation such as MCInt, is good for making an assessment on 1) a specified set of experiments, and 2) where one is looking to distinguish between large and small effects.

The HCInt allowed for a detailed analysis and was able to estimate the variance of performance statistics and measure the effects of covariates on performance. This analysis required not only a large numbers of images and people, but also an appropriate number of errors. If there had only been ten or hundred errors, we would not have been able to perform detailed covariate analysis. In designing very large and extremely large evaluations one needs to state the object of the evaluation and have an idea of the overall accuracy of the biometric being tested. For example, if a biometric has an identification rate of 0.9999 (error rate of one in 10,000), then an evaluation on a data set of 100,000 images would on average produce ten errors. To be able to perform a detailed analysis of performance, such as in HCInt, would require a test set several orders of magnitude larger.

Evaluations of all sizes are needed and have their role in assessing performance of biometrics. Factors effecting the size and design of an evaluation include the evaluations goals and the overall accuracy of a biometric. The greater the accuracy of biometric, the larger the required size of an evaluation. The more detailed analysis needed, the larger the required size of an evaluation. At the other end of the scale, an evaluation with very specific and defined purposes maybe able to meets its goals with a small evaluation.

## 10. Conclusion

At a simple level, FRVT 2002 was an evaluation and comparison of ten face recognition systems. Upon closer examination, FRVT 2002 will have a much broader impact. From an operational perspective, FRVT 2002 results will impact policy, the engineering design of large-scale biometric systems, and how future technology, scenario, and operation evaluations will be designed. From a scientific point of view, FRVT 2002 will have an impact on future directions of research in the computer vision and pattern recognition, psychology, and statistics fields. FRVT 2002 results raise many more questions than they answer.

Before summarizing the findings of FRVT 2002, two potentially important issues need to be addressed:

1) Does face recognition work?
2) Which system is best for my application?

The answers to both of these questions are closely related to one another. Face recognition performance, like other biometric types, is application-dependent. Just as there is no best biometric type for all operational applications, there is no best face recognition system for all operational applications. FRVT 2002 was not designed to be a "buyer's guide for face recognition" –where one looks at graphs or scores and selects the best system for installation. Rather, it is a *technology evaluation* that should assist decision-makers in determining (1) if face recognition technology could potentially meet the performance requirements for an operational application, and (2) which systems should be selected for application-specific scenario evaluations.

In order to determine if face recognition works and which system(s) should be deployed, one first needs to properly define the operational application of interest and operational performance requirements. These requirements need to be as specific as possible because even a small change in operational requirements can sometimes significantly alter anticipated performance. Questions to ask when defining an application include:

- Identification, verification or watch list mode of operation?
- The size of the database for identification or watch list?
- Demographics of the anticipated users (age, sex, etc.)?
- Lighting conditions – indoor/outdoor? Supplemental lighting?
- Is the system to be installed overtly or covertly?
- What is the anticipated user behavior?
- How long has it been since the images in the database were taken?
- What is the required throughput rate?
- How many "exception handling" cases can you handle for a given period of time?
- For each mode of operation, which parameter (identification: rank or identification rate; verification: false alarm or probability of verification; watch list: false alarm or correct alarm) is most vital?
- What are the minimum accuracy requirements?

FRVT 2002 can only provide input to several, but not all of these questions. Questions associated with anticipated user behavior, exception handling, human computer interaction, and how a system is integrated into the business model are not addressed in a technology evaluation such as FRVT 2002. Providing answers to these types of questions are the province of scenario

and operational evaluations. Answers to some of these questions will identify which experiments in FRVT 2002 are relevant to a given application. Results from the relevant experiments will 1) show if face recognition could potentially meet the performance requirements for the application, 2) identify which systems should be selected for follow-up scenario evaluations, and 3) provide a starting point for designing and conducting scenario and operational evaluations for a specific application. Without specifying requirements, implementation constraints, and process models for an application, one cannot accurately determine if face recognition will work or which system should be selected.

FRVT 2002 is the most thorough and comprehensive evaluation of automatic face recognition technology to date. The evaluation has examined many long-standing questions and raised several new questions for further study. These are discussed in detail in Section 9. The conclusions from FRVT 2002 are summarized below:

- Indoor face recognition performance has substantially improved since FRVT 2000.
- Face recognition performance decreases approximately linearly with elapsed time gallery and probe images.
- Better face recognition systems do not appear to be sensitive to normal indoor lighting changes.
- Three-dimensional morphable models substantially improve the ability to recognize non-frontal faces.
- Normalization improves verification and watch list performance.
- On FRVT 2002 imagery, recognition from video sequences was not better than from still images.
- Males are easier to recognize than females.
- Younger people are harder to recognize than older people.
- Outdoor face recognition performance needs improvement.
- For identification and watch list tasks, performance decreases linearly in the logarithm of the gallery size.

One of the goals of a technology evaluation is to identify future directions of research. Among the research directions identified by FRVT 2002 are:

- Recognition from outdoor facial images.
- Recognition from non-frontal facial images.
- Recognition at low false accept/alarm rates.
- Understanding why males are easier to recognize than females.
- Greater understanding of the effects of demographic factors on performance.
- Development of better statistical methods for understanding performance.
- Develop improved models for predicting identification performance on very large galleries.
- Effect of algorithm and system training on covariate performance.
- Integration of morphable models into face recognition performance.
- Understanding the video sequences in FRVT 2002 did not improve performance.

Other major FRVT 2002 accomplishments include the evaluation protocol developed for this test and the associated scoring suite. The evaluation protocol and scoring suite are XML-based. They were designed to be applicable to general biometric evaluations, not just restricted for use in face recognition evaluations.

Face recognition and processing are important research problems spanning numerous fields and disciplines. This is because face recognition, in addition to having numerous practical applications, is a fundamental human behavior that is essential for effective communications and interactions among people. Researchers are interested in how people process faces, and scientists and engineers are working on techniques to replicate the human face processing functions. Research advances along two intertwined paths. One path has an application orientation and the other, a scientific orientation. Advances on both paths reinforce each other, with FRVT 2002 providing

research directions for both paths. In the 1990's, the FERET evaluations stimulated research in face recognition technology and in doing so helped advancing automatic face recognition during its infancy. With the numerous questions it raises, FRVT 2002 is poised to play a similar role in stimulating future face recognition and processing research.

## Acknowledgements

The organizers of the FRVT 2002 gratefully acknowledge the Defense Advanced Research Projects Agency, Department of State, Federal Bureau of Investigation, National Institute of Justice, National Institute of Standards and Technology, and Transportation Security Administration as evaluation sponsors and the ONDCP Counterdrug Technology Assessment Center, United States Customs Service, Department of Energy, Drug Enforcement Administration, Immigration and Naturalization Service, U.S. Secret Service, Technical Support Working Group, Australian Customs, Canadian Passport Office, and United Kingdom Biometric Working Group as evaluation supporters.

The authors extend their thanks to:

The Department of State, specifically Travis Farris for allowing NIST to use the Mexican nonimmigrant visa images for FRVT 2002; John Atkins and Rasool Azad for their invaluable assistance with the images, meta-data, and background information on the images' origins and properties.

Volker Blanz and Thomas Vetter at the University of Freiburg for supplying us with the three-dimensional morphable images. Volker Blanz expeditiously provided us with frontal reconstructions obtained from their 3D Morphable Model implementation.

Alice O'Toole at The University of Texas at Dallas, for supplying us with the video sequence database.

Karen Marshall at NIST for manually classifying the sex of the individuals in the HCInt subset. Also, for the flawless preparation of the media used for conveying the images and similarity files to and from the vendors during the testing phase.

Sam Trahan at NIST for the maintenance of the alternative scoring code branch, and for authoring the sex classification tool.

Tom Gandy and Cathy Schott for their general assistance and their assistance in editing and proofreading the FRVT 2002 reports.

Michael Garris at NIST for his careful review and comments on the XML specification.

Charlie Wilson at NIST for his general assistance and insightful comments.

Kevin Bowyer, Travis Farris, and Russ Neuman for their comments on preliminary drafts of the report.

# Appendix A

## A.1 FRVT 2002 Evaluation Protocol Details

The FRVT 2002 evaluation protocol is a general framework for conducting technology biometric evaluations. The details presented are for FRVT 2002, but the framework is applicable to biometrics in general. The protocol encodes the philosophy and design of the evaluation, specifies the properties of the input to and output from a system, and dictates how performance statistics and results should be computed. The properties of the protocol are:

- All participants are tested in the same manner.
- Training is completed prior to the start of the evaluation.
- Biometric signatures are completely general.
- The input to a system is a target and query set.
- The output is a complete similarity matrix.
- All similarity scores $s(q,t)$ are a function of a query/target pair $(q,t)$.
- A gallery can be any arbitrary subset of the target set with at most one biometric signature per person in the gallery.
- Normalization is a post-processing function.
- Normalization is a function of a gallery and a probe.
- All performance scores are computed from sub-matrices of the complete similarity matrix.

In the FRVT 2002 evaluation protocol, a system is given two sets of biometric signatures: a *target* and a *query* set. A target set $T$ contains the set of signatures that are known to a system. A query set $Q$ contains signatures of unknown identity that are to be recognized. The term 'recognition' covers identification, verification, and watch list tasks. In FRVT 2002, biometric signatures are either digital still images or digital video sequences. The FRVT 2002 evaluation protocol is general enough to be extensible to general biometrics signatures.

For each pair of query and target signatures $(q,t) \in Q \times T$, a system reports a *similarity score* $s(q,t)$. A similarity score is a measure of how similar two biometric signatures are. The FRVT 2002 evaluation protocol was designed with the assumption that each system has its own similarity measure. Systems that report *distances* can be incorporated into the scoring framework by negating their similarity scores.

The complete set of similarity scores over $Q \times T$ is a *similarity matrix*. The key property of the FRVT 2002 evaluation protocol, which allows for great flexibility in scoring, is that for any target-query pair, $(q,t)$ we have $s(q,t)$. From a full similarity matrix, "virtual" experiments are performed from subsets of $Q$ and $T$. The basic components of such a 'virtual' experiment are a *gallery* and a *probe* set. In traditional experiments, such as those in the FERET evaluations and FRVT 2000, an experiment consisted of a single gallery $G$ and probe set $P$. A gallery $G$ is a subset of a target set $T$ and a probe set is a subset of a query set $Q$. For a virtual experiment, performance scores are computed from the similarity scores in $P \times G$. The scores in $P \times G$ are extracted from $Q \times T$.

Operationally, it was expected (but not required) that a FRVT 2002 participant generated the similarity matrix as follows. Given a query element $q$, the system compared $q$ with all targets $t \in T$ to generate a vector, or single column of the similarity matrix. We denote this vector as $s(q,T)$. This process is repeated for all query signatures to generate the full similarity matrix. We denote this matrix as $s(Q,T)$.

In summary, we have,

- $s(q,t)$, a single scalar value, representing the similarity between query element $q$ and target element $t$,
- $s(q,T)$, a vector of $|T|$ elements, representing the similarity between query element $q$ and each target element $t \in T$, and
- $s(Q,T)$, a matrix of $|Q||T|$ elements, representing the similarity between every pair of target and query signatures, $(q,t) \in Q \times T$.

In constructing galleries from a target set, a gallery must contain only one biometric signature for each person. This rule does not present a limitation for scoring, but may require additional planning in designing target and query sets. The most common objection to this rule is creating a gallery that contains multiple images of an individual. One method to put multiple images of a person in the gallery is to place $n$ images of a person in the gallery as $n$ separate biometric signatures. When a probe is presented to a system, this will produce $n$ similarity scores between the probe and the images of this person in the gallery. During scoring, the question is how to make a decision based on the $n$ similarity scores? Does the scoring algorithm take the smallest, largest, or maybe the mean? The trouble with this approach is that the scoring code resolves the issues, not the system being evaluated. The approach taken by the FRVT 2002 protocol is to create a single biometric signature $t$ that contains the $n$ images. Thus, for a probe, a single similarity score is returned and the system decides how to integrate the information contained in the multiple images. In the FRVT 2002 protocol, a biometric signature is a list of data in a specified format. The data in a biometric signature can be heterogeneous. For example, it could consist of a facial image and a fingerprint.

Using gallery and probes allows us to compute performance measures for different categories of images. For example, FRVT 2002 included: (1) probe and galleries that varied according to the subject's age, (2) measurement of the empirical variation of the verification rate and false alarm rate across disjoint probe and gallery sets, (3) the effect of gallery size within a watch list scenario, and so on. Naturally, the possibilities for different experiments are limitless.

At the request of an FRVT 2002 participant, all participants had the opportunity to provide a custom *normalization* function as a component of their system. Operationally, a normalization function is applied to an appropriate subset of the score matrix, just before the scoring algorithms are applied. In its most general form, normalization is any post-processing transform performed on a subset of the similarity matrix. For an algorithm that uses normalization, the final performance scores are computed over the *transformed* values.

In FRVT 2002, we distinguish between two families of normalization functions. One possible normalization function, say $f_1$, takes each $s(p,G)$ as input, and gives as output, a new set of similarity, or "normalized" scores, $s_1(p,G;G) = f_1(s(p,G))$. Here, we use a semi-colon to emphasize that $s_1(p,G;G)$ is a vector of similarity scores that is dependent or *parameterized* by a particular gallery. Recall that $s(p,G)$ and $s_1(p,G;G)$ are all the similarity scores for a *given probe*. This means that normalization occurs on a probe-by-probe basis, and is not a function of the entire similarity matrix $s(P,G)$. If the results for each probe are joined together, we have a new similarity score matrix $s_1(P,G;G)$, where each column of the matrix has been normalized with respect to the gallery.

The second type of normalization function $f_2$, which also operates on a probe-by-probe basis, incorporates the similarity scores between all pairs of *gallery* images[3] as well as the vector of similarity scores $s(p,G)$. That is, $s_2(p,G;G) = f_2(s(p,G), s(G,G))$. Note that if $|G|$ is large, it may not be practical to use such an algorithm, since the matrix $s(G,G)$ may exceed memory limits.

3. This implies that in order to use such a normalization routine, the target and query sets must share a set of (gallery) images.

## A.2 STATISTICS DETAILS

In FRVT 2002, we evaluate an algorithm on *three* related tasks: *identification, verification,* and *watch list*. Each of these tasks can be mapped to sets of operations on the similarity scores.

For the *identification* task, the operational model for FRVT 2002 is a closed universe. That is, for each probe $p \in P$ there exists one (and only one) gallery signature $g^* \in G$ such that signatures $p$ and $g^*$ are of the same subject. In the remainder of this Appendix, we will use $g^*$ to represent the mate of probe $p$.

During identification, for a probe $p \in P$, a system reports the gallery elements that have the highest similarities to the probe image. If the subject appears in the top n candidate gallery images, then the subject is correctly identified. For each probe $p \in P$, we sort $s(p,G)$. A probe $p$ is said to be of *rank k* if $s(p,g^*)$ is the $k^{\text{th}}$ largest similarity score in $s(p,G)$. The definition of *correct identification* can therefore be parameterized with respect to the rank $k$. Over all probes, let $R_k$ denote the number of probes in the top $k$. Then, $R_k/|P|$ is the fraction of probes in the top $k$.

So far, our scoring algorithm assumed that there are no ties among the similarity scores. To handle cases where $s(p,g^*)$ is tied with other values, the FRVT 2002 scoring software reports the *mean* value of two different ranks—the *optimistic* and *pessimistic* rank. The *optimistic rank* is (one plus) the number of similarity scores in $s(p,G)$ that are *strictly* greater than $s(p,g^*)$. The *pessimistic* rank is the number of similarity scores in $s(p,G)$ that are greater to *or equal* to $s(p,g^*)$. For example, suppose that for a particular probe, an algorithm reported identical similarity scores for each element in the gallery. Then, the optimistic rank would be one, since there would be *zero* scores strictly greater than $s(p,g^*)$. The pessimistic rank would be the same size as the gallery, $|G|$ , since in this case, *all* $|G|$ scores would be equal to or greater than $s(p,g^*)$. The final reported rank would be the average of the two ranks, or $(1 + |G|)/2$.

For the *verification* task, the operational model for FRVT 2002 is as follows. A system performs a *verification* when a probe $p$ is presented to a system along with a claim of identity. The system computes the similarity $s(p,g^*)$ where $g^*$ is the stored signature corresponding to the claimed identity. Naturally, for evaluation purposes, it is assumed that the $g^*$ exists—i.e., a subject does not claim to be someone *not* in the gallery. The claim of identity is *accepted* if $s(p,g^*) \geq \tau$, where $\tau$ is some *a priori* operating threshold.[4]

Performance statistics for the verification tasks are computed from collections of *match* and *non-match* scores. A *match score* is any similarity score generated by comparing a probe and gallery element from the *same* subject. Similarly, a *non-match* score is a similarity measure between signatures of *different* subjects.

A verification experiment requires three sets to be specified. The gallery $G$, a set of probes $P_G$ that contain probes of people who are in the gallery. The set $P_G$ determines the set of match scores—i.e., if $P_G$ represents the set of probes with facial imagery in the gallery, then all the *match scores* in the similarity matrix $s(G,P_G)$ may be used to calculate the verification rate. Let $P_N$ represent the set of probes that are used to generate the *non-match* scores. The false accept rate is computed from all the *non-match* scores in the similarity matrix $s(G,P_N)$. Unlike $P_G$, which, by definition of the verification problem must have corresponding gallery images, there is no such requirement for $P_N$. The number of non-match scores varies according on the overlap among the *subjects* in the gallery and $P_N$. If $P_G$ and $P_N$ are the same set, then the number of non-matches is $|G| \cdot |P_G| - |P_G|$. This was the case in the 'round-robin' method of computing verification performance in Phillips et al (2000). When $P_N$ does not contain any people in the gallery, then the number of non-matches is $|G| \cdot |P_N|$.

In the traditional 'round-robin' evaluations, $P_G$ and $P_N$ are often the same set. From an operational standpoint, this models the case where a subject, already with legitimate access to

4. This inclusive (as opposed to exclusive) method of accepting or rejecting a claim is compatible with a Neyman-Pearson observer. A Neyman-Pearson model maximizes the verification rate for a fixed false accept rate (Bickel and Doksum 1977; Egan 1975).

the system (they are in $P_G$), attempts to gain access to the very same system, *under a different identity*. There may be some specialized scenarios where this is a valid model. However, we prefer to model the situation in which a person who does *not* already have access to the system makes a false verification attempt. In this model, the *people* (not just the signatures) in the probe set $P_N$ are different from the people in the gallery. The people in $P_N$ are sometimes referred to as *true imposters*. The rationale for having the non-match scores generated by *true imposters* is that non-match distributions from $s(G,P_G)$ may be different than those from $s(G,P_N)$.

From the match and non-match scores, the verification and false accept rates are respectively computed. The verification rate is calculated by dividing the number of match scores above the threshold $\tau$, $M_\tau$ by the total number of match scores $M$. For the threshold $\tau$, the verification rate (VR) is $M_\tau/M$. The false accept rate is computed in a similar manner. If the total number of non-match scores is $N$ and the number of non-match scores greater than or equal to $\tau$, then $N_\tau/N$ is the false accept rate (FAR).

We first present a simple algorithm for computing an ROC, and then streamline it into a faster version. In our implementation, we assume that $M \ll N$, and that $N = O(M^2)$. Strictly speaking, an ROC is not a 'curve' but a collection of operating points where the verification or false alarm rate changes.

The first step in the ROC algorithm is to sort all $M + N$ match and non-match scores. In computing the corresponding ROC, one only needs to compute performance at thresholds equal to *unique* values of these $M + N$ scores. Let $\tau_i$ be the ith largest similarity score—this can be either a match or nonmatch score. The verification and false accept rates are computed at each threshold $\tau_i$, by starting with $\tau_1$ (the largest score) and proceeding down the sorted list of similarity scores in ascending order.

The algorithm has some (simple) initial conditions. We add the artificial threshold $\tau_0 = \infty$, and initialize an array of match and non-match counters $M_{\tau_i}$ and $N_{\tau_i}$ where we keep a match and non-match counter for each threshold. This initial condition corresponds to: 1) rejecting all identity claims, and 2) the operating point of both the verification and false accept rate are zero. The algorithm loops over each threshold, $\tau_i$, in order. For each threshold, the number of match and non-match scores with a similarity score in the range $[\tau_i, \tau_{i+1})$ is recorded. If $m_i(n_i)$ is the number of match (non-match) scores with values in this range, then $M_{\tau_i}(N_{\tau_i})$ are simply $M_{\tau_i} = M_{\tau_{(i-1)}} + m_{\tau_i}(N_{\tau_i} = N_{\tau_{(i-1)}} + n_{\tau_i})$.

As stated, this algorithm computes the verification and false accept rate at more points than necessary. To show this, let $\tau_i$ and $\tau_{(i-1)}$ be two adjacent thresholds that each have one non-match score and zero match scores. For both thresholds, the verification rate will be the same, but the false accept rate (FAR) will be *lower* at $\tau_{(i-1)}$ than at $\tau_i$. Therefore, operationally, one would run the system at the operating point associated with $\tau_{(i-1)}$. To generalize this, performance only needs to be computed at the operating points $\tau_i$ corresponding to *match* scores. A scoring algorithm adjusting for this observation proceeds by sorting the unique match scores and using these for the set of thresholds that define the match and non-match score counts.

This algorithm computes an ROC for a single algorithm on a single set of data. In FRVT 2002, we wish to examine how a system performance varies with different gallery or probe sets. To study this variation, it is necessary to combine results over a set of ROCs. One method to accomplish this is to measure the variation on the verification rate for each false alarm rate. This is appropriate for combining ROCs from different systems, because it is not possible (nor is it operationally feasible) to set a uniform threshold across different systems (they would be tuned individually). However, for the same system, it is possible to set one threshold across all galleries and probe sets. For each *threshold*, it is possible to compute the variation in the verification and false alarm rates.

To be able to compute variation in verification and false accept rates across multiple galleries and probe sets, we need to again modify our ROC algorithm. Our goal is to combine the ROCs from $R$ experiments, where each experiment consists of a gallery, probe set for matches, and probe set for nonmatches. The first step in computing the combined ROC is to combine and sort all $R$ sets of *match* scores to use for our thresholds. Here, contiguous thresholds may correspond to match scores from different experiments. This is not a problem, however, since we are using these thresholds to 'oversample' each of the individual ROCs.

In our variance analysis, we select 'evenly spaced' thresholds. Each threshold generates $R$ operating points—there is a (VR,FAR) pair for each of the $R$ experiments. In the FRVT 2002 analysis, these individual points are plotted, along with an error ellipse that traces two standard errors in the TAR and FAR dimensions. From a linear algebra standpoint, the principal axes of the error ellipse are the eigenvectors of the covariance matrix of the $R$ operating points.

The operational model of the *watch list* scenario is as follows. In some sense, the watch scenario is a generalization of both identification and verification. In the watch list scenario, a probe $p$ is presented to a system. The system then compares the probe to a gallery, which plays the role of the *watch list*. The system then: a) determines if the person is on the watch list and b) produces an estimate of the identity of the person. Alternately, the system can report the top $n$ matches if there are sufficient matches above a threshold.

More formally, suppose a watch list system is presented with a probe $p$ of a subject that is on the watch list. For a correct result, a probe must pass a verification and an identification requirement. A similarity threshold parameterizes the verification requirement $\tau$ and a rank $k$ parameterizes the identification requirement. The *identification* requirement is fulfilled if the rank of probe $p$ is $k$ or better (lower). Similarly, the detection requirement is met if $s(p,g^*) \geq \tau$. Both requirements must be met for a correct watch list detection and identification. If the probe is *not* on the watch list, then the probe gives a *false alarm* if there is *any* gallery (a.k.a. watch list) element $g \in G$ such that $s(p,g^*) \geq \tau$. This is akin to someone *not* on the watch list being "similar" enough to someone on the watch list to warrant a 'false alarm.' For a watch list, the *detection and identification* rate is the fraction of probes (also on the watch list) that are detected and identified correctly. The *false alarm* rate is the fraction of probes (*without* corresponding gallery elements) that have a similarity to *any* gallery element that is greater than the threshold.

## A.3 HCINT ADDENDUM

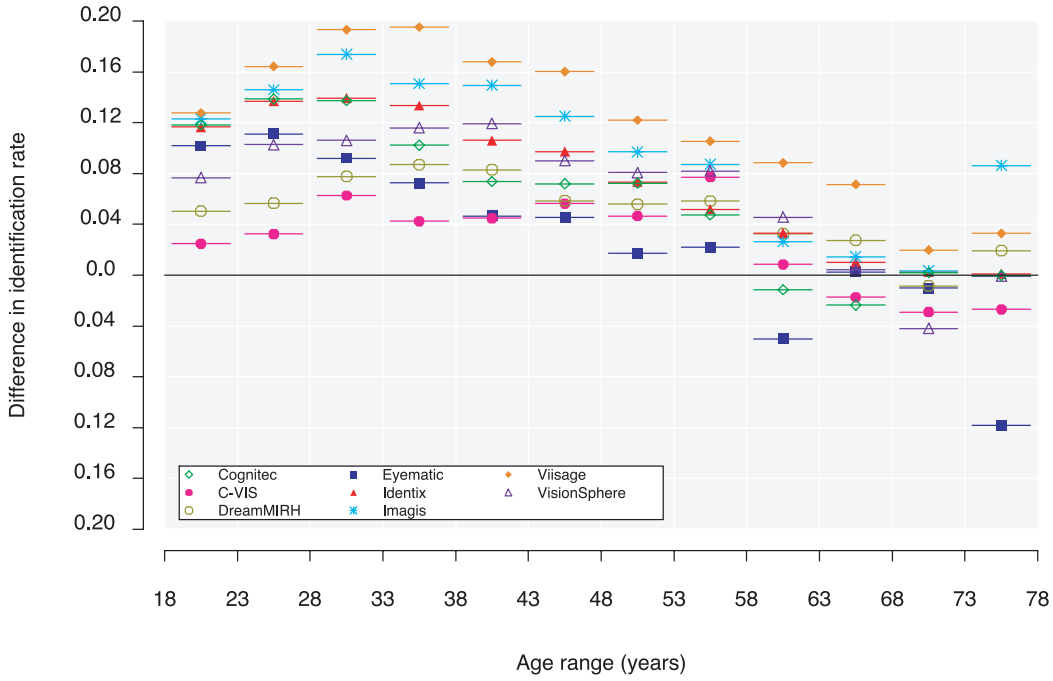


**Fig. 27. Interaction between age and sex for rank 1 identification on HCInt large gallery (for all eight participants). Rank 1 performance for males minus females is plotted for each age bin.**

## A.4 MCINT ADDENDUM

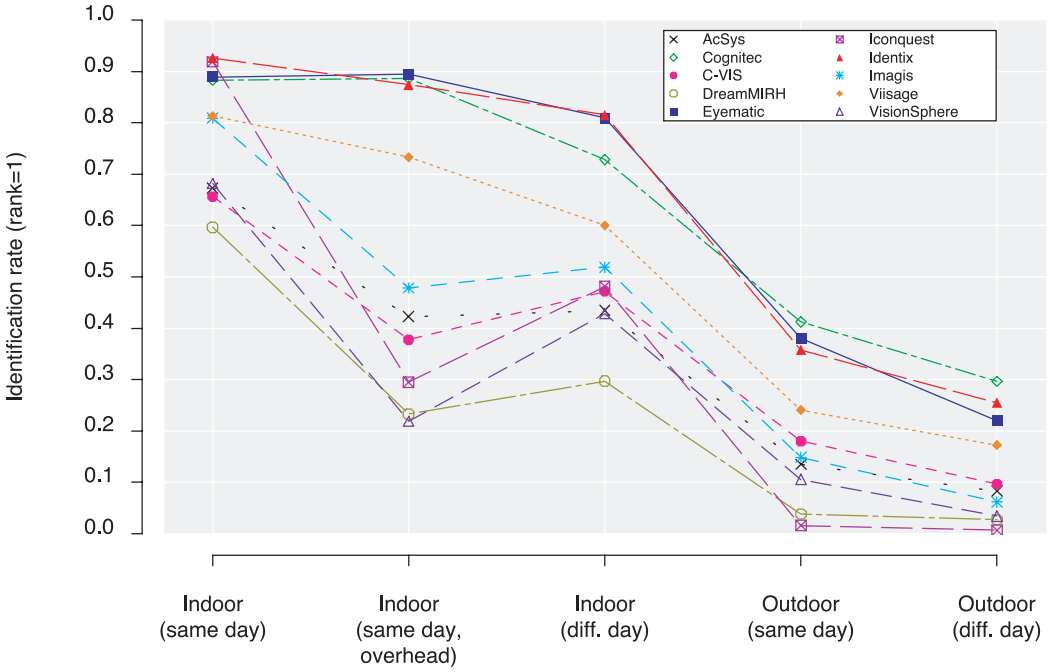


**Fig. 28. Comparison of system performance of different categories of probes. The rank 1 identification rate is plotted.**

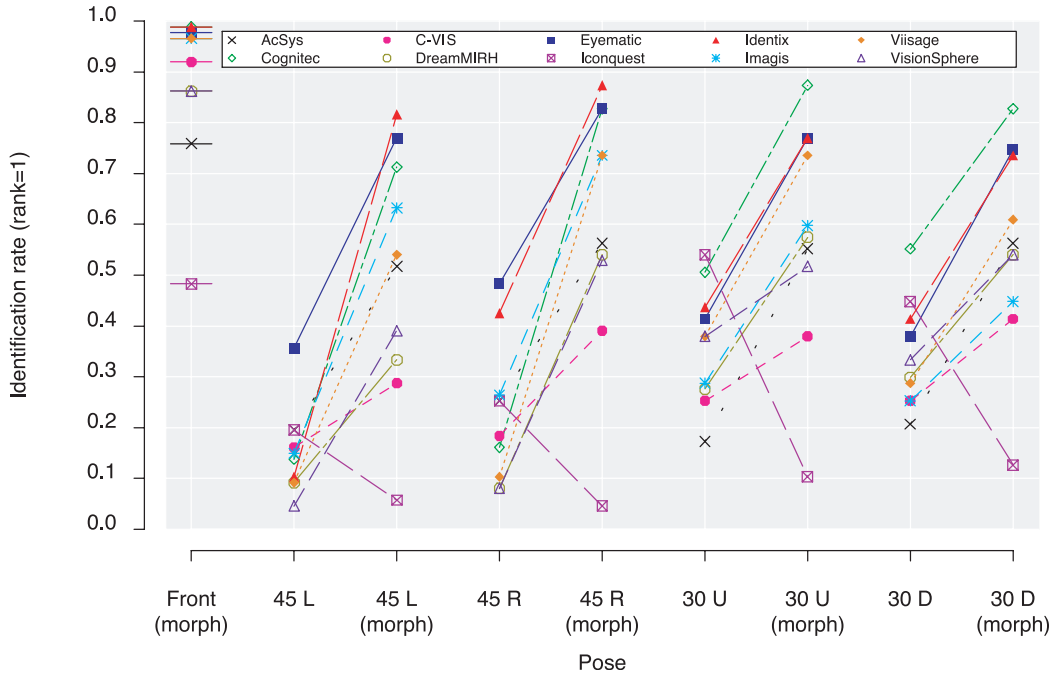


**Fig. 29. The effect of still versus three-dimensional morphable models. The rank 1 identification rate is plotted.**

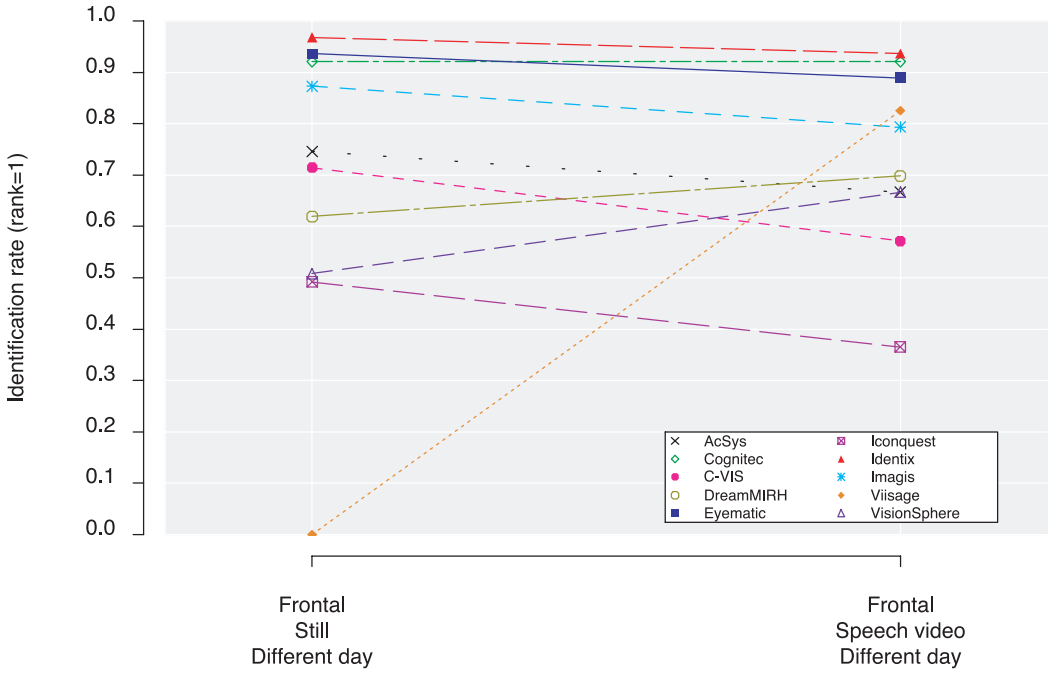


**Fig. 30. Plot showing still versus video recognition for frontal imagery. The rank 1 identification rate is plotted.**

## A.5 PERFORMANCE ON VERY LARGE POPULATIONS

An important question is how recognition performance decreases as the gallery size increases. In HCInt, it was empirically observed rank-one identification rate drops linearly with the logarithm of the gallery size (Figure 11). Our goal was to develop a model that explains the log-linear behavior from the match and non-match distributions. We formalize the model of the log-linear behavior by $P_I(G) = 1 - \alpha \log G$.

Gallery size is denoted by $G$, $P_I(G)$ is rank 1 performance for gallery size $G$, and $\alpha$ is a system performance parameter. The difference among the systems tested is the parameter $\alpha$. Three different models are discussed in this section, and their prediction for the performance for Eyematic is shown in Figure 31. The dashed red line in Figure 31 is Eyematic's actual performance as a function of gallery size.

A lower bound is $P_I(G) = 1/G$, which is the performance of randomly guessing the identity of a probe. Wayman (1999) and Daugman (2002) present models that model identification as $N$ verification attempts. Both present a binomial model[5],[6] $P_I(G) = (1 - P_F(\tau))^{G-1}$.

There are four problems with this model. First, $P_F(\tau)$ is function of $\tau$, and they do not specify how $\tau$ is selected. Figure 31 plots the binomial model for three values of $P_F$: 0.1, 0.01, and 0.001. The binomial model underestimates actual performance dramatically for anything other than small $G$ or very small $P_F$. To obtain the observed performance for the Eyematic system on the HCInt large gallery, i.e., $P_I(37{,}437) = 0.65$, the needed value is $P_F = 1.15 \times 10^{-5}$. In addition, the shape of the curve is not representative. Second, the model does not include the match distribution. The prediction is the same for all match distributions. Third, the model does not fit the observed data. Fourth, asymptotically the random algorithm is a better predictor. The expected performance of randomly guessing the identity of a probe is $P_I(G) = 1/G$. For a given $P_F(\tau)$, there is sufficiently large such for

$$1/G > (1 - P_F(\tau))^{G-1}$$

In this appendix, we present a *moment* model for predicting identification performance as a function of gallery size. The model is a function of both the match and non-match distribution. The moment model is more accurate than proceeding models. The model predicts the log-linear behavior, however, as a rule, it under estimates performance. Before proceeding, we introduce some notation. The probability density function (pdf) of the match distribution is $m(\tau)$ and the cumulative density function (cdf) of the non-match distribution is $N(\tau)$. The model is presented for similarity scores; therefore, we assume that the match distribution is to the right of the non-match distribution. For an operating threshold $\tau$, $N(\tau)$ is the associated false alarm rate.

We will proceed by discussing the case where the size of the gallery is two. For a probe $p$, there is one match score and one non-match score. In the moment model, the match score is sampled from $m(\tau)$ and the non-match score is sampled from $N(\tau)$. We assume that the sampling from the match and non-match distributions is independent. For an operating threshold $\tau$, the probability that the match score will be greater than the non-match score is $N(\tau)$. Since the match score is randomly sampled from $m(\tau)$, the probability that the match score will be greater than the non-match score is

$$P_I(2) = \int_{-\infty}^{\infty} N(\tau) m(\tau) d\tau \qquad (1.1)$$

However, Eq (1.1) is the area under the ROC (AUC) for the verification ROC (see Egan (1975) page 45). In signal detection theory, there is an experiment design called two interval forced choice (2IFC). In this experimental design, an observer is presented with two signals. One signal contains the true signal and the other, noise. The observer must decide which signal contains the true signal and which contains noise. The probability of making the correct decision is the area under the ROC. There is a direct relationship between 2IFC and closed universe identification.

5. See discussion of eq. 13 in Daugman (2003).
6. See discussion around eq. 31 in Wayman (1999). Our penetration rate is 1 and the number of templates is 1.

In identification, the true signal is the correct match, and the noise is the incorrect match. The observer is presented with two matches, and the observer must decide which match is the correct match.

We now proceed to the general case for a gallery of size $G$. For a gallery of size $G$ and an operating threshold $\tau$, the probability that the match score will be greater than all $G-1$ non-match scores is $N(\tau)^{G-1}$. Since the match score is randomly sampled from $m(\tau)$, the probability that all match scores will be greater than the non-match score is

$$P_I(G) = \int_{-\infty}^{\infty} N(\tau)^{G-1} m(\tau)\, d\tau$$

We refer to this as a moment model because for a gallery size $G$, the expected identification rate is a function of the $G-1^{\text{th}}$ moment of $m(\tau)$ about $N(\tau)$.

The 2IFC experimental design can be generalized to *n*IFC. In the generalization, an observer is presented with $n-1$ noise signals and one true signal. The observer must than decide which of the $n$ signals is the true signal. Identification from a gallery of size $G$ is equivalent to a *n*IFC experimental design.

We ran two simulations to see how well the moment model predicted empirical performance. The probability $P_I(G)$ is computed from empirical match and non-match distributions. In our simulations, the match distribution was estimated from 6,000 match scores and the non-match distribution was estimated from 18 million non-match scores. Simulations were run on both non-normalized and normalized scores from Eyematic. The results are in Figure 31. The non-normalized simulation is labeled *non-normalized moment* and the normalized is labeled *normalized moment*. The non-normalized simulation underestimates the empirical slope. Similar results were found for the other participants. However, the normalized simulation fits quiet well with the empirical results. For the other participants that submitted non-normalization functions, the normalized scores were a better fit to the observed data. However, they did not fit the data.

One of the assumptions made in the moment model is that all similarity scores are sampled independently. This is the reason that our model underestimates the identification performance. Interesting, however, the fit is much better for normalized scores. This suggests that one of the features of normalization is that it increases independence among the match and non-match scores.

Next we generalize the moment one more time, to predict performance at rank $k$. Putting the appropriate Bernoulli coefficients into the model, the probability that a probe is at rank $k$ is

$$\int_{-\infty}^{\infty} \binom{G-1}{k-1} (1-N(\tau))^{k-1} N(\tau)^{G-1} m(\tau)\, d\tau,$$

and the performance for gallery size $G$ for rank $k$ is

$$P(G,k) = \sum_{i=1}^{k} \int_{-\infty}^{\infty} \binom{G-1}{i-1} (1-N(\tau))^{i-1} N(\tau)^{G-i} m(\tau)\, d\tau.$$

The availability of a large database allowed a more thorough investigation into the effect of gallery size on performance. To predict the performance, we introduced the moment model and showed its connection with classical signal detection theory. In turn, this connection established a link between verification and identification performance. In addition, the model provides insight into how normalization works. Further research is needed to correct the underestimation in the moment model.

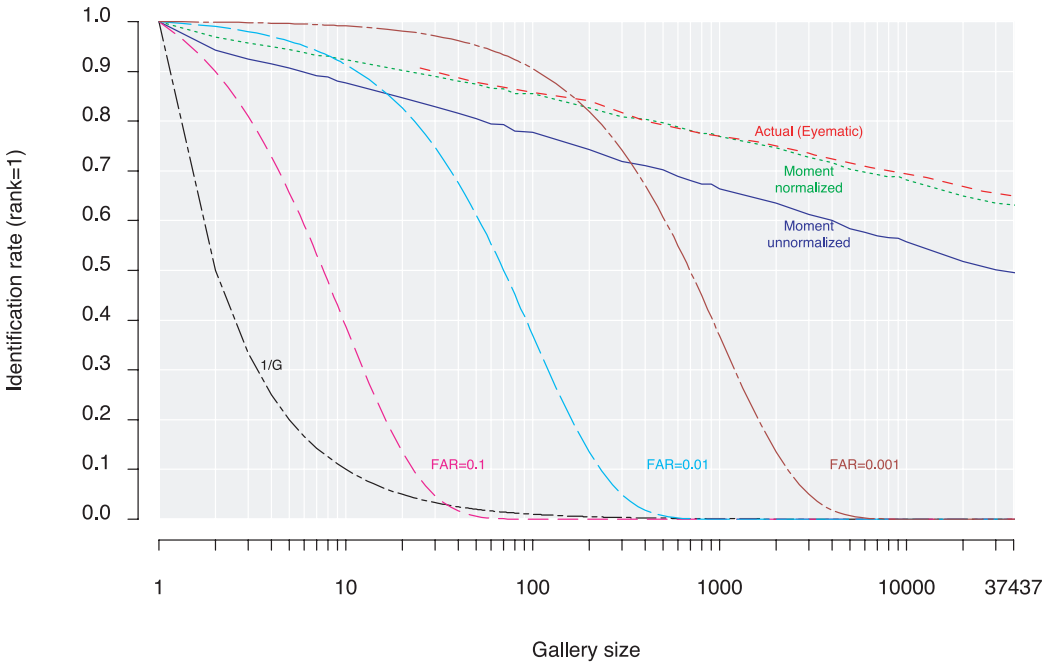


**Fig. 31. Study of identification performance as a function of gallery size.**

## References


Beveridge, J. R., K. She, B. A. Draper, and G. H. Givens. 2001. "A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp 535-542.

Blackburn, D. M., J. M. Bone, and P. J. Phillips. 2001. *FRVT 2000 Report*, Technical Report, http://www.frvt.org.

Blanz, V., and T. Vetter. 1999. "A morphable model for the synthesis of 3D faces," *Computer Graphics Proceeding SIGGRAPH* '99, pp. 187-194.

Blanz,V., S. Romdhami, and T. Vetter. 2002. "Face identification across different poses and illuminations with a 3D morphable model," *Proceedings of International Conference on Automatic Face and Gesture Recognition,* pp. 202-207.

Bickel, P. J., and K. A. Doksum. 1977. *Mathematical Statistics,* Oakland, CA: Holden-Day.

Bone, J. M., and D.M. Blackburn. 2002. "Face Recognition at a Chokepoint - Scenario Evaluation Results," Technical Report, http://www.dodcounterdrug.com/facialrecognition.

Bolle, R. M., N. K. Ratha, and S. Pankanti. 2000. "Evaluating authentication systems using bootstrap confidence intervals," In *Proceedings of 15th International conference on pattern recognition,* pp. 835-841.

Daugman, J. 2003. "The importance of being random: statistical principals of iris recognition," *Pattern Recognition,* Vol. 36, No. 2, pp 279-291.

Egan, J. P. 1975. *Signal detection theory and ROC analysis,* New York: Academic Press.

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

Givens, G., J. R. Beveridge, B. A. Draper, and D. Bolme. 2002. "A statistical assessment of subject factors in PCA recognition of human faces," NIPS Workshop on Statistics for Computational Experiments.

Gross, R., J. Shi, and J. F. Cohn. 2001. "Quo vadis face recognition," *Proceedings of Third workshop on empirical evaluation methods in face recognition.*

MacMillian, N. A., and C. D. Creelman. 1991. *Detection theory: A user's guide,* Cambridge: Cambridge University Press.

Maio, D., D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. 2002a. "FVC 2000: fingerprint verification competition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 24, No. 3, pp. 402-412.

Maio, D., D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain. 2002b. "FVC 2002: second fingerprint verification competition," *Proceedings 16th International Conference on Pattern Recognition,* Vol. 3, pp. 811-814. http://bias.csr.unibo.it/fvc2002/

Mansfield, T., G. Kelly, D. Chandler, and J. Kane. 2001. "Biometric Product Testing Final Report," Technical Report, CESG, http://www.cesg.gov.uk/technology/biometrics/index.htm.

Mansfield, T., and J. Wayman. 2002. "Best practices in testing and reporting performance of biometric devices Version 2.01," Technical Report, National Physical Laboratory, UK, http://www.cesg.gov.uk/technology/biometrics/index.htm.

Martin, A., and M. Przybocki. 2000. "The NIST 1999 speaker recognition evaluation—An overview," *Digital Signal Processing* Vol. 10, pp. 1-18.

Martinez, A. R., and R. Benavente. 1998. *The AR face database,* Technical Report 24, Computer Vision Center, Barcelona, Spain.

Micheals, R. J., and T. Boult. 2001. "Efficient evaluation of classification and recognition systems," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,* Vol. 1, pp 50-57.

Moghaddam, B., and M.-H. Yang. 2002. "Learning gender with support faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 24, No. 5, pp. 707-711.

Moon, H., and P.J. Phillips. 2001. "Computational and performance aspects of PCA-based face recognition algorithms," *Perception,* 30, pp. 301-321.

NIST. 2002. *Use of Technology Standards and Interoperable Databases with Machine-Readable, Tamper-Resistant Travel Documents*. http://www.itl.nist.gov/iad/894.03/fing/fing.html

Phillips, P. J., P. J. Rauss, and S. Der. 1996. "FERET (face recognition technology) recognition algorithm development and test results," Army Research Laboratory technical report, ARL-TR-995. http://www.frvt.org.

Phillips, P. J., H. Wechsler, J. Huang, and P. Rauss. 1998. "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing,* Vol. 16, No. 5, pp. 295-306.

Phillips, P. J., A. Martin, C. L. Wilson, and M. Przybocki. 2000. "An introduction to evaluating biometric systems," *Computer,* Vol. 33, pp. 56-63.

Phillips, P. J., H. Moon, S. Rizvi, and P. Rauss. 2000. "The FERET Evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 22, No. 10.

Phillips, P. J., D. Blackburn, P. Grother, E. Newton, and J. M. Bone. 2003. "Methods for assessing progress in face recognition," In *Biometric Systems: Technology, Design and Performance Evaluation* (eds.) J.L. Wayman, A.K. Jain, D. Maltoni and D. Maio, Surrey, UK: Springer-Verlag London.

Shakhnarovich, G., P Viola, and B. Moghaddam. 2002. "A unified learning framework for real time face detection and classification," *Proceedings Fifth IEEE International Conference on Automatic Face and Gesture Recognition,* pp. 14-21.

Wayman, J. L. 1999. "Error-rate equations for the general biometric system", *IEEE Robotics & Automation Magazine,* Vol. 6, No. 1, pp. 35-48.

Wayman, J. L. ed. 2000. *National biometric test center collected works 1997-2000, Version 1.3,* Technical Report, San Jose State University, http://www.engr.sjsu.edu/biometrics.