FACE RECOGNITION VENDOR TEST 2002

Overview and Summary

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: Overview and Summary

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

[1]DARPA
3701 North Fairfax Drive
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 Road, 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 dataset—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 two techniques for improving face recognition: three-dimensional morphable models, and face recognition from video sequences. Results show that three-dimensional morphable models increases performance, and that face recognition from video sequences offers only a limited increase in performance over still images. For FRVT 2002, a new XML-based evaluation protocol was developed. 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.

1. Introduction & Executive Summary

The Face Recognition Vendor Test (FRVT) 2002 was a large-scale evaluation of automatic face recognition technology. The primary objective of FRVT 2002 was to provide performance measures for assessing the ability of automatic face recognition systems to meet real-world requirements. FRVT 2002 measures performance of the core capabilities of face recognition technology. It provides an assessment of the potential for face recognition technology to meet the requirements for operational applications. However, it does not address many application specific issues and, therefore, is not a "buyer's guide" to face recognition.

FRVT 2002 was an independently administered technology evaluation. Ten participants were evaluated under the direct supervision of the FRVT 2002 organizers at a U.S. Government facility in Dahlgren, Virginia in July and August 2002. Participants were tested using data that they had not previously seen.

The heart of the FRVT 2002 was the high computational intensity test (HCInt). The HCInt consisted of 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[2].

The most likely application of face recognition technology would use images taken indoors. FRVT 2002 results show that normal changes in indoor lighting do not significantly affect performance of the top systems. Approximately the same performance results were obtained using two indoor data sets, with different lighting, in FRVT 2002. In both experiments, the best performer had a 90% verification rate at a false accept rate of 1%.

- For the best face recognition systems, the recognition rate for faces captured *outdoors,* at a false accept rate of 1%, was only 50%. Thus, face recognition from outdoor imagery remains a research challenge area.

- The FRVT 2002 database also consisted of images of the same person taken on different days. The performance results in this case, using indoor imagery, shows improvement in the capabilities of the face recognition systems over the last two years. Compared with similar experiments conducted two years earlier in FRVT 2000, the results of FRVT 2002 indicate there has been a 50% reduction in error rates[3] .

A very important question for real-world applications is the rate of decrease in performance as time increases between the acquisition of the database of image and new images presented to a system. FRVT 2002 found that for the top systems, performance degraded at approximately 5% points per year.

One open question in face recognition is: How does database and watch list size effect performance? Because of the large number of people and images in the FRVT 2002 data set, we were able to report the first large-scale results on this question. For the best system, the top-rank identification rate was 85% on a database of 800 people, 83% on a database of 1,600, and 73% on a database of 37,437. For every doubling of database size, performance decreases by two to three overall percentage points. In mathematical terms, identification performance decreases linearly with respect to the logarithm of the database size.

---

[2] See Section 4 for an overview of the verification, identification and watch list tasks.
[3] D. M. Blackburn, J. M. Bone, and P. J. Phillips (2001), *FRVT 2000 Report*, Technical Report, http://www.frvt.org.

A similar effect was observed for the watch list task. As the watch list size increases, performance decreases. For the best system, the identification and detection rate was 77% at a false alarm rate of 1% for a watch list of 25 people. For a watch list of 300 people, the identification and detection rate was 69% at a false alarm rate of 1%. In general, a watch list with 25 to 50 people will perform better than a larger size watch list.

Previous evaluations have reported face recognition performance as a function of imaging properties. For example, previous reports compared the differences in performance when using indoor versus outdoor images, or frontal versus non-frontal images. FRVT 2002, for the first time, examined the effects of demographics on performance. Two major effects were found. First, recognition rates for males were higher than females. For the top systems, identification rates for males were 6% to 9% points higher than that of females. For the best system, identification performance on males was 78% and for females was 79%. Second, recognition rates for older people were higher than younger people. For 18 to 22 year olds the average identification rate for the top systems was 62%, and for 38 to 42 year olds was 74%. For every ten years increase in age, on average performance increases approximately 5% through age 63. All identification rates were computed from a database of 37,437 individuals.

Since FRVT 2000, new techniques and approaches to assist face recognition have emerged. FRVT 2002 looked at two of these new techniques. The first was the three-dimensional morphable models technique of Blanz and Vetter. Morphable models are a technique for improving recognition of non-frontal images. We found that Blanz and Vetter's technique significantly increased recognition performance. The second technique is recognition from video sequences. We found that, using FRVT 2002 data sets, recognition performance using video sequences was the same as the performance using still images.

In summary, the key lessons learned in FRVT 2002 were:

- Given reasonable controlled indoor lighting, the current state of the art in face recognition is 90% verification at a 1% false accept rate.
- The use of morphable models can significantly improve non-frontal face recognition.
- Watch list performance decreases as a function of size – performance using smaller watch lists is better than performance using larger watch lists.
- In face recognition applications, accommodations should be made for demographic information since characteristics such as age and sex can significantly affect performance.

These findings are discussed in detail in the other FRVT 2002 documents. The complete FRVT 2002 report has three volumes: *Summary and Overview, Evaluation Report,* and *Technical Appendices* (all three are available at http://www.frvt.org). This document, the *Summary and Overview*, briefly presents the key results from the FRVT 2002. The *Evaluation Report* is a detailed description of FRVT 2002 procedures, experiments, and results. The *Technical Appendices* provide supplementary material, detailed documentation of the FRVT 2002 evaluation protocol, participant system and software descriptions, and participant responses to the FRVT 2002 (pre-release versions of the *Evaluation Report* and *Technical Appendices*).

2. FRVT 2002 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 - the high computational intensity (HCInt) test and medium computational intensity (MCInt) test. Each sub-test was 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. The HCInt had to be performed on the equivalent of three high-end workstations, the MCInt on a single workstation. Technical specifications for each participant's HCInt and MCInt systems are provided in the Face Recognition Vendor Test 2002 *Technical Appendices*. Participants were given 11 days to complete each test.

FRVT 2002 was announced on 25 April 2002 and was open to all developers and providers of core face recognition technology. This included academia, research laboratories, and commercial companies. Participants could take the HCInt, MCInt, or both. The participants and the tests they took are provided in Table 1[4]. FRVT 2002 was administered at the U.S. Naval base at Dahlgren, Virginia between 10 July and 9 August 2002.

Table 1. 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.

3. Image Data Sets

This section describes the data sets used in FRVT 2002. A common aspect of all images used in FRVT 2002 is that they contained the face of exactly one individual. 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. The HCInt data set consisted of 121,589 images of 37,437 individuals with at least three images of each person. The images are of good quality and were gathered in a consistent manner. The background is universally uniform. The names of the individuals were encoded to protect the privacy of the subjects. Due to privacy considerations, representative, not actual, images of the data set are shown in Figure 1.

[4] 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.

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

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 at 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 set is from The University of Texas at Dallas and consists of video sequences and still images taken in 2001. The NIST-NSWC-USF data set is comprised of images taken indoors and outdoors. The images were taken over more than three years at three sites. The images in Figure 2 are from NIST, the NSWC, and the USF dataset. The outdoor stills are characterized by changing background and directional sunlight illumination. Figure 3 shows selected frames from 150-frame UT Dallas "facial speech" videos. The two rows show the same subject gathered on different occasions.

Figure 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.

Figure 3. Selected frames sampled from two University of Texas at Dallas video sequences.

4. Overview of Results

FRVT 2002 addressed several important face recognition topics. Performance on frontal faces was examined under different environmental conditions for three tasks – verification, identification and watch lists. Newly developed technologies were examined to understand recent progress in the field and to identify promising avenues of future research. Finally, demographics of the subject population were examined. The three primary face recognition tasks are listed below:

- Verification: "Am who I say I am?"
- Identification: "Who am I?"
- Watch list: "Are you looking for me?"

In a verification task, a person presents their biometric and an identity claim to a face recognition system. The system then compares the presented biometric with a stored biometric of the claimed identity. Based on the results of comparing the new and stored biometric, the system either accepts or rejects the claim. From an evaluation point of view, there are two types of system users. The first is a legitimate user. The second is a person who attempts to impersonate a legitimate user. Verification performance is characterized by two performance statistics. The two statistics characterize the success rate of the two types of users. The first is the rate at which legitimate users are granted access. This is the *verification rate*. The second is the rate at which imposters are granted access. This is the *false accept rate*. The ideal system would have a verification rate of 100% and a false accept rate of 0%. Unfortunately, such a system does not exist. In real-world systems, there is a trade-off between verification and false accept rates.

It is critically important to consider the false accept rates and verification rates together in order to understand the performance capabilities of a face recognition system. It is easy to build a system that always grants access to a subject. This system will have a 100% verification rate since access will always be granted in response to a legitimate user's request. Conversely, this system will also have a 100% false accept rate because it also grants access to imposters. The best system is one that balances the verification rate with a false accept rate in a manner consistent with operational needs.

Examples of this trade-off can be viewed at Figure 4. Verification performance results are shown for the top three systems. For each system, verification performance is reported for three false accept rates: 0.1%, 1% and 10%. Cognitec and Identix have verification rates of 82% with a false accept rate of 0.1%. With a false accept rate of 1%, their verification rates are both 90%. With a false accept rate of 10%, their verification rates are 96%. This illustrates the trade-off between the verification rate and false alarm rate.

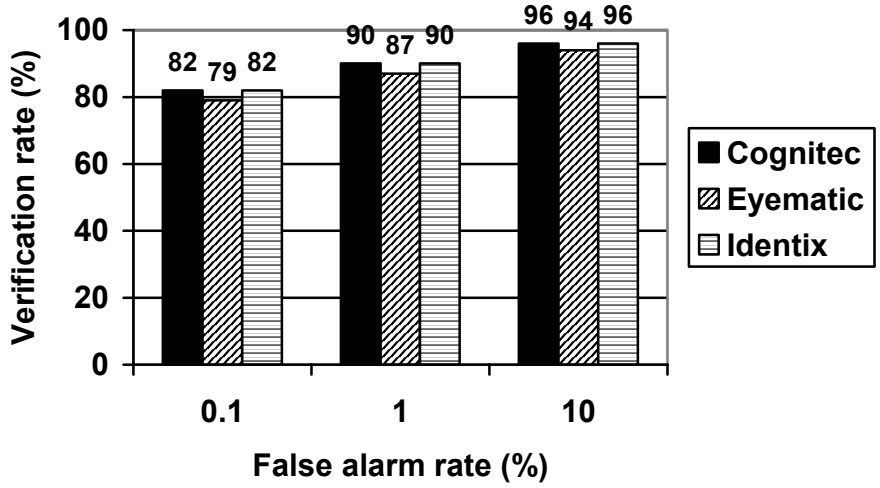

Figure 4. Verification performance is shown for the top three systems on HCInt visa images. Verification rates are reported for false accept rates of 0.1%, 1%, and 10%.

The results in Figure 4 were generated using the HCInt visa dataset (37,437 individuals). The database images consisted of multiple images of the same individuals taken on different days. There is up to three years difference between some pictures in the database. The verification performance from the HCint visa images provides a reliable estimate of performance because of the large number of images in the data set. Figure 5 depicts all FRVT 2002 participants' verification rates (at a 1% false accept rate) using the HCInt visa dataset.

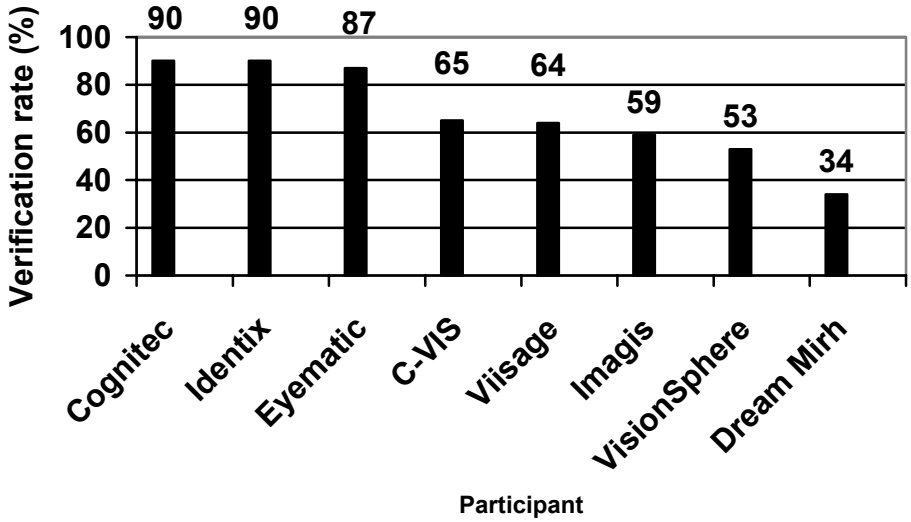

Figure 5. Verification performance is reported for all participants on the HCInt visa dataset. Verification performance is reported at a false accept rate of 1%.

Recognition from frontal images is an important capability of face recognition systems. FRVT 2002 examined recognition performance using frontal images taken under varied conditions. The analysis included the results using the HCInt and MCInt data sets. In all cases, the database consists of images taken indoors under good lighting conditions. Results are reported for five different conditions (Figure 6):

- Matching images of a person taken indoors on the same day with the same illumination.
- Matching images of a person taken indoors on the same day with different illumination.
- Matching images of a person taken indoors on different days with the same illumination.
- HCInt visa images (matching images of a person taken on different days).
- Matching of an image of a person taken indoors with an image taken outdoors. Both images were taken on the same day.

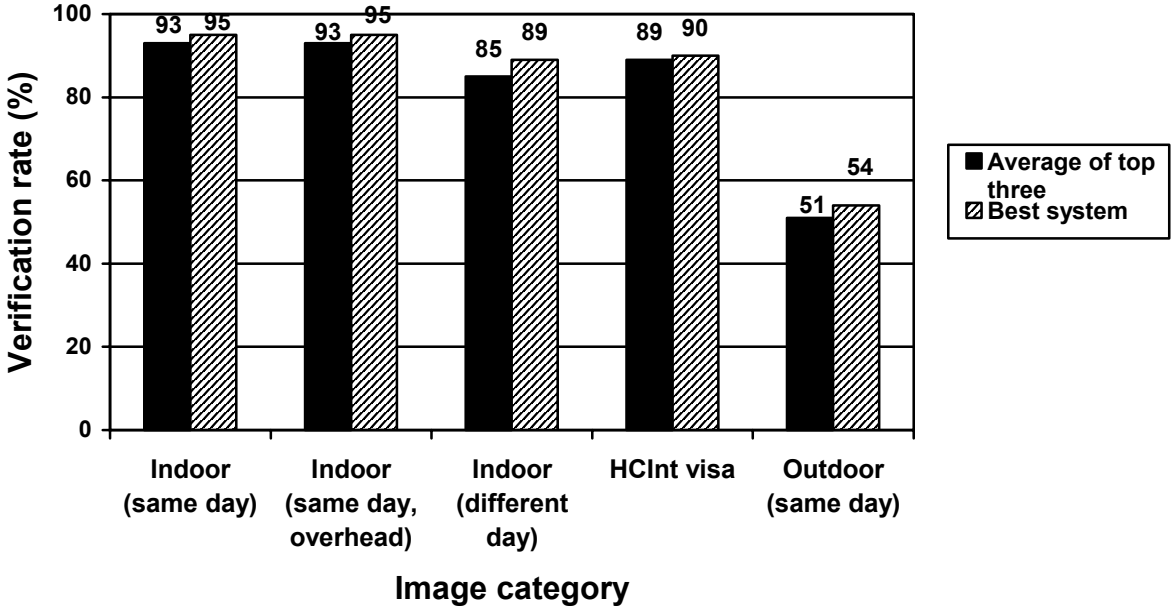

Figure 6. Verification performance is reported for five categories of frontal facial images. Performance is reported for the best system and average of the top three systems in each category. The verification rate is reported at a false accept rate 1%.

Figure 6 shows performance of the best system in each category and average performance for the three best systems in each category. Our analysis shows that reasonable changes in indoor lighting do not affect performance (the two indoor same categories) for top systems. These results are consistent across multiple databases for these systems (indoor different day and HCInt visa categories). Invariance to normal changes in indoor illumination is an issue for some of the other systems that participated in the FRVT 2002. This suggests that recognition from indoor images is reasonably stable. Performance using images of subjects that were taken outdoors, even on the same day, was drastically reduced. This is a trend that was also apparent in FRVT 2000, indicating that the variation and structure of outdoor lighting has a drastic affect on performance. For example, for the best systems, verification performance drops from 95% to 54% - a 40% drop - going from indoors to outdoors.

In the identification task, an image of an unknown person is provided to a system. (In the identification task, we assume that through some other method we know the person is in the database.) The system then compares the unknown image to the database of known people. The results of this comparison are then presented by the system to an operator in a ranked listing of the top *n* 'candidates' (typically anywhere from one to 50). If the correct image is somewhere in the top *n*, then the system is considered to have performed the identification task correctly.

Figure 7 depicts the top three participants' identification rate performance at three different 'ranks'. In this test, the database consisted of 37,437 different images. A rank of '1' is the rate at which the system's 'best' or 'most likely' candidate was correct; the rank 10 results depicts the rate at which the correct identity was

within the top ten candidates. Notice that the performance goes up for higher ranks. Our analysis has also shown that a system's identification performance capability is dependent upon the size of the database. For the best system, the top-rank identification rate was 85% on a database of 800 people, 83% on a database of 1,600, and 73% on a database of 37,437. For every doubling of database size, performance decreases by 2% to 3% points. In mathematical terms, identification performance decreases linearly with respect to the logarithm of the database size.

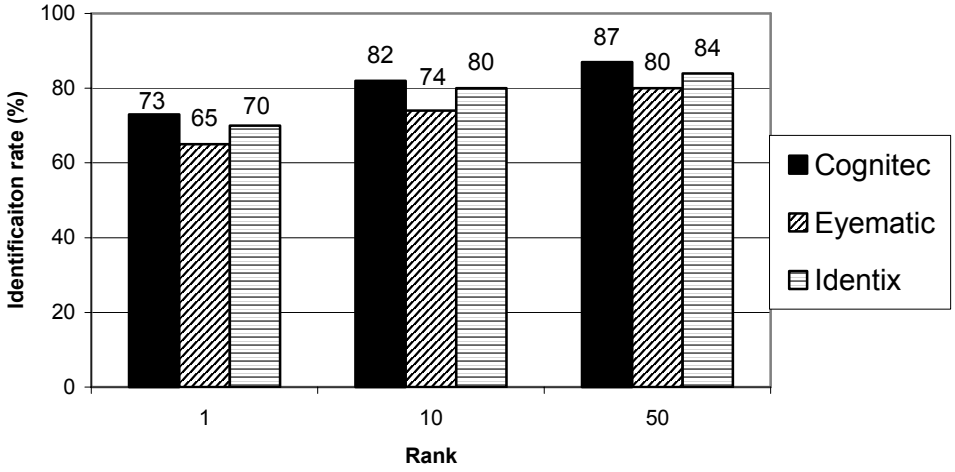

Figure 7. Identification performance for the three best system on the HCInt visa dataset. The database consisted of 37,437 individuals. Identification rates are reported for ranks 1, 10, and 50.

FRVT 2002 also evaluated face recognition technology with respect to a watch list task. In the watch list task, a face recognition system must first detect if an individual is, or is not, on the watch list. If the individual is on the watch list, the system must then correctly identify the individual. The statistic for correctly detecting and identifying an individual on the watch list is called the *detection and identification rate*. In some instances, the system may incorrectly alarm on an individual that is not on the watch list. The rate at which an individual that is not on the watch list is incorrectly alarmed is called the *false alarm rate*.

Typically, the watch list task is more difficult than the identification or verification tasks alone. Figure 8 shows detection and identification rates for varying watch list sizes at a false alarm rate of 1%. For the best system using a watch list of 25 people, the detection and identification rate is 77%. Increasing the size watch list to 3,000 people, decreases the detection and identification rate to 56%. Figure 8 also indicates that the systems achieve better performance for a smaller watch list. If the impetus of the watch list application is to detect and identify the "most wanted" individuals, the watch list should be kept as small as possible. Increasing the size of the watch list reduces the probability that an individual on the watch list is correctly detected and identified when presented to the system.

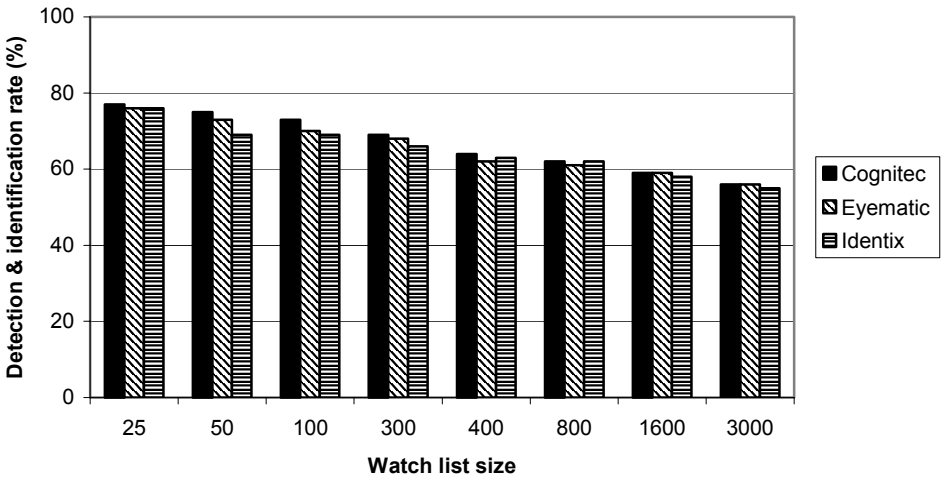

Figure 8. Watch list performance is shown for the three best systems on the HCInt visa dataset. Detection and identification rates are reported at a false alarm rate of 1% for eight watch list sizes.

So far we have concentrated in recognition performance from frontal images. The general face recognition problem requires recognition from non-frontal images. FRVT 2000 and FRVT 2002 show that one of the more difficult tasks for modern face recognition systems is recognizing faces in non-frontal imagery. Most face recognition systems perform well when all of the images are frontal. But, as a subject becomes more and more off angle (both horizontally and vertically), performance decreases.

One potential solution to this problem is the use of 'morphable models.' Although the details of how a morphable model works are quite complex, their use with respect to any face recognition system is straightforward. A morphable model takes a facial image taken from any angle as input, and outputs what that subject might look like if they were facing forward (the model leaves images already facing forward mostly unchanged). If the morphable model does this prediction well, then a face recognition system could use the output from the morphable model as a substitute for the original, off-angle image.

The efficiency of a morphable model was evaluated in FRVT 2002. Baseline performance was measured on a set of non-frontal images. The set contained people looking left, right, up and down. Performance was also measured on the same set of imagery after they had been processed by the morphable model. The transformed non-frontal faces in the FRVT 2002 were generated by the techniques of Blanz and Vetter[5]. As shown in Figure 9, there was a dramatic improvement in performance using the imagery provided from the morphable models. For example, on non-frontal images rotated either to left or right, Identix's performance increases from 26% on the original non-frontal images to 84% on the morphed images. FRVT 2002 participants were not informed prior to testing that the provided input would include morphable model imagery. Therefore, it is likely that with better vendor integration, morphable models might be able to boost performance even more than the preliminary, exploratory data presented here.

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

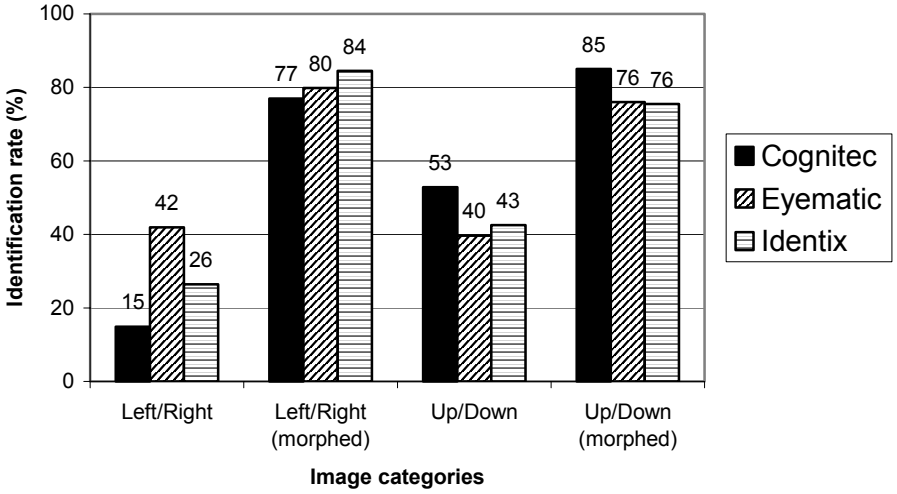

Figure 9. Identification performance is shown on non-frontal and morphed non-frontal images. The left/right and up/down categories are top identification rates for the original non-frontal images. The left/right (morphed) and up/down (morphed) categories are top identification rates for the morphed non-frontal images. Performance is on a database of 87 individuals.

Another important component of FRVT 2002 was the use of *still* versus *video* imagery. Our analysis has determined that there is little recognition benefit gained by using video images in lieu of still images. For the top three performers, performance drops 3% points when using video images instead of still images. The results indicate that for pure *recognition*, video does not appear to make any difference in performance. FRVT 2002 did not address the issue of detecting faces in video. Using video to detect faces could greatly assist a face recognition system.

FRVT 2002 investigated several demographic aspects of face recognition. Specifically, we examined how sex, age, and time-delay affect recognition. Males are easier to recognize than females. For the top systems, identification rates for males were 6% to 9% points higher than that of females. For the best system, identification performance on males was 78% and for females was 79%. Older people are easier to identify than younger people. Figure 10 depicts the average identification rate for the top three performers, as a function of *age*. In Figure 10, each bar is the identification performance for a five-year range of performance. For example, the first two bins are 18 to 22 year olds and 23 to 27 year olds. For 18 to 22 year olds the average identification rate for the top systems was 62%, and for 38 to 42 year olds was 74%. For every ten years increase in age, on average performance increases approximately 5% points. Identification results are computed from a database of 37,437.

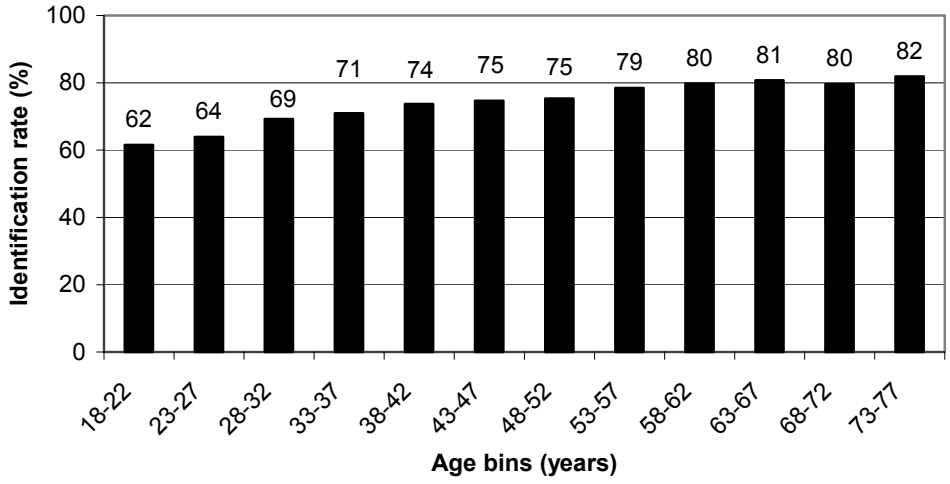

Figure 10. Identification performance is shown broken out by age of an individual. Each bin is labeled by the age range it contains (five year intervals). Identification rate is the average for the top three systems. Performance is on a database of 37,437.

Although it is less of a 'demographic' than sex or age, the elapsed time between database and new images, also affects performance. Figure 11 shows this difference with respect to 60 day 'bins.' As the elapsed time between the database and new images increases, performance decreases. For the better systems, identification performance decreases at 5% points per year. This makes sense intuitively, since we expect that as the face changes over time, face recognition algorithms cannot as easily model these facial variations.

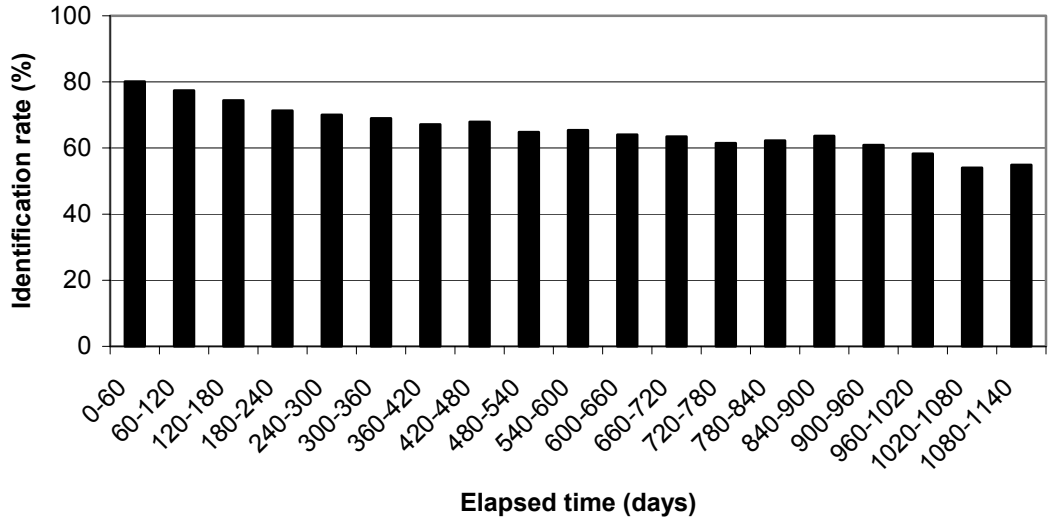

Figure 11. Identification performance is reported broken out by elapsed time between database and new image. Performance is reported in 60 day intervals. The average rank one identification rate for the top three systems is reported on a database of 37,437 individuals.

5. 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 a 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 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 in 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 database and new 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.
- 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 database or watch list size.

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