DHS S&T Technical Paper Series

A Scenario Evaluation of High-Throughput Face Biometric Systems: Select Results from the 2019 Department of Homeland Security Biometric Technology Rally

Jacob A. Hasselgren
John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility

Arun R. Vemury

*The U.S. Department of Homeland Security,
Science and Technology Directorate,
Biometric and Identity Technology Center*

Keywords: Face Recognition, High-throughput Systems, Scenario Testing, Commercial Systems, Acquisition Systems, Matching Systems

August, 2020

Homeland Security
Science and Technology

Executive Summary

OVERVIEW: : This study was conducted by the United States Department of Homeland Security (DHS) at the Maryland Test Facility as part of a series of biometric technology evaluations, known as the DHS Biometric Technology Rallies (“Rallies”). Biometric technologies use people’s physical features, such as facial characteristics, to confirm their identity. They can be used in the air travel environment to make security screening faster and more reliable. The Rallies are one of the only large-scale, scenario evaluations of complete, commercially available biometric systems. These are the results from the second such test, which took place in the Spring of 2019.

MOTIVATION: Most studies of biometric technologies are lab tests, which focus on specific parts of the technology, such as the matching algorithm. A scenario test, such as the one performed in this study, evaluates a full biometric system in all of its parts, including cameras and algorithms, otherwise known as acquisition systems and matching systems. Lab tests are useful to track the performance of those parts over time and to inform developers. Scenario tests better predict the performance of a full operational system.

WHAT WE DID: We tested biometric systems from an international group of commercial vendors in a designed scenario test. Ten face acquisition systems were used by 430 diverse human test subjects. For each of these acquisition systems, we measured efficiency, effectiveness, and user satisfaction. We also tested eight face matching systems. We measured the robustness of each matching system by evaluating whether each matching system could correctly match photos from each of the acquisition systems.

MAJOR TAKEAWAYS: The major findings of this study are as follows. First, most acquisition systems tested in the 2019 Rally were fast and satisfying for test subjects to use. However, these systems still struggled with effectiveness, with only one system reliably able to identify all test subjects. These levels of effectiveness were not well anticipated by the commercial providers of these systems. Second, the most prominent source of error was related to acquiring a suitable face photo. Matching two acquired photos was less challenging. This suggests renewed focus on human-machine interaction and testing that includes acquisition systems offer the most direct path forward for improving operational systems. Finally, the performance of some matching systems varied significantly, depending on the acquisition system used to capture images. We provide recommendations for choosing a matching system depending on the operational use.

A Scenario Evaluation of High-Throughput Face Biometric Systems: Select Results from the 2019 Department of Homeland Security Biometric Technology Rally

Jacob A. Hasselgren, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury

Abstract—International biometric testing standards distinguish scenario evaluations from technology evaluations. Scenario evaluations measure the performance of an end-to-end system, in a simulated real-world environment, using live human participants. Executing biometric scenario evaluations is challenging, but they provide important insights that technology evaluations cannot, such as the simulated performance of the full system and the ability to attribute errors to specific system components. These insights are crucial for assessing what systems should be selected for an operational biometric deployment. The U.S. Department of Homeland Security Biometric Technology Rallies are a series of scenario evaluations of commercial biometric systems designed to operate in high-throughput environments. They are one of the only large-scale, scenario evaluations of complete, commercially available biometric systems. The 2019 Biometric Technology Rally tested the performance of ten face acquisition systems and eight face matching systems with a sample of 430 diverse human subjects. The 2019 Rally found that most (6/10) face acquisition systems maintained average transaction times under five seconds and that half (5/10) received satisfaction ratings in excess of 95% positive. However, less than half (4/10) of the acquisition systems were able to reliably acquire images from 99% of the tested participants and only a single (1/10) system produced images suitable for identifying all 430 participants. These levels of effectiveness were not well anticipated by commercial providers of these acquisition systems, meaning if system owners used vendor provided estimates of performance to plan an operational deployment, serious deficiencies, potentially requiring costly reworks or program cancellation, could have occurred. Results from the 2019 Rally also led to two additional findings. First, the most prominent source of errors in high-throughput face biometric systems were related to acquiring a suitable face biometric sample, not matching two suitable face biometric samples. A renewed focus on user interaction during image acquisition (camera placement, camera adjustment, informative signage, etc.) offers significant room to improve the performance of high-throughput face biometric systems. Second, when matching systems were tested in combination with acquisition systems, half showed statistically significant levels of variation in performance across acquisition systems. The remaining half worked well ($>$ 95% true identification rate) only on some acquisition systems, with one matching system working well only on images from a single acquisition system. We propose a matching system taxonomy (robust, brittle, and specialist) to describe this variation and discuss the impact of matching system choice on operational error rates.

Index Terms—Face Recognition, High-throughput Systems, Scenario Testing, Commercial Systems, Acquisition Systems, Matching Systems

1 Introduction

1.1 Biometric System Testing

THE increasing adoption of biometric technologies, particularly in the public sphere, has highlighted the need to understand the performance characteristics of these systems. The testing of biometric systems is defined by ISO/IEC 19795-2, ”Biometric performance testing and reporting, Part 2: Testing methodologies for technology *and scenario evaluations*” [1] (emphasis ours). As the title suggests, this international standard outlines two distinct kinds of biometric tests; technology and scenario evaluations. Technology evaluations involve isolating particular biometric system components, such as a matching algorithm, and ”conduct[ing] exploratory testing” on static datasets, often for the purpose of improving an engineering process. Scenario evaluations, by contrast, measure the performance of end-to-end systems, in real-time, on human participants. Scenario evaluations are designed to be externally valid, meaning that the ”simulated performance” measured is designed to estimate real-world performance. This makes performance data from scenario evaluations more applicable to the task of selecting which biometric systems should be considered for operational deployment.

Both varieties of biometric evaluations are useful. The outcomes of technology evaluations inform the technical staff developing biometric systems as to the theoretical limitations of those systems and where future development efforts may show promise. They can also be used to track progress of a biometric system component, such as a matching algorithm, over time on a static dataset. The outcomes of scenario evaluations inform the technical staff working to deploy biometric systems as to the practical limitations of those systems and to critical performance characteristics,

- *J. Hasselgren, J. Howard, Y. Sirotin, and J. Tipton work at the Maryland Test Facility in Upper Malboro, Maryland.*
- *A. Vemury works at the United States Department of Homeland Security, Science and Technology Directorate in Washington, DC.*
- *Authors listed alphabetically. E-mail correspondence should be sent to info@mdtf.org*

such as anticipated error rate and what factors are likely to impact that error rate. Because scenario testing involves full systems, it can also estimate the relative contribution of each system component to the observed, overall error rate.

Despite the uniqueness and relevancy of scenario evaluations to operational performance, the *vast* majority of current biometric testing protocols fall into the category of technology evaluations. Perhaps the most prominent biometric technology evaluations are a series of tests from the U.S. National Institute of Standards and Technology (NIST), including the Fingerprint Vendor Technology Evaluation [2], the Iris Exchange X evaluation [3], and the Face Recognition Vendor Test [4]. There are also a plethora of publicly available biometric datasets that organizations can use to conduct their own technology evaluations. For example, MegaFace [5] and the IARPA Janus datasets [6] [7] [8] can be used to explore the overall accuracy of different matchers or the effect of parameter tuning on a broad swath of face images. Other publicly available datasets can be used to investigate specific issues, such as the ability of different face recognition algorithms to overcome make-up [9] or disguises [10].

Scenario evaluations are comparatively rare and when such studies have been executed, they have tended to focus on niche issues, such as contactless fingerprint collection [11] [12], and biometric spoofing [13]. Very few evaluations of full, commercial biometric systems have been conducted to date, despite a study from NIST calling for such testing in 2008 [14], and more recent calls for such testing from privacy groups [15]. The NIST study noted that it is critical to not only test biometric systems prior to deployment, but to separately consider the efficiency, effectiveness, and user satisfaction of image acquisition systems. The effectiveness of a biometric system refers to the ability of the system to successfully perform all tasks with all users and includes failures-to-acquire biometric samples and errors in matching acquired samples [16]. System efficiency refers to the speed with which users can perform all actions required to complete a system-level transaction, with more efficient systems having shorter transaction times. Finally, satisfaction refers to whether users have a positive perception of their interaction with the system, such that systems with a higher satisfaction scores should be more likely to be broadly accepted by the public as part of a larger work flow. Importantly, these metrics apply to the *total* biometric system and thus can only be calculated via the scenario evaluation model. These metrics can not be calculated using the more ubiquitous technology evaluations.

1.2 High-throughput Biometric Systems

High-throughput biometric systems are an emerging class of commercial biometric offering. High-throughput systems perform the same general functions as traditional biometric systems but at a notably different scale. High-throughput biometric systems can perform thousands of biometric transactions per hour and, because of these volumes, typically operate unstaffed or intentionally under-staffed (i.e. one staff member responsible for multiple systems) [17]. One use case for such systems is at aircraft boarding gates where high-throughput biometric systems can be used to verify the identity of thousands of passengers daily. Indeed, starting in 2018, such a system is being piloted in the United States for expedited boarding of outbound international flights [18].

However, the combination of large volumes and staffing constraints creates several unique challenges. First, in high-throughput systems, even seemingly small error rates can result in serious operational issues. A high-throughput system processing 100,000 people in a day with a 1% failure rate will inconvenience a thousand users with delays or alternate processing requirements. This could result in discontent amongst systems users and increased staffing costs for system operators and owners. Second, high-throughput systems place a greater emphasis on efficiency than systems designed for lower volume or staffed operation. However, optimizing a system to achieve efficiency can be challenging, particularly under space constraints, requiring a focus on system usability and human factors. Finally, satisfaction in a high-throughput environment can be decreased by the increased pace of operations and by the lack of human interaction.

1.3 The 2019 Biometric Technology Rally

This paper presents the results of the 2019 Biometric Technology Rally (“2019 Rally”), a large-scale scenario evaluation of current high-throughput commercial biometric systems with 430 diverse users, sponsored by the U.S. Department of Homeland Security (DHS), Science and Technology Directorate (S&T). This evaluation is unique in that it includes ten acquisition systems and eight matching systems, all commercially available in 2019, for a total of 80 acquisition/matching system combinations. This report quantifies the efficiency, effectiveness, and user satisfaction of the tested acquisition systems using the same matching system as the 2018 evaluation [17]. Additionally, the methodology used in executing this scenario evaluation allows acquisition errors and matching errors to be calculated separately and their magnitudes compared, building on recent evidence that image acquisition, not matching, is the main factor constraining the overall biometric system performance for operational systems [17] [19] [20]. Finally, this report evaluates the impact of using different matching systems on the samples acquired from each acquisition system, showing robustness of matching system to acquisition system and vice versa.

2 Methods

A call for participation in the 2019 Rally was issued to commercial providers of both biometric acquisition systems and biometric matching systems in November 2018. This call included the full testing plan and performance objectives as outlined below. No information was withheld from providers, and only those who believed they could meet the objectives were encouraged to apply. Providers had one month to submit an application package. Acceptance notifications were sent to selected providers in February 2019 and the test was held in May, giving the providers roughly three months to optimize their acquisition or matching system to meet the objectives of the 2019 Rally. We believe the results presented here are representative of the current state of commercial high-throughput biometric systems because of the commercial nature of these systems, the lead time given to optimize these systems around the test constraints, and the broad sample of acquisition and matching systems tested as part of this evaluation[1].

2.1 Acquisition Systems

Acquisition systems were required to collect face images suitable for biometric matching within a 10 second time constraint and were required to function autonomously, without a human operator. Acquisition systems were responsible for directing all aspects of the user interaction with the system. Furthermore, acquisition systems were required to collect, process, and submit face images within the period of time in which the user was interacting with the system (i.e. no batch or offline processing). Acquisition systems were required to, at a minimum, provide one face biometric sample per individual but could provide up to three samples. Acquisition system providers were encouraged to configure their system to submit biometric samples according to a speed-accuracy tradeoff: biometric samples were to be submitted as quickly as possible and additional samples should be submitted only if they are of superior quality relative to prior samples [21]. Finally, acquisition systems were required to fit within a 6 by 8 foot floorspace. Other than these requirements, no restrictions were placed on the general form factor (kiosk, walk-through, etc.) of acquisition systems.

2.2 Matching Systems

Matching systems were required to generate templates for acquired face biometric samples (i.e. process) and return biometric similarity scores given two face biometric templates. Matching systems were evaluated based on their ability to match biometric probe images gathered by various acquisition systems to a gallery of previously acquired images. Specifically, matching systems were evaluated based on their ability to identify each probe image against a preset gallery. Each probe image was compared serially against each gallery image and the match was determined as the gallery identity at rank-1 score above a specified threshold. Matching systems were evaluated both on their ability to correctly identify subjects enrolled in the gallery (in-gallery subjects) and to report probe images of subjects who are not enrolled in the gallery (out-of-gallery subjects) as unidentified (see Section 2.5).

2.3 Subjects and Sample Size

A sample of 430 diverse paid volunteer subjects was recruited from the general public via online advertising to serve as users of the acquisition systems. During study enrollment, subjects self-reported their demographics including their race, ethnicity, age, gender, height, weight, and use of eyewear. Age was limited to individuals over 18 years of age for Institutional Review Board (IRB) purposes. Race was defined in accordance to the U.S. Census categories [22]. Subject demographic distributions for these 430 subjects are shown in Fig. 1.

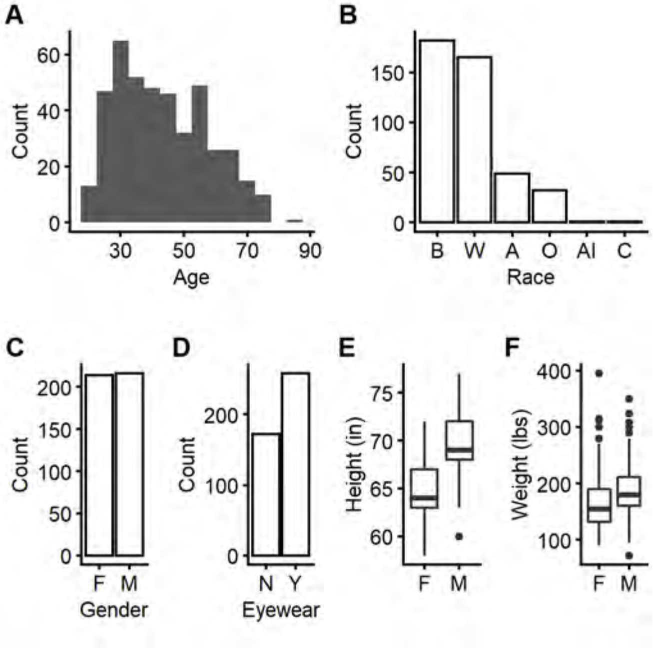

Fig. 1. Distributions of the demographic variables self-reported by test subjects. **A.** Distribution of test subject ages. **B.** Counts of subjects identifying with each racial category: (B) Black or African-American; (W) White; (A) “Asian”; (O) “Other Race”; (AI) “American Indian or Alaska Native”; (C) “Aboriginal peoples of Canada”; (H) “Native Hawaiian or other Pacific Islander”. Groups A, O, AI, C, and H are grouped in to a general “Other” category during analysis. **C.** Counts of subject gender: (F) Female; (M) Male. **D.** Subject response to whether or not they wear eyewear: (N) No; (Y) Yes. **E-F.** Boxplots of subject height and weight by gender.

2.4 Test Process

The 2019 Rally took place at the Maryland Test Facility, an S&T affiliated biometrics testing laboratory. The test process for the 2019 Rally used the same framework for evaluating the performance of generic biometric systems as the 2018 Rally [17]. Each acquisition system was installed in a standard Rally test station (Fig. 2). Each test station was arranged side-by-side but separated from other stations by grey cloth-covered walls, 7.5 feet high, to avoid activity in one station from impacting the processes taking place at another.

Testing took place in separate morning and afternoon sessions over a five-day period (10 sessions total). Each session included ∼45 volunteers, broken into three treatment groups of ∼15. Following informed consent, subjects were briefed as to the purpose of the scenario test and told that biometric systems were going to acquire their images for the purpose of performing a biometric identification. They were asked to comply with all instructions presented by the systems but were not specifically instructed regarding the mechanistic details of the individual acquisition systems.

[1] The 2019 Rally tested high-throughput face, finger, and iris acquisition and matching systems. The majority of systems selected for participation, and the best performing systems, were face systems. For brevity, the results presented in this report include only those from face acquisition and matching systems. Select results for iris and fingerprint systems can be found at https://mdtf.org/Rally2019/Results2019

Following the briefing, subjects proceeded to biometric enrollment. Here, each subject was identified using face, finger, and iris biometrics to confirm their associated subject ID given during an informed consent process. Subjects wore these associated IDs on a wrist-band during the entire test. The ID was used as the ground-truth link between the subject, the biometric transaction, and the acquired images.

To test each acquisition system, a treatment group of subjects was queued at the test station in which the acquisition system was installed. Subjects entered the station one at a time to use the system after their ground-truth identity was recorded by test staff. To mitigate habituation and carry-over affects, the order in which each treatment group encountered each acquisition system was counterbalanced so every system was encountered in each serial position, and every system followed every other system an equal number of times. All acquisition systems operated autonomously and were completely unstaffed during testing. Image submissions were made by each station in real time via a common web-based application programming interface (API). Following their interaction with each acquisition system, subjects were asked to provide a satisfaction score that rated their overall experience (see Fig. 2). Acquisition systems were given five minutes to process the entire treatment group of ~15 volunteers. Discounting the average time required for scanning wrist-bands and rating satisfaction, this left, on average, ~10 seconds for each subject to use an acquisition system.

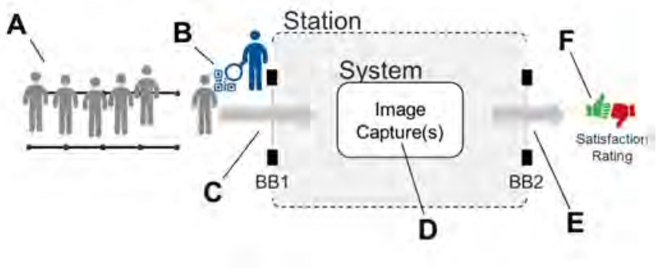

Fig. 2. The test process performed at each test station during the 2019 Rally. Commercial face capture systems (System) were installed within a dedicated test station (Station). **A.** Test subjects queued at each station. **B.** Test staff established the ground-truth identity of each subject by scanning a QR code printed on the subject's wristband. **C.** Subjects entered the test station, triggering a beam break (BB1). **D.** Subjects interacted with the face capture system, which submitted images (biometric samples) for storage. **E.** Subjects exited the test station, triggering a beam break (BB2). The duration of each subject's interaction with the system was measured as the difference in time between BB2 and BB1. **F.** Subjects rated the test station based on their level of satisfaction using a kiosk.

Following the acquisition portion of the evaluation, acquired biometric samples were processed by the matching systems included in the 2019 Rally. Matching systems conformed to a common, simplified biometric API that consisted of two function calls. The first accepted an image in the form of a base64 encoded byte string and returned a biometric template in the form of a byte array. The second function call accepted two biometric templates and returned a similarity score. Probe biometric samples from each acquisition system and historic gallery biometric samples were processed in this fashion to create sets of comparison scores that were used to evaluate effectiveness of each acquisition/matching system combination.

2.5 Biometric Galleries

Biometric samples collected by acquisition systems were matched back to a "historic" gallery of 1,958 face images from 500 unique people. These images were acquired over the course of five years using a variety of face acquisition devices. Of the 430 subjects who participated in the 2019 Rally, 354 had images in the historic gallery. There were 76 out-of-gallery subjects that participated in the test and 146 subjects who were enrolled in the gallery but did not participate in the test (i.e. distractor subjects). "Same-day" enrollment biometric samples were also captured by a trained staff member for all test subjects. The same-day samples are not used in the calculation of the results in this report.

3 Results

To comply with information sharing agreements between S&T and the 2019 Rally technology providers, all acquisition and matching system names are aliased in this report. This section will first introduce performance results for the ten face acquisition systems tested in the 2019 Rally (aliased AS.1 - AS.10) and then explore acquisition system performance across the eight matching systems (aliased MS.1 - MS.8). The ten acquisition systems and eight matching systems were selected by expert review from a pool of 26 and 14 applications, respectively. The final selected systems represent commercial offerings from companies headquartered in six different countries, across three continents.

3.1 Satisfaction

Satisfaction was measured using a rating kiosk positioned at the exit of each Rally station [23]. Subjects were asked to rate their experience using a four-level scale. Fig. 3A plots satisfaction scores obtained for each 2019 Rally face acquisition system. The aggregate metric ("Satisfaction Score") quantifies the percentage of positive satisfaction scores ("happy" or "very happy") out of the total. 2019 Rally participants were told the minimum acceptable satisfaction for a system was 90%. The objective was to achieve an aggregate satisfaction of $>$ 95%. This objective range is highlighted green in Fig. 3A and 2019 Rally systems that met this target are denoted with a filled red point (●). Overall, five of ten face acquisition systems met the 95% satisfaction objective.

3.2 Efficiency

The key measure of efficiency in the 2019 Rally was transaction time, computed as the amount of time each test volunteer spent between the entry (BB1) and exit beam breaks (BB2, see Fig. 2). The maximum acceptable average transaction time was ten seconds and the objective was for systems to maintain an average transaction time of under five seconds. Fig. 3B shows the mean transaction times for each 2019 Rally face acquisition system with the objective range highlighted in green. Acquisition systems that met the five second objective are denoted with a filled red point (●). Overall, six of ten face acquisition systems met the five second efficiency objective.

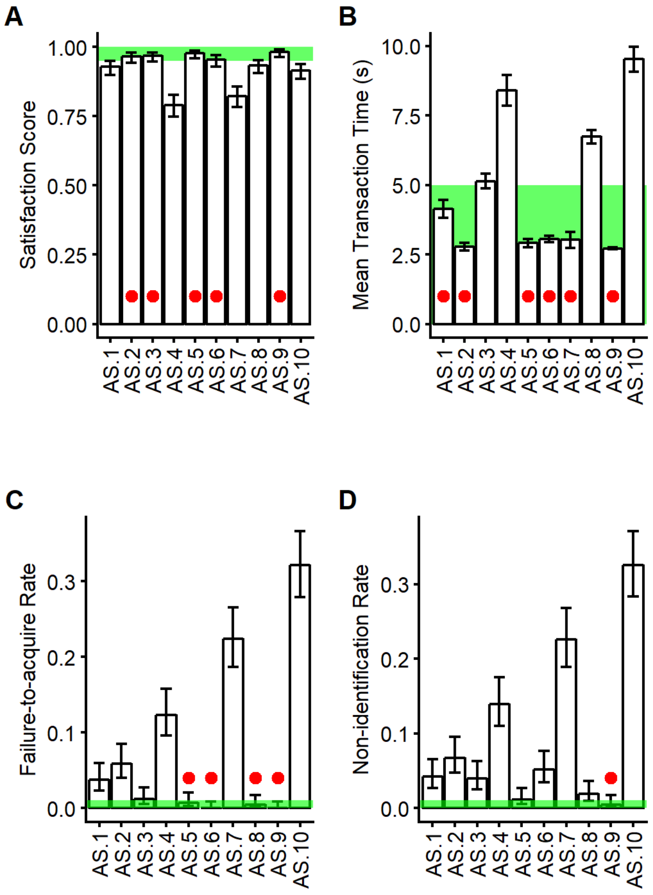

Fig. 3. Performance of ten face acquisition systems in the 2019 Rally (AS.1 - AS.10). Objective ranges are highlighted green. Error bars denote 95% confidence intervals. 2019 Rally systems that met the objective target of each measure are denoted with a filled red point (●). **A.** Satisfaction score with the system as rated by subjects. Calculated at the proportion of “happy” or “very happy” subject ratings acquired immediately after system use. **B.** Efficiency of the system measured as the mean time subjects spent interacting with the system, including entry and exit into the test station. **C.** System failure-to-acquire rate (FtAR) measured as the proportion of subject transactions for which the system failed to submit face images suitable for matching. **D.** System non-identification rate (NIR) measured as the proportion of subject transactions for which the correct subject identity could not be established, inclusive of FtAR.

3.3 Effectiveness

Effectiveness in performing the tasks required in the 2019 Rally was quantified using two metrics: failure-to-acquire rate (FtAR) and non-identification rate (NIR). FtAR was defined as the proportion of subjects for whom the acquisition failed to collect a sample that could be successfully processed into a template by a single commercial matching algorithm. NIR was defined as the percentage of subjects for whom a 2019 Rally system was unable to capture biometric samples that generated an appropriate match determination. Samples for the 354 in-gallery subjects had to produce a rank-1, above threshold, match back to the correct ground-truth identity in the historic gallery whereas images of the 76 out-of-gallery subjects could not produce a match result above threshold for any subject in the historic gallery (see Section 2.5). We note that the scale of this matching exercise, 430 probe images matched to a gallery of $\sim$ 2,000 images of 500 people, are roughly similar to biometric deployments at aircraft boarding gates, such as [18].

Generation of templates, calculation of match scores, and threshold (corresponding to a false match rate of $\sim$1:1000) needed to compute FtAR and NIR for this section was computed using the commercial algorithm and threshold from the 2018 Rally evaluation [17]. The minimum acceptable NIR for the 2019 Rally was 5% and consequently, the maximum acceptable FtAR was 5%. The objective of the 2019 Rally was for systems to achieve an FtAR and a NIR of $<1\%$. Figs. 3C & D show the FtAR and NIR effectiveness metrics for each acquisition system. In these figures, acquisition systems that met the NIR and FtAR objectives are denoted with a filled red point (●). Overall, four of the ten face acquisition systems met the 1% FtAR effectiveness goal but only one subsequently met the 1% NIR objective.

3.4 Acquisition System Effectiveness: Estimated versus Actual Error Rates

Section 3.1 - 3.3 shows that, in general, face acquisition systems included in the 2019 Rally were more likely to meet satisfaction and efficiency measures than those of effectiveness. Moreover, this failure to meet effectiveness goals was not well anticipated by system providers. In preparation for the 2019 Rally, acquisition system providers were asked to estimate the likely performance of their technology within the high-throughput unattended use-case described by the 2019 Rally test plan. Following the evaluation, we compared these estimated levels of performance to those observed during the 2019 Rally. Figs. 4A & B summarize the measured and estimated NIR and FtAR of the 2019 Rally acquisition systems. This figure shows that four of ten acquisition system providers significantly underestimated FtAR and another four of ten providers significantly underestimated NIR. All measures of statistical significance in Sections 3.5 and 3.6 use a p value of 0.05 and a Bonferroni corrected Fisher’s exact test.

3.5 Acquisition System Effectiveness: Failure-to-acquire versus Failure-to-match

Fig. 4C shows that failure-to-acquire was the dominant source of subject non-identification for at least half of 2019 Rally acquisition systems, outstripping failure-to-match by nearly three fold. Failing to acquire a sample capable of creating a template was a significantly greater source of error than all other sources of error combined for five of ten of the acquisition systems. On these five systems, failure-to-acquire rates ranged from 5% to 30%, significantly higher than the failure-to-match rates observed in our study, or the false non match rates reported in technology evaluations such as [4]. The reverse pattern (i.e. other sources of error greater than failure-to-acquire) was observed on only one acquisition system. Taken in conjunction with the results from Section 3.4, these results highlight that not only a significant number of vendors under-estimate the collection challenges associated with high-throughput biometric systems (Fig. 4C) but that many of them are not aware they are under-estimating this crucial performance variable (Fig. 4A).

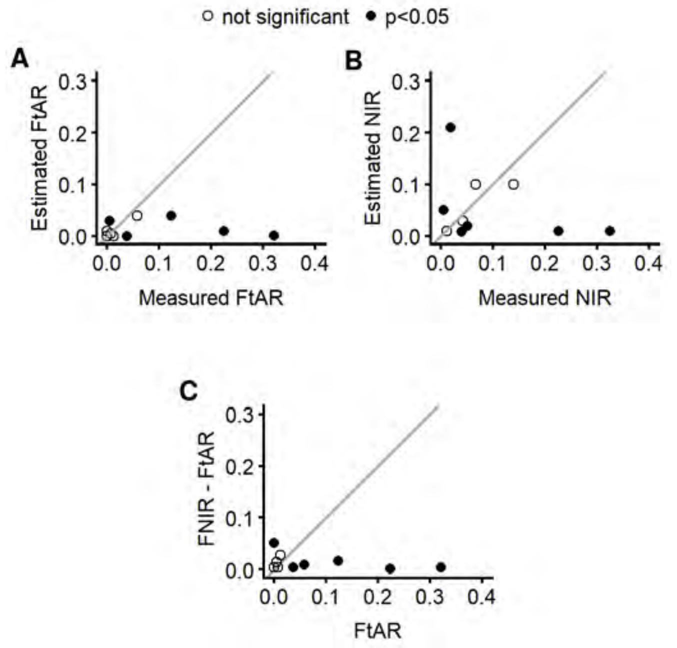

Fig. 4. Expected (estimated) and actual (measured) errors dominating acquisition system performance. Rates that are significantly different (Bonferroni corrected p$<$ 0.05, Fisher's exact test) are denoted with a filled point (●) **A.** Failure-to-acquire rates (FtAR) estimated by vendors plotted as a function of FtAR rates measured during the evaluation. **B.** Non-identification rates (NIR) estimated by vendors plotted as a function of NIR measured during the evaluation. **C.** False non-identification rate (FNIR) minus FtAR are the error rates for errors not related to failure-to-acquire (FNIR-FtAR). FNIR-FtAR is plotted as a function of FtAR for each system. Note FtAR is the dominant source of error.

3.6 Interaction of Acquisition and Matching Systems

Sections 3.3 and 3.5 evaluated acquisition systems using a single reference, commercial matching system. This section analyzes the performance of the eight matching systems included in the 2019 Rally (see Section 2.2) on face samples collected by the ten 2019 Rally acquisition systems.

Fig. 5 shows radar plots of matching system performance (MS.1 - MS.8) across acquisition systems discussed in the previous sections (AS.1 - AS.10). The distance from the origin on each radar plot represents the true identification rate (TIR) of a single matching/acquisition system pair. This TIR is expressed both inclusive of failure-to-acquire (cyan curve), which represents the performance of the matching/acquisition system pair as a whole, and exclusive of failure-to-acquire (pink curve), which represents the performance of the matching system only on face samples it was able to template. This matching analysis was done at a threshold that produced a false match rate of $\sim$1:1,000,000.

Fig. 5 shows that the performance characteristics varied substantially across matching systems. However, despite these systems being built by different commercial entities, often in different parts of the world, these performance variations appeared to follow a discernible pattern. Visual inspection of these patterns suggests a taxonomy of matching system robustness that characterizes these systems into three general categories:

- **Specialist** (MS.1) - Matching systems that worked well with only a single acquisition system as characterized by a single peak in Fig. 5.
- **Brittle** (MS.2, MS.3, and MS.4) - Matching systems that performed well with some, but poorly with other acquisition systems: AS.8 and AS.6 as characterized by two prominent dimples in Fig. 5.
- **Robust** (MS.5, MS.6, MS.7, and MS.8) - Matching systems that had generally stable, and high, performance across all acquisition systems as characterized by nearly perfect circles in Fig. 5.

The existence of these robustness classifications was confirmed using the following statistical approach. For each matching system, the TIR (excluding failure-to-acquire) of the best acquisition system was identified. In Fig. 5, this is the point on the pink curve furthest from the origin and is denoted by a red filled point (●), or points (●, ●, ...) in case of an exact tie. The TIR of the best matching/acquisition system pair(s) was compared with the performance of the other acquisition/matching system pairs, for a given matching system, using pairwise tests (Bonferroni corrected p$<$ 0.05, Fisher's exact test). If the system pair performance was found to be significantly different from the best, that system pair is denoted by a filled black point (●). System pair performance that was not significantly different from the best system pair is denoted by an open black point (○).

Using this approach, we can see the TIR of all matching/acquisition system pairs for matching systems MS.7 and MS.8 were not significantly different from the best. For MS.5 and MS.6, only system AS.6 had significantly lower performance than the best matching/acquisition system pair. These matching systems appear capable of reliably matching face samples from a number of diverse acquisition systems. For MS.2, MS.3, and MS.4, five acquisition systems had performance significantly lower than the best system pair. Finally, for MS.1, all nine system pairs had significantly lower performance than the best matching/acquisition system combination. This indicates acquisition systems MS.1, MS.2, MS.3, and MS.4 may only work with face samples from select acquisition systems.

This analysis also shows that face acquisition systems, despite operating on the same individuals within the same use-case and environment, produce samples with different qualities and characteristics. To illustrate this point, Fig. 6 shows face samples from four test volunteers that participated in the 2019 Rally across the ten different acquisition systems. The differing qualities and characteristics of face samples are harder for *some* matching systems to match. For instance, acquisition system AS.6 had significantly lower performance for all matching systems with the exception of MS.7 and MS.8. Acquisition systems AS.2, AS.3, and AS.4 had significantly lower performance on specialist and brittle matching systems. This suggests that the quality of the images from system AS.6 and, to a lesser extent, AS.2, AS.3, and AS.4 is more challenging for some modern, commercial biometric matching systems tested as part of the 2019 Rally. Face samples acquired on these acquisition systems may yield inconsistent performance when used with different matching systems for different purposes.

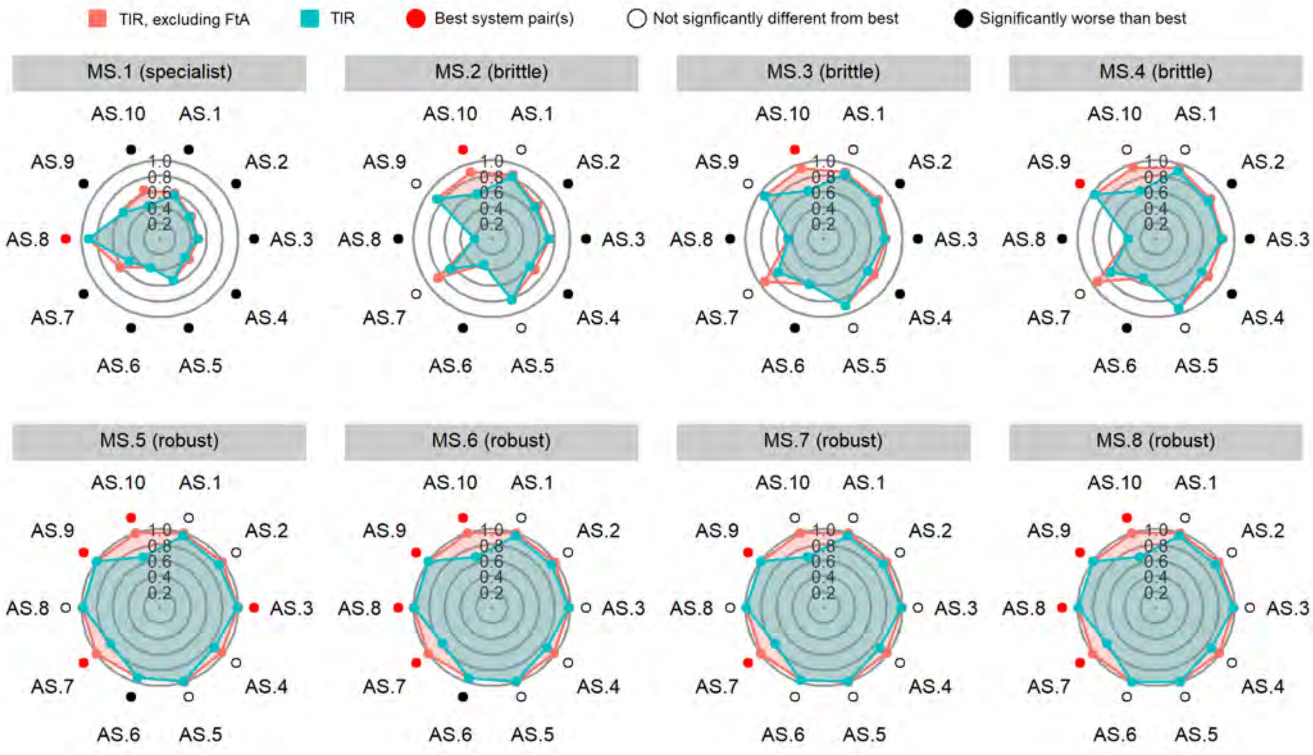

Fig. 5. Matching system performance characteristics across acquisition systems. Radar plots show the true identification rate (TIR) of each face matching system in combination with each of ten face acquisition systems at a false match rate threshold of ∼1:1,000,000. Distance from the origin indicates better performance. Cyan curve: TIR performance inclusive of all sources of error (total performance focus). Pink curve: TIR performance excluding failure-to-acquire (matching focus). Points around the perimeter indicate matching system robustness as follows: Red filled points (●) - the best matching/acquisition system pair or a series of pairs tied for best performance; Open points (○) - matching/acquisition system pairs that had performance not significantly different from the best; Filled points (●) - matching/acquisition system pairs that had performance significantly lower than the best (Bonferroni corrected $p < 0.05$, Fisher's exact test). Note three different families of matching system performance characteristic curves: specialist, brittle, and robust.

4 Discussion

4.1 On the Importance of Scenario Evaluations

Scenario evaluations characterize the simulated performance of an end-to-end biometric systems with live subjects. The error rates and error conditions observed during a scenario evaluation are more directly relatable to how a system will perform if deployed operationally than the error rates observed during technology testing. For example, Figure 5 shows a non-identification rate between 1 and 7% for most acquisition systems in our evaluation. Conversely, the error rates of the best performing systems in a technology evaluation, such as [4], are usually measured in exceedingly small decimals (a false non-match rate of 0.0001, for example). Clearly, if a face recognition system were deployed to a location, such as an airport, we would expect error rates more in line with the numbers suggested by the scenario evaluation model.

However, despite being the topic of half an international standard [1] and a general increase in the pace of operational biometric deployments, scenario evaluations of biometric technology are still relatively rare. New capabilities to perform scenario evaluations of biometric systems are being developed, specifically in Germany [24] and the United States [25]. However, the overall ratio of scenario evaluations to the number of planned or in-progress operational deployments remains low. We believe increasing the volume and variety of biometric scenario evaluations is crucial to ensure accurate, high-performing, and equitable biometric systems.

4.2 The Current State of High-throughput Biometric Face Systems

This research describes the results of a scenario evaluation of commercial biometric systems using live subjects within an operationally relevant high-throughput scenario. In general, we found that high-throughput face acquisition systems are fast and well accepted by users (Sections 3.2 & 3.1). Over half of the acquisition systems tested were able to acquire a biometric sample in under five seconds. Five acquisition systems received positive satisfaction ratings from over 95% of tested subjects. Four others received positive satisfaction ratings from over 90% of tested subjects. However, biometric systems struggle to maintain effectiveness in high-throughput operations. While four face acquisition systems were able to acquire samples that generated templates for over 99% of tested subjects, only one was able to successfully match all tested subjects (Section 3.3). Importantly, these levels of effectiveness were not well anticipated by commercial providers of these acquisition systems, meaning if system owners used vendor provided estimates of

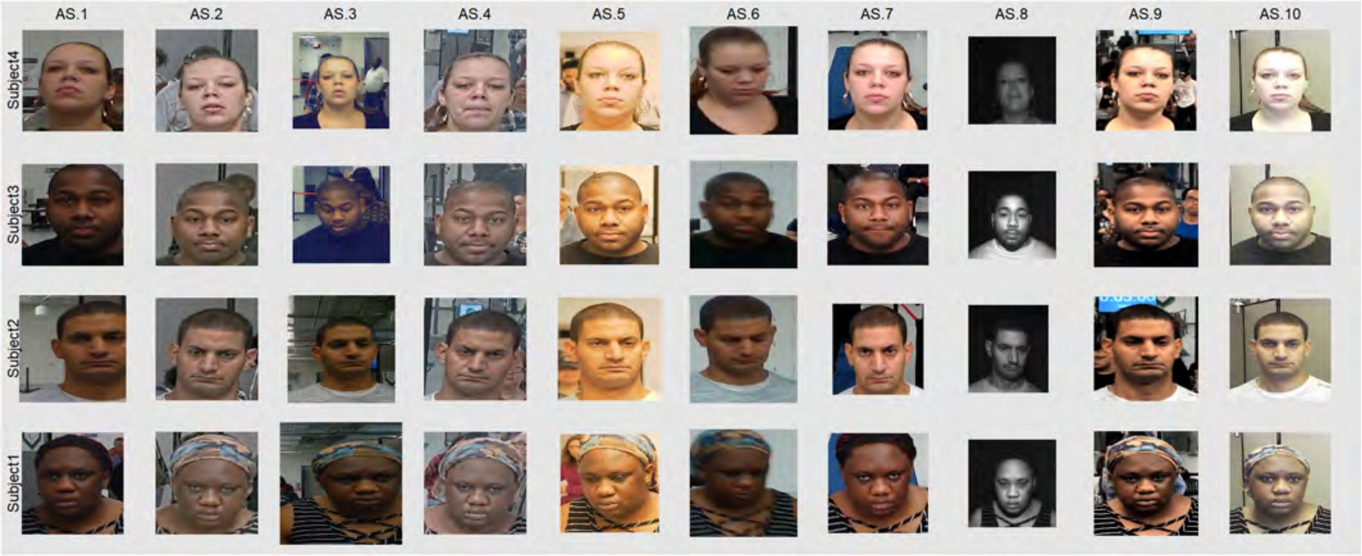

Fig. 6. Face biometric samples from four test volunteers (rows) that participated in the 2019 Biometric Technology Rally across the ten face acquisition systems that were evaluated in this study (columns).

performance to plan an operational deployment, serious deficiencies, potentially requiring costly reworks or program cancellation, could have occurred (Section 3.6).

High-throughput biometric systems will play an important and increasing role in the identification of individuals at land borders, airports, and similar use-cases. Results of the 2019 Biometric Technology Rally documented here indicate broad improvements in acquisition system performance when compared to the results of the 2018 Rally [17]. However, much continued design work is needed since the majority of acquisition systems failed to meet effectiveness goals. We believe smart, face-aware capture sub-systems [26] will become crucial to obtaining high quality, matchable, biometric samples in the future.

4.3 Performance is Dominated by Acquisition Errors

The ability to independently evaluate the relative contribution of acquisition errors and matching errors is a unique component of our test methodology. We show that acquisition errors can dominate the performance of a high-throughput biometric system, especially of those systems using robust matching algorithms (Figure 4 & Section 3.5). This finding suggests that a renewed focus on user interaction during image acquisition (camera placement, camera adjustment, informative signage, etc.) offers significant room to improve the performance of most, but not all, high-throughput, public facing, face biometric systems. This is in contrast to many face biometric technology evaluations that seek an ever-expanding performance envelope (pose, lighting, size, angle, etc.) in which face matching systems can be successful. These technology evaluations are likely the most direct route to improve the raw performance of face biometric algorithms. Appropriate control of user interaction during face image capture can not only control variation in the captured image but can also help ensure that images are captured and matched only for appropriate individuals, an important consideration in crowded environments. This becomes ever more critical to guarantee privacy as matching performance envelopes are expanded.

4.4 Choice of Acquisition System can Impact Matching System Performance

In addition to acquisition systems, we tested eight commercial face matching systems across samples collected by the acquisition systems in our study. Half of those matching systems were classified as either specialist or brittle. On these systems, the acquisition system used to acquire images to arrive at estimates of performance matters greatly. For example, had matching system MS.4 been tested with only images from AS.6, its observed TIR, excluding failure to acquire, would have been 52.8% (Figure 5). Had this same matching system been tested with images from AS.10, its observed TIR would have been 95.5%, a difference of over 40%. However, on systems classified as robust, the image source mattered little. For example, the observed TIR using MS.8 on the same images from AS.6 and AS.10 was 99% and 100%, respectively. Brittle or even specialist matching systems may provide acceptable performance in use-cases where acquisition can be carefully controlled. However, use-cases where acquisition systems or environment varies across sites require robust matching systems to ensure consistent performance. These results highlight the importance of considering matching performance across a range of acquisition systems.

4.5 Future Work

Finally, much future work in the area of high-throughput biometric system testing remains. First, this report does not investigate differences in accuracy rate across demographic variables, but such efforts, similar to [19] and [27], are a necessary part of the evaluation process for any public facing biometric system. Second, while the 2018 and 2019 Rallies were the first of their kind scenario tests to measure whole system biometric effectiveness, they did so at the individual level, processing a single individual at a time. Many use-cases for high-throughput biometric systems require processing groups of people, such as families. Future tests should consider the effectiveness of these systems in processing groups. Finally, the 2019 Rally demonstrated that it is possible for high-throughput biometric systems to be extremely effective, with one system capturing and matching nearly all subjects that interacted with the system. While desirable from an evaluation standpoint, this level of capture effectiveness raises concerns that these systems will capture indiscriminately, inadvertently capturing samples from individuals outside of the biometric process. Additional work is needed to ensure that biometric systems appropriately capture images for individuals that have chosen to interact with a biometric process and explicitly avoid capturing images for individuals that have not. The authors hope these issues will receive more attention in the near future.

5 Acknowledgments

This research was sponsored by the Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034. The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers. The data were acquired using the IRB protocol “Development and Evaluation of Enhanced Screening” number 120180237, approved by New England IRB. The paper authors acknowledge the following author contributions: All authors conceived the work and executed the evaluation, J. Howard, Y. Sirotin, and J. Hasselgren designed the scenario test, performed statistical data analysis, and wrote the paper. Y. Sirotin developed the statistical technique for categorizing matching algorithms across acquisition systems. J. Tipton and A. Vemury edited the paper. Authors are listed alphabetically. Correspondence regarding this work can be sent to info@mdtf.org.

The authors thank the staff of the SAIC Identity and Data Sciences Laboratory: Jeffrey Chudik for initial data wrangling; Andrew Blanchard and Kirsten Huttar for providing software engineering support; Laura Rabbitt and Nelson Jaimes for human factors support; Frederick Clauss and Jeffrey Chudik for providing integration engineering support; Rebecca Rubin for technical document support and editing; as well as Colette Bryant, Rebecca Duncan, Patty Hsieh, and Kevin Slocum for support in executing the 2019 Biometric Technology Rally.

References

[1] “ISO/IEC 19795-2:2007 Information technology–biometric performance testing and reporting–part 2: Testing methodologies for technology and scenario evaluations,” Standard, 2007.

[2] C. Watson, G. Fiumara, E. Tabassi, S. Cheng, P. Flanagan, and W. Salamon, “Fingerprint vendor technology evaluation, nist interagency/internal report 8034: 2015.”

[3] G. W. Quinn, P. J. Grother, M. L. Ngan, and J. R. Matey, “IREX IV: Part 1, Evaluation of iris identification algorithms,” Tech. Rep., 2013.

[4] P. Grother, M. Ngan, and K. Hanaoka, “Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification,” National Institute of Standards and Technology, Tech. Rep., Apr 2018, https://www.nist.gov/sites/default/files/documents/2018/04/03/frvt_report_2018_04_03.pdf, last accessed on 06/07/18.

[5] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, “The Megaface benchmark: 1 million faces for recognition at scale,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4873–4882.

[6] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, “Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[7] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, “Iarpa janus benchmark-c: Face dataset and protocol,” in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 158–165.

[8] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, “Iarpa janus benchmark-b face dataset,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 90–98.

[9] K. Kotwal, Z. Mostaani, and S. Marcel, “Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.

[10] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa, “Recognizing disguised faces in the wild,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 2, pp. 97–108, 2019.

[11] G. Fiumara, “Nail to nail (n2n) fingerprint capture challenge,” 2017.

[12] S. M. Furman, S. M. Furman, B. C. Stanton, M. F. Theofanos, J. M. Libert, and J. D. Grantham, *Contactless fingerprint devices usability test*. US Department of Commerce, National Institute of Standards and Technology, 2017.

[13] I. ODNI, “Iarpa-baa-16-04 (thor)(2016),” *URL https://www.iarpa.gov/index.php/research-programs/odin/odin-baa*.

[14] M. Theofanos, B. Stanton, and C. A. Wolfson, “Usability & biometrics: Ensuring successful biometric systems.” [Online]. Available: https://www.nist.gov/sites/default/files/usability_and_biometrics_final2.pdf

[15] C. Garvie, “Statement of clare garvie before the u.s. house of representatives committee on oversight and reform hearing on facial recognition technology (part 1): Its impact on our civil rights and liberties,” Center on Privacy and Technology at Georgetown Law, Washington, D.C., Tech. Rep., 2019.

[16] “ISO/IEC 19795-1:2006 Information technology–biometric performance testing and reporting–part 1: Principles and framework,” International Organization for Standardization, Standard, 2006.

[17] J. J. Howard, A. A. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. Vemury, “An investigation of high-throughput biometric systems: Results of the 2018 Department of Homeland Security Biometric Technology Rally,” in *2018 Nineth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. IEEE, 2018.

[18] C. Manaher. (2018) Privacy impact assessment for the traveler verification service. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/privacy-pia-cbp056-tvs-january2020_0.pdf

[19] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, “Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems,” *Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, 2019.

[20] Y. B. Sirotin, “Usability and user perceptions of self-service biometric technologies.” International Biometric Performance Conference, Gathersburg, MD, 2016, https://www.nist.gov/sites/default/files/documents/2016/12/06/07_ibpc_usability_20160414.pdf, last accessed on 06/07/18.

[21] J. J. Howard, A. A. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. Vemury, “On efficiency and effectiveness tradeoffs in high-throughput facial biometric recognition systems,” in *2018 Nineth*

IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS). IEEE, 2018.

[22] U. Census, "Race and ethnicity," United States Census Bureau, Tech. Rep., Jan 2017, https://www.census.gov/mso/www/training/pdf/race-ethnicity-onepager.pdf, last accessed on 06/07/18.

[23] L. R. Rabbitt, J. A. Hasselgren, C. Cook, and Y. B. Sirotin, "Measuring satisfaction with standard survey instruments and single-button responses on kiosks," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 1429–1433.

[24] R. Breithaupt, "New developments in biometric security testing and certification." International Face Performance Conference, Gathersburg, MD, 2018, https://nigos.nist.gov/ifpc2018/presentations/20_breithaupt_2018.11.28_IFPC_2018_NIST_V07.pdf, last accessed on 03/17/20.

[25] The Maryland Test Facility. [Online]. Available: https://www.dhs.gov/science-and-technology/maryland-test-facility

[26] "ISO/IEC 24358 Face-aware capture subsystem specifications," International Organization for Standardization, https://www.iso.org/standard/78489.html, Approved Work Item, 2019.

[27] J. J. Howard, Y. B. Sirotin, , and A. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 Tenth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. IEEE, 2019.

Jacob Hasselgren. Jacob Hasselgren holds a Master of Science degree from Purdue University with a thesis that characterizes habituation and learning effects on iris recognition system performance. He actively contributes to research in biometrics, computer vision, human factors (particularly the human machine interaction with identification devices), test engineering, and data science. He also serves as the project editor for an ISO technical report discussing differential impact of demographics on biometric systems. He currently works as the Test Director and Principal Engineer of the SAIC Identity and Data Sciences Lab which provides third party, independent assessment of identification technologies at the Maryland Test Facility.

John Howard. Dr. Howard received his Ph.D. in Computer Science from Southern Methodist University. His thesis was on pattern recognition models for identifying subject specific match probability. His current research interests include biometrics, computer vision, machine learning, testing human machine interfaces, pattern recognition, and statistics. He has served as the principal investigator on numerous R&D efforts across the intelligence community, Department of Defense, and other United States Government agencies. He is a member of the SAIC Identity and Data Sciences Lab and currently the Principal Data Scientist at the Maryland Test Facility.

Yevgeniy Sirotin. Dr. Sirotin holds a Ph.D. in Neurobiology and Behavior from Columbia University and has diverse research interests in behavior and human computer interaction. His past research spans mathematical psychology (cognitive modeling), neurophysiology (multi-spectral imaging of the brain), psychometrics (mechanisms of visual and olfactory perception), biometrics (design and testing of identity systems), and human factors (usability). He currently works as Principal Investigator and Manager of the Identity and Data Sciences Laboratory at SAIC which supports applied research in biometric identity technologies at the Maryland Test Facility.

Jerry Tipton. Jerry Tipton is the Program Manager and Director of SAIC's Identity and Data Sciences Lab. He has over 20 years experience in the biometric industry with over 15 years managing research portfolios in support of various United States Government agencies. He currently supports the S&T at the Maryland Test Facility.

Arun Vemury Arun Vemury received his Master of Science in Computer Engineering from George Washington University. His current research interests include biometrics, pattern recognition, machine learning, and operations research. He serves as the Director of the Biometrics and Identity Technology Center for S&T.