Appeared in Proceedings of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Los Angeles, USA, October 2018

An Investigation of High-Throughput Biometric Systems: Results of the 2018 Department of Homeland Security Biometric Technology Rally

John J. Howard Andrew J. Blanchard
Yevgeniy B. Sirotin Jacob A. Hasselgren
The Maryland Test Facility
{john, andrew, yevgeniy, jacob}@mdtf.org

Arun R. Vemury
Department of Homeland Security, Science and Technology Directorate
arun.vemury@hq.dhs.gov

Abstract

The 2018 Biometric Technology Rally was an evaluation, sponsored by the U.S. Department of Homeland Security, Science and Technology Directorate (DHS S&T), that challenged industry to provide face or face/iris systems capable of unmanned, traveler identification in a high-throughput security environment. Selected systems were installed at the Maryland Test Facility (MdTF), a DHS S&T affiliated biometrics testing laboratory, and evaluated using a population of 363 naive human subjects recruited from the general public. The performance of each system was examined based on measured throughput, capture capability, matching capability, and user satisfaction metrics.

This research documents the performance of unmanned face and face/iris systems required to maintain an average total subject interaction time of less than 10 seconds. The results highlight discrepancies between the performance of biometric systems as anticipated by the system designers and the measured performance, indicating an incomplete understanding of the main determinants of system performance. Our research shows that failure-to-acquire errors, unpredicted by system designers, were the main driver of non-identification rates instead of failure-to-match errors, which were better predicted. This outcome indicates the need for a renewed focus on reducing the failure-to-acquire rate in high-throughput, unmanned biometric systems.

1. Introduction

High-throughput biometric systems differ from the traditional biometric collection and matching application in several regards. First, high-throughput systems are designed to process hundreds to hundreds of thousands of individuals in a short time span. Examples include automated admission to a major sporting event or facilitating immigration at a major airport. Because of these volumes, high-throughput systems emphasize the speed with which an identification or verification operation can be achieved. Second, with high volumes, even sub percentage error rates can result in a significant number of individuals experiencing delays or requiring alternate processing. Consequently, high-throughput systems must achieve high biometric accuracy while also minimizing processing times per person. Finally, these systems are often optionally manned or purposefully understaffed (one monitor for several systems) and as such must be intuitive to the naive user.

At a high level, all biometric systems have two generalized failure points; failure-to-acquire (error in the Data Capture or Signal Processing Subsystem) and failure-to-verify/identify (error in the Matching or Decision Subsystem) [6]. There is copious work documenting failure-to-verify/identify performance across various algorithms and datasets [1, 8, 7]. However, less attention has been focused on the problem of reducing failure-to-acquire rates, despite the evidence that, in operational biometric deployments, this error rate can be the dominant factor in overall false non-recognition rate [3, 4, 9, 10, 11].

A failure-to-acquire error occurs when a biometric system is unable to capture a sample that passes the system's internal quality standards. Root causes can range from the uncorrectable, such as a missing digit, to the easily mollified, such as an obstruction from glasses or an off-axis pose. Where failure-to-acquire numbers are reported (typically in commercial marketing material) there is little consistency in how they are calculated. For example, two systems can estimate failure-to-acquire rates based on different populations (demographics, habituation, size) and collection conditions (manned vs. unmanned, time constraints). Moreover, some biometric systems still report failure-to-acquire rates for loosely defined sub-groups, such as "uncooperative" subjects, which have little relevance when applied to public-facing systems. Consequently, it is difficult to use these data points to 1) compare different biometric systems or 2) anticipate the failure-to-acquire rate a system will experience if deployed in an operational setting.

The 2018 Biometric Technology Rally ("Rally") was de-

978-1-5386-7180-1/18/$31.00

Appeared in Proceedings of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Los Angeles, USA, October 2018

signed to provide a repeatable test methodology with which to measure the state of the biometric industry, specifically regarding throughput, capture capability, matching capability, and user satisfaction metrics. This report documents the performance of biometric systems tested as part of the Rally. Section 2 describes the required and optional components of systems under test during the Rally, as well as the evaluation metrics. Section 3 presents the quantitative results of the Rally. Finally, Section 4 presents the conclusions of this research.

2. Methodology

Systems under consideration to participate in the 2018 Biometric Technology Rally ("Rally systems") were required to collect and provide facial biometric imagery capable of supporting identification operations. Rally systems were also required to be unmanned and physically constrained to a 7 by 8 foot space. Inside this space, Rally systems were free to use any combination of form factor, hardware, software, etc. to meet the goals of the Rally. Rally systems were solely responsible for automatically directing all aspects of human subject interaction necessary to perform a collection operation (i.e. instructions, feedback, etc.). Finally, Rally systems were required to collect, process, and submit data within the period of time in which the user was interacting with the system (i.e. no batching, offline processing).

At a minimum, Rally systems were required to provide a single face image. Optionally, Rally systems could provide up to three face images, up to three iris pairs and up to three identification results during each user transaction. To support on-board identifications, a gallery of historical images (1848 images from 525 unique subjects) was provided prior to testing. The rationale for allowing multiple sample/identification submissions per subject was to encourage Rally systems to attempt capture/identification operations as quickly as possible to reveal any trade-offs between acquisition time and biometric accuracy within and across systems [5].

The Rally design called for a minimum of 320 subjects to interact with each Rally system. Of these, 90% were expected to be in the historic face gallery. Consequently, Rally systems performing on-board identifications were responsible for reporting "out-of-gallery" for some percentage of the test population. This concept, known as open-universe identification, was also documented in [2] and is decidedly more difficult than identification evaluations where a mated sample exists in the gallery for all probes such as [7].

Eleven commercial companies participated in the Rally ("Rally participants"), ranging from mid-sized businesses to large multi-national corporations. Six systems collected face images only. The remaining five collected both face and iris imagery. Additionally, eight of the eleven performed on-board face matching against the historic gallery and provided identification results (six of the face only systems and two of the face/iris systems). Broadly, Rally systems fell into one of two interaction categories; those that instructed test subjects to temporarily stop to allow for collection (*pause-and-go* systems), and those that intended test participants to proceed unabated (*walk-through* systems).

Rally participants were briefed regarding the test design, evaluation metrics, system requirements, and operating environment starting in the Fall 2017. All information was made available via a series of webinars with supporting material publicly hosted. All aspects of the Rally experimental design were transparent to all Rally participants (i.e. no information was withheld as an experimental control). Rally participants were given two full days to install and configure their systems at the MdTF prior to test execution. It was the hope of the design team that both this long lead and install time would allow Rally participants to optimize their systems for the Rally tasking and environment.

Rally systems were tested with 363 naive users over a six day period in March 2018. The test population was evenly split in terms of gender, roughly one-third Caucasian and two-thirds African American, and ranged in age from 20 to 85. All subjects were briefed regarding the purpose of the Rally and that Rally systems were intended to perform biometric identifications. Subjects were given general directions to comply with presented instructions but were otherwise unfamiliar with the specifics of individual Rally systems.

Following the in-brief, face and iris enrollment images were collected from each subject at a manned station, operated by trained MdTF personnel. Each test subject was assigned a unique ID, which was worn on a wrist-band during the test. The ID was used to link the subject to their enrollment and subsequent transactions. Next, subjects were organized into groups of fifteen and each group was queued at a Rally station. Each subject in a group entered the Rally station one-at-time after a scan of their ID wrist-band. After entering the Rally station, subjects were given no additional direction from MdTF personnel - all system interaction instructions were provided by the Rally system. Image and identification results were submitted via a common web-based application programming interface (API). Following their interaction with each Rally system, test subjects were asked to provide a satisfaction score that rated their overall experience (see Figure 1). Rally systems were given five minutes to process the entire group of fifteen. Discounting the average time required for scanning wrist-bands and rating satisfaction, this left on average 10 seconds for each subject to use a Rally system. Each group of subjects used all eleven Rally systems. To mitigate habituation and carry-over affects, station order was fully counterbalanced.

Rally systems were evaluated on both user-centric [12]

Appeared in Proceedings of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Los Angeles, USA, October 2018

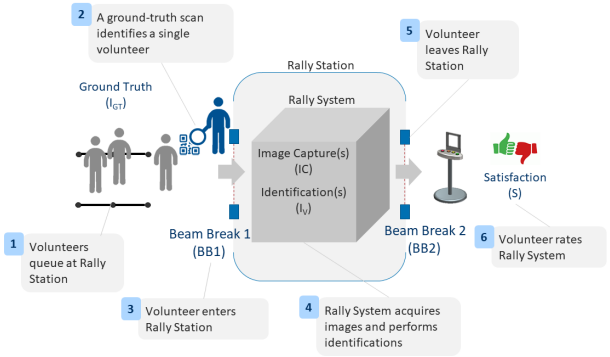

Figure 1. 2018 Biometric Technology Rally Test Protocol

and technical performance [6] criteria. First, user interaction time with each Rally system was measured via a series of beam breaks. Second, the user's satisfaction score for each transaction was recorded. Third, the ability of each system to acquire face and iris (optionally) imagery was measured. Fourth, the ability of this imagery to match using a MdTF matching engine (algorithm & threshold[1]) was measured. Finally, for Rally systems performing on-board matching, the ability of that matching engine to report the correct identity was measured. Acquisition and matching metrics were recorded as a function of time (e.g. failure-to-acquire by 5 seconds). Prior to the Rally, threshold and objectives were established based on results of prior biometric technology testing at the MdTF (see Table 1). These were purposefully designed to be aggressive to encourage innovation. However, we also believe these levels are representative of the rates required for high-throughput systems to be successful if ever widely adopted.

Table 1. 2018 Biometric Technology Rally Metrics

Metric	Threshold Level	Objective Level
Efficiency	10 s	5 s
Satisfaction	$> 90\%$	$> 95\%$
Face Failure to Acquire	$< 5\%$ @ 20 s	$< 1\%$ @ 20 s $< 5\%$ @ 5 s
Iris Failure to Acquire	$< 5\%$ @ 20 s	$< 1\%$ @ 20 s $< 5\%$ @ 5 s
Face Vendor True Identification Rate	$> 95\%$ @ 20 s	$> 99\%$ @ 20 s $> 95\%$ @ 5 s
Face MdTF True Identification Rate	$> 95\%$ @ 20 s	$> 99\%$ @ 20 s $> 95\%$ @ 5 s
Iris MdTF True Identification Rate	$> 95\%$ @ 20 s	$> 99\%$ @ 20 s $> 95\%$ @ 5 s

We believe the test methodology described above is a first-of-its-kind in regards to biometric testing. Aspects that distinguish this test design from past evaluations are:

- Providing a consistent test population and test methodology for calculating comparable performance metrics across different commercial biometric systems, specifically in regards to failure-to-acquire rate.
- Challenging commercial biometric system providers to meet ambitious performance and timing objectives and to do so in an unmanned operating mode.
- Encouraging commercial biometric system providers to think holistically regarding test subject instruction and feedback, not simply sensor and algorithm design.
- Advocating that commercial biometric systems consider the trade-off between throughput and accuracy when designing systems to meet the demands of high-throughput, high-security environments [5].

3. Results

This section describes the results of the Rally. To comply with information sharing agreements between the DHS S&T and the various Rally participants, Rally system names are aliased in the remainder of this report.

3.1. Anticipated Results

Prior to the Rally, Rally participants were asked to estimate their anticipated failure-to-acquire and true identification rates with respect to their system's facial biometric components (face biometrics being common across all Rally systems). Estimates of failure-to-acquire rate ranged from 0% to 1.5%, while estimates of true identification rate ranged from 78% to 100%. Notably, only five of the eleven participants were able to provide an expected failure-to-acquire rate. Table 2 shows the metrics anticipated by the system designers.

3.2. Efficiency Results

The key measure of efficiency in this evaluation was transaction time, which was quantified as the amount of time a test subject spent between the entry and exit beam breaks (*BB1* and *BB2*, see Figure 1). This time is inclusive of walk-up time (after breaking *BB1*), interactions with the Rally system, and walk-out time (before breaking *BB2*).

Meeting the threshold average transaction time of 10 s was required for a Rally system to reliably process all subjects within each test group. The objective level (*5 s*) challenged Rally participants by setting an aggressive target not typically achieved by commercial biometric systems.

Figure 2 shows the distributions of transaction times measured for each Rally system. All face systems were

[1]Algorithm threshold was set to a elicit an FMR of between 1/5000 and 1/10000. Corresponding FNMR is roughly 0.9% on controlled, high-quality samples

Table 2. 2018 Biometric Technology Rally Anticipated Metrics

System Alias	Anticipated Face Failure to Acquire Rate	Anticipated Face True Identification Rate
System 1	0.0150	0.980
System 2	0.0000	1.000
System 3	NA	NA
System 4	NA	NA
System 5	0.0000	1.000
System 6	NA	0.950
System 7	NA	0.990
System 8	NA	0.950
System 9	0.0003	0.991
System 10	NA	0.780
System 11	0.0000	0.970

able to maintain a mean transaction time under the threshold value of 10 seconds, with the exception of *System 6* ($\mu = 10.59s$). Three face systems also met the objective value of 5 seconds. Face/iris systems took longer to collect with two face/iris systems exceeding the threshold requirement; *System 2* ($\mu = 11.18\,s$) and *System 4* ($\mu = 10.82\,s$). *System 5* was the only face/iris system to maintain a mean transaction time under the objective value. The sample size for each distribution in Figure 2 was 363 transactions with the exception of *System 4* ($n = 357$), *System 6* ($n = 359$), *System 8* ($n = 361$), and *System 2* ($n = 362$).

3.3. Satisfaction Results

Satisfaction was measured using a rating kiosk positioned at the exit of each Rally station. Subjects were asked to rate their experience using a 4-level "happiness scale". Figure 3 shows the counts of recorded satisfaction scores by Rally system. The aggregate metric (S) quantifies the percentage of positive satisfaction scores ("Happy" or "Very Happy") out of the total. Sample sizes are identical to the numbers presented in Section 3.2. The aggregate satisfaction score for all systems was in the high eighties or nineties with the exception of *System 6* ($S = 0.69$). However, only two systems, *System 8* ($S = 0.96$) and *System 9* ($S = 0.97$) met the objective requirements of the Rally.

3.4. Acquisition and Matching Results

Biometric performance of Rally systems was measured cumulatively. First, failure-to-acquire errors (FtA) were recorded whenever a system failed to produce a templatable image within the indicated time interval. Second, overall true identification rates (TIR) were computed for both the MdTF matching engine (mTIR) and for Rally system matching engines (vTIR, optional). These rates are inclusive of any FtAs, false non *and* false positive identifications (i.e. this is the percentage of the population that transited each system and was correctly identified). Table 3 and 4 outline the matching results at 5 and 20 seconds after subject beam break (*BB1*, See Figure 1), respectively. The sample sizes used to calculate the rates shown in these tables was 363 with the exception of *System 8* ($n = 361$).

Four Rally systems were able to meet the FtA rate (FtAR) thresholds for face at the 5 second mark. Two of these were also able to meet the TIR requirements at 5 seconds using both the MdTF and on-board matching engines. Interestingly, two systems were able to meet the more stringent 20 second FtAR objective ($< 1.0\%$) in under 5 seconds. However, neither was able to meet the 99% TIR objectives at 20 seconds, indicating that these Rally systems were not able to significantly improve their capture/identification performance given additional time.

Table 3. 2018 Biometric Technology Rally Matching Results at 5 seconds

System Alias	Face FtAR	Iris FtAR	Face mTIR	Face vTIR	Iris mTIR
System 1	0.091	0.521	0.904	0.713	0.477
System 2	0.981	0.997	0.019	0.008	0.003
System 3	1.000	1.000	0.000	NA	0.000
System 4	0.625	0.777	0.372	NA	0.220
System 5	0.157	0.152	0.810	NA	0.000
System 6	0.405	NA	0.595	0.344	NA
System 7	0.047	NA	0.826	0.634	NA
System 8	0.006	NA	0.978	0.981	NA
System 9	0.008	NA	0.978	0.967	NA
System 10	0.022	NA	0.948	0.915	NA
System 11	0.129	NA	0.851	0.813	NA

Table 4. 2018 Biometric Technology Rally Matching Results at 20 seconds

System Alias	Face FtAR	Iris FtAR	Face mTIR	Face vTIR	Iris mTIR
System 1	0.025	0.127	0.970	0.763	0.862
System 2	0.113	0.113	0.882	0.879	0.840
System 3	0.102	0.132	0.887	NA	0.815
System 4	0.110	0.140	0.884	NA	0.815
System 5	0.124	0.096	0.826	NA	0.000
System 6	0.245	NA	0.755	0.457	NA
System 7	0.014	NA	0.887	0.645	NA
System 8	0.003	NA	0.975	0.989	NA
System 9	0.008	NA	0.978	0.970	NA
System 10	0.022	NA	0.948	0.915	NA
System 11	0.129	NA	0.851	0.813	NA

3.5. Comparison of Acquisition and Match Errors

Using the values in Table 3 or 4 we can calculate the false non-identification rate (FNIR) per station by taking

Appeared in Proceedings of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Los Angeles, USA, October 2018

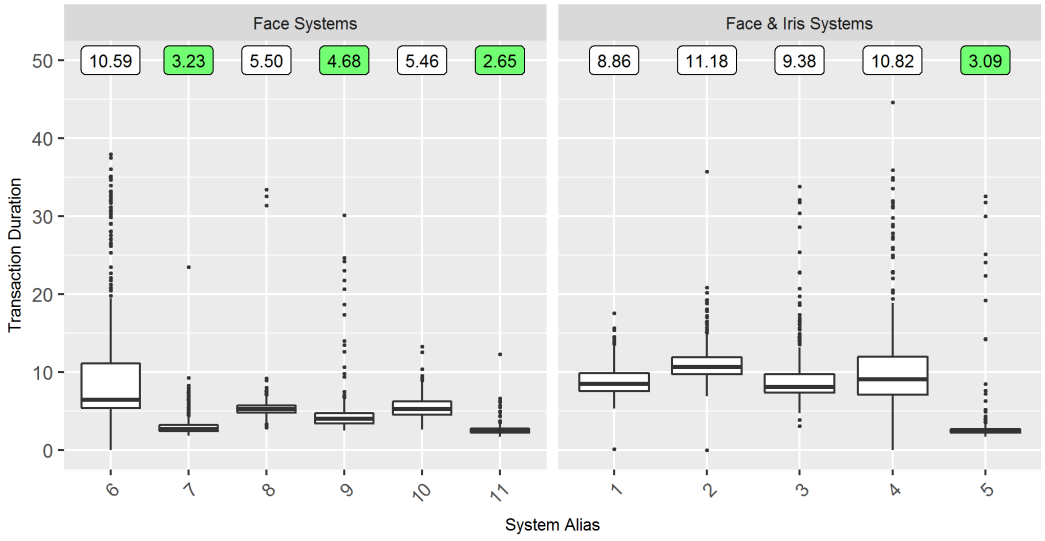

Figure 2. 2018 Biometric Technology Rally Efficiency Metrics

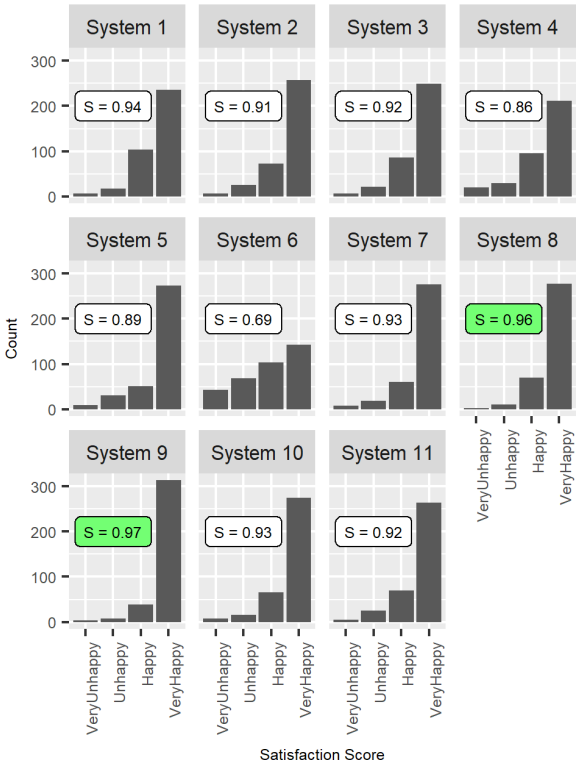

Figure 3. 2018 Biometric Technology Rally Satisfaction Results

1 − mTIR. Subtracting FtAR from FNIR yields the observed false non-match rate, or non-FtAR errors. This is the percentage of transactions that resulted in templatable images but ultimately failed to match. Figure 4 plots the non-FtA error rate (FNIR-FtAR) versus the FtAR for face per Rally system at 20 seconds. Points above the identity line ($y = x$) indicate a larger portion of the errors from that system stemmed from a failure-to-match an acquired sample. Points with statistically different non-FtAR than FtAR are filled ($p < 0.05$; Wilson score interval with continuity correction). Points below identity indicate a larger portion of errors at that system stemmed from being unable to acquire a templatable image. Six of the eleven Rally systems had a greater incidence of FtAR than other errors (points below identity). Only two had greater non-FtAR errors (points above identity), stemming from poor image quality.

4. Conclusions

4.1. On the Feasibility of Rapid, High-Quality Face Capture

It is important to contextualize the Rally objective of a 1% failure rate for a biometric system. In a high-throughput environment where thousands of users may utilize a biometric system in the course of a single day, a 1% failure rate corresponds to dozens of people experiencing delays or requiring alternate processing. To meet similar volumes, modern IT systems typically provide service reliability measured in the far fractions of a percent (e.g. 99.99...% uptime). We believe biometric systems must meet similar standards of reliability to be feasible under comparable volumes.

Tellingly, none of the Rally systems were able to meet the 99.0% true identification rate objective. Additionally, there was a clear divide between the minority of systems that were able to achieve greater than or approximately 95% identification rates and the majority of the Rally systems, whose true identification rate was in the 70-80% range. We

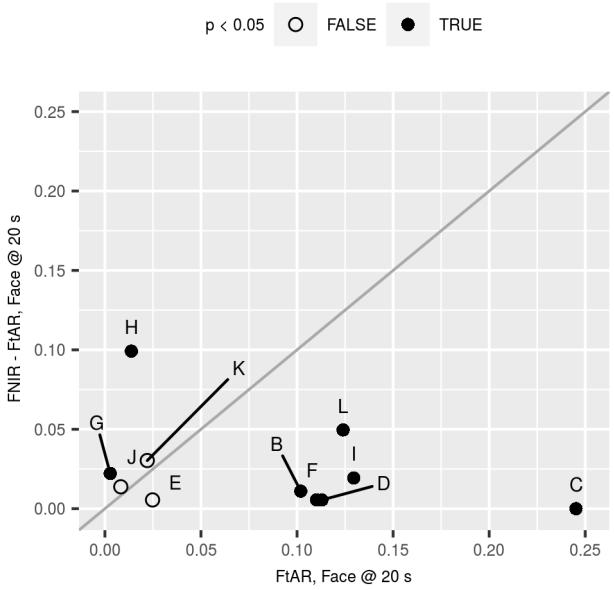

Figure 4. Failure to Acquire versus Failure to Match Error Rates

believe this is strong evidence of the difficulty associated with high-throughput, high-reliability biometric identification and highlights the need for improved system designs.

For example, of the four Rally systems with strong matching and acquisition performance at the 5 second mark (*Systems 7, 8, 9,* and *10*), only one showed notable improvement when considering the full 20 second window. The noted improvement was at *System 7*, which moved from a 4.7% failure-to-acquire rate to a 1.4% failure-to-acquire rate given the extra time allotment. However, even this improvement was not due to any explicit speed accuracy trade-off. In fact, few Rally systems had methods of handling non-optimal user behavior or to recall/hold users that did not produce a quality image. These approaches should have improved overall accuracy, despite increasing the duration of the transaction, especially considering all test subjects were successfully enrolled by a human operator. Innovative user interface design and process flow in the future may begin to drop failure rates below Rally objectives.

4.2. Failure to Acquire as a Major Source of Error

Figure 4 shows that the dominant cause of test subject non-identification for seven of the eleven stations that participated in the Rally was failure-to-acquire an image. In six of these stations this evidence was very strong ($p > 0.05$). This indicates that, while biometric algorithm accuracy continues to warrant investigation, failure-to-acquire errors can be frequently and overwhelmingly more impactful, outstripping failure-to-match errors by six-fold or more. We conclude that the primary means of performance improvement for at-least half of the Rally systems, would be to focus on sensor placement, sample acquisition, and general human factors. These data points also indicate that acquisition errors may have been the dominant form of error were these solutions deployed in a high-throughput environments. Biometric system vendors who focus on holistic system design, inclusive of human factors considerations, may be able to reduce their failure-to-acquire rate to levels that make them feasible in high-throughput, unstaffed, high-security environments.

4.3. Discrepancies Between Anticipated and Measured Error Rates

Comparing the anticipated and measured error rates (Tables 2 and 4, respectively) shows a clear divide between the expectations of biometric system vendors and the realities of high-throughput, unmanned biometric collection. Six of the eleven Rally participants elected not to provide failure-to-acquire estimates, indicating this metric may be poorly understood or documented from an industry perspective. Measured failure-to-acquire rates were uniformly higher than those anticipated by the Rally participants and the median failure-to-acquire across all systems was $> 10\%$ for both face and iris modalities.

Estimates of identification rate were slightly better. Of the nine Rally systems who estimated true identification rate, two had measured metrics that exceeded their predications; *System 8* (Anticipated $= 0.95$; Measured $= 0.975$) and *System 10* (Anticipated $= 0.78$, Measured $= 0.948$). Five others were off by greater than 10%. Had these vendor-provided, anticipated error rates been used to plan the details of an operational deployment, such as expected throughput, staffing requirements, etc., costly redesigns would have likely been required. Furthermore, the test population used in the Rally was compliant, cooperative, undistracted, unencumbered, and paid for their efforts. In other words, the error rates experienced during the Rally are likely a lower bound on what could be expected should these technologies be implemented in operational, high-throughput environments. This suggests that the main determinants of system performance are poorly understood by many system designers. It also illustrates that inclusive commercial evaluations, which measure acquisition and matching performance using consistent populations and methodologies, are important to the overall progress of the biometric industry.

Acknowledgements

This research was funded by the Department of Homeland Security, Science and Technology Directorate on contract number W911NF-13-D-0006-0003. The views presented here are those of the authors and do not represent those of the Department of Homeland Security or of the U.S. government.

Appeared in Proceedings of IEEE 9th International Conference on Biometrics: Theory, Applications and Systems (IEEE BTAS), Los Angeles, USA, October 2018

References

[1] P. J. Grother, M. L. Ngan, and K. Hanaoka. Face recognition vendor test (FRVT) ongoing. 2018. 1

[2] P. J. Grother, M. L. Ngan, and G. W. Quinn. Face in video evaluation (five) face recognition of non-cooperative subjects. Technical report, 2017. 2

[3] J. A. Hasselgren. Scenario tests for immigration exit. International Biometric Performance Conference, Gathersburg, MD. NIST, 2016. 1

[4] J. A. Hasselgren. Measuring usability at the Maryland Test Facility. Federal Identity Summit. AFCEA, 2017. 1

[5] J. J. Howard, A. A. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. R. 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. 2, 3

[6] I. ISO. IEC 19795-1: Information technology–biometric performance testing and reporting-part 1: Principles and framework. *ISO/IEC, Editor*, 2006. 1, 3

[7] 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*, pages 4873–4882, 2016. 1, 2

[8] G. W. Quinn, P. J. Grother, M. L. Ngan, and J. R. Matey. IREX IV: Part 1, evaluation of iris identification algorithms. Technical report, 2013. 1

[9] Y. B. Sirotin. Efficient test design for biometric exit scenarios. NIST, International Biometric Performance Conference, Gathersburg, MD, 2016. 1

[10] Y. B. Sirotin. Usability and user perceptions of self-service biometric technologies. NIST, International Biometric Performance Conference, Gathersburg, MD, 2016. 1

[11] Y. B. Sirotin, J. A. Hasselgren, and A. Vemury. Usability of biometric iris-capture methods in self-service applications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):2019–2023, 2016. 1

[12] M. Theofanos, B. Stanton, and C. A. Wolfson. Usability and Biometrics: Ensuring successful biometric systems. Technical report, NIST, 2013. 2