Operational Tradeoffs in the 2018 Department of Homeland Security Science and Technology Directorate Biometric Technology Rally

1st Jacob A. Hasselgren
Test Director
Identity and Data Sciences Lab
The Maryland Test Facility
Upper Marlboro, MD, USA
jacob@mdtf.org

2nd John J. Howard, Ph.D.
Data Scientist
Identity and Data Sciences Lab
The Maryland Test Facility
Upper Marlboro, MD, USA
john@mdtf.org

3rd Yevgeniy B. Sirotin, Ph.D.
Human Factors Scientist
Identity and Data Sciences Lab
The Maryland Test Facility
Upper Marlboro, MD, USA
yevgeniy@mdtf.org

4th Andrew J. Blanchard
Software Engineer
Identity and Data Sciences Lab
The Maryland Test Facility
Upper Marlboro, MD, USA
andrew@mdtf.org

5th Arun Vemury
Program Director
Biometric Technology Center
Department of Homeland Security
Science and Technology Directorate
Washington D.C, USA
arun.vemury@hq.dhs.gov

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

This research documents the operational tradeoffs between these three measures of system performance. Specifically, we perform two tradeoff analyses: efficiency versus effectiveness and satisfaction versus both efficiency and effectiveness. These tradeoff analyses allow us to determine how and if these three performance measures are related in the various kinds of biometric systems we tested. For example, are higher throughput systems also more effective? Do people prefer systems that are faster or more effective? Our results show there is no clear relationship between how quickly a system can process a user and how well it can identify the user. Furthermore, there was also no significant relationship observed between how quickly a system can process a user and how satisfied the user is with the system. However, there was a strong relationship between how well a system identifies a user and how satisfied the user is with the system. These outcomes suggest that some systems could benefit by leveraging additional collection time to collect a higher quality image. Users did not tend to prefer faster systems but did prefer a system they thought was working as intended. Finally, these results also show that in regards to public acceptance, systems designers should focus on correctly identifying larger populations of users rather than how quickly a given user can be processed.

***Index Terms*—biometrics, biometric technology rally, , effectiveness, efficiency, facial recognition, satisfaction, system design, tradeoffs, usability**

I. Introduction

High-throughput biometric systems are an emerging class of biometric technology that has particular application to use cases at the U.S. Department of Homeland Security (DHS) [3], [4], [8]–[10], [12]. High-throughput biometric systems are designed to process large numbers of people in a short amount of time [1], [5]. Examples include high traffic border crossings and screening crowds at a major sporting event. At these volumes, even error rates that would typically be considered acceptable for a biometric system (one to three percent) could cause hundreds to thousands of non-identification exceptions, meaning high-throughput systems must be extremely accurate. Additionally, these systems have only a limited amount of time to process individual users and as such must be able to achieve these high accuracies while also considering throughput. Finally, in order to scale, high-throughput biometric systems are typically unmanned or understaffed (one manager for several systems) and consequently need to be intuitive and provide a satisfying user experience [1], [5], [6]. A crucial step in designing a high-throughput biometric system is determining how much focus should be on what human-centered design calls “usability goals” [7], [11]. Those goals are:

- Effectiveness - a measure of how well a biometric system can collect and identify a user’s biometric signature

This research was funded by the Department of Homeland Security Science and Technology Directorate on contract number W911NF-13-D-0006-0003 and 70RSAT18CB0000034. 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.

- Efficiency - a measure of how quickly a biometric system can collect and identify a user's biometric signature
- Satisfaction - a measure of the user's postive attitudes and perceptions of the biometric system

Ideally, a high-throughput biometric system would score well in respect to all three goals. However, system designers often make choices that prioritize certain attributes [2]. Placing emphasis on one or more of the usability goals may impact the observed performance in respect to the remaining goals. For example, a system designed to focus on fast transaction times may sacrifice image quality and thus matching capability. Another system might be designed for longer transaction times, accepting that it might impact people's perceptions of the process. Understanding the relationship, or lack thereof, between these three measures of performance is important as DHS and other government stakeholders seek to influence the biometric community to design systems that better suit their needs, particularly in the challenging and relatively novel high-throughput scenario.

The DHS Science and Technology (S&T) Directorate 2018 Biometric Technology Rally ("rally") was an evaluation designed to measure the state of the industry in regards to high-throughput biometric systems. Specifically, the rally was designed to measure the efficiency (throughput), effectiveness (capture capability, matching capability), and user satisfaction of biometric systems that have an average transaction time of ten seconds or less. The results of the rally cataloged the performance of systems under these conditions. They have been presented elsewhere [5], but an important corollary to these numbers, and the topic of this research, is the relationship between the aforementioned usability goals of efficiency, effectiveness and satisfaction.

II. Test Methodology

Eleven commercial companies participated in the rally ("rally participants"), which took place at the Maryland Test Facility (MdTF), a DHS S&T affiliated laboratory, in March of 2018. Each rally participant was required to install a system ("rally system") at the MdTF that collected facial biometric samples capable of supporting identification operations. Rally systems were also required to be unmanned and physically constrained to a 7 ft. by 8 ft. space. The area around each space was instrumented with beam break sensors, configured by the MdTF and placed at the entrance and exit of each station. This instrumentation allowed the collection of transaction times. The MdTF also setup and configured a standard, four button, satisfaction kiosk ("Very Happy", "Happy", "Unhappy", and "Very Unhappy") at the exit of each station, which allowed the collection of satisfaction scores.

Inside their space, rally systems were free to use any combination of form factor, hardware, software, etc. to meet the objectives 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 via an application programming interface (API) within the period of time in which the user was interacting with the system (i.e. no batching, offline processing). Images submitted via this API were saved and cataloged by the MdTF. A single, common, commercial, matching algorithm was then used by the MdTF to identify these images against a gallery of other facial samples that had previously been acquired by the MdTF. This rate, given the moniker the MdTF true identification rate (mTIR)[1], is our measure of system effectiveness.

The rally was announced in November of 2017 and executed in March of 2018, giving rally participants four months to design and configure their systems. The details of the high-throughput workflow, as discussed in Section I were explained to each rally participant. Additionally, threshold and objective levels for efficiency and effectiveness were defined by DHS S&T. The threshold for participation was to provide imagery that could be used to identify 95% of test volunteers within 10 seconds. The objective goal was to provide imagery that could be used to identify 99% of test volunteers within 5 seconds. These levels were designed to be aggressive but are also what we believe to be representative of the level of performance required to be successful in high-throughput environments.

During the execution of the rally, three hundred and sixty-three (363) human subjects ("test volunteers") were recruited from the general public. Test volunteers were split into groups of 15 and were processed over the course of six days. Test volunteer groups queued at each rally system and entered the system sequentially. Each test volunteer was fitted with a wristband with a printed QR code containing that test volunteer's test ID. Before each test volunteer entered each station, an MdTF staff scanned the volunteer's wristband to establish a groundtruth identity. The order in which each group experienced each rally system was counterbalanced to avoid habituation and carry-over effects. Figure 1 demonstrates the overall process at each rally system.

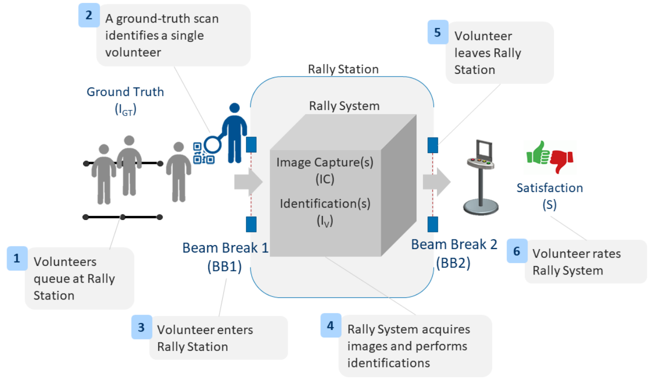
2 A ground-truth scan identifies a single volunteer
5 Volunteer leaves Rally Station
Rally Station
Rally System
Ground Truth (I_{GT})
Image Capture(s) (IC)
Identification(s) (I_V)
Satisfaction (S)
Beam Break 1 (BB1)
Beam Break 2 (BB2)
6 Volunteer rates Rally System
1 Volunteers queue at Rally Station
3 Volunteer enters Rally Station
4 Rally System acquires images and performs identifications

Fig. 1. Rally Collection Process

[1] As part of our test, rally participants were also given the option of performing an "onboard" identification with an algorithm of their choosing. This metric was called vendor true identification rate, or vTIR. However, an investigation of vTIR is out of scope in this research.

At a minimum, rally systems were required to provide a single face image. Optionally, rally systems could provide up to three face images and up to three iris pairs during each user transaction. The rationale for allowing multiple sample submissions per test volunteer was to encourage rally participants to attempt capture operations as quickly as possible to reveal any trade-offs between acquisition time and biometric accuracy both within and across systems [6]. When a system did provide multiple samples per test volunteer, the true identification rate was calculated using only the last submitted image.

III. Results

This section presents the results of the tradeoff studies between efficiency, effectiveness, and user satisfaction at the various rally systems. To comply with information sharing agreements between DHS S&T and the various rally participants, rally system names are aliased for the remainder of this research.

A. *Effectiveness vs. Efficiency*

First, we examined the relationship between our measure of effectiveness, mTIR, and our measure of efficiency, transaction times. This allows us to answer questions such as "Were rally systems that took longer more successful at identifying people?" In this context, mTIR is transactionally inclusive of failure-to-acquire rate (FtAR), meaning it is equivalent to the percentage of subjects who transited through a rally system and were subsequently identified by that system. Transaction time is defined as the delta between a test volunteer entering and exiting a rally station, measured via beam breaks (see Figure 1). Figure 2 plots the relationship between efficiency (average transaction time, x-axis) and effectiveness (facial mTIR at 20 seconds, y-axis). Triangles denote rally participants that provided iris images in addition to facial samples. The shaded yellow region corresponds to rally threshold requirements and the shaded green region corresponds to rally objective requirements, as discussed in Section II.

Figure 2 suggests no clear monotonic relationship between rally system efficiency and effectiveness. For example, the top three most efficient stations had average transaction times below four seconds and the three least efficient stations had average transaction times above ten seconds. Yet these two subsets of rally systems had a very similar ability to collect matchable faces, with average mTIR performance in the 85% - 90% range.

Interestingly, effectiveness of stations with *mid-range* efficiency appeared to be much higher than the extremes. Indeed, rally systems that met both efficiency and effectiveness thresholds (Crestone, Castle, and Elbert) had average transaction times between 4.5 and 9 seconds. The other rally system that had markedly higher effectiveness numbers, Lincoln, also had average transaction times in this range. On one extreme, it appears that capturing face images too rapidly could present challenges in terms of face image quality (e.g. motion blur or issues with pose control). On the other extreme, it appears that performing face capture as a step in a larger iris acquisition effort both significantly increases transaction times and degrades overall face capture performance, the latter being largely due to increased facial failure-to-acquire rates.

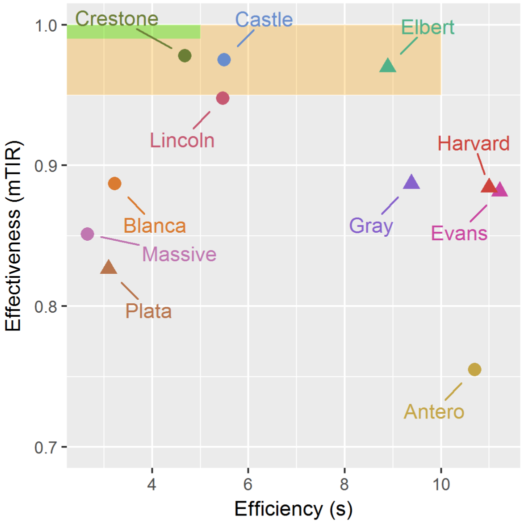

Fig. 2. Tradeoff Between Efficiency and Effectiveness

B. *Satisfaction vs. Effectiveness and Efficiency*

Next, we examine the tradeoff between the user satisfaction metric and both efficiency and effectiveness. Whereas efficiency and effectiveness are easier to estimate in an engineering or laboratory environment, satisfaction can only be measured through user testing. Since user testing can be costly, it is important to understand how satisfaction with biometric systems is related to that system's efficiency and effectiveness. For example, are satisfaction scores lower for systems with slower performance or for systems with a higher failure rate? We quantified user satisfaction by aggregating satisfaction scores collected by after each rally system and calculating the percentage that were "Very Happy" or "Happy" divided by the total number of scores.

Figure 3 plots the relationship between efficiency and satisfaction (left panel), and effectiveness and satisfaction (right panel) across rally systems. The black lines and shaded regions on the plots show linear regression slopes and 95% confidence regions respectively. Although satisfaction tended to be lower for slower systems, this effect was not statistically significant ($p = 0.144$). On the other hand, satisfaction was strongly related to system effectiveness, with system effectiveness explaining 70% of the variability in average satisfaction ($p < 0.05$). This strong relationship is surprising given that rally systems, and therefore the test volunteers, were not required to perform different actions based on their identification outcome. Indeed, rally systems did not even know at the time of collection the result of the identification operation using that image. Thus, this relationship appears to exist between *perceived outcome*, i.e. people think the system is working for them, and satisfaction. Furthermore, the relationship between satisfaction and effectiveness but not efficiency indicates that highly effective systems may be better suited to use-cases where satisfaction is important, even if efficiency is sacrificed. In other words, systems that compromise effectiveness for efficiency may be less satisfying to use. We suspect this effect would have been more pronounced given realistic exception processing for those that were not successfully identified, which was out of scope for the rally. Regardless, in the context of high-throughput systems, these results suggest that a system user is much more likely to notice how well a system works, rather than how quick a system works.

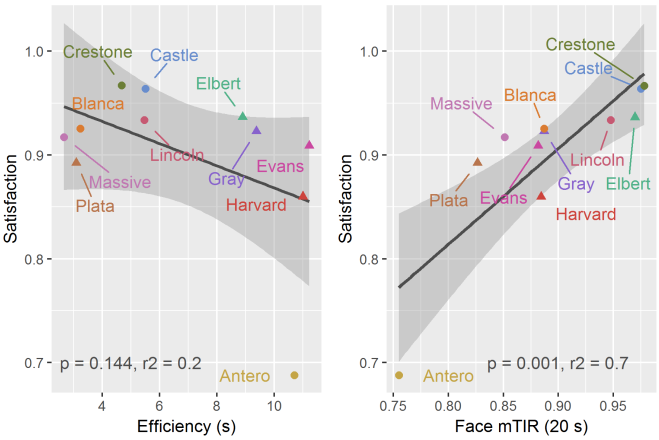

Fig. 3. Relating Efficiency (left) and Effectiveness (right) to Satisfaction

IV. Conclusions

In any engineering effort, it is important to consider the relationship between measures of system success during the design process. For biometric systems, these measures of success have been established as efficiency, effectiveness, and satisfaction [11]. However, until now there has been little work in understanding how these measures are interconnected in modern, commercial biometric systems. For example, does increasing efficiency, i.e. decreasing average transaction time, results in less effective systems by lowering observed identification rate? Here, we show that while rapidly collected images (under 4 seconds) may occasionally result in lower quality images due to blur, no significant relationship exists between transaction times and identification rates. Systems that were both very fast, and relatively slow, tended to have very similar identification rates. Systems with mid-range transaction times performed the best, suggesting a measured approach to biometric sample acquisition may result in the most efficient system designs.

Another important outcome of this work is the relationship between user satisfaction and the efficiency and effectiveness of a biometric system. We showed a strong relationship existed between user satisfaction and effectiveness while no significant relationship existed between user satisfaction and efficiency. This, again, points to the notion that expediting image acquisition at the expense of image quality may result in sub-optimal system performance, not from a lowered identification rate, but from a lowered public acceptance.

It is important to understand the efficiency, effectiveness and user satisfaction of all biometric systems. However, with the unique demands of public facing, high-throughput biometric systems, it becomes important to understand these metrics not only in isolation but also the relationship they have with each other. We believe this work and the 2018 DHS S&T Biometric Technology Rally establish a methodology for calculating these relationships across systems. The results suggest designers of these systems should focus on the cybernetics of the human-machine interaction to ensure additional acquisition time actually increases sample quality and thus reduces error rates. They also suggest this additional time, if properly utilized, is unlikely to effect public acceptance. These results could lead to better design of high-throughput systems in the future.

Acknowledgment

This research was funded by the Department of Homeland Security Science and Technology Directorate on contract number W911NF-13-D-0006-0003 and 70RSAT18CB0000034. 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. The authors would like to thank the staff at the MdTF, our colleagues at the DHS, and the commercial entities who participated in our test for their efforts while executing the 2018 DHS S&T Biometric Technology Rally.

References

[1] T. Dunstone and N. Yager. *Biometric System and Data Analysis: Design, Evaluation, and Data Mining*. Springer, 2009.

[2] K. N. Faddis, J. R. Matey, and J. Stracener. Improving tactical biometric systems through the application of systems engineering. *IET Biometrics*, 2:1–9, 2012.

[3] J. A. Hasselgren. Scenario tests for immigration exit. IBPC. NIST, 2016.

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

[5] J. J. Howard, A. J. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. R. Vemury. An investigation of high-throughput biometric systems: Results of the 2018 Department of Homeland Security Biometric Technology Rally. The 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems. IEEE, (in press) 2018.

[6] J. J. Howard, A. J. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. R. Vemury. On efficiency and effectiveness tradeoffs in high-throughput facial biometric recognition systems. The 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems. IEEE, (in press) 2018.

[7] S. ISO. 13407: 1999. *Human-centred design processes for interactive systems*.

[8] Y. Sirotin. Efficient test design for biometric exit scenarios. IBPC. NIST, 2016.

[9] Y. Sirotin. Usability and user perceptions of self-service biometric technologies. IBPC. NIST, 2016.

[10] Y. Sirotin, J. A. Hasselgren, and A. Vemury. Usability of biometric iris-capture methods in self-service applications. HFES 2016 Annual Meeting, pages 2019–2023. HFES, 2016.

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

[12] A. Vemury. Biometric Concepts of Operation in the Airport Environment. IBPC. NIST, 2016.