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

On Efficiency and Effectiveness Tradeoffs in High-Throughput Facial Biometric Recognition Systems

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

This research discusses the evaluation of biometric systems that are designed to process hundreds to tens of thousands of individuals in short time spans. We propose a method for evaluating a system's performance across capture attempts for the purpose of identifying characteristics that are advantageous in these high-throughput environments. We also present a novel modification to the traditionally accepted biometric performance metrics of failure-to-acquire, and true-match rate. Namely, this paradigm shift holds that these metrics are a function of time and, as such, vary with the time available for a biometric system to interact with a user. This research demonstrates the utility of these time-based metrics in evaluating the performance of multiple, commercially available, high-throughput systems. We show that different biometric systems have notably different time-based performance curves using a corpus of data collected during the 2018 Department of Homeland Security, Science and Technology Directorate (DHS S&T) Biometric Technology Rally. These curves and the deviations between them are useful when quantifying the suitability of a technology, evaluated via scenario testing, for deployment in an operational environment where the throughput of the target population is a key performance parameter.

1. Introduction

The performance testing of modern biometric systems is based on the ISO/IEC 19795 Standard. Specifically, Part 1 addresses testing principles and frameworks [2], Part 2 addresses testing methodologies for technology and scenario evaluation [3], and Part 6 addresses testing methodologies for operational evaluation [4]. Performance metrics, such as failure-to-enroll, failure-to-acquire, false-non-match rate, and false-match rate are introduced in `Section 4.6` of [2]. For example, failure-to-acquire rate is defined as the "proportion of verification or identification attempts for which the system fails to capture or locate an image or signal *of sufficient quality*".

The definitions in `Section 4.6` do not mention an affiliation between these metrics and time or number of capture attempts. However, `Section 5.4` of [2] implies a relationship when it introduces the concepts of presentations, attempts, and transactions, noting that "Biometric systems often process a sequence of samples in a single attempt, *... collecting samples until one of sufficient quality is obtained, or the system times out.*" `Section 5.5.2` of [2] highlights two factors that impact this collection time, noting that system throughput is "...based on both computational speed and human-machine interaction" and that "...adequate throughput rates [are] *critical for the success of any biometric system*" (all emphasis added by the authors).

From a reading of these sections we believe the following statements to be consistent with the standard. First, quality impacts failure-to-acquire and failure-to-match rate. Second, quality varies with sample acquisition time and sample acquisition attempt. Third, the acquisition time per person (the inverse of throughput) is based on a number of factors and is critical to the success of a biometric systems.

However, in their current form, no biometric testing standard offers a methodology for measuring the impact of collection time or collection attempt number on sample quality and thus failure-to-acquire/match rate. For example, `Section 8` of [2] details the specifics of how to calculate and analyze failure-to-acquire rate. This section acknowledges this rate will "... depend on thresholds for sample quality, as well as the allowed duration for sample acquisition", but simply recommends setting and reporting these values along with the observed failure-to-acquire rate. This recommendation treats collection time as a property fixed by the design of the biometric system test and reported independently of measures of system acquisition and matching rates. However, in practice, setting quality/acquisition-time thresholds in a consistent way across numerous, diverse biometric systems is often not possible, making this recommendation difficult to implement.

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

In this research we propose a different and novel approach that instead treats collection time and collection attempts as 1) factors incorporated into the design of the experiment and 2) closely integrated with reported acquisition and matching rates. This approach allows these rates to be more directly compared across systems and reveals both between system and within system trade-offs between speed and accuracy. These concepts are particularly useful when comparing the performance of high-throughput biometric systems that must maintain high match rates but have a limited amount of time to interact with individual users. Section 2 of this research presents the characteristics that distinguish high-throughput biometric systems from their traditional counterparts. It also outlines the concept of capture and time-based performance metrics in more detail and presents the specifics of a biometric system evaluation that was designed and executed with the intent of quantifying these metrics. Section 3 then shows capture and time-based performance graphics for eleven biometric systems and documents how these metrics are useful for identifying characteristics that are advantageous in a high-throughput environment and for comparing performance between different high-throughput systems. Finally, Section 4 presents the overall conclusions of this research.

2. Methodology

2.1. High-throughput Biometric Systems

High-throughput biometric systems differ from traditional biometric collection and matching applications in several regards. First, high-throughput systems are designed to process hundreds to tens of thousands of individuals in a short time span. Example use cases 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 a biometric 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. Therefore, high-throughput systems must achieve very high biometric accuracy while simultaneously minimizing processing times per person. For context, modern information technology (IT) systems that service similar volumes typically provide reliability measured in the far fractions of a percent (e.g. 99.99...% uptime). Finally, in order for high-throughput systems to scale, they must be optionally manned or purposefully understaffed (one monitor for several systems) and as such need to be intuitive to the naive user without human intervention.

Because of these key differences, optimal high-throughput systems should implement work-flows that are more advanced than traditional biometric applications in three specific criteria areas:

1. To achieve shortened processing times, high-throughput systems should have a strategy for acquiring a sample of "good-enough" quality quickly and recognize when that condition has been met.

2. To maintain high biometric accuracy, high-throughput systems should adjust when good-enough quality samples are not being acquired. This adjustment can take the form of either external modification, such as altering subject feedback, or internal modification, such as increased lighting, lens focus, etc.

3. To allow for scalability, high-throughput systems should perform collections with minimal operator intervention and be intuitive to the untrained user.

While these concepts may be novel when applied to biometrics systems, the notion of stopping criteria (list criteria one) and operator survival (list criteria two) are well researched topics in autonomous system design [6, 5] and key conclusions of the DARPA Robotics Challenge [7].

2.2. Capture-Based Performance Metrics

To measure a high-throughput biometric system's capacity to meet the criteria outlined in Section 2.1 we introduce the notion of capture-based performance metrics. Given multiple opportunities to capture/submit a biometric sample, systems that are meeting criteria one, should have a decreasing number of overall captures per opportunity. This signals that a biometric system is recognizing when a good-enough quality sample has been acquired. Similarly, systems that are meeting criteria two, should see an increase in the relative quality of samples acquired in latter collection opportunities. This signals that a biometric system is adjusting itself or the subject in a way that results in an increased probability of a match (see Figure 1).

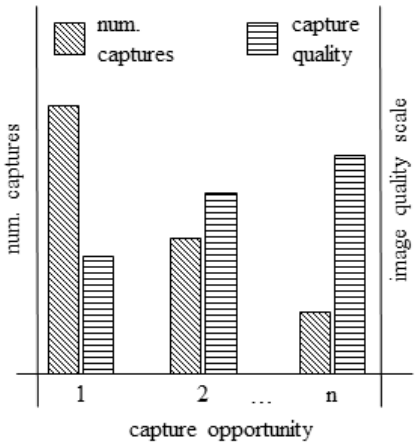

Figure 1. Capture-Based Performance Conceptual

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

2.3. Time-Based Performance Metrics

To measure high-throughput biometric performance across systems, we introduce the notion of time-based performance metrics (see Figure 2). Intuitively, this concept is straightforward. At the time a biometric system becomes aware of a user, i.e. $t = 0$, failure-to-acquire rate is 100% for any system under test. Equivalently, true match/identification rate is 0%. At some time $t = t_s > t = 0$ the system acquires and attempts to match/identify a sample for subject s. This t_s time is dependent (much like throughput) on computational speed and human-machine interaction, and as such is different for each subject s. There exists some subject who had the smallest collection time ($t_{s_{min}}$) and some subject who had the longest collection time ($t_{s_{max}}$). The aggregate failure-to-acquire rate at $t = t_{s_{max}}$ is the failure-to-acquire rate that exists despite a system being allowed more time to collect on any subject. We coin the term $ftar_{\infty}$ for this rate and believe it to be equivalent to the traditional definition of failure-to-acquire rate, i.e. the percentage of subjects from a test population for whom images cannot be acquired because of presentation, feature extraction, or quality control issues [2].

By measuring t_s times across a test population we can create graphics similar to those presented in Figure 2 that show the acquisition and match performance of a given biometric system as a function of time. By fixing this test population across multiple biometric systems, we can create a comparable picture of how different systems acquire and match samples within a given time window. This type of analysis is particularly useful for the evaluation of high-throughput biometric systems where both time and matching rate are critical to the overall performance of the system.

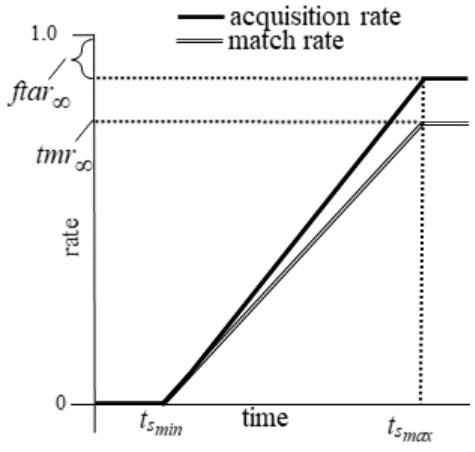

Figure 2. Time-Based Performance Conceptual

2.4. Test Design

To test the concepts outlined in Sections 2.2 and 2.3, a novel biometric system evaluation was proposed by DHS S&T and given the moniker the 2018 Biometric Technology Rally ("Rally"). Eleven industry organizations were selected to participate in the Rally ("Rally participants"). Each Rally participant was allowed to install a biometric acquisition system ("Rally system") at the Maryland Test Facility (MdTF), a DHS S&T affiliated biometric testing laboratory. Rally systems were required to fit in a 7 by 8 foot space and be capable of capturing and submitting face imagery in support of identification operations. Images were submitted via a common web-based application programming interface (API) that saved imagery data for offline processing and recorded the time associated with each submission. Rally systems were required to be unmanned during collection and were solely responsible for directing all aspects of test subject interaction necessary to perform a collection (i.e. instructions, feedback, etc.). Finally, Rally systems were required to collect and submit biometric imagery (via the API) before each test subject left the immediate Rally system area.

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. The rationale for allowing multiple sample submissions per subject was to encourage Rally participants to attempt capture operations as quickly as possible to reveal any trade-offs between acquisition time and biometric accuracy *within* transactions from a single system (see Section 2.2). Rally participants were advised that they would not be penalized for submitting a poorly matching image early in a transaction if a strongly matching image was provided later. They were also advised that they should only submit additional samples if they judged those samples were likely to improve the matching performance of their system. Since all Rally systems were required to submit face images and each did so using a custom work-flow/configuration, the effect of acquisition time on biometric accuracy could be explored *across* systems as well (see Section 2.3).

Six of the eleven Rally systems collected face images only. The remaining five collected both face and iris[1]. Rally systems were tested with a population of 363 users recruited from the general public ("subjects"). Subjects were briefed as to the purpose of the Rally and that Rally systems were intended to perform biometric identifications. They were asked to comply with presented instructions but were otherwise naive to individual Rally systems. Subjects were organized into groups of 15 and queued at a Rally station. The ground truth identity of each subject was established and

[1]Iris acquisition and matching performance is out of scope for this report. However, the fact that some systems were collecting face and iris samples impacted their time-based face performance and as such is noted.

they were directed to enter the Rally station one-at-time. After entering the Rally station, all system interaction guidance was provided by the unmanned Rally system. Rally systems were instructed to maintain an average transaction time of ten seconds or less per test subject. This transaction time was measured by a series of beam breaks at the entrance (*BB1*) and exit (*BB2*) of each Rally station (see Figure 3). Each group of subjects used all eleven Rally systems. To mitigate habituation and carry-over affects, station order was fully counterbalanced.

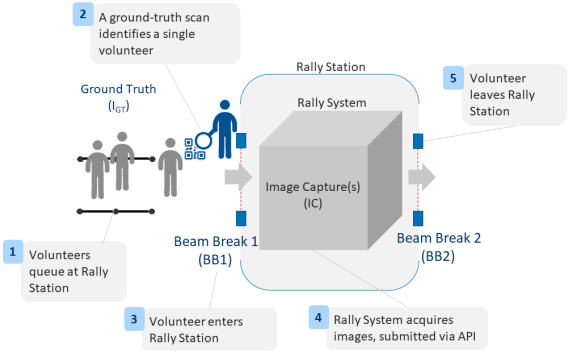

Figure 3. 2018 Biometric Technology Rally Test Protocol

3. Results

To comply with information sharing agreements between the DHS S&T and the Rally participants, Rally system names are aliased in the remainder of this report. A high-level summary of the modalities and interaction model used by each Rally system is provided below. Walk-though systems intended test subjects to proceed unabated. Pause-and-go systems intended test subjects to stop temporarily to allow for collection. *Systems 5* and *11* are walk-through systems. *System 5* collected face and iris while *System 11* collected face-only. *Systems 6, 7, 8, 9,* and *10* are face-only pause-and-go systems. *Systems 1, 2, 3,* and *4* are face and iris pause-and-go systems. The general results of the Rally are presented in [1]. This remainder of this report focuses on the results-oriented application of the concepts discussed in Sections 2.2 and 2.3.

3.1. Capture-Based Performance within Rally Systems

Six of the eleven Rally systems provided multiple facial biometric samples per subject. Investigating the characteristics of these samples, across capture opportunity, allows for an understanding of if systems are meeting the desirable criteria of high-throughput systems as discussed in Section 2.1. Specifically, we are looking to see if systems are recognizing when a good-enough sample has been acquired, as evidenced by a reduction in the number of captures as the

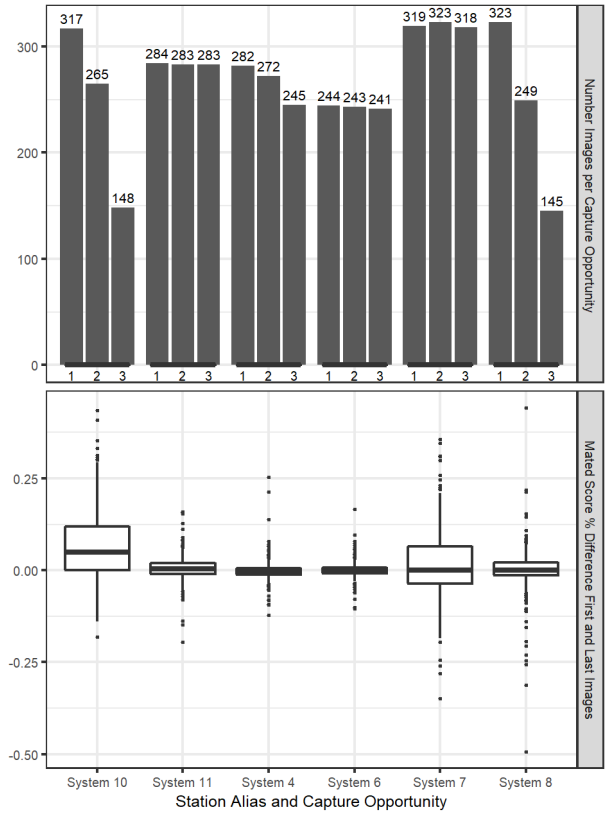

Figure 4. Capture-Based Performance within Rally Systems

number of capture opportunities progresses. Furthermore, if systems are adjusting to the acquisition of poor quality samples, we should see a gradual increase in the rank-one, in-gallery identification score (our measure of capture quality) for images captured in the latter capture opportunities.

The lower graphic in Figure 4 shows the distributions of *differences* in rank-one, in-gallery identification score from the first image submitted to the last image submitted for Rally systems that submitted multiple face images. The upper graphic in Figure 4 shows the number of images submitted in each capture slot (1-3). Only two of the six Rally systems, (*System 8* and *System 10*), appear to have attempted to curtail subsequent collection attempts based on the properties of previously acquired samples. The remaining four systems uniformly collect and submit close to three images per subject across the majority of the test population, regardless of the quality of the first or subsequent acquisitions (upper graphic). Of these two, only *System 10* demonstrated an ability to reliably submit images of increasing quality in latter capture opportunities. The mean differences in in-gallery identification score between the first and last images from all other systems is zero or nearly zero, indicating no tendency toward image quality improvement as capture opportunity progresses (lower graphic).

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

3.2. Time-Based Performance Across Rally Systems

This section applies the time-based performance metrics discussed in Section 2.3 to the dataset collected during the Rally. Figure 5 shows the probability of acquisition (*Pr(Acquired)*) and identification (*Pr(Identified)*) as a function of time after the first beam break (*BB1*, see Figure 3). Identification rate is shown using a MdTF matching engine built around a top-tier commercial algorithm at two different match thresholds. *Pr(Identified, t1)* is a less restrictive threshold, representative of what would be used in closed-set, small-gallery identification operations. *Pr(Identified, t2)* is a higher threshold, more appropriate for Class-N identifications or where larger gallery sizes are expected.

The time-based performance curves in Figure 5 offer several observations that would not have been obvious given the traditional failure-to-acquire and false-non-match rate metrics. First, Rally systems have markedly different time performance profiles. Some systems, such as *6* and *4* have gradual rises in their acquisition rate over time, indicating they are spending a variable amount of time with each subject. Others, such as *Systems 8* and *11* are over 95% of the way to their, very different, $ftar_{\infty}$ plateaus in under three seconds and have fully reached those marks in five seconds. Allowing additional capture time per subject is unlikely to benefit these systems. These profiles have a noticeable impact on system's ordinal ranking at different times, as shown in Table 1 and 2.

Table 1. Identification Performance (t1) Rankings by Time

Rank	$t = 1$	$t = 5$	$t = 10$	$t = 15$
1	System 9	System 8	System 8	System 8
2	System 7	System 9	System 1	System 9
3	System 10	System 10	System 9	System 1
4	System 8	System 1	System 10	System 10
5	System 11	System 7	System 7	System 7
6	System 1	System 11	System 11	System 2
7	System 2	System 5	System 2	System 11
8	System 3	System 6	System 5	System 3
9	System 4	System 4	System 4	System 4
10	System 5	System 2	System 3	System 5
11	System 6	System 3	System 6	System 6

Second, the face/iris systems, most notably *Systems 2* and *3*, started submitting face images later than their face only counterparts. Because of their iris capture component, these systems spent longer positioning/instructing subjects. However, despite this long lead time, the face images from these systems are of excellent quality, as evidenced by the relatively minor drop in identification rate with the *t2* threshold. In fact, the identification performance of the faces from four of the face/iris systems was the second to fifth best in the Rally at fifteen seconds with the *t2* threshold, out performing every face only system except *System 8* (see Table 2). Conversely, some pause-and-go face only stations saw large reductions in identification rate between the two thresholds. For example, *System 7* had relatively high acquisition ($> 97\%$) and identification ($> 90\%$) rates with the *t1* threshold. However, identification rate drops below 50% when the more restrictive *t2* threshold was used. *Systems 9* and *10* showed a similar pattern.

Table 2. Identification Performance (t2) Rankings by Time

Rank	$t = 1$	$t = 5$	$t = 10$	$t = 15$
1	System 9	System 8	System 8	System 8
2	System 8	System 1	System 1	System 1
3	System 7	System 11	System 2	System 2
4	System 10	System 9	System 11	System 3
5	System 11	System 5	System 9	System 4
6	System 1	System 10	System 5	System 11
7	System 2	System 6	System 3	System 9
8	System 3	System 7	System 4	System 5
9	System 4	System 4	System 10	System 10
10	System 5	System 2	System 6	System 6
11	System 6	System 3	System 7	System 7

4. Conclusions

4.1. On the Need for Intelligent Capture Control

Rally systems were given the opportunity to demonstrate their ability to intelligently select when to continue capture operations. Only half of the participants elected to participate in this aspect of the Rally. Of those that did participate, only one system demonstrated an ability to continue capture/image submission operations only when previously acquired/submitted samples were of a lower quality. Despite this poor showing, the ability to understand when an image of sufficient quality has been collected from a given subject is a desirable criteria that can significantly effect performance of high-throughput systems. Notably, no Rally system managed to achieve $> 99.0\%$ identification rate, even with a controlled, compliant, and compensated test population [1]. These results indicate that biometric system developers seeking to enter the high-throughput market should adopt a renewed focus on quantifying biometric sample quality and modifying system process flows based on these outcomes. Efforts in this area may begin to align high-throughput biometric system error rates with those of other high-volume IT systems.

4.2. On the Utility of Time-Based Performance Evaluations

Because of their unique requirements, some aspects of high-throughput biometric systems evaluation will differ from the traditional biometric system testing protocols outlined in [2, 3, 4]. These evaluations must take into account

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

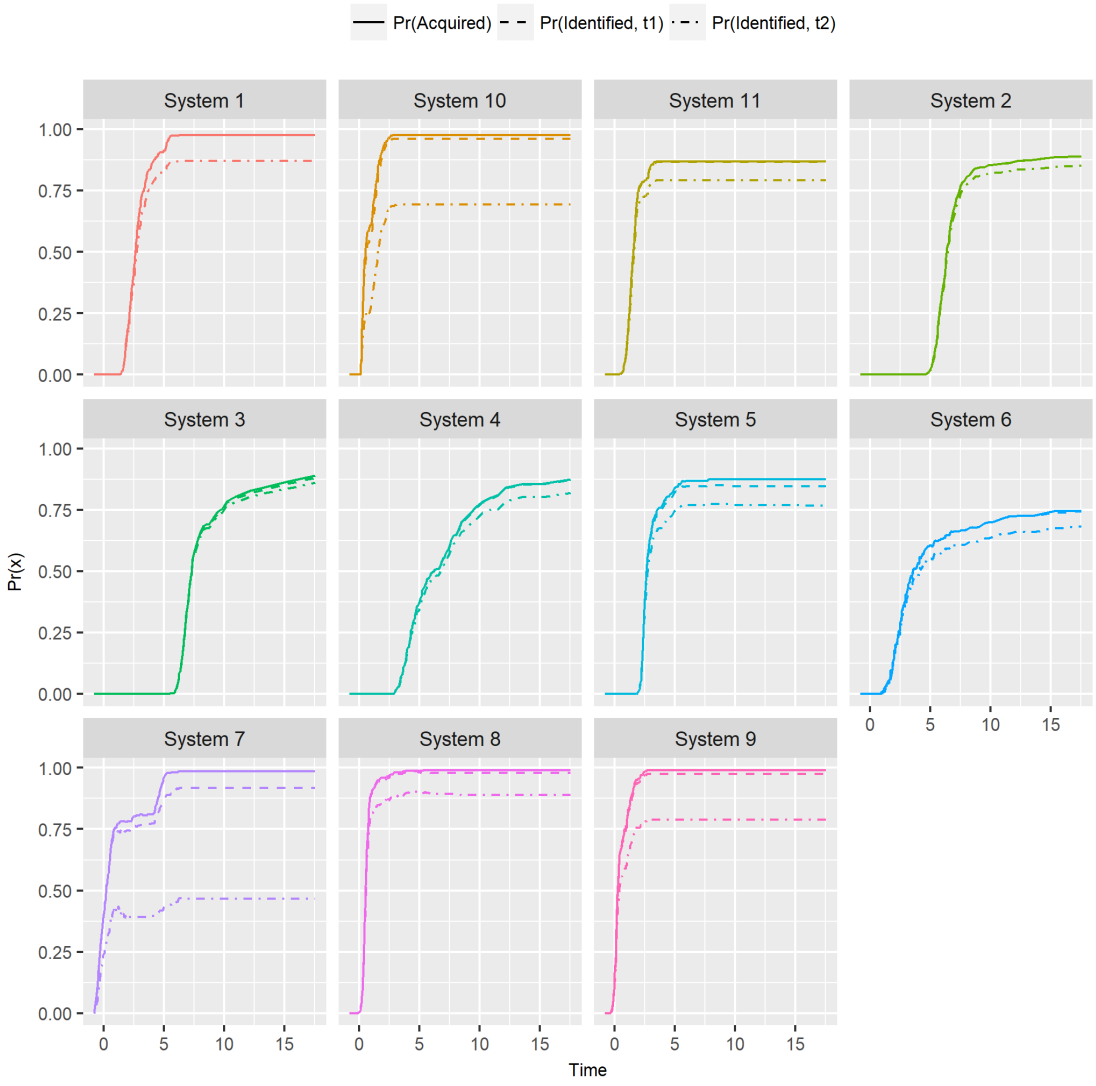

Figure 5. Time-Based Acquisition and Matching Performance across Systems

not simply the matching and acquisition performance of a given system but also the time with which these rates can be achieved. The time-based performance metrics introduced by this research provide one method for adapting the traditional biometric system testing protocols to the needs of high-throughput systems. They allow system evaluators to answer questions such as "What system has the best performance in under X seconds?" and "Could system Y perform better given more time per subject". These are key considerations when determining which high-throughput systems to field in operational deployments, how those systems should be configured, and what work-flows they can support. We believe that automated, unmanned, high-throughput biometric systems have significant potential in a variety of everyday use cases. However, traditional biometric system performance metrics may not be descriptive enough to allow for decisions on the optimal system for a given high-throughput environment. The time-based performance metrics discussed here are a first-step in adapting these traditional metrics for high-throughput applications.

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] J. J. Howard, A. A. 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. In *2018 Nineth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*. IEEE, 2018. 4, 5

[2] I. ISO. IEC 19795-1: Information technology–biometric performance testing and reporting-part 1: Principles and framework. *ISO/IEC, Editor*, 2006. 1, 3, 5

[3] I. ISO. IEC 19795-2: Information technology–biometric performance testing and reporting-part 2: Testing methodologies for technology and scenario evaluation. *ISO/IEC, Editor*, 2007. 1, 5

[4] I. ISO. IEC 19795-6: Information technology–biometric performance testing and reporting-part 6: Testing methodologies for operational evaluation. *ISO/IEC, Editor*, 2012. 1, 5

[5] E. Menegatti, N. Michael, K. Berns, and H. Yamaguchi. *Intelligent Autonomous Systems 13: Proceedings of the 13th International Conference IAS-13*, volume 302. Springer, 2015. 2

[6] O. Omidvar and P. van der Smagt. *Neural systems for robotics*. Academic Press, 1997. 2

[7] H. A. Yanco, A. Norton, W. Ober, D. Shane, A. Skinner, and J. Vice. Analysis of human-robot interaction at the darpa robotics challenge trials. *Journal of Field Robotics*, 32(3):420–444, 2015. 2