IBIA

International
Biometrics+Identity
Association

Understanding the Performance of Facial Recognition Algorithms

The International Biometrics + Identity Association (IBIA) is the leading voice for the biometrics and identity technology industry. It advances the transparent and secure use of these technologies to confirm human identity in our physical and digital worlds. #identitymatters

Understanding the Performance of Facial Recognition Algorithms

Executive Summary

This paper addresses the performance of facial recognition algorithms, an issue that has emerged as a major point of contention during the current policy debates about the use and limits of facial recognition.

The thrust of the argument to limit the use of facial recognition is that the technology is not yet ready for prime time. The primary arguments are that facial recognition algorithms are basically too imperfect because they are 'discriminatory' against people with dark skin tones and display low levels of matching performance.

The purposes of this paper are to:

- Demonstrate these performance arguments are not supported by the evidence documented in recent National Institute of Standards and Technology (NIST) testing, the world's premier standards and testing body. NIST shows stunningly high levels of accuracy and clear superiority of the technology compared to human recognition systems, both in terms of accuracy rates and performance across a range of skin tones. This is supported by the latest academic research conducted by a group of the preeminent scholars on facial recognition.
- Explain the factors that affect the performance differences of facial recognition algorithms, including the application, the rest of the system, variations in quality of the algorithms
- Summarize the many benefits of facial recognition
- Highlight the challenges in the use of facial recognition that remain and address the work in progress to further improve the technology

> *The field of research today known as Artificial Intelligence traces its origins to a workshop at Dartmouth College in 1956. Attendees became the founders and leaders of the field and were, with the benefit of hindsight, unrealistic about the likely course of progress. For example, Herbert Simon predicted, “machines will be capable, within twenty years, of doing any work a man can do.” Marvin Minsky agreed, writing "within a generation ... the problem of creating 'artificial intelligence' will substantially be solved." What AI research has delivered are highly specialized tools which approximate or improve upon human performance in narrow areas, yet exhibit no generalized behavior that humans would recognize as intelligence. Deep Learning is another such term that implies an ongoing process similar to that employed by humans; whereas what actually occurs is a highly sophisticated, one time, training on substantial amounts of carefully annotated data. Thereafter the system works well with information similar to the training data but does not adapt to new data until a subsequent training period.*

Let's Stop Using Semantically Loaded Terms like 'Discriminatory'

- Let's dispense with this term so we can focus on the essential facts about performance of facial recognition systems, including accuracy and systemic errors, instead of extraneous and emotional issues
- 'Discriminatory' is a semantically loaded term because it implies intent
- However, facial recognition is performed by a machine, and machines have no intent
- The argument that algorithm developers exhibit racial/ gender blindness producing algorithms that perform less effectively for other than white males is **not supported** by the facts
- NIST has active test and analysis effort to assess this issue
- Recent (12 April 2019) results for verification algorithms (i.e. 1: 1 search) show the top 20 performing algorithms, with elapsed time between images from 2 – 16 years, are most effective for blacks with black females often the most accurate
- The test results for identification (i.e. 1: N search) are expected during 2Q 2019
- The most appropriate composition of test datasets, to insure effective testing, is still somewhat of an unsettled issue

> *Cost of new dataset development for effective large-scale testing is a significant issue, beyond the resources of all but government and the largest companies. It may be feasible to continue to employ existing facial recognition datasets, by recharacterizing their metadata to more accurately reflect subject demographics, once there is consensus on what changes, if any, are needed.*

Performance Differences of Algorithms

- All algorithms have some performance differences across different demographic groups, genders, and age cohorts
- These differences are being addressed and there has been rapid improvement, which is ongoing
- For verification applications (fraud detection, access control, etc.), in the latest NIST testing, the top performing algorithms are more accurate with black males and females than with whites and have less than 1% false non-match rates for all groups at 0.1% false match rate
- For investigative applications, progress has been dramatic with a major update report expected from NIST during the 2nd quarter of 2019

Facial Recognition and Facial Classification are Different and Should Not Be Confused

- Face recognition seeks to identify an individual from their face image
- Facial classification seeks to classify a face by estimating, for example, gender, age, or race
- The algorithms are built and trained separately
- The process of classification estimation involves one image, while facial recognition involves comparison of pairs
- An MIT study, which is a large part of the "facial recognition is biased narrative", only examined facial classification, specifically for gender
- A joint FIT/Notre Dame study provides a more complete and accurate view, as do the NIST tests

Algorithms Are Just Part of a Facial Recognition System

- The performance of a facial recognition system depends on a number of factors; the algorithm is one such factor. The camera, its resolution, positioning, distance, and lighting set an upper limit on performance. Subject pose and expression can also influence performance
- Camera resolution and distance matter; humans require about 25 pixels per meter resolution to detect the presence of humans, but can recognize motion at lower resolutions
- Ambient or artificial lighting has an enormous impact on system performance
- In other words, all the components of the facial recognition system must perform properly, in addition to using a high-performance algorithm, and these elements can be adjusted easily
- Knowing all this, some facial recognition applications employ human facial examiners who make the final match/no match decision after the facial matching algorithm selects a list of potential matches; they use applications specifically designed for facial examinations

The Application Matters

- Facial verification and facial identification systems, until quite recently, have been designed to match portrait style (mugshot, driver license, visa, passport) images
- With good lighting, pose, and expression control, performance can be stunningly good and good mugshot accuracy conforms to photography standards adopted by NIST for the FBI further developed by ISO
- Matching of "in the wild" images (a reference to image quality -- candid, unposed, not portrait style images) has matured dramatically in the past 5 years, with verification accuracy of top algorithms now at 99%. An update on investigation applications is expected to show comparable progress and further maturation is expected in the near term

> *Some algorithms are much better than others, as in everything else. In golf, there is Tiger Woods and then there is the rest of us.*

Not all Algorithms are Alike

- Market entry is relatively easy and the number of algorithm providers has expanded from about 10 in 2010 to about 100 today, with many offering multiple algorithms
- Some algorithms are much better than others, as would be expected. Objective testing like that performed by NIST reveals the differences.
- Algorithm performance for a selfie, social media, or a commodity web camera is considerably different from an algorithm used for security or law enforcement applications

NIST Has Tested More Than 170 FR Algorithms, with Wide Variations in Performance Observed

- Six (6) algorithms are less accurate than a coin toss
- Most are more accurate than human observers, including those trained and employed to do recognition
- The top performing algorithms are much better performing than humans
- Many algorithms match blacks more accurately than whites
- Algorithm matching of females is frequently less accurate than males
- Algorithm performance is less accurate for most applications involving children
- The application makes a difference
- Portrait style 1: N and 1: 1 matching is extraordinarily accurate (considerably more accurate than fingerprint technology circa mid 2000's when FBI went to partial "lights out" fingerprint matching)

> *Nothing is perfect and no system performs perfectly. The real question is whether automated facial recognition is better than the current systems. And under this criterion, data clearly demonstrates superior performance of automated facial recognition.*

- In the wild ("candid, unposed, non-portrait images"), matching is less accurate but quite suitable for lead generation, typically with stalled investigations
- Likewise, matching is less accurate for poor quality images
- Notwithstanding exceptional algorithm accuracy, validation has not been performed to allow "lights out" use of facial recognition technology when there are potential adverse consequences to the subjects. Human review is required
- Algorithms are not commoditized as performance varies greatly, from the best identifying 99.4% of individuals in a gallery of 12 million subjects to below 40% for the worst

Demand for Perfection of Algorithms is Not a Performance Standard for the Real World

- No system – or human – performs perfectly
- The real question is whether automated facial recognition is better than other systems or humans. And under this criterion, data clearly demonstrate superior performance of automated facial recognition
- For family, friends, professional acquaintances, and celebrities, human recognition works well
- For unfamiliar persons, few individuals perform well at face recognition or matching
- Skilled passport examiners are only about 80% accurate when unaided by automation
- The top performing algorithms outperform the mean performance of all human groups including skilled forensic face examiners with unlimited time and the best automated tools; (although a few humans in the more skilled groups outperform circa 2017 top algorithms)
- Machines can memorize millions of faces, and humans only thousands, enabling machines to do things unaided that humans cannot, including to:
  - Identify missing children who do not know their names
  - Identify exploited children in dark web pornography
  - Identify disoriented adults (e.g. with amnesia, Alzheimer's)
  - Flag likely driver license application fraud for human review
  - Identify likely Visa fraud for human review
  - Identify likely Passport fraud for human review
  - Provide leads for further investigation when a surveillance photo is the only information
  - Detect border (and other) fraudulent use of stolen identity documents

People are Comfortable with Face Recognition

- Following the iPhone X introduction on November 3, 2017, tens of millions of Americans have become familiar and entirely satisfied with facial recognition technology for personal use
- The 2019 Center for Data Innovation public opinion survey found that only 1 in 4 Americans think the government should strictly limit the use of facial recognition technology
- The technology is widely used worldwide, and adoption is growing
- DHS pilot projects at several airports, dispensing with boarding passes and ID cards in favor of facial recognition for international flights, have been enthusiastically greeted by the traveling public
- Frequent international travelers already hope for domestic adoption
- Technology advancement is inexorable, and each generation has the responsibility to decide how to balance the benefits of new technology with privacy and appropriate uses

The IBIA is the leading voice for the biometrics and identity technology industry. It advances the transparent and secure use of these technologies to confirm human identity in our physical and digital worlds. Visit us at **www.ibia.org.**

ENDNOTES

[1] Grother, et al., 'Ongoing Face Recognition Vendor Test (FRVT)', www.nist.gov, 2019, p. 106 Figure 81, https://www.nist.gov/sites/default/files/documents/2019/04/15/frvt_report_2019_04_12.pdf .

[2] Loc. cit. and p. 39 Figure 16.

[3] Boulamwini and Gebru, 'Gender Shades: Intersectional Accuracy Disparities In Commercial Gender Classification', proceedings.mlr.press, 2018, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[4] Krishnapiya, et al., 'Characterizing the Variability in Face Recognition Accuracy Relative to Race', arxiv.org , 2019, https://arxiv.org/abs/1904.07325.

[5] Grother, et al., Op. cit.

[6] Axis Communications, 'Perfect Pixel Count', www.axis.com, 2014, p. 4 Table 3, https://www.axis.com/files/feature_articles/ar_perfect_pixel_count_55971_en_1402_lo.pdf.

[7] Grother, et al., Op. cit., p.45 Figure 22.

[8] Grother, et al., Op. cit.

[9] Phillips, et al., 'Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms', www.pnas.org, 2018, https://www.pnas.org/content/115/24/6171.

[10] Castro and McLaughlin, 'Survey: Few Americans Want Government to Limit Use of Facial Recognition Technology, Particularly for Public Safety or Airport Screening', www.datainnovation.org, 2019, https://www.datainnovation.org/2019/01/survey-few-americans-want-government-to-limit-use-of-facial-recognition-technology-particularly-for-public-safety-or-airport-screening/.

#identitymatters

IBIA

International
Biometrics+Identity
Association

1325 G Street, NW, Suite 500
Washington, DC 20005

202.888.0456 | IBIA.ORG