IBIA
International
Biometrics+Identity
Association

NIST Report on Facial Recognition: A Game Changer

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

Summary

The recent NIST report on the performance of facial recognition algorithms across different demographics is a game-changer.[1] It provides new and comprehensive data on the performance of algorithms across demographic groups. This data serves to debunk the semantically-loaded misleading arguments on facial recognition performance that privacy activists have pushed in their zeal to ban a technology that enhances public safety and security.

The testing showed wide variations in performance, ranging from algorithms that are less accurate than a coin toss, to high-performing algorithms that are overwhelmingly accurate with virtually undetectable demographic differences. These latter algorithms are 20 times more accurate than the most highly-skilled human groups.[2]

Unfortunately, media coverage drew generalizations from a few algorithms and focused on the low-performing algorithms in the study, not the range of algorithms tested, and certainly not the significant number of high-performing algorithms.

This paper summarizes the key NIST findings of high-performing algorithms (ignoring those algorithms that no prudent organization would use); how the data serves to debunk the fundamental misrepresentations of the activist arguments that the algorithms are biased and not good enough; and serves to reaffirm the benefits of facial recognition and the serious risks to public safety and national security of a blanket ban on the technology.

NIST key findings on demographic differentials

- NIST tested 189[3] algorithms from laboratories and vendors around the world (a large number because NIST allows anyone to submit algorithms for testing).
- The test results, as expected, show wide variations in algorithm performance with respect to demographic differentials, and NIST explicitly states that it is not accurate to draw generalizations about algorithm performance.[4] Some perform very well; others do not.
- The most accurate **high-performing identification algorithms** (a one-to-many search in which the technology uses an image to search a database of images to find potential matches) display virtually 'undetectable' differences among demographic groups; [5] more than 30 of the 189 identification algorithms NIST tested have false non-match rates (misses) less than three per thousand, [6] providing far greater accuracy than humans could ever achieve.
- The most accurate **high-performing verification algorithms** (a one-one verification search where 2 images are compared to determine similarities of the faces) display both low false positives and false negatives; more than 50 tested algorithms have false non-match rates (misses) less than three per thousand, [7] and false match rates (erroneous matches) less than one per hundred thousand, [8] again, greater accuracy than humans could ever achieve.
- The low performing algorithms show significant performance differences among demographic groups and are not usable by government or industry.

[1] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR 8280, (pp. 1–79). doi: 10.6028/nist.ir.8280 Re

[2] Private communication with James Loudermilk, Senior Director, National Security Solutions, an IDEMIA company and IBIA member organization, who did the analysis in an as yet unpublished paper.

[3] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR 8280, (p.1). doi: 10.6028/nist.ir.8280 Re

[4] Op. cit. (p.10)

[5] Op. cit. (pp. 3, 8)

[6] Op. cit. (pp. 64, 65)

[7] Op. cit. (pp. 54, 58)

[8] Op. cit. (pp. 56, 57)

> *Differences in algorithm performance most likely result from natural variations among people in facial bone structures, skin tones, and image capture. The NIST testing shows researchers have made significant progress reducing performance variation across the board, and ongoing efforts will continue this trend.*

Performance variations does not mean 'bias' has been introduced into facial recognition algorithms

- NIST uses the term 'demographic differences' (not 'bias') to describe performance variations, confirming that variation is technical and scientific.
- Differences in algorithm performance most likely result from natural variations among people in facial bone structures, skin tones, and image capture. The NIST testing shows researchers have made significant progress reducing performance variation across the board, and ongoing efforts will continue this trend. There is no reason to believe that computer vision technology is yet approaching performance boundary conditions.
- This is precisely what happened with fingerprint matching of Asian women.[9]
  - With smaller surface area, thinner skin, and more closely spaced and thinner ridge structure in their fingerprints, it was difficult to capture and match those fingerprints, a fact about which the researchers were unaware, a shortcoming in human knowledge.
  - When these natural variations became known, researchers fine-tuned the algorithms to address and resolve the issue, confirming the value of continuing research to improve algorithms and for ongoing NIST testing to spur improvement and to identify flaws.
- That developer 'bias' connotes unfounded prejudice is highly unlikely.
  - Machines do not have emotions and do what they are programmed to do.
  - Commercial entities in this space, especially the more successful ones, are international entities offering their products all over the world.
  - To be successful those products need to work well with every demographic.
  - Many leading algorithm developers in both academia and industry are themselves minorities, as are many in their management.

The argument that algorithms are not perfect (not 'good enough') cannot be taken seriously

- No machine system – or human – is perfect, and certainly humans are not likely to be transformed into perfect beings.
- In the real world, the pertinent question is whether automated facial recognition, which augments human decision-making, is better than the alternative of human recognition only.
- For many critical public safety activities, it is simply not acceptable to limit performance to human capability, or alternatively to not perform the activity at all.

[9] Mitra, S., & Gofman, M. (2017). Biometrics in a Data Driven World: Trends, Technologies, and Challenges. Boca Raton, FL: CRC Press.

> *The top performing algorithms outperform mean performance of all human groups including skilled forensic face examiners. Algorithm performance for the high performers, across the board, is more than 20 times better than skilled professional examiners.*

Automated facial recognition is indisputably more accurate than current human recognition only systems

- Here the evidence is indisputable – that the use of automated facial recognition, referred to human decision makers, is significantly more accurate than the use of human recognition alone.
  - Measured accuracy of human visual passport inspection is notoriously low, determined by some to be in the range of 80% or less (for example, Passport Officers' Errors in Face Matching).[10]
  - The top performing algorithms outperform mean performance of all human groups including skilled forensic face examiners.
  - Nearly every story that accepts the flawed portrayal of facial recognition cites as so-called "history" a single 2012 study by Brendan Klare et. al., Face Recognition Performance: Role of Demographic Information.[11] The study's principal author has recently made clear that paper is not a basis for the claims of bias others have made.[12]
  - Even if the claims were accurate seven (7) ago when it was written, the rapid pace of technological advancement underscores the paper's inherent weakness as a piece of documentary evidence. The latest 2019 Department of Commerce National Institute of Standards and Technology (NIST) reports and several academic studies demonstrate the obsolescence of Klare's paper, as he himself has said.[13]
  - Algorithm performance for the high performers, across the board, is more than 20 times better than skilled professional examiners.
- NIST's January 2020 FRVT Verification Report lists five algorithms, under suitable conditions with good photos, lighting etc, have an accuracy rate of 99.9% or better. Otherwise, the accuracy, for high performing algorithms is in the 98-99% range, and algorithm performance continues to rapidly improve. [14]

[10] White D, Kemp RI, Jenkins R, Matheson M, Burton AM (2014) Passport Officers' Errors in Face Matching. PLoS ONE 9(8): e103510. https://doi.org/10.1371/journal.pone.0103510

[11] Klare, B. F., Burge, M. J., Klontz, J. C., Bruegge, R. W. V., & Jain, A. K. (2012). Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security, 7(6), 1 (pp. 789–1801). doi: 10.1109/tifs.2012.2214212. http://openbiometrics.org/publications/klare2012demographics.pdf

[12] Klare, B. (2019, September 12). Race and Face Recognition Accuracy: Common Misconceptions. https://blog.rankone.io/2019/09/12/race-and-face-recognition-accuracy-common-misconceptions/

[13] Op. cit.

[14] "Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification," Grother P., Ngan M., and Hanoka K., 2020/01/22, Pp 26-29

> *A ban on facial recognition will preclude its use in forensic analysis, severely limiting the capability of law enforcement officials to solve crimes, identify missing and abused children, and apprehend human traffickers, to name just a few of the vital missions that are enhanced by the use of facial recognition.*

Automated facial recognition can do things that humans cannot do

- Machines can memorize millions of faces, humans only thousands; this enables machines to do things unaided that humans cannot, including:
  - Identifying missing children who do not know their names
  - Identify exploited children in dark web pornography
  - Identifying disoriented (amnesia, Alzheimer's, etc.) adults
  - Flagging likely driver license application fraud for human review
  - Flagging likely visa fraud for human review
  - Flagging likely passport fraud for human review
  - Providing leads for further investigation when a surveillance photo is the only information
  - Border (and other) fraudulent use of stolen identity documents

Banning facial recognition in law enforcement poses serious risks to public safety

- A ban on facial recognition will preclude its use in forensic analysis, severely limiting the capability of law enforcement officials to solve crimes, identify missing and abused children, and apprehend human traffickers, to name just a few of the vital missions that are enhanced by the use of facial recognition.
- Facial recognition is also critical in real time in cases of mass shootings, bombings, and other disasters. In the case of the Boston Marathon bombing, in 2013, facial recognition was not at its current level of sophistication. The FBI and other law enforcement agencies spent countless hours reviewing photos and videos before the two brothers were determined to be suspects and in-depth investigation could begin. Since the Boston Marathon bombing, the technology has improved by orders of magnitude and facial recognition now is a crucial element in counterterrorism and law enforcement around the country and the world.
- NIST results show that demographic differences of high-performing facial recognition algorithms are virtually unmeasurable[15] and these are the algorithms that government should be using.
  - Supporters of a ban or a moratorium on facial recognition in law enforcement argue it is 'biased' against dark-skinned people and will result in greater incarcerations.
  - The reality is that the NIST results show that demographic differences of high-performing facial recognition algorithms are virtually unmeasurable, and these are the algorithms that government should be using. This is one of the values that the Government (and industry) derives from NIST testing.

[15] Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. NISTIR 8280, (p. 2). doi: 10.6028/nist.ir.8280 Re

> *When law enforcement uses facial recognition, it does so to produce leads, where there may be none. The result is a 'gallery' of potential matches that provide potential leads that must be supplemented by additional evidence sufficient to rise above evidentiary standards.*

  - This argument also assumes that the current system of human recognition is accurate and unbiased. In fact, as described, human recognition alone is far less accurate than when augmented by automated facial recognition, and eyewitness testimony is notoriously biased.

- Banning facial recognition will only result in foregoing improvements in our flawed existing law enforcement system and, in some cases, it may be tantamount to deciding not to investigate crime.
- If security was predicated on 100% perfection, we would most certainly be discouraged from doing anything to enhance it.

Facial recognition use in law enforcement is only to generate leads

- Contrary to some commentary, facial recognition is not used for positive identification of an individual to establish guilt in a court of law.
- When law enforcement uses facial recognition, it does so to produce leads, where there may be none. The result is a 'gallery' of potential matches that provide potential leads that must be supplemented by additional evidence sufficient to rise above evidentiary standards.
- Most jurisdictions configure their systems to return either none, or multiple candidates, to reinforce to detectives that these are leads and not affirmative indicators of guilt.
- Many jurisdictions also use human examiners to review the gallery as an additional check on the system.

Conclusion

The recent facial recognition NIST testing on demographic differentials and overall accuracy show massive performance improvements. The facts are out: automated facial recognition is more accurate and less biased than human recognition alone, putting to rest the argument that gender and racial bias and poor performance justify a ban on the use of facial recognition. To the contrary, the evidence is clear that racial and gender bias and poor performance are not justifications to ban the technology.

The state of facial recognition, however, is still open for improvement on gender, age, and skin color. Further improvements will be impossible without continued use and refinement of the technology. We need to ensure the United States leads the research and development effort on this critical technology.

Bans and moratoriums by the United States Government will only hurt the United States. Other countries will continue their research and countries, like China, will gain an unfair advantage in the development of this critical technology.

As government, academia and industry go forward to improve facial recognition technology, it is important to understand that developers are not able to conduct such tests without access to volumes of diverse data only held by government. Increasingly, modern biometric algorithms employ machine learning. Existing machine learning methodologies are critically dependent on large volumes of training data representative of the operational environment for which they are intended. Only government has, or is likely to have, such data and, it must find a way to provide researchers with access to such data.

With facts and evidence, the NIST report informs the policy debate on facial recognition, making possible an open and transparent process with a careful balancing of benefits and appropriate uses, instead of a simplistic reaction to ban the technology.

#identitymatters

IBIA

International
Biometrics+Identity
Association

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

202.888.0456 | IBIA.ORG