ITIF | INFORMATION TECHNOLOGY & INNOVATION FOUNDATION

The Critics Were Wrong: NIST Data Shows the Best Facial Recognition Algorithms Are Neither Racist Nor Sexist

BY: MICHAEL MCLAUGHLIN AND DANIEL CASTRO | JANUARY 2020

A close look at data from a new NIST report reveals that the best facial recognition algorithms in the world are highly accurate and have vanishingly small differences in their rates of false-positive or false-negative readings across demographic groups.

INTRODUCTION

The National Institute of Standards and Technology (NIST) recently released a report that examined the accuracy of facial recognition algorithms across different demographic groups. The NIST report found that the most accurate algorithms were highly accurate across all demographic groups. But NIST tested nearly 200 algorithms from vendors and labs around the world—it allows anyone to submit an algorithm for testing—and since many of the algorithms it tested displayed some bias, several news outlets and activists have misleadingly concluded that facial recognition systems are racist and sexist.[1] But a close look at the data reveals a different picture.

Facial recognition technology compares images of faces to determine their similarity, which the technology represents using a similarity score. The technology often performs one of two types of comparisons. The first comparison is known as a one-to-many or identification search, in which the technology uses a probe image to search a database of images to find potential matches. The second comparison is known as a one-to-one or verification search as the technology compares two images to determine the similarity of the faces in them. In many cases, the faces in images are considered a match if their similarity score meets or exceeds the match threshold, a number the operator assigns that represents a minimum acceptable similarity score. The technology has many commercial and non-commercial uses, and will likely be integrated into more products and services in the future to enhance security, improve convenience, and increase efficiency, such as by helping find victims of human trafficking, expediting passengers through airport security, and flagging individuals using forged identification.[2]

NIST assessed the false positive and false-negative rates of algorithms using four types of images, including mugshots, application photographs from individuals applying for immigration benefits, visa photographs, and images taken of travelers entering the United States. NIST's report reveals that:

- The most accurate *identification* algorithms have “undetectable” differences between demographic groups[3]
- The most accurate *verification* algorithms have low false positives and false negatives across most demographic groups[4]
- Algorithms can have different error rates for different demographics but still be highly accurate

KEY FINDINGS

As detailed below, NIST found that the most accurate algorithms—which should be the only algorithms used in government systems—did not display a significant demographic bias. For example, 17 of the highest performing verification algorithms had similar levels of accuracy for black females and white males: false-negative rates of 0.49 percent or less for black females (equivalent to an error rate of less than 1 in 200) and 0.85 percent or less for white males (equivalent to an error rate of less than 1.7 in 200).[5]

While the most accurate algorithms did not display a significant demographic bias, it is also true that the majority of the algorithms NIST tested generally performed better on men and individuals with lighter skin tones. However, it is important to recognize that there is a stark difference between the best and worst algorithms. In comparison to the false-negative rates under 1 percent for black females and white males among the highest-performing algorithms, the lowest-performing algorithms had false-negatives rates, for blacks and whites, as high as 99 percent.[6] This wide range of accuracy is not surprising considering that NIST allows anyone to submit an algorithm for testing, ranging from large companies with production systems to small research groups whose algorithms have not left the lab—algorithms are tested even if they are not incorporated into a commercially available product.

The Most Accurate Identification Algorithms Have Undetectable Differences Between Demographics

NIST found that some highly accurate algorithms had false-positive demographic differentials that were so small as to be “undetectable” for one-to-many searches.[7] Moreover, for most algorithms, black men had lower false-negative rates than white men, and several of the top algorithms had better false-negative rates for white women than white men.8 Several algorithms also provided uniform similarity scores across demographic groups, meaning that the algorithms provided similar match and non-match scores regardless of race and gender.[9] The uniform scores indicate that these algorithms would have small demographic differentials if an operator applied a threshold. But different thresholds can affect demographic differentials. For example, at least six of the most accurate identification algorithms had higher false-positive rates for black men than white men at one threshold, but lower false-positive rates for black men than white men at another threshold.[10]

The Most Accurate Verification Algorithms Have Low False Positives and Negatives Across Most Demographics

The most accurate verification algorithms have low false positives and negatives across most demographics. For example, when NIST applied thresholds so that the algorithms had false-positive rates of 0.01 percent for white males, more than half of the 17 most accurate algorithms had false-positive rates of 0.03 percent or better for black males, Asian men, and white women.[11] This equates to the algorithms falsely matching these individuals three times or less per every 10,000 comparisons to an imposter compared to one per every 10,000 for white males. At another threshold, seven of the top algorithms displayed no false-positive bias between white men, black men, Asian men, and white females.[12] At this threshold, several algorithms also had false-positive rates of 0.003 percent or less for black women or Asian women while white males had false-positive rates of 0.001 percent.[13]

False negatives were also low for the most accurate verification algorithms. Five of the 17 most accurate algorithms had false-negative rates of less than one percent across all demographic groups when NIST applied a threshold that set false-positive rates at 0.01 percent.[14] Similarly, the best verification algorithms had less than one percent false-negative rates across countries and demographic groups. For example, the algorithm Visionlabs-007 had below a 1 percent false-negative rate for nearly all countries and demographic groups for border crossing application images. There were two exceptions—Somalian and Liberian women under 45. Nonetheless, the algorithm had a false-negative rate below 1.4 percent for each of these groups.

Algorithms Can Have Different Error Rates for Different Demographics But Still Be Highly Accurate

Some algorithms perform differently on one group compared to another, but still maintain true positive and true negative accuracy rates greater than 99 percent for all races and sexes.[15] Because these algorithms have very low error rates, differences that are small in absolute terms may seem large if expressed in relative terms. For example, an algorithm from Dutch firm VisionLabs, Visionlabs-007, had a false-negative rate four times higher for the nationality it performed poorest on (Somalian) than the nationality it performed best on (Salvadoran).[16] Nonetheless, the algorithm only had a false-negative rate of 0.63 percent for individuals from Somalia. Another example is the performance difference of a verification algorithm from Camvi, a firm based in Silicon Valley, for white males and American Indian females. At one particular threshold, the algorithm had a false-positive rate that was 13 times higher for American Indian females than white men.[17] But at this threshold, the algorithm had barely more than one false match of American Indian females for every 10,000 imposter comparisons to other American Indian females. It is also true that most verification algorithms had higher false-negative rates for women than men. But NIST notes that this "is a marginal effect—perhaps 98 percent of women are still correctly verified—so the effect is confined to fewer than 2 percent of comparisons where algorithms fail to verify."

PUTTING NIST'S DATA IN CONTEXT

Recent reporting on how law enforcement in San Diego used facial recognition from 2012-2019 can also help put NIST's data in context. In 2018, various law enforcement entities made 25,102 queries to a database of 1.8 million mugshot images.[18] Law enforcement officials uses of the technology included attempts to determine whether an individual had a criminal record and attempts to discover the identity of individuals who lacked identification. These use cases were likely one-to-many searches. Law enforcement did not track the success of the program, making it unclear how many false positives or false negatives the system registered as well as how many mated or non-mated searches—a search in which an image of the individual was not in San Diego's database—they performed.

But we can consider a few scenarios to make a rough estimate of how the most accurate algorithms might perform in a city like San Diego, assuming San Diego's images and hardware were of similar quality to NIST's.[19] Under the first scenario, let us assume that all 25,102 probe images law enforcement used had a match in the database of 1.8 million mugshot images (an unlikely event), and that law enforcement did not apply a threshold to limit false positives or negatives (also unlikely). NEC-2, the best identification algorithm NIST tested in an earlier 2019 report, failed to rank the correct candidate as the most likely match only 0.12 percent of the time when performing a search of a database containing images of 3 million individuals.[20] At this rate, the technology would have succeeded in listing the correct individual in the San Diego search as the most likely match 24,970 times out of the 25,000 searches and failed 30 times.

Under a second scenario, let us assume law enforcement applied a threshold that allowed for one false positive every 1,000 non-mate searches. At this rate, NEC-3 had a false-negative rate of 0.26 percent. We also assume that half of the more than 25,000 probe images had a match in the database and that half did not have a match. In this scenario, the algorithm would have registered 13 false positives and 33 false negatives.

CONCLUSION

Developers and users of facial recognition technology, law enforcement, and lawmakers can take several actions to promote the development and responsible use of facial recognition technology. First, developers should continue to improve accuracy rates across different demographics, including by diversifying their datasets.[21] Second, the government should set standards for the accuracy rates of the systems it deploys. Third, law enforcement should have standards for the quality of images it uses in a facial recognition search, which can affect the accuracy of facial recognition algorithms.[22] Fourth, the users of facial recognition technology should carefully choose which match threshold is appropriate for their goal. Lastly, lawmakers should consider how law enforcement typically uses the technology and the different implications of false positive and false negatives when developing regulations. In most law enforcement scenarios, law enforcement is using facial recognition technology to return a list of possible suspects that humans review. And there are different implications when algorithms incur false positives or false negatives. In many cases, a subject can make a second attempt at recognition when a facial recognition system produces a false negative. This implication differs from the possible effects of false positives, which could allow an individual access to a facility they should not enter.

Finally, while there is no place for racial, gender or other types of discrimination in societies, to ban facial recognition unless it performs exactly the same across every conceivable group is impractical and would limit the use of a societally valuable technology. Many critics of facial recognition technology complain that the technology is not accurate enough, but refuse to give specifics on what they would consider sufficient—refusing to set a clear goal post for industry—which suggests they are not serious about wanting to improve the technology and oppose it for other reasons.

Reasonable people may disagree on when it is appropriate to use facial recognition, but the facts are clear that the technology can be highly accurate. As previous NIST reports have shown, many of the algorithms have accuracy rates that exceed 99 percent, and as the new report shows, the differences across demographics are minimal for the best algorithms.[23]

About the Authors

Michael McLaughlin is a research analyst at the Information Technology and Innovation Foundation. He researches and writes about a variety of issues related to information technology and Internet policy, including digital platforms, e-government, and artificial intelligence. Michael graduated from Wake Forest University, where he majored in Communication with Minors in Politics and International Affairs and Journalism. He received his Master's in Communication at Stanford University, specializing in Data Journalism.

Daniel Castro is vice president of ITIF and director of ITIF's Center for Data Innovation. His research interests include health IT, data privacy, e-commerce, e-government, electronic voting, information security, and accessibility. Before joining ITIF, Castro worked as an IT analyst at the Government Accountability Office, where he audited IT security and management controls at various government agencies. He has a B.S. in foreign service from Georgetown University and an M.S. in information security technology and management from Carnegie Mellon University.

About ITIF

The Information Technology and Innovation Foundation (ITIF) is a nonprofit, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy. Recognized as the world's leading science and technology think tank, ITIF's mission is to formulate and promote policy solutions that accelerate innovation and boost productivity to spur growth, opportunity, and progress.

For more information, visit us at www.itif.org.

ENDNOTES

1. Tom Higgins "'Racist and Sexist' Facial Recognition Cameras Could Lead to False Arrests," *The Telegraph*, December 20, 2019, https://www.telegraph.co.uk/technology/2019/12/20/racist-sexist-facial-recognition-cameras-could-lead-false-arrests.

2. Tom Simonite, "How Facial Recognition Is Fighting Child Sex Trafficking," *Wired*, June 19, 2019, https://www.wired.com/story/how-facial-recognition-fighting-child-sex-trafficking/; "Face Recognition Nabs Fake Passport User at US Airport," *VOA News*, August 24, 2018, https://www.voanews.com/silicon-valley-technology/face-recognition-nabs-fake-passport-user-us-airport.

3. We defined the most accurate identification algorithms as the twenty algorithms that had the lowest false-negative identification rates for placing the correct individual at rank one when searching a database that had images of 12 million individuals in NIST's September 2019 identification report. NIST provided error characteristics data by race and sex for ten of these algorithms in its recent report. Consequently, we analyzed the performance of NEC-2, NEC-3, VisionLabs-7, Microsoft-5, Yitu-5, Microsoft-0, Cogent-3, ISystems-3, NeuroTechnology-5, and NTechLab-6; Patrick Grother, Mei Ngan, and Kayee Hanaoka, Face Recognition Vendor Test (FRVT) Part 2: Identification (Washington, DC: National Institute of Standards and Technology, September 2019), 47, https://www.nist.gov/system/files/documents/2019/09/11/nistir_8271_20190911.pdf#page=49.

4. We defined the most accurate verification algorithms as those that rank in the top 20 on NIST's FRVT 1:1 leaderboard on January 6, 2020. NIST has since updated the leaderboard. Not all of these algorithms were tested in NIST's most recent demographics report. We analyzed the performance of algorithms that NIST provided data for in Annexes 6, 13, 15, and Figure 22. These algorithms are visionlabs-007, everai-paravision-003, didiglobalface-001, imperial-002, dahua-003, tevian-005, alphaface-001, ntechlab-007, yitu-003, innovatrics-006, facesoft-000, intellifusion-001, anke-004, hik-001, camvi-004, vocord-007, and tech5-003; National Institute of Standards and Technology, FRVT 1:1 Verification (FRVT 1:1 leaderboard, accessed January 6, 2020), https://pages.nist.gov/frvt/html/frvt11.html.

5. The high-performing algorithms include visionlabs-007, everai-paravision-003, didiglobalface-001, imperial-002, dahua-003, tevian-005, alphaface-001, ntechlab-007, yitu-003, innovatrics-006, facesoft-000, intellifusion-001, anke-004, hik-001, camvi-004, vocord-007, and tech5-003.. Comparisons made at the same match threshold.

6. The low performing algorithms, according to their performance for false non-match rate on Figure 22 of the NIST demographics report, include shaman_001, isap_001, ayonix_000, amplifiedgroup_001, saffe_001, videonetics_001, and chtface_001.

7. Patrick Grother, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (Washington, DC: National Institute of Standards and Technology, December 2019), 3, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf#page=6.

8. To compare white men and white women, we analyzed false-negative rates for ranking the correct matching image as the top potential match. Algorithms that had lower false-negative rates for white women than white men include NEC-2, NEC-3, and VisionLabs-7; Patrick Grother, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (Washington, DC: National Institute of Standards and Technology, December 2019), 63https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf#page=66; National Institute of Standards and Technology, Ongoing Face Recognition Vendor Test (FRVT) (part 3: demographic effects, annex 16: identification error characteristics by race and sex), https://pages.nist.gov/frvt/reports/demographics/annexes/annex_16.pdf.

9. Patrick Grother, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (Washington, DC: National Institute of Standards and Technology, December 2019), 66, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf#page=69.

10. These algorithms include VisionLabs-7, Microsoft-5, Yitu-5, Microsoft-0, ISystems-3, and NeuroTechnology-5; National Institute of Standards and Technology, Ongoing Face Recognition Vendor Test (FRVT) (part 3: demographic effects, annex 16: identification error characteristics by race and sex), https://pages.nist.gov/frvt/reports/demographics/annexes/annex_16.pdf.

11. One of these algorithms, for example, is visionlabs-007; National Institute of Standards and Technology, Ongoing Face Recognition Vendor Test (FRVT) (part 3: demographic effects, annex 6: cross-race and sex false match rates in United States mugshot images), https://pages.nist.gov/frvt/reports/demographics/annexes/annex_06.pdf.

12. An example of such an algorithm is anke-004.

13. An example of such an algorithm is yitu-003.

14. These algorithms are visionlabs-007, everai-paravision-003, didiglobalface-001, alphaface 001, and intellifusion-001; National Institute of Standards and Technology, Ongoing Face Recognition Vendor Test (FRVT) (part 3: demographic effects, annex 15: genuine and imposter score distributions for United States mugshots), https://pages.nist.gov/frvt/reports/demographics/annexes/annex_15.pdf.

15. These algorithms include visionlabs-007 and everai-paravision-003.

16. In this case, NIST set the threshold to “the lowest value that gives FMR ≤0.00001.”

17. National Institute of Standards and Technology, Ongoing Face Recognition Vendor Test (FRVT) (part 3: demographic effects, annex 15: genuine and imposter score distributions for United States mugshots, 19), https://pages.nist.gov/frvt/reports/demographics/annexes/annex_15.pdf#20.

18. DJ Pangburn, "San Diego's Massive, 7-Year Experiment With Facial Recognition Technology Appears to Be a Flop," *Fast Company*, January 9, 2020, https://www.fastcompany.com/90440198/san-diegos-massive-7-year-experiment-with-facial-recognition-technology-appears-to-be-a-flop.

19. In each of the scenarios, we are assuming that the racial and gender makeup of San Diego's mugshot database is similar to NIST's mugshot database.

20. Patrick Grother, Mei Ngan, and Kayee Hanaoka, Face Recognition Vendor Test (FRVT) Part 2: Identification (Washington, DC: National Institute of Standards and Technology, September 2019), 47, https://www.nist.gov/system/files/documents/2019/09/11/nistir_8271_20190911.pdf#page=49.

21. Chinese developers often had lower false positives for Chinese faces, suggesting that increasing the representation of minority faces in training data may reduce bias.

22. For example, although the false-negative rates were frequently lowest for black individuals in mugshot images, false-negative rates were relatively high for black individuals in images taken at border crossings. This difference could result from inadequate exposure in the latter photos; Patrick Grother, Mei Ngan, and Kayee Hanaoka, *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects* (Washington, DC: National Institute of Standards and Technology, December 2019), 54, https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf#page=57.

23. For example, a previous NIST report revealed that the top 20 identification algorithms failed to place the correct individual as the top potential match when searching a database containing images of 12 million individuals less than one percent of the time; Patrick Grother, Mei Ngan, and Kayee Hanaoka, Face Recognition Vendor Test (FRVT) Part 2: Identification (Washington, DC: National Institute of Standards and Technology, September 2019), 47, https://www.nist.gov/system/files/documents/2019/09/11/nistir_8271_20190911.pdf#page=49.