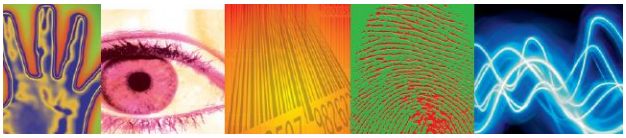


Top 10 takeaways: NIST demographic effects report

The Biometrics Institute has put together a list of top 10 takeaways from the very detailed National Institute of Standards and Technology (NIST) *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects Report*, published 19 Dec 2019. The first page comprises our key takeaways while the subsequent pages give more detail on each of these points and references them to the original NIST report.

1. **Terminology is important.** Neither face estimation nor face classification is the same as face recognition.

2. **Accuracy of *face recognition* algorithms must not be confused with results from *face* classification studies.** NIST draws attention to two studies where poor accuracy results from *face classification* algorithms – classifying gender and skin type – have been widely cited in discussions of bias in *face recognition* algorithms.

3. **Not all face recognition algorithms perform the same way**. The best face recognition algorithms give false non-match rates that are "absolutely low".

4. **Women produce higher false non-match rates.** However, this is a "marginal effect" - 98% of women are still correctly verified. The effect is confined to fewer than 2% of comparisons where algorithms fail to verify.

5. **Contemporary face recognition algorithms exhibit demographic differentials of various magnitudes.** False positive differentials are much larger than false negative differentials and exist broadly, across many, but not all, algorithms tested. Variances are significantly smaller or negligible in the most accurate algorithms.

6. **False positive rates are highest in West and East African and East Asian people and lowest in Eastern Europeans.** Using higher quality application photographs from a global population of applicants for immigration benefits, false positives are also higher in women than in men and elevated in the elderly and in children.

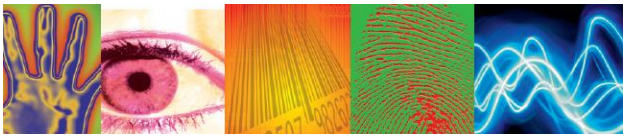
7. **Differing false negative rates exist because of varying degrees of image quality and lighting.** With US domestic mugshots collected using a photographic setup specifically to produce high-quality images, false negatives are higher in Asian and American Indians. But using lower-quality US border crossing images, false negatives are higher in people born in Africa or the Caribbean.

8. **Demographic differentials in *one-to-one verification* algorithms are usually, but not always, present in *one-to-many search* algorithms**. In some highly accurate identification algorithms, false positive differentials are undetectable.

9. **Often algorithms behave how humans intuitively behave.** For example, the probability of getting a false match across people of the same sex, age and ethnicity is higher than it is across a uniform population.

10. **Know your algorithm!** There are about 200 out there, some perform very well, others do not. It is not accurate to draw generalisations about algorithm performance overall, especially when evaluating in a test environment. Therefore, system owners should consider measuring operational algorithm accuracy, perhaps using a biometrics testing laboratory.

Email: manager@biometricsinstitute.org | Web: www.biometricsinstitute.org



The Top 10 takeaways in more detail

1. **Terminology is important.** Neither face estimation nor face classification is the same as face recognition.

   Face *recognition* compares images to determine if they are of the same person. Face *estimation* labels faces in, for example, age ranges, gender or ethnic group. Face *classification* determines a characteristic of a face like whether a person is happy or sad, wears glasses or no glasses. The accuracy of one bears no correlation to the others.

   NIST report page 4*:*

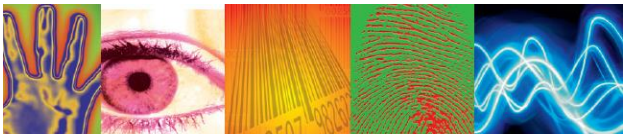
   > "We use the term **face analysis** as an umbrella for any algorithm that consumes face images and produces some output. Within that are **estimation** algorithms that output some continuous quantity (e.g., age or degree of fatigue). There are **classification** algorithms that aim to determine some categorical quantity such as the sex of a person or their emotional state. Face classification algorithms are built with inherent knowledge of the classes they aim to produce (e.g.,happy, sad). Face **recognition** algorithms, however, have no built in notion of a particular person."

2. **Accuracy of *face recognition* algorithms must not be confused with results from *face classification* studies.** NIST draws attention to two studies where poor accuracy results from *face classification* algorithms – classifying gender and skin type – have been widely cited in discussions of bias in *face recognition* algorithms.

   Most of the media discussion around facial recognition bias cites two studies* that have shown poor accuracy of face gender classification algorithms on black women. Those studies didn't look at popular commercially available facial recognition algorithms, but they did say that the results on the algorithms they used showed that black women were male 35% of the time. That is being widely cited to negatively criticise face recognition accuracy per se. But this is face classification – it is really asking "does the software classify that this image is male or female, in terms of likely gender?" – it is not asking the face recognition question of "is this person John Doe?"

   *References:
   - Joy Buolamwini. *Gender shades: Intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. Technical report, MIT Media Lab, 01 2017.
   - Inioluwa Raji and Joy Buolamwini. *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products*. In Conference on AI, Ethics and Society, pages 429– 435, 01 2019.

3. **Not all face recognition algorithms perform the same way**. The best face recognition algorithms today give false non-match rates (FNMR) that are "absolutely low".



NIST report page 54*:*

> "FNMR is absolutely low: In one-to-one verification of mugshots, **the best algorithms give FNMR below 0.5% at the reasonably stringent FMR criterion of 0.00001. FNMR is generally below 1%** with exceptions discussed below. For the more difficult application-border crossing comparisons, the best algorithm almost always gives FNMR below 1%. **These error rates are far better than the gender-classification error rates that spawned widespread coverage of bias in face recognition."**

In layman's terms, this means that for an FMR of "one in a hundred thousand", the best algorithms will find the true correct match at least 99.5% of the time.

4. **Women produce higher false non-match rates (FNMR).** However, this is a "marginal effect" - 98% of women are still correctly verified. The effect is confined to fewer than 2% of comparisons where algorithms fail to verify.

   NIST report page 56:

   > "Women give higher FNMR: In most cases, algorithms give higher false non-match rates in women than men. **Note that this is a marginal effect - perhaps 98% of women are still correctly verified - so the effect is confined to fewer than 2% of comparisons where algorithms fail to verify. It is possible that the error differences are due to relative prevalence some unknown covariate.** There are some exceptions, however: In Kenya, Nigeria, Jamaica men give higher FNMR. This applies in Haiti and Ghana also but only for people aged 45 or over."

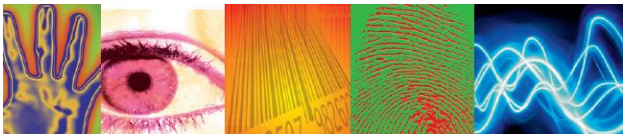
5. **Contemporary face recognition algorithms exhibit demographic differentials of various magnitudes.** False positive differentials are much larger than false negative differentials and exist broadly, across many, but not all, algorithms tested. Variances are significantly smaller or negligible in the most accurate algorithms.

   NIST report page 2*:*

   > "Contemporary face recognition algorithms exhibit demographic differentials of various magnitudes. Our main result is that false positive differentials are much larger than those related to false negatives and exist broadly, across many, but not all, algorithms tested. Across demographics, false positives rates often vary by factors of 10 to beyond 100 times. False negatives tend to be more algorithm-specific, and vary often by factors below 3."

   NIST report page 6:

   > "The accuracy of algorithms used in this report has been documented in recent FRVT evaluation reports [16, 17]. These show a wide range in accuracy across algorithm developers, with the most accurate algorithms producing many fewer errors than lower-performing variants. More accurate algorithms produce fewer errors, and will be expected therefore to have smaller demographic differentials."



6. **False positive rates are highest in West and East African and East Asian people and lowest in Eastern Europeans.** Using higher quality application photographs from a global population of applicants for immigration benefits, false positives are also higher in women than in men and elevated in the elderly and in children.

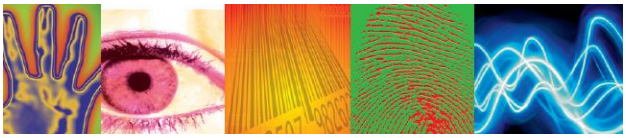
   NIST report page 2:

   > "False positives: Using the higher quality Application photos, false positive rates are highest in West and East African and East Asian people, and lowest in Eastern European individuals. This effect is generally large, with a factor of 100 more false positives between countries. However, with a number of algorithms developed in China this effect is reversed, with low false positive rates on East Asian faces. With domestic law enforcement images, the highest false positives are in American Indians, with elevated rates in African American and Asian populations; the relative ordering depends on sex and varies with algorithm.
   >
   > We found false positives to be higher in women than men, and this is consistent across algorithms and datasets. This effect is smaller than that due to race. We found elevated false positives in the elderly and in children; the effects were larger in the oldest and youngest, and smallest in middle-aged adults."

7. **Differing false negative rates exist because of varying degrees of image quality and lighting.** With US domestic mugshots collected using a photographic setup specifically to produce high-quality images, false negatives are higher in Asian and American Indians. But using lower-quality US border crossing images, false negatives are higher in people born in Africa or the Caribbean.

   NIST report page 3:

   > "False negatives: With domestic mugshots, false negatives are higher in Asian and American Indian individuals, with error rates above those in white and African American faces (which yield the lowest false negative rates). However, with lower-quality border crossing images, false negatives are generally higher in people born in Africa and the Caribbean, the effect being stronger in older individuals. **These differing results relate to image quality: The mugshots were collected with a photographic setup specifically standardized to produce high-quality images across races; the border crossing images deviate from face image quality standards.**
   >
   > **In cooperative access control applications, false negatives can be remedied by users making second attempts."**



NIST report page 7:

> "When comparing high-quality application photos, error rates are very low and measurement of false negative differentials across demographics is difficult. This implies that better image quality reduces false negative rates and differentials."

NIST report page 54:

> "FNMR in African and African American subjects: In domestic mugshots, the lowest FNMR in images of subjects whose race is listed as black. However, when comparing high-quality application photos with border-crossing images, FNMR is often highest in African born subjects. We don't formally measure contrast or brightness in order to determine why this occurs, but inspection of the border quality images shows under exposure of dark skinned individuals often due to bright background lighting in the border crossing environment. In mugshots this does not occur. In neither case is the camera at fault."

8. Demographic differentials in *one-to-one verification* algorithms are usually, but not always, present in *one-to-many search* algorithms. In some highly accurate identification algorithms, false positive differentials are undetectable.

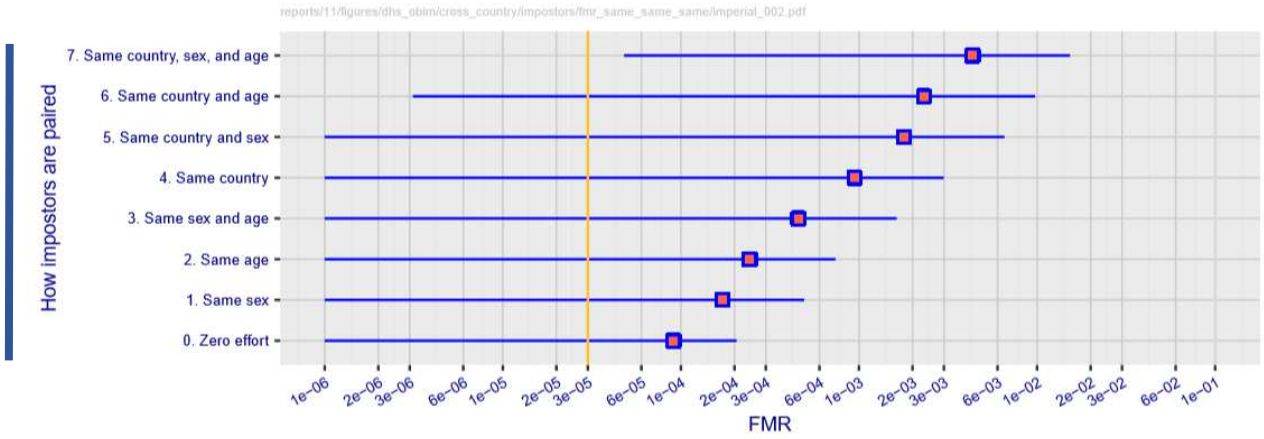
   NIST report page 8:

   > "Identification Algorithms: The presence of an enrollment database affords one-to-many algorithms a resource for mitigation of demographic effects that purely one-to-one verification systems do not have. We note that demographic differentials present in one-to-one verification algorithms are usually, but not always, present in one-to-many search algorithms. See Section 7. One important exception is that some developers supplied identification algorithms for which false positive differentials are undetectable."

9. Often algorithms behave how humans intuitively behave. For example, the probability of getting a false match across people of the same sex, age and ethnicity is higher than it is across a uniform population.

   The graph below shows how an imposter who comes from the same country, who is of the same sex and about the same age has a better chance of looking like the person they are pretending to be, than somebody who has a different sex or comes from a different country, or is of a different age – which is intuitively what you would expect.

NIST report page 30:



The red point in the plot shows the mean of false match rates over particular sets of demographic groups.

10. **Know your algorithm!** There are about 200 out there, some perform very well, others do not. It is not accurate to draw generalisations about algorithm performance overall, especially when evaluating in a test environment. Therefore, system owners should consider measuring operational algorithm accuracy, perhaps using a biometrics testing laboratory.

    NIST report page 3:

    > "Operational implementations usually employ a single face recognition algorithm. Given algorithm-specific variation, **it is incumbent upon the system owner to know their algorithm.** While publicly available test data from NIST and elsewhere can inform owners, it will usually be informative to specifically measure accuracy of the operational algorithm on the operational image data, perhaps employing a biometrics testing laboratory to assist. **Since different algorithms perform better or worse in processing images of individuals in various demographics, policy makers, face recognition system developers, and end users should be aware of these differences and use them to make decisions and to improve future performance."**

Definition references

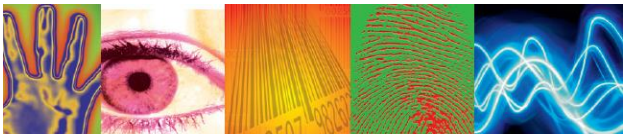
- NIST defines face analysis, classification, estimation and recognition on page 4 of their report
- A definition of common biometric vocabulary is available in ISO/IEC 2382 Part 37

Contact

This document has been compiled with the support of the Biometrics Institute Future Direction Group. To find out more, contact manager@biometricsinstitute.org.

Disclaimer

The Biometrics Institute provides guiding material as a tool to help its members conduct due diligence. While the Institute has used reasonable care to ensure the accuracy of the material, due to the content and variable inputs during and after the process of implementing biometrics, the institute cannot



be held accountable for outcomes or compliance. The material has been prepared for informational purposes only and is not intended to provide legal or compliance advice. Organisations should consult industry experts should they require advice on the technical, legal or compliance aspects of the material.

Last updated: May 2020

Email: manager@biometricsinstitute.org | Web: www.biometricsinstitute.org