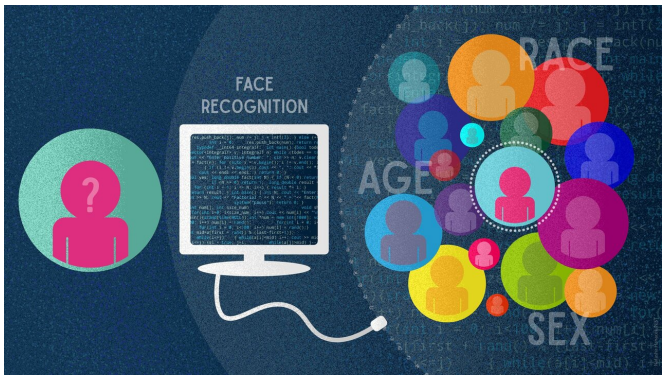
Study evaluates effects of race, age, sex on face recognition software

28 January 2020, by Chad Boutin



A new NIST study examines how accurately face recognition software tools identify people of varied sex, age and racial background. Credit: N. Hanacek/NIST

How accurately do face recognition software tools identify people of varied sex, age and racial background? According to a new study by the National Institute of Standards and Technology (NIST), the answer depends on the algorithm at the heart of the system, the application that uses it and the data it's fed—but the majority of face recognition algorithms exhibit demographic differentials. A differential means that an algorithm's ability to match two images of the same person varies from one demographic group to another.

Results captured in the report, Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects (NISTIR 8280), are intended to inform policymakers and to help software developers better understand the performance of their algorithms. Face recognition technology has inspired public debate in part because of the need to understand the effect of demographics on face recognition algorithms.

"While it is usually incorrect to make statements across algorithms, we found empirical evidence for the existence of demographic differentials in the majority of the face recognition algorithms we studied," said Patrick Grother, a NIST computer scientist and the report's primary author. "While we do not explore what might cause these differentials, this data will be valuable to policymakers, developers and end users in thinking about the limitations and appropriate use of these algorithms."

The study was conducted through NIST's Face Recognition Vendor Test (FRVT) program, which evaluates face recognition algorithms submitted by industry and academic developers on their ability to perform different tasks. While NIST does not test the finalized commercial products that make use of these algorithms, the program has revealed rapid developments in the burgeoning field.

The NIST study evaluated 189 software algorithms from 99 developers—a majority of the industry. It focuses on how well each individual algorithm performs one of two different tasks that are among face recognition's most common applications. The first task, confirming a photo matches a different photo of the same person in a database, is known as "one-to-one" matching and is commonly used for verification work, such as unlocking a smartphone or checking a passport. The second, determining whether the person in the photo has any match in a database, is known as "one-to-many" matching and can be used for identification of a person of interest.

To evaluate each algorithm's performance on its task, the team measured the two classes of error the software can make: false positives and false negatives. A false positive means that the software wrongly considered photos of two different individuals to show the same person, while a false negative means the software failed to match two photos that, in fact, do show the same person.

Making these distinctions is important because the

class of error and the search type can carry vastly different consequences depending on the real-world application.

"In a one-to-one search, a false negative might be merely an inconvenience—you can't get into your phone, but the issue can usually be remediated by a second attempt," Grother said. "But a false positive in a one-to-many search puts an incorrect match on a list of candidates that warrant further scrutiny."

What sets the publication apart from most other face recognition research is its concern with each algorithm's performance when considering demographic factors. For one-to-one matching, only a few previous studies explore demographic effects; for one-to-many matching, none have.

To evaluate the algorithms, the NIST team used four collections of photographs containing 18.27 million images of 8.49 million people. All came from operational databases provided by the State Department, the Department of Homeland Security and the FBI. The team did not use any images "scraped" directly from internet sources such as social media or from video surveillance.

The photos in the databases included metadata information indicating the subject's age, sex, and either race or country of birth. Not only did the team measure each algorithm's false positives and false negatives for both search types, but it also determined how much these error rates varied among the tags. In other words, how comparatively well did the algorithm perform on images of people from different groups?

Tests showed a wide range in accuracy across developers, with the most accurate algorithms producing many fewer errors. While the study's focus was on individual algorithms, Grother pointed out five broader findings:

1. For one-to-one matching, the team saw higher rates of false positives for Asian and African American faces relative to images of Caucasians. The differentials often ranged from a factor of 10 to 100 times, depending on the individual algorithm. False positives might present a security concern to the system owner, as they may allow access to impostors.
2. Among U.S.-developed algorithms, there were similar high rates of false positives in one-to-one matching for Asians, African Americans and native groups (which include Native American, American Indian, Alaskan Indian and Pacific Islanders). The American Indian demographic had the highest rates of false positives.
3. However, a notable exception was for some algorithms developed in Asian countries. There was no such dramatic difference in false positives in one-to-one matching between Asian and Caucasian faces for algorithms developed in Asia. While Grother reiterated that the NIST study does not explore the relationship between cause and effect, one possible connection, and area for research, is the relationship between an algorithm's performance and the data used to train it. "These results are an encouraging sign that more diverse training data may produce more equitable outcomes, should it be possible for developers to use such data," he said.
4. For one-to-many matching, the team saw higher rates of false positives for African American females. Differentials in false positives in one-to-many matching are particularly important because the consequences could include false accusations. (In this case, the test did not use the entire set of photos, but only one FBI database containing 1.6 million domestic mugshots.)
5. However, not all algorithms give this high rate of false positives across demographics in one-to-many matching, and those that are the most equitable also rank among the most accurate. This last point underscores one overall message of the report: Different algorithms perform differently.

Any discussion of demographic effects is incomplete if it does not distinguish among the fundamentally different tasks and types of face recognition, Grother said. Such distinctions are important to remember as the world confronts the

broader implications of face recognition technology's use.

More information: Patrick Grother et al. Face recognition vendor test part 3:, (2019). DOI: 10.6028/NIST.IR.8280

Provided by National Institute of Standards and Technology