Facial Recognition Bias

Vincent White

2020 was an eventful year for biometrics, and Facial Recognition Technology ('FRT') in particular. The advent of COVID changed many behaviours and inadvertently accelerated the uptake and application of facial recognition in some areas, and also presented an unexpected new challenge in the identification of individuals wearing masks. However, on balance, the year was overshadowed by controversy, cases of false arrest, doubts about the accuracy of facial recognition, and allegations of 'bias', as measured by error rates compared across different demographic groups.

The media, perhaps unintentionally, was responsible for some inaccurate reporting on the subject, quoting error rates without context or interpretative analysis, and misconstruing their significance. These figures all too frequently are seized upon as evidence of FRT's inaccuracy, and blurred with an objection to a particular use case, for example in applications of surveillance by law enforcement. (Note that, conversely, many applications are viewed by the general public as benign, or at least necessary, for example identity verification used to unlock one's phone, or a biometric scan at an airport for security reasons).

Kjell Carlsson, an analyst at Forrester, succinctly described the problem:

"The moves reflect a lack of popular understanding of the technology–the public is conflating facial recognition with body recognition and tracking, facial analysis, facial detection, gender/age/ethnicity recognition, biometric validation, etc. as well as misunderstanding the difference between the use case and the technology".

In December 2019, the US National Institute of Science and Technology continued their ongoing benchmarking and assessment of multiple FRT algorithms with an analysis of demographic effects, to specifically examine the assertions that such differences could experience variations in accuracy, amounting to potential bias[1].

Despite their authoritative standing on the subject of FRT performance, the detail of NIST's findings was diluted, condensed, and certain media sources and groups staunchly opposed to FRT in most circumstances, such as civil liberties organisations, selectively cited the research as proving inherent bias in [all] FRT technology.

The truth is that accuracy in FRT has improved in leaps and bounds, as evidenced in the NIST benchmarking over a period of years. The sweeping conclusion that there are always significant false-positives differentials is unfounded; it depends on the algorithm and use case.

- In simple terms, there are two measures of error, false positive and false negative. The significance and implications of these error rates depends very much on the context or use case.
- Differentials are not the same in one-to-one matching and one-to-many matching. In KYC, identity verification is one-to-one, and screening is one-to-many. For identity verification, a false-negative might merely be an inconvenience, as a repeat matching effort can be made, whereas a false-positive could have serious consequences (i.e., someone else unlocks my phone, or accesses my bank account). For screening it is the converse: false negatives are very undesirable, and false positives are an inconvenience.

[1] https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf

- Note also that attention may be focused on false-positives, exaggerating their severity, on the misapprehension that this would necessarily entail an automated decision of some kind (for example, falsely accusing a suspected criminal), whereas it typically represents rather a burden on human intervention, i.e., to scrutinise and adjudicate a potential match.
- 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. In the context of KYC we would expect reasonably high quality, portrait images on identity documents that must meet certain quality criteria.
- "False negative error rates vary strongly by algorithm, from below 0.5% to above 10%. For the more accurate algorithms, false negative rates are usually low with average demographic differentials being, necessarily, smaller still".

On the last point, the finding is that there is a high degree of variance, with superior algorithms performing better according to many measures, including the effect of demographic differentials. As with any such distribution of performance, it is problematic to make an overall generalisation.

"We note that demographic differentials present in one-to-one verification algorithms are usually, but not always, present in one-to-many search algorithms. One important exception is that some developers supplied identification algorithms for which false positive differentials are undetectable. Among those is Idemia, who publicly described how this was achieved".

NIST Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects

As informative as the results of the NIST studies may be, they are somewhat scientifically clinical, and perhaps less accessible or comprehensible to the general public. To this end, perhaps more compelling findings have come from researching hybrid interaction and performance of humans and machines.

While the state-of-the-art in facial recognition algorithms exceeded human performance many years ago, a study from the US National Academy of Sciences[2] specifically examined collaborative performance, and where we get the best results, with almost unmeasurable error rates, are from combining the latest algorithms with the highest performing humans, which may in turn be a combination of natural aptitude (the 'super-recognisers') and training (that is, professional facial examiners).

Interestingly, this fusion is superior to any two algorithms or two individuals – and this is supported by computational theory that fusing systems works better when their search strategies differ.

The hybrid human-machine combination delivers the best performance, and this is precisely the ideal that we strive for in many other areas of human-computer interaction (e.g. automation in aviation), which exhibit several desirable qualities in terms of design and the outcomes produced, and therefore also the suitability of certain applications or use cases:

- To harness the superior performance of machine in terms of accuracy and/or fallibility, particularly with routine or mundane tasks, or those that are computationally complex;
- To ensure there is not a single factor dependency on machine alone, to avoid particular types of fallibility specific to rules-based systems;
- To maintain humans as the ultimate arbiter or adjudicator in cases of conflict - it's critical we have a human adjudicator, we need this to make determinations of what are false-positives, and in some cases, to make nuanced judgements;
- To benefit from the synergy of different approaches by human and machine.

[2] https://www.pnas.org/content/115/24/6171

Applying FRT to watchlist screening in KYC is precisely this scenario of getting the best results from a collaborative human-machine effort; leveraging an algorithm's superior capacity to identify similarity, without excluding human oversight and necessary decision-making in adjudication.

In my previous article, I argued that facial recognition is a wholly more suitable and successful approach to watchlist screening, given all that we have learned about the limitations and shortcomings of screening using alphanumeric data.
In this piece, I sought to challenge some of the misconceptions about accuracy and bias that may unfairly be hindering the adoption of facial recognition as a superior method of watchlist screening. Where screening a database of considerable size and a top-performing matching algorithm is used, false-positive differentials from demographic factors can be undetectable.

The views, thoughts, and opinions expressed in this article are my own, and not necessarily those of my employer.