More Trust. Less Crime. High Standards.

The Metropolitan Police Service Facial Recognition Technology: Understanding accuracy and demographic differences

A. BACKGROUND AND CONTEXT

1. The Metropolitan Police Service (the Met) uses Facial Recognition Technology (FRT) to assist officers in identifying persons of interest. Whilst the MPS LFR Policy Document provides the reference point for how each technology is referred to by the Met, in policing, the technology can be split into three operational use cases:
    - **Live Facial Recognition (LFR)** compares a live camera feed of faces against a predetermined watchlist to find a possible match that generates an alert.
    - **Retrospective Facial Recognition (RFR)** is a post-event use of facial recognition technology, which compares still images of faces of unknown subjects against a reference image database in order to identify them.
    - **Operator Initiated Facial Recognition (OIFR)** is a near-real-time use of facial recognition technology, where an officer takes a photograph of a subject via a device and submits it for immediate search against a reference image database.

2. An accurate LFR system requires that:
    (i) Whenever an individual on the watchlist passes the system the LFR system should generate an alert and
    (ii) Whenever an individual who is not on the watchlist passes the LFR system, the LFR system should not generate an alert and in line with the current MPS LFR Documents, this should then automatically delete all biometric data relating to that individual.

3. While an accurate RFR / OIFR system requires that,
    (i) If the person in the still image (or being photographed for OIFR) also has a face image in the image reference dataset, this 'mated' image should be among the top list of candidate images returned.

4. In addition to having a good level of accuracy such that policing can expect to achieve its purposes for using FRT, it is also important that the FRT performs well in terms of minimising algorithmic / system biases. For example, performance differences between different demographic groups would disadvantage one demographic group in comparison to another. This could undermine public trust in the use of the technology. System bias may also prevent policing from achieving objectives they set out to achieve when using FRT. The Court of Appeal in the *Bridges* case identifies race and gender as being particularly relevant considerations for FRT.[1]

[1] *R (on the application of Bridges) v the Chief Constable of South Wales Police* [2020] EWCA Civ 1058 at (amongst others) paras 164, 176, 181

B. HOW TO UNDERSTAND LFR ACCURACY

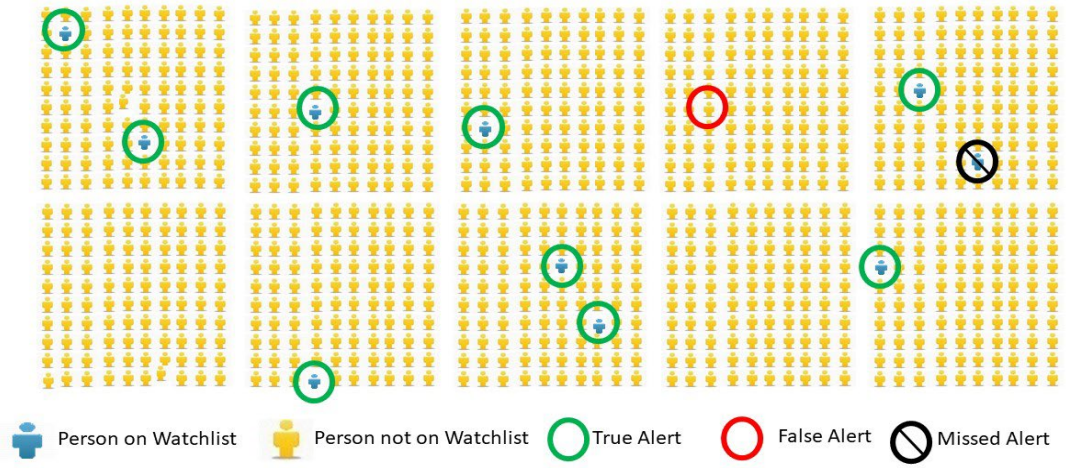
It is incorrect to describe the 'accuracy' of a LFR system by a single figure (for example 98% (in) accurate). Instead, the internationally accepted standards to assess overall system accuracy are determined based on two measures:

(i) the True-Positive Identification Rate (TPIR) and
(ii) the False-Positive Identification Rate (FPIR)

5. The diagram below shows a simplified example of how a Live Facial Recognition system calculates these measures.

<table>
<tr><th></th><th>True Positive Identification Rate (TPIR)</th><th>False Positive Identification Rate (FPIR)</th></tr>
<tr><td>What is it?</td><td>Describes:
• the total number of times an individual(s) on a watchlist who is known to have passed the LFR system and correctly generate an alert;
• <u>as a proportion of</u> the total number of times those individuals[2] pass the LFR system, regardless of whether an alert is generated by the LFR system or not.

The TPIR is also known as the True Recognition Rate.</td><td>Describes:
• the number of individuals who pass the LFR system, but who are <u>not</u> on the watchlist and who incorrectly generate an alert
• <u>as a proportion of</u> the total number of occasions people[3] pass the LFR system.

The FPIR is also known as the False Alert Rate.</td></tr>
<tr><td>Worked Example *</td><td colspan="2"> 
</td></tr>
<tr><td></td><td>The True Positive Identification Rate would be 90% if 10 people on the watchlist each pass the LFR system, and a Correct alert is generated for 9 out of 10 of those people (with no alert being generated against the 10th person – Missed alert).</td><td>The False Positive Identification Rate would be 0.1%, if for every 1,000 people that passed the LFR system, an alert was generated against one person who was not on the watchlist.</td></tr>
<tr><td></td><td colspan="2">*Simplified to demonstrate the concept of TPIR & FPIR</td></tr>
</table>

[2] Given the need to know that an individual has passed the LFR system, this needs to be established using a controlled watchlist of known individuals and a record kept of when they pass the LFR system. This list is known as a 'bluelist' and those on it are seeded into the passing crowd to establish if the LFR system generates an alert against them or not.

[3] This is a separate measure to the TPIR and does not need to use a 'bluelist' of known individuals. This is a measure of how often the LFR system incorrectly generates an alert against an individual who passes the LFR system and that individual is not on the watchlist. If a person is on a watchlist or not can be established following an alert.

C. HOW ACCURATE IS THE MET'S FRT COMPARED TO OTHERS?

6. The National Institute of Standards and Technology (NIST) is a physical sciences laboratory and a non-regulatory agency of the United States Department of Commerce. NIST have run open, large-scale Face Recognition Vendor Tests (FRVT) to assess the accuracy of facial recognition algorithms since 2000[4]. NIST Tests are normally run on very large 'data sets' of still images (typically between 1.6 and 12 million). These tests allow the Met to compare the baseline accuracy of different algorithms from different vendors. They have also allowed the Met to monitor the improvement of facial recognition accuracy over time.

What does NIST tell us about the Met's FRT?

7. The Met's current facial recognition systems use an algorithm supplied by NEC, one of the top performing vendors in NIST FRVT evaluations. The NIST Test report published in 2019 evaluated over 200 algorithms for their accuracy. Its findings state that:

> *"NEC, which had produced broadly the most accurate algorithms in 2010, 2013, submitted algorithms that are substantially more accurate than their June 2018 versions and on many measures are now the most accurate".*

8. NEC's facial recognition algorithm earned one of the top rankings from FRVT 2022, NIST results show the NEC algorithm performance has continued to improve with each year.

D. WHAT ABOUT DEMOGRAPHIC DIFFERENCES IN THE MET'S FRT COMPARED TO OTHERS?

9. Consideration of biometric system performance variation across demographic groups involves measurement of the overall system accuracy to establish the statistical significance of any differences in performance for different demographic groups. The *Bridges* decision identifies gender and race as being particularly relevant to LFR.

10. From 2019, the NIST 'face recognition vendor tests' started to assess whether demographics such as gender or race cause FR Identification system accuracy to vary. Tests were run on a 2.6 million-image dataset where images were balanced with respect to representation of gender and race. The NIST results demonstrate that not all algorithms show uniform accuracy levels across the different demographics. However, NEC, the vendor used by the Met was found to perform well, with NIST saying that NEC had:

> *"…provided an algorithm for which the false positive differential was undetectable"* and the NEC-3 algorithm "*is on many measures, the most accurate [NIST] have evaluated*".

11. This provides assurance to the Met as the tests showed that the variation in accuracy between male, female, black and white individuals in the NEC face comparison algorithm is imperceptible.

How do the algorithms tested by NIST relate to the Met's FRT?

12. The NIST Tests provide the Met with robust, transparent, independent and comparable information on how different algorithms including the NEC algorithms perform, in terms of both accuracy and demographic differences. However, the NIST Tests do not directly replicate the conditions found when deployed to support law enforcement use cases; for example in a live operational environment. These include factors such as environmental conditions, the number and density of subjects passing the LFR system, how subjects behave when passing the LFR system and occlusion given the uncontrolled environment. Occlusions can be anything that may obscure the person's face from the camera when passing the system, such as sunglasses. This can be compounded further in a busy

[4] https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt

environment, when other people in the crowd can often obfuscate others when passing the camera. NIST recognises this and recommends that end users should 'know their algorithm' [in the context of their system and Concept of Operation].

E. WHAT HAS THE MET DONE TO UNDERSTAND THEIR FRT IN AN OPERATIONAL ENVIRONMENT?

13. The NIST Tests can only take the Met so far, and by their nature, factors relevant to an operational environment can only be effectively tested in operational use. Further controlled testing would not accurately reflect operational conditions, particularly the numbers of people who need to pass the LFR system in a way that is necessary to provide the Met with further assurance.

14. As a result, in August 2021 the Met was awarded Home Office Science, Technology, Analysis & Research (STAR) funding to undertake testing of the accuracy and equitability of FRT in an operational environment for LFR, OIFR and RFR.

15. In collaboration with South Wales Police (SWP), this work was awarded to the National Physical Laboratory (NPL) at the end of 2021. The NPL is a prestigious world-leading centre of excellence that provides cutting-edge measurement science, engineering and technology to underpin prosperity and quality of life in the UK. In order to deliver on the objectives of the research, it was necessary to use LFR in the operational use cases of UK Policing. Data collection for the evaluation took place in July and August alongside five operational deployments of LFR, four in London and one in Cardiff.

16. A cohort of volunteers were selected to take part in the study who were of varying age, gender and race, the volunteers were seeded into the crowd passing the LFR system at each deployment so as to appear in the LFR video footage.

17. The data was then evaluated 'post event' with a balanced watchlist and mated facial photographs taken of the volunteers in a variety of settings to realistically replicate the use cases for LFR, RFR and OIFR.

18. The full results are presented in the National Physical Laboratory's commissioned report 'Facial Recognition Technology in Law Enforcement Equitability Study'

What does this study tell us about accuracy of the Met's FRT?

19. The NPL report gives us an impartial, scientifically underpinned, evidence-based robust analysis of the performance of the Met's facial recognition technologies in operational conditions in terms of (i) accuracy and (ii) equitability (bias) related to subject demographics.

20. For RFR and OIFR the performance metrics are measured in terms of True-Positive Identification Rate: TPIR - the proportion of 'mated' identification searches (i.e., where the subject has a record in the reference image database) that include the mated reference among the candidates returned.

21. For **every** probe image submitted for RFR or OIFR in the evaluation (based on a gallery of 178,000 images), the correct match was returned at Rank 1 (i.e., as the top match). With a TPIR of **100%, this is the best performance possible.**

22. In terms of LFR, for summarising operational performance, NPL have provided performance figures for two different watchlist sizes: (i) a watchlist of 10,000 reference images, which is broadly in line with those used on the Met's LFR operational deployments to date and (ii) a watchlist of 1000 reference images a size more typical for SWP LFR deployments.

23. The performance figures use industry standard measures;
    (i) True-Positive Identification Rate (TPIR) – the rate of successful recognition when subjects on the watchlist pass through the zone of recognition
    (ii) False-Positive Identification Rate (FPIR) – the rate of incorrect recognition (i.e., false positives or false alerts) when subjects not on the watchlist pass through the zone of recognition.

24. The table below shows the results of combined data from all five deployments:

Watchlist size 10000			**Watchlist size 1000**		
Metric	**Threshold Setting**	**Result**	**Metric**	**Threshold Setting**	**Result**
TPIR	0.60	= 89 %	TPIR	0.60	= 89 %
FPIR	0.60	≈ 0.017 % (1 in 6000)	FPIR	0.60	≈ 0.002 % (1 in 60,000)

Did the study find any demographic differences in the Met's FRT?

25. Results for RFR and OIFR showed the best possible performance. For every probe submitted, the correct match was returned at Rank 1 (this is the top match) resulting in a TPIR of 100%. This TPIR is identical for all demographic subgroups and these results show that **performance is the same, regardless of race, gender or age.**

26. In relation to LFR, NPL found that at a face-match threshold of 0.60, any differences in TPIR by gender, by race, or by race/gender combined were not statistically significant. This means that the systems performance is not biased towards any race or gender. The study has shown that at face match thresholds of 0.60, 0.62 and 0.64 the number of subjects with a false positive is very small and there is **no statistically significant imbalance between demographics**. The system performance is the same for race and gender.

27. In relation to age, the NPL found that at a threshold of 0.62 the observed differences in TPIR were not statistically significant. At a face-match threshold of 0.60, the observed variation in TPIR did show statistical significance with TPIR improving with subject age. This means that the system is slightly more likely to locate those sought by the Met as they age, but not more likely to inconvenience those of younger age, as the FPIR is found to be equitable between gender, race, and age. There is **no statistically significant imbalance between demographics.** In relation to trying to locate those of younger age, the NPL recognised;

    "……..the lower performance of the under 20s is therefore assessed to be due to both demographic and environmental factors, these being a combination of subject age and as a result subject height, and crowdedness in the zone of recognition……."

28. Where the Met is particularly seeking to locate those of younger age and whom maybe of a shorter stature, consideration should be given of how busy the area is. The risk that subjects may be shielded from the camera by a taller person walking in front of them and blocking the camera's view must be taken in to account. Therefore, deployment locations and camera positioning should form part of the technical optimisation process.

29. Reflective of the need for continuous improvement, the Met will continue to monitor its FRT performance, in terms of both overall system accuracy and demographic differential performance going forward.

F. WHAT ABOUT THE VIEWS OF OTHER EXPERTS BEYOND THE MET?

30. The Met's Technology, Research and Innovation unit has considerable expertise in relation to facial recognition technology. Met personnel are members of the Organisation for Scientific Area Committees for Forensic Science Facial Identification Subcommittee[5] and hold an executive position on the Facial Identification Scientific Working Group [6]

31. To support public confidence and to ensure the Met undertakes resilient analysis in discharging its Public Sector Equality Duty, the Met also recognises the value in seeking the views of others with recognised expertise.

32. The Met's approach outlined in this document has therefore benefited from peer review from the National Physical Laboratory, the UK's national metrology institute responsible for developing and maintaining the national primary measurement standards with recognised expertise in testing biometric systems. It has also sought input from the Defence Science and Technology Laboratory (DSTL), an executive agency of the Ministry of Defence whose stated purpose is to deliver high impact science and technology for the UK's defence, security and prosperity.

Protective marking:	Official
Publication scheme Y/N:	No
Title:	Understanding the Metropolitan Police Service FRT System's Accuracy and Demographic Differences
Version:	Version 3.0
Summary:	A published summary to assist the public understanding the steps the Metropolitan Police Service has taken to understand the accuracy of its algorithm its performance in relation to demographic issues.
Branch/ OCU:	MPS FRT
Review date:	Feb 2024

[5] https://www.nist.gov/organization-scientific-area-committees-forensic-science/facial-identification-subcommittee
[6] https://www.fiswg.org/