Live Facial Recognition: Algorithm Performance, Testing and Threshold Setting

Summary

Live Facial Recognition (LFR) is used as a policing capability to help forces prevent crime, protect the public, and locate individuals who are lawfully sought by the police or who may be at risk of harm. The use of LFR provides a proportionate tactical option to locate individuals more efficiently, enabling officers and staff to be redeployed to other policing priorities. In many circumstances, LFR represents a less intrusive means of identification when compared with alternative policing tactics.

The software operates using CCTV cameras to scan crowds of people specifically looking for individuals on a pre-determined list (watchlist) who we seek to locate. They can be people who are wanted for crime or missing and or at risk of harm.

Essex Police has been regularly deploying Live Facial Recognition (LFR) since August 2024, delivering strong operational outcomes. These include over 120 arrests for a variety of crimes including robbery, serious sexual offences and for offenders who have breached court orders. During deployments we have safeguarding several members of the public who were missing and or at risk of harm.

With each deployment, we aim to correctly identify as many wanted people as we can and protect as many people as possible who are at risk of harm, while minimising the chance that we will end up stopping innocent people.

We recognise that algorithmic systems are not always equitable at identifying people from different backgrounds correctly and equally. However, with rigorous testing, comparative analysis, and mitigation of any observed performance differentials we are able to adjust the settings of the software so that it has minimal impact but maximises the opportunity to locate people swiftly.

In conjunction with the Home Office / National Police Chiefs' Council (NPCC), Essex Police commissioned two independent studies to analyse and assess the performance of the algorithm used by the force. These studies were conducted by the University of Cambridge and the National Physical Laboratory (NPL).

Why Test?

Essex Police are committed to our legal and moral duty to ensure that we treat all citizens fairly and that we do not discriminate. To ensure that policing in Essex serves the whole community, the force took the decision to subject the algorithm to independent testing through two separate scientific studies.

This decision to test was framed by national policing guidance on the deployment of Live Facial Recognition (LFR), as set out in the College of Policing Authorised Professional Practice (APP), and informed by relevant case law,

The two studies were conducted independently of each other by Cambridge University and then the National Physical Laboratory (NPL).

In particular, the testing undertaken by the NPL mirrored earlier independent testing conducted for the Metropolitan Police Service and South Wales Police in 2022 and utilised the exact same operational footage. This approach enabled consistent, comparable, and collaborative testing, supporting a rigorous and in-depth scientific assessment of algorithm performance in real-world operational environments.

The test plan was specifically designed to help identify any impact this technology may have on people, in particular race, age and sex.

How to Understand LFR Accuracy

The term ‘accuracy’ of a LFR system can be difficult to describe and cannot be explained 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)

This means how well the algorithm identifies people (TPIR) and how often the system incorrectly generates an alert for someone who is not on the watchlist (FPIR).

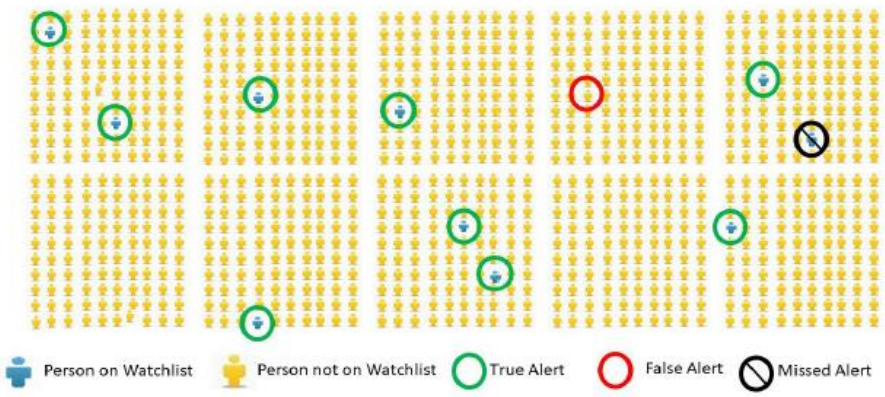
The accuracy of the LFR algorithm is also informed be the ‘sensitivity threshold’, the sensitivity threshold setting allows for the software to be adjusted to try and identify as many people as possible on the watchlist balanced against reducing the risk of falsely identifying members of the public. When the threshold setting is adjusted it can have an impact on the level of disparity between certain groups (ethnicity, age or gender). The basic principle is that if the threshold setting is raised then you are less likely to identify as many people as possible but will reduce the risk of false positives. Lowering the threshold increases identification but also raises the risk of false positives, while

potentially reducing bias in positive identifications. Once the threshold level has been set then the software will only alert to potential identifications above that score.

The diagram below shows a simplified example of how an LFR 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>

The LFR algorithm is not designed to target individuals. The watchlist is made up of individuals who are wanted for crimes, missing and or at risk of harm. The watchlists are populated based on crime data and intelligence which determines the individuals that go on the watchlist. The Live Facial Recognition (LFR) algorithm itself then seeks to identify individuals by comparing live images against the pre-defined watchlist; therefore, the algorithm does not 'target' individuals.

Other than when specific deep-dive analysis is undertaken, Essex Police does not collect demographic data relating to the composition of watchlists. This is because watchlists are populated solely based on crime data and intelligence considerations, and not by reference to any protected or other personal characteristics.

What were the results of the Testing?

The studies both showed that the algorithm used performs at a high level. In summary they tell us the following.

- Both studies presented performance results across a range of identification thresholds, and the National Physical Laboratory (NPL) also assessed performance against different watchlist sizes (180,000, 18,000 and 1,800). This evidence provided an important basis for informing and supporting decisions on the appropriate operational threshold setting.
- Neither study found evidence that the algorithm disproportionately or excessively falsely identifies individuals. This has been the central national concern in relation to the use of Live Facial Recognition, and the findings do not support that concern.
- Both studies identified performance disparities in the algorithm's ability to correctly identify certain demographic groups. In particular, the findings indicated that the algorithm was more effective at identifying Black males than any other demographic group. The Cambridge study found this disparity to be statistically significant, while the NPL study other did not consider the disparity as statistically significant.
- The NPL study found that at sensitivity thresholds of 55, 60, and 65 the algorithm was equitable with no statistically significant bias identified in the True Positive Identification Rate and False Positive Identification Rate.
- NPL also found that the rate of false positives was substantially lower than the national requirement of fewer than one false alert per 1,000 faces scanned.
- Both studies showed that the environmental conditions have an impact on how well the algorithm performs in terms of identifying people.

What has Essex Police done following the receipt of both reports.

Essex Police received the Cambridge report first, with the NPL Report being received three months later.

Following receipt of the Cambridge study, we took the decision to pause our deployment whilst we awaited the findings from the NPL study and investigated the Cambridge results. This process included the following:

- Engagement with our software provider who sought to make improvements to their technology considering the findings. The provider has since confirmed that they have made advancements to reduce disparity and are working on further improvements to the algorithm.
- Stakeholder Engagement including the Essex Data Ethics Committee, Police Chief Scientific Advisor, and the NPCC lead for facial recognition.
- The Force also conducted an assessment on the outcome data from our previous deployments; this assessment does not indicate that there is a disparity following an alert, intervention, or arrest across any demographic.
- Upon receipt of the NPL study, Essex Police sought advice from Professor Paul Taylor, the Police Chief Scientific Advisor, to assist with interpreting the results of the two reports.
    - Professor Taylor, provided advice on the Cambridge and NPL Studies, offering an interpretation of the results, in respect of specific questions we had around thresholds set at below 50, above 55 and 51.
    - He advised that the analysis suggested we should not use a threshold below 50 unless there was exceptional operational need to maximise identifications.
    - He further noted that the results did not support adopting a threshold above 55, because of the risk of not identifying offenders who are a threat to the public. The threshold could be lowered to 51, but [according to NPL's analysis] the gain of 1.7% in identifications [which] is not necessarily worth increasing the proportion of false alarms by .23%.
- Essex Police also engaged with the Essex Data Ethics Committee (EDEC), an independent advisory body who provide ethical oversight, scrutiny, and guidance for data-driven projects undertaken by public services across Essex.
    - The EDEC made several recommendations to the force, which we have adopted.
    - Their prevailing guidance was to not rely on or favour one report but instead use the findings and recommendations from both to further inform our approach to the use of LFR.

- The Force have responded to this advice with the results from both reports informing the revision of our policies and procedures. We have not sought to rely on or favour one report – we have used the finding from both to further enhance our use, oversight and effectiveness of LFR.

What do the results tell us?

Having analysed and interpreted the results of the academic testing, engaged with our supplier, and sought and acted on the views of various individuals and bodies, we are now able to better understand the demographic performance of our Live Facial Recognition (LFR) system.

As outlined, the intention of Essex Police is to ensure that its use of Live Facial Recognition (LFR) seeks to maximise the identification of individuals on the watchlist, whilst minimising the risk of misidentification and ensuring there is no unfair or disproportionate impact on any particular group.

Importantly, the studies found that our LFR technology is unlikely to incorrectly identify people who are not on the watchlist - this has been the central national concern in relation to the use of Live Facial Recognition, so it is important to note the performance of the algorithm in that regard.

Based on the findings from the studies, the interpretation we have sought from the Police Chief Scientific advisor, the assessment of the data from our previous deployments, and the activity we have progressed gives us reassurance that there are algorithm settings at which the system can be operated safely, lawfully and equitably.

We are therefore confident that operating the system at a threshold of 55 delivers an equitable policing response, with no evidence of demographic disparity in outcomes. This position will continue to be actively monitored through ongoing review of positive alerts and operational outcomes to ensure that this assessment remains valid over time.