National Physical Laboratory

NPL REPORT MS 43

FACIAL RECOGNITION TECHNOLOGY IN LAW ENFORCEMENT EQUITABILITY STUDY FINAL REPORT

DR TONY MANSFIELD

MARCH 2023

Facial Recognition Technology in Law Enforcement
Equitability Study
Final Report

Dr Tony Mansfield
Data Science Department

ISSN 1754-2960

National Physical Laboratory
Hampton Road, Teddington, Middlesex, TW11 0LW

Approved on behalf of NPLML by
Kevin Lees, Group Leader SED/DS/DATAAM

CONTENTS

1 SUMMARY

1.1 ABOUT THE EVALUATION

In August 2021 the Metropolitan Police Service (MPS) was awarded Home Office Science Technology Analysis & Research funding to undertake testing of the accuracy and equitability of Facial Recognition in an operational environment for three policing use cases:

- Live Facial Recognition (LFR)
- Retrospective Facial Recognition (RFR)
- Operator Initiated Facial Recognition (OIFR)

The National Physical Laboratory (NPL) was invited to submit a proposal and plan to conduct such an evaluation for MPS and South Wales Police (SWP), and a contract was awarded at the end of 2021.

The NPL test strategy [1] for an evaluation conformant with the standards ISO/IEC 19795-1 [2] and ISO/IEC 19795-2 [3] for biometric testing and reporting was agreed in April 2022. Noteworthy aspects addressed in the evaluation include:

1:N identification: The operational use cases evaluated use facial recognition for 1:N identification (rather than 1:1 verification).

Face in video: Live Facial Recognition involves identification of faces in live video.

Real time processing: Live Facial Recognition must operate in real time; the recognition decision must be given within seconds of the subject being videoed. To achieve this when there are many people in the field of view to be recognised, the algorithms may need to limit the number of faces processed per video frame or limit the number of video frames being processed.

Operational environment and settings: The evaluation uses face image and video data collected in the operational environment under operational settings.

Accuracy, demographic variation & equitability: The evaluation measures the accuracy of facial recognition, the variations in accuracy between different demographics, and assesses equitability, i.e., whether outcomes are broadly equivalent across demographics under operational use case settings.

Under 18's: The age range of the policing applications extends to those below age 18, and it is important to know whether performance is different for this demographic.

Large demographically balanced datasets*:* The testing of low error rates in a statistically significant manner requires large datasets. To achieve the required scale, the evaluation uses a supplementary reference image dataset of 178,000 face images (Filler dataset). This is an order of magnitude larger than the typical watchlist size of an operational Live Facial Recognition deployment. To avoid introducing a demographic bias due to reference dataset composition, a demographically balanced reference dataset was used, with equal numbers in each demographic category. For assessment of equitability under operational settings, the results from the large dataset are appropriately scaled to the size and composition of watchlist or reference image database of the operational deployment.

Data collection for the evaluation took place in July and August 2022 alongside five MPS and SWP operational deployments of Live Facial Recognition in London and Cardiff. A set of facial photographs were taken of a Cohort of subjects in a variety of settings. These Cohort subjects were seeded into the crowd flow to appear in the LFR video. Cohort subjects were not included on the operational deployment watchlists: an imperative to avoid interrupting the policing operation (and to reflect the safeguards of the data protection impact assessments regarding accurate data processing). Instead, the operational LFR video footage featuring the Cohort was saved to be re-played offline against the evaluation's Cohort and Filler watchlists at a later date.

The facial recognition technology and version tested is **NEC Neoface V4** [1] using **HD5 Face Detector**. These facial detection and recognition algorithms are those currently used by MPS and SWP for Live Facial Recognition, Retrospective Facial Recognition and Operator Initiated Facial Recognition.

The offline running of the data using the Neoface system emulating operational use, and our analyses of performance have taken place from September to November 2022.

This report sets out the findings of our evaluation, and is organised as follows:

- In the remainder of this section, we highlight some of the key findings of the study.
- Section 2 summarises the objectives of the evaluation.
- Section 3 provides details of the demographics assessed in the evaluation.
- Section 4 elaborates on criteria for equitability, and statistical significance of performance differences.
- Section 5 outlines the methodology, the image and video data collected for the evaluation and how it was used.
- Section 6 gives evaluation results for Retrospective Facial Recognition.
- Section 7 gives evaluation results for Operator Initiated Facial Recognition.
- Section 8 gives evaluation results for Live Facial Recognition.
- Section 9 provides further discussion on some of the findings and notes aspects that may be worthy of further testing or analysis.
- Section 10 provides a glossary of terms and abbreviations.

1.2 KEY FINDINGS – RETROSPECTIVE FACIAL RECOGNITION

Retrospective Facial Recognition 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. For each identification search the system returns a candidate list of the records in the reference image database that best match the submitted probe image. The top R matches are returned for a pre-specified value R.

Recognition accuracy for Retrospective Facial Recognition is measured in terms of:

- True-Positive Identification Rate: TPIR(N, R, 0) - 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.

TPIR depends on the number of candidates to be returned (R), the number of records in the reference image database (N). No face-match threshold is used in RFR.

[1] https://www.necsws.com/solutions/evidence-insights/facial-recognition-software/

1.2.1 In the evaluation, RFR always returned the correct reference identifier at Rank 1

For every probe image submitted for RFR in the evaluation, the correct reference identifier was returned at Rank 1 (i.e., as the top match). This is the best performance possible.

$$\mathrm{TPIR}_{\mathrm{RFR}\ 178400,\,1,\,0} = 100\ \%.$$

It follows that TPIR is identical for all demographic subgroups, and with no demographic performance variation in TPIR, performance is equitable.

1.3 KEY FINDINGS – OPERATOR INITIATED FACIAL RECOGNITION

Operator Initiated Facial Recognition is a near-real-time use of facial recognition technology, where an officer takes a photograph of a subject via a mobile device and submits it for immediate search against a reference image database. For each identification search the system returns a short candidate list of the records in the reference image database that best match the submitted probe image.

Recognition accuracy for Operator Initiated Facial Recognition is measured in terms of:

- True-Positive Identification Rate: TPIR(N, R, 0) – the proportion of mated identification searches (where the subject has a record in the reference image database) that include the mated reference among the candidates returned.

1.3.1 In the evaluation, OIFR always returned the correct reference at Rank 1

For every probe image submitted for OIFR in the evaluation, the correct reference identifier was returned at Rank 1 (i.e., as the top match). This is the best performance possible.

$$\mathrm{TPIR}_{\mathrm{OIFR}\ 178400,\,1,\,0} = 100\ \%.$$

It follows that TPIR is identical for all demographic subgroups, and with no demographic performance variation in TPIR, performance is equitable.

1.4 KEY FINDINGS – LIVE FACIAL RECOGNITION

Live Facial Recognition (LFR) compares a live camera video feed of faces against a predetermined watchlist to find a possible match that generates an alert.

The recognition accuracy of Live Facial Recognition is measured in terms of:

- True-Positive Identification Rate (TPIR) – the rate of successful recognition when subjects on the watchlist pass through the zone of recognition
- 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.

TPIR is sometimes referred to as the True Recognition Rate, and FPIR as the False Alert Rate.

TPIR and FPIR depend on the face-match threshold setting of the LFR system. FPIR and, to a lesser extent TPIR, also depend on the number of face images on the watchlist – the number of facial comparisons per subject passing through the zone of recognition, and the potential for a false positive, increases with watchlist size.

1.4.1 A substantial improvement in Live Facial Recognition accuracy

Our tests encompass a range of watchlist sizes and face-match thresholds. For summarising operational performance, we use a face-match threshold of 0.6 which is the default setting of the Neoface facial recognition software. We provide performance figures for two different watchlist sizes: (i) a watchlist of 10,000 reference images, which is close to the size of that previously used in MPS LFR deployments and (ii) a watchlist of 1000 reference images, a size more typical of SWP LFR deployments.

At these settings, combining the data from all five deployments:

Watchlist size 10,000
- $TPIR_{10000,\,1,\,0.6} = 89\ \%$
- $FPIR_{10000,\,0.6} \approx 0.017\ \%$ (1 in 6000)

Watchlist size 1000
- $TPIR_{1000,\,1,\,0.6} = 89\ \%$
- $FPIR_{1000,\,0.6} \approx 0.002\ \%$ (1 in 60,000)

These accuracy levels are a considerable improvement on that reported [4] for previous versions of the Neoface software. At that time, with watchlist size between 2000 and 4000, averaged over four deployments, TPIR ≈ 72 % and FPIR ≈ 0.1 % (1 in 1000).

1.4.2 The range of variation in True Positive Identification Rate due to demographic effects was the same as that due to environmental effects

At a face-match threshold of 0.6, the variation in TPIR due to demographics of the Test Cohort ranged from a TPIR of 83 % to TPIR of 93 %. The extent of this variation was the same as that due to environmental effects (ranging from TPIR of 83 % to 94 %).

1.4.3 The observed variation in True Positive Identification Rate across gender and ethnicity was not statistically significant

At face-match threshold 0.6, the ethnicity-gender group with the best TPIR was the Asian-Female group, and the poorest TPIR was for the Black-Female group. However, the observed differences in TPIR by gender, by ethnicity, and by ethnicity-gender combined were not statistically significant at the 0.05 significance level. (Statistical significance quantifies whether the observed performance difference is likely due to chance, or due to some underlying factor. Following convention, a 0.05 significance level was set prior to evaluation and analysis of results. Section 4.1 provides further detail on statistical significance.)

1.4.4 Variation of True Positive Identification Rate between age groups

At face-match threshold 0.6, the observed variation in TPIR for the different age groups was statistically significant, TPIR improving with subject age.

The TPIR of 93 % for the oldest quartile (age 42 and over) is significantly higher than the TPIR of 89 % for those in the 20-to-41 age group.

The TPIR of 83 % for the youngest quartile (the under 20's) is significantly lower than the TPIR of 90 % for those aged 20 or over. However, it should be noted that the under-18 portion of the Cohort all attended on the busiest day of the LFR deployments and that, when the zone of recognition was crowded, the TPIR worsened. The lower performance for the under 20's is assessed to be due to both subject and environmental factors, these being a combination of subject age and as a result subject's height, and crowdedness in the zone of recognition leading to shorter subjects being hidden by others from the camera's field of view.

1.4.5 Demographic variation in False Positive Identification Rate is dependent on the face-match threshold

The LFR false positive cases against the 178,000 Filler image watchlist were analysed to determine any significant variation in FPIR between the gender, ethnicity and age demographic groups.

At face-match threshold of 0.64 or higher, there were no false positive identifications, thus at this threshold the FPIR is identically 0.0 for all demographic groups.

At face-match threshold of 0.62, only one Cohort subject had false positive identifications, and, at face-match threshold of 0.60, seven Cohort subjects had false positive identifications. In neither case is the imbalance between demographics statistically significant.

False positive identifications increase at lower face-match thresholds of 0.58 and 0.56 and start to show a statistically significant imbalance between demographics with more Black subjects having a false positive than Asian or White subjects.

1.4.6 Equitability

Under the criteria we have set for equitability (see Section 4):

- TPIR of the system at face-match threshold 0.6 is equitable across gender and ethnicity groups.
- FPIR is equitable between gender and ethnicity and age at face-match threshold 0.6 and above.
- At face-match thresholds lower than 0.6 FPIR equitability will depend on settings of the operational deployment, including size and composition of the watchlist, and the number of crowd subjects passing through the zone of recognition during the deployment.
  Given our observations on the demographic variation in FPIR, we would recommend, where operationally possible, the use of a face-match of 0.6 or above to minimise the likelihood of any false positive and adverse impact on equitability.

2 EVALUATION OBJECTIVES

This report provides the results of an evaluation of accuracy and equitability of facial recognition technology in three operational use-cases:

- **Live Facial Recognition**[2] (LFR) compares a live camera video 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 mobile device and submits it for immediate search against a reference image database.

[2] In the study the processing of face image data to evaluate LFR, RFR and OIFR was all carried out retrospectively, and omitting any operator involvement in submitting individual images for an identification search, or in adjudication of candidate matches returned by the system. Nevertheless, we shall use the terms LFR and OIFR even though the processing was not ‘live’, or ‘operator initiated’

The objectives of the evaluation are to assess the performance of facial recognition technology in an operational setting in terms of accuracy and equitability related to subject demographics, to add to Law Enforcement's understanding on how their facial recognition algorithms perform, and to provide information on how best to configure FR technology for effective and fair deployment on operational use cases.

The evaluation determines for each operational use case:

- What is the accuracy of the facial recognition algorithms?
  Accuracy of LFR is measured in terms of the True Positive Identification Rate (TPIR) and False Positive Identification Rate (FPIR) as a function of the face-match threshold. Accuracies of RFR and OIFR are measured in terms of the True Positive Identification Rate (TPIR) as a function of the number of top matches returned.
- Equitability is assessed through the consideration of the variations in accuracy for different demographic groups
  - What is the variation in accuracy between the demographic groups?
  - Are the variations in accuracy large enough to be statistically significant?
  - How do demographic variations in accuracy affect outcomes in the operational settings?
- Are demographic performance variations similar over the different operational use cases?
- Are variations in accuracy affected by environmental factors (such as location, crowd density, etc.) and system factors (such as algorithmic thresholds, and composition of watchlist or reference database)?

The evaluation has also collected a ground-truth dataset the UK Law Enforcement Community can use for future testing of other facial recognition algorithms.

3 DEMOGRAPHICS

3.1 DEMOGRAPHIC CATEGORIES

The demographic factors addressed in this evaluation are: (Self-defined) Ethnicity, Gender, Age and Height.

3.1.1 Ethnicity

The self-defined ethnicity of Cohort and Filler subjects was classified in accordance with the ONS 5+1 high-level ethnic groups [5]. For sourcing of Cohort subjects, for selection criteria of face images for the Filler reference dataset, and for analysis of performance the grouping of ethnicities used is:

- 'Asian or Asian British' (Bangladeshi, Chinese, Indian, Pakistani, Other Asian)
- 'Black, Black British' (African, Caribbean, Other Black)
- 'White' (English, Welsh, Scottish, Northern Irish, Irish, Gypsy, Roma, Other White)

We shall refer to these high-level groups as 'Asian', 'Black' and 'White' in this report.

These three groups are the largest in the national population (White: 81%, Asian, 9.6%, Black 4.2% [5]) and consequently the largest in the MPS and SWP custody records and of greatest relevance for policing operations. The remaining two high-level groups 'Mixed or Multiple' and 'Other ethnic group' were intentionally omitted in favour of having higher numbers of Cohort and Filler data for the demographic groups studied.

3.1.2 Gender

Gender[3] of the Cohort and Filler subjects was self-declared as Male or Female.

3.1.3 Age

To align with the Policing operational use cases for Facial Recognition, the study addressed a broad range of ages. Ages of Cohort subjects were self-declared.

The agencies approached to provide Cohort subjects for the trial were asked (in addition to requirements for gender and ethnicity) to recruit from across the age range 18–65 plus, and to avoid over-representation in the older age ranges. To include under-18's in the Cohort, NPL accepted the offer by MPS to provide Police Cadet volunteers of ages 12 to 18 for one of the deployments.

The selection of Filler data drawn from MPS records is assumed to be representative of the overall subject age profile in the MPS custody records. In the final datasets, the age profile of Cohort and Filler was quite similar in terms of interquartile ranges. For analysis of the differences in performance between age groups, we divided Cohort subjects into quartiles by age, from the youngest quartile to the oldest quartile.

3.1.4 Height

Height data was recorded for the Cohort as it was thought possible that occlusion of one subject behind another could adversely affect face detection, and this would be a more likely occurrence for the shortest subjects than for the tallest.

4 ASSESSING EQUITABILITY

The evaluation is assessing facial recognition accuracy, variations in performance for different demographics, and equitability between demographics in operational systems. Demographic variation in accuracy performance may be more easily observable at settings outside the normal operational parameters, hence our use of a Filler dataset to allow a watchlist much larger than those of typical LFR deployments. However, equitability must be assessed at typical operational settings.

For consideration of equitability in operational LFR deployments we hypothesise two watchlists of size and demographic composition to be typical for MPS and SWP operational deployments. The first has 10,000 face images, the second 1000 face images, and in both cases the composition of the watchlist can be expected to be proportional to the number of arrests in 2020/2021 by MPS and SWP respectively [6].
To scale results of larger watchlist (size $c \times N$) to a smaller watchlist (size N) we note that, provided FPIR(N, T) is small:

$$\text{TPIR}(c \times N, 1, T) \approx \text{TPIR}(N, 1, T) \text{ and}$$
$$\text{FPIR}(c \times N, T) \approx c \times \text{FPIR}(N, T).$$

Equitability between demographics requires that, in the operational setting, the outcomes for the subjects (i.e., recognition rates and false alert rates) should be broadly equivalent for demographics considered.

[3] **Gender**: classification as male, female or another category based on social, cultural or behavioural factors. (Gender is generally determined through self-declaration or self-presentation and may change over time.)

We cannot require exact equivalence as, even when there is no demographic variation in performance, due to the statistical nature of biometrics small deviations in observed performance must be expected. Thus, in assessing whether a system is equitable, criteria are needed for broad equivalence of performance figures.

In this study we use the following criteria for determining equitability:

a) The system is performing equitably if there is no variation in performance between demographic groups (e.g., when there are no false positives at the face-match threshold applied).

b) The system is performing equitably if the variation in performance between demographic groups is not statistically significant.

c) The system is performing equitably if the variation in performance between demographics is inconsequential in the operational system (e.g., if there is virtually no difference in outcomes for the different demographics).

4.1 STATISTICAL SIGNIFICANCE

Statistical significance quantifies whether the observed performance difference is likely due to chance, or due to some underlying factor of interest. For testing of statistical significance in the evaluation, we use the conventional significance level 0.05 (5%).

The significance level relates to the probability of falsely rejecting the ‘null’ hypothesis that there is no underlying difference in performance rates.

Note that testing for performance variation over different demographic attributes involves multiple hypothesis tests. A multiplicity of tests each at 5% significance level can increase the probability of rejecting the null hypothesis to a value much higher than 5%. There are methods to address this issue by requiring stricter significance thresholds for each individual test (however this increases the chance of missing a demographic effect). Such methods were NOT applied in this evaluation.

5 OUTLINE OF METHODOLOGY AND DATA

The tests conducted emulate the operational LFR, RFR and OIFR use cases.

A set of facial photographs were taken of a Cohort of subjects in a variety of settings including ‘Custody-style’ images captured in accordance with the Police Standard for Still Digital Image Capture of Facial Images [7]. The custody-style images were used for enrolling custody subjects onto a LFR watchlist, and facial image reference dataset for RFR and OIFR.

Video featuring Cohort subjects was collected alongside an operational LFR deployment and so reflects typical operational conditions. The behaviour of Cohort subjects seeded into the crowd was also consistent with normal crowd behaviour (other than being under instruction to ensure that they did pass through the zone of recognition, and to have a barcode scanned after going through the zone of recognition to log timings, and to walk through the zone of recognition the requisite number of times).

The following differences between the evaluation conditions and normal operational use should be noted.

a) For purposes of the evaluation, the watchlist used for testing of Live Facial Recognition was an order of magnitude larger than typical for an MPS LFR deployment. The watchlist contained nearly 180,000 face images, which is about 20 times the size of any watchlist used operationally to date. The increased watchlist size was to help ensure there would be sufficient data on true and false positive alerts to draw meaningful (statistically significant) conclusions regarding demographic differences in performance.

b) To address the purposes of the evaluation in assessing demographic variation in performance, the Filler and Cohort watchlists comprised an approximately equal number of face images for males and females of Asian, Black, and White ethnicities. In operational deployments the demographic balance of the watchlist would be different, and more likely to reflect the demographic balance in society, or of images in the MPS or SWP Custody Image Systems.

c) The facial recognition algorithms were configured to run in a bulk processing mode without the need for operator involvement or interaction during the process. The volume of data makes it impractical to involve the operator in processing each image, though this might be the case in operational use of RFR or OIFR.

d) In particular, the settings of the systems were not dynamically adjusted by test staff during running as conditions of facial images and video footage varied.

e) There was no operator or test staff adjudication of candidate matches returned by the system; the effects of such adjudication are outside the scope of the study.

f) The LFR video was not ‘live stream’ video but had been saved as mp4 files. At times, artefacts of compression were noticeable in the videos. Such artefacts can affect face image quality and may have had a negative impact on the accuracy of facial recognition compared with uncompressed video used for live deployments.

5.1 COHORT DATA SUBJECTS

A Cohort of test subjects meeting the demographic requirements for the study were recruited via acting extras agencies and an additional under-18 Cohort of test subjects were provided from MPS Police Cadet volunteers.

We have used the demographic details that subjects provided when they completed our demographics form; in some cases this was different to that stated by the supplying agency. A summary of the details is shown in Table 1. The number and demographic mix of the Cohort data subjects is sufficient to test equitability between Male/Female gender, Asian/Black/White ethnicity, and Age groups.

Table 1 — Summary of Cohort composition

Cohort composition: 405 Data subjects			
		Female	**Male**
Self-defined ethnicity based on ONS 5+1 codes	Asian (A1, A2, A3, A4, A9)	53	45
	Black (B1, B2, B9)	60	51
	White (W1, W2, W3, W9)	84	86
	Mixed and Other	8	16
Age	Age range	12 - 76 years	
	Lower Quartile	20 years	
	Median	30 years	
	Upper Quartile	42 years	

5.2 FACIAL IMAGES OF COHORT

A set of facial photographs was taken of each Cohort subject alongside the MPS and SWP deployments of Live Facial Recognition. The photographs were taken in a variety of settings as summarised in Table 2. Measurement of RFR and OIFR accuracy is based on the Cohort probe and reference images as listed in the table.

Table 2 — Summary of types of reference and probe images used in the study

Image type	Uses	Camera / Location	Number of images
Custody style[4] Image of the full head with all hair, neck, shoulders and ears. Subject facing square to the camera, looking directly at camera. Diffuse lighting to provide uniform illumination across the face without hot spots or shadows. Plain flat background with an 18% shade of grey.	Enrolment of reference image for watchlist or reference image database Probe images for RFR	Canon EOS850D Indoor	686
OIFR A facial image taken on a mobile phone (same model as used by SWP for OIFR). In cases where the image is out-of-focus, there is motion blur, or the subject's eyes were closed a further image would be taken.	Probe images for OIFR Probe images for RFR	Samsung XcoverPro Indoor	567
		Samsung XcoverPro Outdoor Some images at LFR location	576
Selfie Mobile phone image taken by Cohort subject. Subject permitted to pose as they wished for the photo. A neutral expression was not required, and the subject did not need to pose 'square to the camera'	Probe images for RFR	Motorola G(60)s Indoor	570
		Motorola G(60)s Outdoor	453
Ad hoc digital camera image Camera image of the subject taken either inside or outside. Uncontrolled conditions and background. No flash or supplementary lighting. Some of the ad-hoc outdoor photos were taken at the LFR location.	Probe images for RFR	Nikon D40 Indoor	506
		Nikon 1 J5 Outdoor Some images at LFR location	582
Note: Images taken with Cohort subject wearing facemasks are excluded. (Performance and equitability for subjects wearing facemasks not assessed in this study).			

[4] Conformant to "Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images" Version 2, 2007

5.3 VIDEO FROM LIVE FACIAL RECOGNITION DEPLOYMENTS

Measurement of LFR accuracy is based on the Cohort identification transactions (the 'recognition opportunities' as Cohort subjects walk through the zone of recognition of the LFR system). Details of LFR deployments are shown in Table 3.

Table 3 — Summary of LFR deployments with seeded Cohort subjects

Date	Location	Duration	Estimated number of crowd subjects
7 Jul 2022	London, Oxford Street	7 hours	24000
14 Jul 2022	London, Oxford Street	8 hours	35000
16 Jul 2022	London, Oxford Street	8 hours	38000
28 Jul 2022	London, Piccadilly Circus	7 hours	28000
13 Aug 2022	Cardiff, Queen Street	4.5 hours	7000

Weather conditions were quite similar during the five deployments, precluding analysis of weather effects. All days were bright and sunny, and several Cohort subjects wore sunhats and sunglasses in keeping with the conditions. August 13 was exceptionally hot (peak temperature: 33 °C).

Video was collected from two camera systems spanning the zone of recognition of the LFR deployment except for the afternoon of 13 August when a single camera was used.

The number of crowd subjects was estimated by taking a one-minute sample of each 30-minute video and combining contemporaneous footage from both cameras. Staff counted the number of subjects walking towards the camera in each sample video to estimate the number of crowd subjects seen by the LFR camera per minute. Multiplying by 30 gives the number of crowd subjects for the pair of videos, and the daily estimate sums the numbers for that day's videos.

The estimated number of subjects per minute is also used as a measure of how congested the zone of recognition was at the times the Cohort recognition opportunities were recorded. (The actual numbers in the zone of recognition would need to take into account (i) that the crowd walks both towards and away from the camera, and (ii) the time duration it takes for someone in the crowd to move through the zone of recognition. When the zone of recognition was very busy, it took more time for subjects to walk through.

5.4 FILLER FACIAL IMAGES

A Filler dataset was used to provide a large watchlist or reference dataset for evaluating LFR, RFR and OIFR performance. The demographic composition of the dataset is shown in Table 4.

Table 4 — Summary of Filler facial image set

Filler dataset from MPS holdings: ~ 178,000 facial images from ~ 116,000 individuals			
		Female	**Male**
Self-defined ethnicity based on ONS 5+1 codes	Asian (A1, A2, A3, A4, A9)	~ 30 000	~ 30 000
	Black (B1, B2, B9)	~ 30 000	~ 30 000
	White (W1, W2, W3, W9)	~ 30 000	~ 30 000
Age at date image taken	Age Range[5]	Approx. 12 - 76 years	
	Age Lower Quartile	22 years	
	Age Median	31 years	
	Age Upper Quartile	41 years	

The Filler dataset has approximately 180,000 face images: 30,000 for each of the Gender/Ethnicity groups shown. Approximately 2000 of the supplied images could not be enrolled at the default settings (e.g., those with a profile rather than frontal face image). We did not attempt to adjust the settings to enrol the failed cases; in previous deployments it has been observed that poor quality watchlist images were prone to increasing the incidence of false alerts.

Each Filler image corresponds to a custody record, and the Filler dataset contains multiple images for some individuals who have multiple custody records. Guidance on watchlist composition [7] suggests that if multiple different images of a subject are available, consideration be given to including these in the watchlist to improve the likelihood of a match. Thus, the inclusion of multiple images for some individuals is not atypical of the operational use case.

6 EVALUATION RESULTS – RETROSPECTIVE FACIAL RECOGNITION

6.1 METHODOLOGY

6.1.1 Comparison program

To automate the testing of RFR and OIFR, NEC provided a programme to NPL's specification that enabled batch processing of identification searches using the Neoface V4 algorithm without requiring operator involvement. The program uses the Neoface facial recognition server to perform an identification search of each probe image in a specified directory against a reference image database or watchlist, and logs the identifiers of the returned candidates, and the corresponding comparison scores.

6.1.2 Reference image database and Probe images

The reference image dataset for RFR combines (i) an enrolled watchlist of Custody-style Cohort images, and (ii) the enrolled Filler dataset of custody images.

[5] Age range for Filler dataset taken as the 0.1% - 99.9% percentiles

The probe images used for RFR were the full set of facial images of the Cohort subjects collected in conjunction with the LFR deployments. The image types are listed in Table 2.

To facilitate the use of the matching programme, the collected images were cropped and downsized. Cropping was necessary as the programme for batch processing cannot handle cases where multiple faces appear within an image. The Neoface system itself can handle this situation, but the process requires operator interaction in selecting the faces of interest. Some of the face images were cropped quite tightly to avoid other nearby faces in the background.

Given the size of the Filler dataset, there was a chance that a Cohort subject could also feature in the Filler dataset. This could risk a correct match between Cohort subject and their Filler image being counted as a false positive. (It would also complicate handling of any Cohort requests to have their data removed from the evaluation datasets.) The issue was addressed by checking cases where a Cohort-to-Filler comparison score was beyond the typical range for 'non-mated comparisons' and within the range for 'mated comparison'. In such cases if the Cohort and Filler metadata agree the Filler image would be removed from the Filler datasets. Six checks were made, and four Filler subjects were removed.

6.1.3 Running RFR

For RFR the matching process was configured to return for each probe the comparison score and candidate ID of the top 200 matches in the reference image database. (This would allow reporting of the RFR TPIR for the correct reference being returned with the top R matches for R = 1 to R = 200 as per the test strategy).

6.2 ACCURACY

For every probe image submitted for RFR, the correct reference was returned at Rank 1 (i.e., as the top match). This is the best possible performance[6].

$$TPIR_{RFR\ 178400,\ 1,\ 0} = 100\ \%.$$

It follows that TPIR is identical at 100 % for all demographic subsets of the submitted probe images and, with no demographic variation in TPIR, the performance is equitable.

It also follows that $TPIR_{178400,\ 1,\ 0} = 100\ \%$ for all the different types of probe image, and in particular the images taken with the OIFR device.

The result also suggests that in the operational use of RFR, the number of top-matching candidates returned for operator adjudication could be somewhat smaller than 200 (say 10 rather than 200).

6.3 LOW SCORING MATED COMPARISONS & HIGH-SCORING NON-MATED COMPARISONS

Excluding matches between identical twins (addressed in Section 6.4), in the full set of RFR identification searches the highest non-mated comparison score was 0.653, and there were ten non-mated comparison scores above 0.60 (arising from probes images of four Cohort subjects). The probes of one Cohort subject had matches above score 0.6 against five different Filler subjects. This particular Cohort subject is clearly an outlier, but we have not

[6] All probes are recognised at Rank 1, and therefore no breakdown is given by subset for different image types, different Cohort demographics, different rank values.

been able to isolate a specific reason behind this anomaly. (In biometric systems it is not uncommon that a few subjects are more prone to false-matches than others [8]).

Thirteen RFR identification searches (out of approximately 4000) returned mated comparison score below the 0.653 highest non-mated score. The probe images in these cases generally met configured face detection parameters but were less than ideal. Typical issues being reflections in glasses, shadows over the eyes, glasses' frames over the eyes. Despite these issues the subjects were still correctly recognised at Rank 1.

There is no fixed face-match threshold at which in all mated identification searches the correct reference is returned at Rank 1, and any non-mated identification search would return zero matches.

6.4 IDENTICAL TWINS

In the Cohort there was a pair of identical twins; identical twins are known to be a challenging case for facial recognition. In the running of RFR and OIFR each twin was always correctly recognised at Rank 1, and their sibling was returned as the Rank-2 candidate. The non-mated comparison scores between the identical twins were (unsurprisingly) higher than that usual for non-mated comparison scores, and within the range 0.7 to 0.8 of typical for mated comparison scores.

In running of LFR, both of the identical twins were enrolled onto the Cohort watchlist; they were always recognised correctly and there were no cases of one of the twins being incorrectly recognised as the other.

There was also a pair of fraternal twins in the Cohort. RFR and OIFR non-mated comparison scores between the fraternal twins were within the usual range for non-mated scores. Neither twin featured as a Rank-2 candidate for their sibling.

7 EVALUATION RESULTS – OPERATOR INITIATED FACIAL RECOGNITION

7.1 METHODOLOGY

The methodology for evaluating OIFR is very similar to that for RFR. The differences being that:

a) Only the 'OIFR-style' images taken with the Samsung XcoverPro mobile phone were used as Probe Images. The reference image database combines the enrolled Filler dataset, and the enrolled 'Custody-style' Cohort images as in the case for RFR.

b) The NEC matching program was configured to return just the top 6 matches.

7.2 ACCURACY

For every probe image submitted for OIFR the correct reference was returned at Rank 1 (i.e., as the top match). This is the best possible performance.

$TPIR_{OIFR,\ 178\ 400,\ 1,\ 0} = 100\ \%$.

It follows that the OIFR TPIR is identical at 100 % for all demographic groups and there is no demographic variation in TPIR and the OIFR TPIR performance is equitable.

7.3 LOW SCORING MATED COMPARISONS & HIGH-SCORING NON-MATED COMPARISONS

Excluding matches between identical twins, in the full set of OIFR identification searches the highest non-mated comparison score was 0.653. Five OIFR identification searches returned the correct reference image at Rank 1, though with a mated comparison score below this value (0.55-0.64). In four of the cases, features of the face and eyes were obscured by reflections in the subjects' glasses. In the fifth case due to the angle of the sun there were dark shadows in the eye-sockets. In the operational OIFR use-case the operator taking the photograph might seek to take a further image of the subject when such face-image quality issues arise.

The OIFR use case currently operates without a face-match threshold, but had a face-match threshold of 0.66 been set, then with our OIFR probe images, and with the Cohort enrolled, all but these 5 cases would have been recognised at Rank 1, and if the Cohort subject is not enrolled onto the reference image database, no candidates would have been returned for the subject (except in the case of the identical twins).

8 EVALUATION RESULTS – LIVE FACIAL RECOGNITION

8.1 METHODOLOGY

8.1.1 Collection of LFR video featuring Cohort

Cohort subjects were seeded into to the Crowd flow over the course of the operational LFR deployments. These deployments were running with an operational watchlist, which did not feature the Cohort.

At the street location of the LFR system, start and endpoints for the repeat walks by Cohort subjects through the zone of recognition (Recognition Opportunities) were selected such that a round trip through the zone of recognition back to the start point should take at least one minute. To prevent multiple alerts from a single recognition opportunity, the Neoface algorithm is configured such that once an individual is alerted, a second alert within 30 seconds is ignored.

To count and log timings of the Cohort recognition opportunities, each Cohort subject had been issued with a lanyard and badge showing their unique reference number (URN) as a number and a barcode. Scanning the barcodes logged the URN and time of scan to a spreadsheet.

At the LFR location Cohort subjects were briefed:

a) To ensure that they walked through the zone of recognition of the LFR systems.

b) To walk as they would normally; there was not a need to look directly at the camera, but not to be looking down at their mobile phone while walking towards the camera.[7]

c) To avoid bunching into one large group. It was suggested that they might walk in pairs and allowed to converse.

[7] While it is operationally realistic that some of the crowd intentionally or accidentally avoid showing their face to the LFR system, the focus of the evaluation was on the accuracy and equitability of the LFR for those faces processed by the system.

d) Cohort subjects were instructed to have their lanyard-badge barcode scanned for each walk through the zone of recognition.

We did not prohibit or suggest the wearing of caps, sunhats or sunglasses, glasses. The wearing of sunglasses and hats was typical for the non-cohort public passing the LFR system.

8.1.2 Enrolment of Cohort and Filler watchlists

In conjunction with the LFR deployment, face images of the types listed in Table 2 were collected from the Cohort. These included ‘custody image’ style photographs. For subjects that normally wear glasses, photographs were taken both with and without glasses. Enrolment of Cohort subjects onto a Cohort watchlist used the custody-style photographs without glasses. These photographs were taken at a high resolution (6000 x 4000 pixels, with approximately 700 pixels between the eyes) and were cropped and downsized (to approximately 200 pixels between eyes) to meet requirements for watchlist enrolment (max 300 pixels between eyes).

A Filler watchlist was enrolled of the full Filler dataset. (Filler images averaged approximately 130 pixels between eyes.)

8.1.3 Offline running of LFR

The LFR video footage was provided after the deployment as a series of 30 minute .mp4 files from each camera. This allowed the video to be run retrospectively on the Neoface V4 system replicating the live video stream but using the Cohort and Filler watchlists.

The collected LFR videos were run against Cohort and Filler watchlists at a range of face-match thresholds 0.56 to 0.64 spanning the Neoface default setting of 0.60.

In addition to the face-match threshold, the algorithm has configurable settings controlling face detection, (based on proprietary measures for face reliability, face quality score, face frontal score, minimum and maximum pixels between eye centres). Good quality frontal face images of sufficient size give the most accurate identification results, but the strictest settings may mean that some faces go undetected by the system. Relaxing the criteria for face detection increases the number of faces to be processed for comparison which may improve the true recognition rate but may also increase the false match rate. The increased computational workload may mean that fewer frames of video can be processed in real-time operation.

8.1.4 Face detection parameters

In this evaluation we first ran the LFR videos without adjusting face detection thresholds from the Neoface software default settings. Our results at these settings showed an anomaly in the recognition performance at the Cardiff deployment. The extreme heat on 13th August caused the LFR cameras to malfunction. In order to continue the deployment with minimum delay, the failing system was replaced with a second single camera system but without the normal processes of optimising camera and zone of recognition settings. Our initial results showed a significantly lower True Recognition Rate for the replacement system, and inspection of the video footage showed that face images were smaller and less ‘frontal’ than before the change. Accordingly, the affected LFR videos were re-run with face detection settings adjusted to allow for face images of a smaller size and with greater divergence from a frontal pose. The change improved accuracy to the levels prior to the camera malfunction.

Investigation into the effects of altering face detection settings was also carried out for processing the LFR video of the four London deployments. Here the potential values for ‘Face reliability score’ and ‘Face frontal score’ were limited by requirements for real-time operation on the servers used. The changes made led to an improvement in TPIR. However, we observed that some of the additional faces found are of poorer quality and more liable to falsely match, increasing the FPIR.

8.1.5 Determination of performance

For determination of the False Positive Identification Rate, LFR videos were run against the Filler watchlist emulating the Live Facial Recognition process but using the saved videos in place of the live video feed. The resulting ‘Match Details Report’ produced by the system lists all the Crowd and Cohort matches against the Filler watchlist at the threshold setting. The Match Details Report was inspected to determine which of the matches are of a Cohort subject matched to a Filler reference image. This determination was aided by:

a) the barcode scanned timings of Cohort recognition opportunities,

b) timing information derived from the Match Details Report,

c) images of the Cohort taken on the street,

d) the green lanyard with URN badge normally visible in the captured face image shown in the Match Details Report.

Matches between Crowd and Filler were disregarded in the evaluation, as no ‘ground truth’ is available to establish which of these matches are correct.

For determination of the True Positive Identification Rate, LFR videos were run against the combined Filler and Cohort watchlists. The recognitions logged in the Match Details Report were then compared against the log of barcode scans to determine how many of Cohort subject’s recognition opportunities resulted in their correct recognition, and how many times they were missed.

8.2 ACCURACY

Table 5 — TPIR and FPIR by threshold setting

Face-match threshold	Face detection settings	Observed TPIR	Observed FPIR	FPIR anticipated under operational measures: Watchlist 10k	FPIR anticipated under operational measures Watchlist: 1k
T		$TPIR_{178400, 1, T}$	$FPIR_{178000, T}$	$FPIR_{10000, T}$	$FPIR_{1000, T}$
0.64	(a)	79 %	0.00 %	< 0.004 %	< 0.001 %
0.62	(a)	82 %	0.05 %	< 0.004 %	< 0.001 %
0.60	(a)	85 %	0.25 %	0.014 %	0.002 %
0.60	(b)	89 %	0.30 %	0.017 %	0.002 %
0.58	(a)	88 %	0.48 %	0.027 %	0.003 %
0.56	(a)	94 %	1.15 %	0.065 %	0.007 %

Notes:

$TPIR_{178400, 1, T}$ based on 4000 mated recognition opportunities

$FPIR_{178000, T}$ based on 4000 non-mated recognition opportunities

$FPIR_{10000, T}$ & $FPIR_{1000, T}$ estimates based on scaling results on the 178k watchlist to 10k and 1k

Face detection settings (a)

London deployments: Face Reliability 0.80, Face Frontal Score: 0.4, Eyes Min Distance 60, Face Roll 30

Cardiff deployment: Face Reliability 0.80, Face Frontal Score: 0.3, Eyes Min Distance 40, Face Roll 30

Face detection settings (b)

London deployments: Face Reliability 0.75, Face Frontal Score: 0.35, Eyes Min Distance 60, Face Roll 30

Cardiff deployment: Face Reliability 0.80, Face Frontal Score: 0.3, Eyes Min Distance 40, Face Roll 30

Table 5 shows the observed True Positive and False Positive Identification Rates, aggregated over the video from all five LFR deployments, and at face-match thresholds ranging from 0.56 to 0.64. Over this threshold range the True Recognition Rate ranges from 79% to 94%.

At a face-match threshold of 0.64, the software produced no false positives against the watchlist of 178,000 Filler images. At threshold 0.62, two non-mated recognition opportunities (by the same individual) gave a false positive. At threshold 0.60, ten non-mated recognition opportunities generated false positives (increasing to twelve with the revised face detection settings).

The evaluation watchlist of 178,000 images is much larger than that used in operational deployments: all MPS LFR deployments to date have used watchlists smaller than 10000. With a watchlist of 10,000 rather than 178,000 face images the FPIR should reduce by a factor of 17.8. Thus, in an operational setting with a watchlist of 10,000 face images and a face-match threshold of 0.60, the anticipated False Alert Rate ($FPIR_{10000, 0.6}$) is 0.017 %[8] (approximately 1 in 6000), and at face-match threshold of 0.62 the anticipated False Alert Rate ($FPIR_{10000, 0.62}$) is 0.004 % (approximately 1 in 25000). This would equate to one or two false alerts per LFR deployment day were the crowd numbers similar to those of the deployments used in this study.

These accuracy levels are a considerable improvement on that reported for previous versions of the Neoface software (with watchlist size between 2000 and 4000, averaged over four deployments, TPIR ≈ 72 % and FPIR ≈ 0.1 % (1 in 1000). [5]).

[8] Sometimes it is easier to comprehend FPIR when expressed as a ratio e.g., “1 in 6000” than when expressed as a percentage, in this case “0.017%”. x% = 1 in (100 ÷ x)

8.3 DEMOGRAPHIC VARIATION IN FALSE POSITIVE IDENTIFICATION RATE

Table 6 — Number of Cohort subjects with false positive by Gender, Ethnicity & Age

Face-match threshold Face-detection settings	FPIR	Female	Male	Asian	Black	White	Age <21	Age 21-30	Age 31-42	Age >42
0.64 (a)	0.00 %	0	0	0	0	0	0	0	0	0
0.62 (a)	0.05 %	1	0	0	1	0	0	1	0	0
0.60 (a)	0.25 %	2	4	2	4	0	0	5	1	0
0.60 (b)	0.30 %	2	5	4	3	0	0	7	0	0
0.58 (a)	0.48 %	7	8	4	11	0	2	9	3	1
0.56 (a)	1.15 %	16	17	8	22	3	7	18	7	1
Recognition opportunities: gender, ethnicity & age balance		51%	49%	26%	29%	45%	26%	26%	24%	24%
Notes: Watchlist size:178,000 Recognition opportunities: 4000										

Table 6 shows the demographic of the Cohort subjects falsely matched to Filler images at the tested face-match thresholds.

At face-match thresholds of 0.64 and higher there were no false positives.

At face-match thresholds of 0.62 and 0.60 the number of subjects with a false positive is small, and a statistically significant imbalance between demographics is not shown.

At face-match thresholds of 0.58 and 0.56 we observe that false positives are not uniform between demographic groups. In particular, the number of false positives is disproportionately higher for Black subjects than for Asian or White subjects; this demographic variation in $FPIR_{178000,\ 0.56}$ is statistically significant ($p < 0.01$).

8.4 DEMOGRAPHIC VARIATION IN TRUE POSITIVE IDENTIFICATION RATE

Table 7 — True Positive Identification Rate by demographics and other factors

Demographic	TPIR	Statistically Significant	*p*-value
All Cohort subjects	89 %		
Gender			
Female	87%	no	
Male	90 %	no	
Ethnicity			
Asian	91 %	no	
Black	86 %	no	$p = 0.144$
White	89 %	no	
Ethnicity & Gender			
Asian – Female	91 %	no	
Asian – Male	90 %	no	
Black – Female	83 %	no	$p = 0.051$
Black – Male	90 %	no	
White – Female	89 %	no	
White – Male	88 %	no	
Age			
< 21 years	84 %	Yes	$p = 0.037$
21-30 years	87 %	no	
31-42 years	91 %	no	
> 42 years	93 %	Yes	$p = 0.009$
Height			
< 164 cm	86 %	no	$p = 0.157$
164-170 cm	87 %	no	
171-178 cm	90 %	no	
> 178 cm	91 %	no	$p = 0.116$

Other factors	TPIR	Statistically Significant	*p*-value
LFR Deployment			
Jul 07 – Oxford St	90 %	no	
Jul 14 – Oxford St	91 %	no	
Jul 16 – Oxford St	83 %	Yes	$p = 0.017$
Jul 28 – Piccadilly	89 %	no	
Aug 13 – Cardiff	94 %	Yes	$p = 0.003$
Crowd density			
People walking toward camera per minute			
< 71	91 %	Yes	$p = 0.005$
> 70	85 %	Yes	$p = 0.005$

Notes:

TPIR = TPIR $_{178400,\ 1,\ 0.60}$

Configuration parameter settings:
Face-match threshold: 0.60
Face Reliability: 0.75 (London) 0.8 (Cardiff)
Face Frontal Score: 0.35 (London) 0.3 (Cardiff)
Eyes Min Distance: 60 (London) 40 (Cardiff)
Face Roll: 30

Table 7 shows the observed variation in the True Recognition Rate for different demographic subgroups and other factors of interest at a face-match threshold of 0.6. A *t*-test (Welch's unequal variance *t*-test) was used to determine whether demographic differences in performance are statistically significant at the 0.05 significance level. Table 7 also shows computed *p*-values for cases where the difference (higher or lower TPIR) is statistically significant, cases close to the 0.05 significance threshold, and outliers in each demographic or environmental category.

8.4.1 Ethnicity and Gender

At a face-match threshold of 0.6, the ethnicity-gender group with the best TPIR was the Asian-Female group, and the poorest TPIR was for the Black-Female group. However, the observed differences in TPIR by gender, by ethnicity, and by ethnicity & gender combined were not statistically significant at the 0.05 significance level.

- The demographic variation in TPIR between Black females ($M = 0.83$, $SD = 0.26$) and the other ethnicity-gender demographics ($M = 0.90$; $SD = 0.20$) was not significant ($t(71.7) = 1.98$; $p = .051$)

8.4.2 Age

The observed TPIR was significantly lower for those aged 20 or below (M = 0.84; SD = 0.25), than for those aged 21 and over (M = 0.90; SD = 0.19), (t(143.1) = 2.11; p =.037). However, this age effect on performance is confounded with the effect of congestion in the Zone of Recognition. Most of the subjects in the under-20 age category were Police Cadet volunteers attending the deployment on 16 July. This was also the deployment in which the zone of recognition was the most congested. We discuss this further in Section 8.4.3.

The observed TPIR for those aged over 42 (M = 0.93; SD = 0.17) was significantly higher than for those aged between 21 and 42 (M = 0.89; SD = 0.2), (t(231.9) = 2.11; p = 0.036).

8.4.3 Crowdedness of zone of recognition, and subject height

The observed TPIR when crowd density was above 70 (M = 0.85, SD = 0.25) was significantly higher than that when crowd density was below 70 (M = 0.91, SD = 0.15), (t(315.0) = 2.86, p = 0.005).

It also appears that subject height is an important factor when the zone of recognition is crowded. The supposition is that shorter subjects are more likely than taller subjects to be occluded or partly occluded in the camera field of view when the zone of recognition is crowded.

On the most crowded day 16 July, the observed TPIR for subjects shorter than 170 cm (median height) (M = 0.77, SD = 0.31) was significantly lower than that for subjects taller than 170 cm (M = 0.91, SD = 0.17), (t(86.5) = 2.69, p = 0.009). The lower performance of the under 20's 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.

9 DISCUSSION

9.1 ARE DEMOGRAPHIC PERFORMANCE VARIATIONS SIMILAR FOR LFR, RFR AND OIFR?

The same face matching algorithm is used for LFR, RFR, and OIFR and in the evaluation using the same Cohort individuals and same reference image sets, the findings on demographic performance variation are very similar. Due to the similarity, it may be possible to use the bulk RFR software and a set of Cohort data to assess the FPIR equitability of a specific watchlist prior to LFR deployment.

Table 8 — Demographic of subjects with high non-mated comparison scores in RFR testing

Non-mated comparison score	Female	Male	Asian	Black	White	Age <21	Age 21-30	Age 31-42	Age >42
score ≥ 0.66	0	0	0	0	0	0	0	0	0
score ≥ 0.64	1	0	0	1	0	0	1	0	0
score ≥ 0.62	1	1	1	1	0	0	2	0	0
score ≥ 0.60	1	3	2	2	0	1	3	0	0
score ≥ 0.58	5	4	3	5	1	4	4	0	1
score ≥ 0.56	11	6	4	11	2	5	9	2	1
Identification searches: gender, ethnicity & age balance	51%	49%	28%	26%	46%	21%	28%	26%	25%
Notes: Watchlist size:178,000; Non-mated identification searches: 3,943									

Table 8 shows the demographics for high-scoring non-mated comparison scores arising from RFR testing, and we note that the demographic distribution is very similar to Table 6 showing the demographic distribution of LFR false positives. The majority of non-mated comparison scores at threshold 0.56 arose from probe images and recognition opportunities of subjects of Black ethnicity.

9.2 TPIR VARIATION IN PERFORMANCE & EQUITABILITY

In Section 8.4 demographic variation in performance was examined at the default face-match threshold of 0.6. At different thresholds the TPIR values can change for each demographic subgroup, and the statistically significant cases can vary.

9.3 FPIR EQUITABILITY

Equitability is dependent on the face-match threshold settings and on the size and demographic composition of the LFR watchlist or RFR reference database.

The demographic variation in the non-mated score distribution does not affect equitability if settings are such that the chance of a false alert is very low. However, if settings allow for a higher number of false alerts, these are likely to occur disproportionately within Black and Asian ethnicities.

9.4 EFFECT OF REFERENCE DATASET/WATCHLIST SIZE AND COMPOSITION ON FPIR

The size and composition of the watchlist tested in the study is useful for revealing demographic variation in performance of the algorithm, but equitability should be judged on a watchlist more typical of an operational deployment.

The candidates lists and comparison scores output by the batch RFR facial comparison program provide data to enable an estimation of FPIR of the Cohort probe dataset against subsets of the reference database of a specified size and demographic profile.

In identification systems configured to produce multiple candidates (such as RFR) Selectivity is the average number of candidates returned in a non-mated identification transaction where the comparison score exceeds the face-match threshold. Selectivity and FPIR differ at low thresholds but converge at high thresholds as false positives become rarer.

We can calculate Selectivity separately for each Gender-Ethnicity component of the Filler dataset:

$$\begin{aligned} SEL_{FILLER}(T) = {} & SEL_{Asian\_F}(T) + SEL_{Black\_F}(T) + SEL_{White\_F}(T) + \\ & SEL_{Asian\_M}(T) + SEL_{Black\_M}(T) + SEL_{White\_M}(T) \end{aligned}$$

and then scale each component in accordance with the specified size and demographic profile of the operational reference database, e. g., the scaling factor for $SEL_{Asian\_F}(T)$ is:

$$\frac{\text{number of Asian Female reference images for operational watchlist}}{\text{number of Asian Female reference images in the Filler dataset}}$$

Summing the scaled components then calculates an average of number of candidates returned based on the specified demographic profile.

Applying this process to the RFR results for identification searches of Cohort Probes (of each demographic in turn) against the entire Filler dataset provides an estimate of the FPIR by demographic for the notional MPS watchlist (in Table 9) and SWP watchlist (in Table 10).

Note that when the observed selectivity is zero, or very low, the anticipated FPIR is given as '< 1 in 10,000' or '< 1 in 100,000': there are insufficient data points to support lower estimates.

Table 9 – Selectivity and estimated FPIR by probe demographic for notional MPS Watchlist

	Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Face match threshold: 0.66	0 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000
Face match threshold: 0.64	0 < 1 in 10,000	0 < 1 in 10,000	0.00003 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000	0 < 1 in 10,000
Face match threshold: 0.62	0 <1 in 10,000	0.00011 1 in 9,200	0.00003 < 1 in 10,000	0 <1 in 10,000	0 <1 in 10,000	0 <1 in 10,000
Face match threshold: 0.60	0 <1 in 10,000	0.00022 1 in 4,600	0.00021 1 in 4,700	0.00023 1 in 4300	0 <1 in 10,000	0 <1 in 10,000
Face match threshold: 0.58	0.00001 <1 in 10,000	0.00033 1 in 3,000	0.00048 1 in 2,000	0.00046 1 in 2,100	0 <1 in 10,000	0.00017 1 in 5,800
Face match threshold: 0.56	0.00008 <1 in 10,000	0.00098 1 in 1,000	0.00109 1 in 920	0.00093 1 in 1,000	0.00004 <1 in 10,000	0.00017 1 in 5,800

Notes:
- Notional MPS Watchlist: watchlist size 10,000 with demographic profile based on MPS arrest data
- In the table selectivity is shown as decimal, and estimated FPIR shown as ratio

In Table 9 we see that for the size 10,000 notional MPS watchlist at thresholds above 0.6 the extent of demographic variation in FPIR is quite limited. However, at threshold 0.56 the demographic variation observed for the full Filler dataset persists and, depending on numbers and demographic composition of the crowd, could show a noticeable effect on outcomes.

Table 10 – Selectivity and estimated FPIR by probe demographic for notional SWP Watchlist

	Asian Female	Asian Male	Black Female	Black Male	White Female	White Male
Face match threshold: 0.66	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000
Face match threshold: 0.64	0 <1 in 100,000	0 <1 in 100,000	0.000000 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000
Face match threshold: 0.62	0 <1 in 100,000	0.000002 <1 in 100,000	0.000000 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000
Face match threshold: 0.60	0 <1 in 100,000	0.000003 <1 in 100,000	0.000002 <1 in 100,000	0.000003 <1 in 100,000	0 <1 in 100,000	0 <1 in 100,000
Face match threshold: 0.58	0.000000 <1 in 100,000	0.000005 <1 in 100,000	0.000004 <1 in 100,000	0.000006 <1 in 100,000	0 <1 in 100,000	0.000031 1 in 32,000
Face match threshold: 0.56	0.000012 1 in 86,000	0.000015 1 in 67000	0.000008 <1 in 100,000	0.000012 1 in 84,000	0.000007 <1 in 100,000	0.000031 1 in 32,000

Notes:
- Notional SWP Watchlist: watchlist size 1,000 with demographic profile based on SWP arrest data
- In table selectivity is shown as decimal, and estimated FPIR shown as ratio

Table 10 shows the demographic variation in performance changes for the demographic profile and size of the notional SWP watchlist. Moreover, even at face-match threshold 0.56, the FPIR is now below 1 in 30,000 for each ethnicity-gender, and the demographic variation is likely to be inconsequential unless the number of crowd subjects in the deployment is somewhat in excess of 32,000.

9.5 RETROSPECTIVE FACIAL RECOGNITION PERFORMANCE ON CHALLENGING FACIAL IMAGES

The perfect results from the testing of Retrospective Facial Recognition are very promising. It should be noted that all the face images were taken by test staff, or Cohort in the case of

selfies, and when a facial image taken considered unsatisfactory by the photographer, e.g., out of focus, motion blur, subject eyes shut, generally a second image would be taken. For evaluation of demographic equitability this was appropriate for the images need to be consistent across demographics. It should be acknowledged that using images of lower quality, or lower resolutions, may not achieve the same level of performance.

9.6 RECOGNITION AGAINST SAME-DAY IMAGE

Note that, in the evaluation, images were collected from Cohort subjects over one or two days. This is typical for such evaluations due to limitations on project duration, and the difficulty of obtaining a stable corpus of test subjects who can participate over an extended time-period. Other factors being equal, TPIR rates for facial recognition against a recent photograph are likely to be better than TPIR against historic photographs. In terms of demographic differentials in performance this means that the study has not been able to address the effects of ‘template ageing’.

9.7 SUGGESTIONS FOR FURTHER INVESTIGATIONS

In the report we have made comment some issues where further information may be able to extend knowledge on how facial recognition systems perform in more challenging cases:

- effects of poor-quality video and video compression on Live Facial Recognition performance
- extending testing of Retrospective Facial Recognition to facial images of lower quality or resolution
- consideration of the effects of template ageing on performance and any demographic effects.

Policing may wish to consider how these issues relate to their use cases for RFR and LFR, and the nature of images they plan to use. Further controls or testing may be appropriate for facial images of lower quality or resolution. Some testing may be able to reuse the facial images and video collected for this study, downgrading the images in a controlled manner. The Filler dataset contained some cases where there were two or more time-separated facial images of data subjects. Such data might allow for assessment of demographic effects in template ageing.

9.8 TEST CORPUS ARISING FROM THE STUDY

The facial images and video footage collected in the study are to be provided to the Metropolitan Police Service along with associated ground truth metadata to enable reuse of the data for future testing of facial recognition systems.

10 TERMINOLOGY AND ABBREVIATIONS

Candidate: Image of a person from the watchlist or reference database returned as result of an identification search.

Cohort: Subjects recruited to provide a corpus of facial images and video for recognition in the evaluation.

Comparison score: Numerical value of the similarity between compared probe and reference facial images.

Crowd: Members of the public passing through the zone of recognition of the LFR system.

Equitable: Equitability of an operational deployment requires that differences, where any exist, in the outcomes for the subjects from different demographics should be inconsequential**.**

Face-match threshold: The comparison score value above which the compared images will be considered to match.

FPIR: False Positive Identification Rate (for LFR) is the proportion of recognition opportunities of subjects who are not on the watchlist which return a (false positive) match against a candidate on the watchlist.

$$\mathrm{FPIR}(N, T) = \frac{\text{Num. non-mated recognition opportunities that return a match against a candidate on the watchlist}}{\text{Num. non-mated recognition opportunities}}$$

where N represents the number of images on the watchlist, and T the face-match threshold.

Filler dataset: Dataset drawn from MPS holdings of custody images and used to supplement Cohort reference images to provide large reference dataset for the evaluation.

LFR: Live Facial Recognition

Mated: A mated identification search is one in which the subject in probe image also has a reference image in the reference database. A mated recognition opportunity is one where the subject walking though the LFR zone of recognition has an image in the LFR watchlist. Similarly, a mated comparison score is produced from comparisons of two face images of the same individual.

Non-mated: A non-mated identification search is one in which the subject in probe image does not have a reference image in the reference database. A non-mated recognition opportunity is one where the subject walking though the LFR zone of recognition does not have a facial image in the LFR watchlist. Similarly, a non-mated comparison score is produced from comparison of face images of different individuals.

MPS: Metropolitan Police Service

NEC: NEC Software Solutions – the company providing the Neoface facial recognition technology evaluated in this report

OIFR: Operator Initiated Facial Recognition

Probe image: A facial image that is searched against a watchlist or reference database.

Rank: The rank of a candidate facial image is its position in the set of candidates returned by an identification search listed in decreasing order of similarity to the probe (i.e., the Rank-1 candidate is the best matching candidate)**.**

Reference image: A facial image in the watchlist or reference database.

Recognition opportunity: The period when a subject moves through the zone of recognition of an LFR system with their face visible to the LFR camera.

RFR: Retrospective Facial Recognition

SEL: Selectivity (for RFR) the average number of candidates returned in a non-mated identification transaction for which the candidate comparison score exceeds the face-match threshold.

$$\mathrm{SEL}(N, T) = \frac{\text{Total Num. candidates returned with score above threshold } T \text{ in the set of non-mated identification transactions}}{\text{Num. non-mated identification transactions}}$$

where N represents the number of images on the watchlist, and T the face-match threshold.

SWP: South Wales Police

TPIR: True Positive Identification Rate (for LFR): the proportion of mated recognition opportunities that are correctly identified.

$$\mathrm{TPIR}(N, 1, T) = \frac{\text{Num. mated recognition opportunities correctly identified}}{\text{Num. mated recognition opportunities}}$$

where N represents the number of images on the watchlist, and T the face-match threshold (the '1' denotes that LFR only considers the top match.) In policing this is often referred to as the True Recognition Rate.

TPIR: True Positive Identification Rate (for RFR & OIFR) is the proportion of mated identification searches that include the mated reference among the candidates returned.

$$\mathrm{TPIR}(N, R, T) = \frac{\text{Num. mated identification searches where the mated reference is among the candidates returned}}{\text{Num. mated recognition opportunities}}$$

where N represents the number of images on the watchlist, R the number of best matching candidates returned and T the face-match threshold (T=0 if no threshold is applied).

Watchlist: A set of reference images (of individuals of interest to policing) against which a probe image is searched.

Zone of recognition: Three-dimensional space within the field of view of the Live Facial Recognition camera and in which the imaging conditions for robust facial recognition are met.

BIBLIOGRAPHY

[1] 'Facial Recognition Technology in Law Enforcement Equitability Study - Test Strategy', 2022

[2] ISO/IEC 19795-1:2021 Biometric testing and reporting - Part 1: Principles and framework

[3] ISO/IEC 19795-2: 2007 Biometric testing and reporting - Part 2: Testing methogolgies for technology and scenario evaluation

[4] NPL & MPS, 'Metropolitan Police Service Live Facial Recognition Trials 2016-2019', February 2020

[5] Office for National Statistics, 'Ethnic group, England and Wales: Census 2021'

[6] Gov.uk, 'Arrest Data March 2018 to March 2021'

[7] NPIA, 'Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images', National Policing Improvement Agency, 2007

[8] ISO/IEC 30137-1:2019 Use of biometrics in video surveillance systems - Part 1: System design and specification

[9] N. Yager and T. Dunstone, 'The Biometric Menagerie', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 220-230, 2010

ACKNOWLEDGEMENTS

The work was funded by the Home Office with Science, Technology, Analysis & Research (STAR) funding through the Metropolitan Police Service. (MPS)

MPS support to the project included:

- Provision of server and Neoface Facial Recognition software licence,
- Provision of camera and mobile devices for photographing various types of facial images including taking custody style images consistent with those in operational policing,
- Providing a demographically balanced ‘filler dataset’ of custody images to serve as a large watchlist for the evaluation,
- Volunteer police cadets to provide test subjects between ages 12 and 18,
- Running of operational LFR deployments to collect video footage of our test Cohort in operational conditions,
- Obtaining a venue in the vicinity of the LFR deployment for NPL to brief and photograph the test Cohort.

South Wales Police support to the project included:

- Running of an operational LFR deployments to collect video footage of our test Cohort in operational conditions,
- Providing a venue in the vicinity of the LFR deployment for NPL to brief and photograph the test Cohort.

Ingenium Biometrics assisted NPL in the taking still photographs and video footage of the test Cohort alongside the Policing LFR deployments.

Ingenium Biometrics and The School of Engineering, University of Kent assisted in providing technical review of the evaluation, analyses, results, and report.