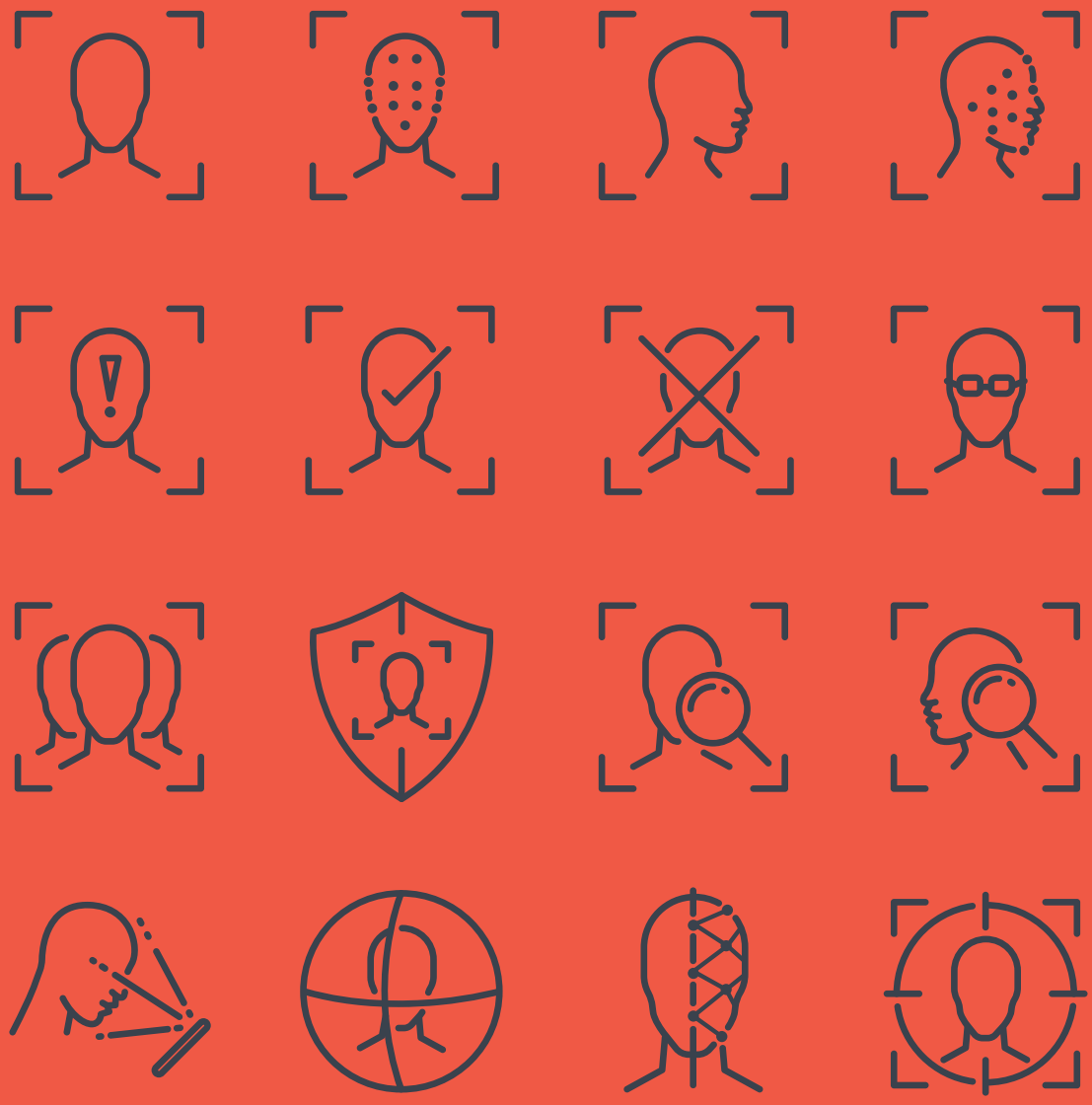
September 2018

AN EVALUATION OF SOUTH WALES POLICE'S USE OF AUTOMATED FACIAL RECOGNITION



Universities' Police Science Institute
Crime and Security Research Institute
Cardiff University

REPORT BY:

Bethan Davies, Martin Innes and Andrew Dawson

Universities' Police Science Institute
Crime & Security Research Institute
Cardiff University
Level 2, Friary House,
Greyfriars Road
Cardiff
CF10 3AE

TABLE OF CONTENTS

Glossary of Key Terms

AFR = Automated Facial Recognition. This technology works by analysing key facial features to generate a mathematical representation of them, and then comparing these against the mathematical representation of known faces in a database, to determine possible matches. This is based on digital images (still or from live camera feeds). In a policing context, AFR is used to help verify the identities of 'persons of interest' to police.

AFR Identify = 'slow-time' application of AFR technology, comparing still images of unknown suspects and persons of interest related to past crimes or incidents, against a custody database of circa 450,000 people, and returning up to 200 results. These results are ranked based on the algorithm generating a 'similarity score'. Possible relevant images are then reviewed by the operator, who decides if a possible 'match' has been generated. If so, information is returned to the investigating officer classified as intelligence.

AFR Locate = 'real-time' deployment of AFR technology, which compares live camera feeds of faces against a predetermined 'watchlist' in order to locate persons of interest. This generates possible matches that are reviewed by the operator(s).

AFR system = refers to the technology that is provided by NEC and operationalised collectively with other networked components (e.g. cameras / laptops).

AFR 1 – 4 = codes used by South Wales Police's AFR Identify team to distinguish between results.

AFR1 = indicates that informed by the AFR system the operator identified match(es) they believe to be accurate and have returned these details to the investigating officer for further investigation. In the report this is referred to as "match generated by the system and confirmed by operator."

AFR2 = indicates that although the 'probe image' (see below) was successfully run through the AFR system and results were returned, the operator did not identify any matches. Referred to in the report as "match generated by the system, but rejected by operator."

AFR3 = indicates that the probe image was not of sufficient quality to be analysed by the AFR system, but was of sufficient quality to be circulated to officers. Referred to in the report as "image rejected by the system (poor quality), but circulated to officers."

AFR4 = indicates that the probe image was not of sufficient quality to be analysed by the AFR system, or to be circulated to officers. Referred to in report as "image both rejected by the system and of too poor quality to be circulated to officers."

Alert = a notification triggered by the AFR Locate system when a possible match has been identified. This alert is displayed to operators.

Amber (also yellow) watchlist = watchlist containing images of individuals identified by police as suspects in ongoing investigations.

Blue watchlist = watchlist containing images of police personnel (officers and staff), designed to provide a baseline test of the system's effectiveness.

CSV files = details of the possible matches made by the AFR Locate system are recorded in CSV files (spreadsheets), which can be exported and subject to further analysis.

Enrolment = adding / uploading an image onto the AFR system, usually via a 'watchlist'.

Faces per frame = manually configurable setting. The number of faces that can be analysed by the AFR system per image frame.

False positive = a possible match suggested by the AFR system that is assessed incorrect by the human operators.

Fixed camera = operators cannot adjust these cameras as their position is fixed. They were used in locations, including the train station and Cardiff Airport, during the Champions League deployment.

Flow = system designed by NEC to count the number of faces detected and analysed (also referred to as 'transactions') by the system during live AFR Locate deployments.

Frames per second = manually configurable setting. The number of frames that can be analysed by the AFR system per second.

Frequent hitter = see Lamb.

Green watchlist = watchlist containing images of 'friendlies' during the Champions League deployment. These individuals were 'known' to police, but posed no immediate threat to safety, and so knowledge of their presence was for intelligence purposes only.

Hit = see Match.

Intervention Team = usually comprises two officers who are allocated to a single AFR Locate deployment location (a van). Operators can send intervention teams to stop individuals where they deem a match generated by the AFR system to be a true positive.

Lamb = someone with a particular face type that generates a lot of false matches. Also referred to as a 'frequent hitter'.

M20 = the second AFR algorithm used by South Wales Police, issued by NEC. In use from October 2017 onwards.

Match = a possible match between two images generated by the AFR system. Sometimes referred to as a 'hit' by operators.

Operator = South Wales Police officer or staff member tasked with operating the AFR Locate system.

Operator logs (logbooks) = used by operators to record details of matches from the AFR system and their decision on whether the matches are correct (true positives) or not (false positives).

Person(s) of interest = this term includes people wanted on warrants and suspects for particular unsolved crimes. It can also potentially include individuals where police professional judgement suggests that they pose a risk of committing a crime, missing persons and vulnerable people (e.g. dementia sufferers).

Pixels between the eyes = manually configurable setting. Relates to the resolution of images (higher resolution images contain more pixels). This setting means that the AFR system will only recognise a face when the set number of pixels are present between the eyes.

Probe image = the still image of an unknown person, which is enrolled onto the AFR Identify system.

Processor = carries out the basic instructions from programs. Some processors have more power than others (i.e. laptop processors are typically less powerful because of the need for them to be small and conserve energy). Where programs are designed to use multiple processors, this is termed 'multi-threading'.

PTZ = 'Pan, Tilt, Zoom' cameras used for AFR Locate deployments. The position and zoom of these cameras is adjustable, using a gaming controller connected via a laptop.

Red watchlist = watchlist containing images of 'high priority' individuals wanted by police on warrant.

S17 = the first AFR algorithm used by South Wales Police, issued by NEC. In use between May and October 2017.

Similarity scoring = manually configurable setting. When conducting comparisons between database images and camera images, the AFR system scores (possible) matches on their similarity. This score falls between 0 and 1 and is based on the likeness of two images – the higher the score, the greater the likeness. For AFR Locate, a minimum similarity score can be set (e.g. 0.59), which means operators will only be alerted to matches scoring above that threshold.

True positive = a possible match generated by the AFR system that is judged correct by the operators.

Watchlist = defined set of facial images made up of persons of interest. Used in AFR Locate deployments, usually containing 600-800 images.

White boxing = visual indication that the AFR system is detecting faces during Locate deployment. If the system detects 'a face' in the CCTV feed, it projects a white box around it.

EXECUTIVE SUMMARY

This report details findings from an evaluation of South Wales Police's deployment of Automated Facial Recognition (AFR) between May 2017 and March 2018, conducted by Cardiff University. It is possibly the first, large-scale, independent academic evaluation of the application of AFR by police conducted in a naturalistic environment. The evaluation report sets out evidence and insights relating to three principal areas:

- The policing outcomes that are attributable to AFR, either wholly or in part;
- Lessons learned about the process of implementing this innovative technology into policing processes and systems, and what is needed to optimise its effectiveness and efficiency;
- Some of the ethical, legal and public permission issues that attend the deployment of such approaches.

Using a multi-method research design, the evaluation examined how the AFR system was used to identify suspects and other persons of interest to the police across a range of contexts and situation, in its two principal modes:

- AFR Locate – involves 'real-time' monitoring of a location with the algorithm checking faces passing through the field of vision of a camera against a defined watchlist of images. Watchlists typically contain between 600-800 individuals (the total number was larger at the Champions League, 1200, due to the scale of the event). When a possible match with an image is made this is reviewed by the operator. If they confirm the possible match then a policing intervention may be undertaken.
- AFR Identify – involves scanning images of unidentified suspects taken from past crime scenes, against a database of circa 450,000 custody images held by police. The system returns a ranked list of possible matched images to the operator, who will make a judgement about whether any of these should be made available as intelligence to the investigating officer.

The research evidence collected suggests that in framing an understanding of how and why AFR technology can be applied in support of policing tasks to generate particular outcomes, it is important to conceptualise it as a socio-technical system. It is not a 'plug and play' technological solution, but rather an application that requires adoption and adaptation if it is to be usefully integrated within police routines. Reflecting this, the evaluation is structured around three key themes that are understood to be interacting and recursively positioned in relation to each other in terms of generating useful insights and evidence:

- Organisational performance – pivots around how organisational processes and system are both shaping of and shaped by the ways the technology is implemented;
- System performance – concerns how the hardware and software that together comprise the system actually function across different situations and contexts;
- Operator performance – is focused upon how the behaviours and decisions made by the human users influence the contribution that the system makes.

Attending to how these factors collectively shape the policing outcomes delivered leads us to conclude that in a policing context the technology should be reconceptualised from 'Automated Facial Recognition' to 'Assisted Facial Recognition'. This is on the grounds that this more accurately describes how police were using it. The label of 'automated' suggests the algorithm is autonomously deciding whether an image is matched or not. This is not what was happening. In both Locate and Identify modes the system was generating suggestions about possible matches to the human operators, who were responsible for confirming or refuting these. So whilst the label of 'Automated Facial Recognition' is a descriptor of its application in other contexts (for example border security and passport control) this is not how it was observed to be used by police.

In contrast to other research into facial recognition technologies, which have traditionally been controlled in laboratory environments or under fairly 'controlled' experimental conditions, this evaluation has been based on real-world, real-time deployments in 'uncontrolled' settings. Positioned in this way, questions such as 'does the technology work?' and 'was it a success?' are rather simplistic given the complexities of integrating and operationalising an innovative technology such as AFR into police organisational routines and systems. The 'headline finding' of this research is that AFR technologies can certainly assist police to identify suspects and persons of interests, to both solve past crimes and prevent future harms. However, it is not a simple or straight forward undertaking and requires considerable organisational investment and effort. Framed in this way the tables below summarise the outcomes achieved by South Wales Police during the evaluation period.

Summary Outcomes

Evaluation period: June 2017-March 2018

LOCATE

DEPLOYMENTS (EVENTS FOR WHICH WE HAVE COMPLETE DATA)	
	TOTAL
Number of deployments	11
Number of S17 deployments	3
Number of M20 deployments	9*
Time deployed (minimum)	Approx. 55 hours

Note: blue matches were removed from the analysis of events following Champions League, because recording became flawed as the implementation continued, and the results therefore skew the data.

*Includes M20 laptop deployed at Joshua
**Match generated by the system is confirmed true or false based on operator decisions recorded in logbooks.

OUTCOMES FROM S17 'OLD' ALGORITHM		
	Total	**% of total**
Matches	2710	100%
Confirmed** true positives	94	3%
Confirmed** false positives	1962	72%
Unknown	654	24%
Arrests	4	<1%

OUTCOMES FROM M20 'NEW' ALGORITHM		
	Total	**% of total**
Matches	146	100%
Confirmed** true positives	38	26%
Confirmed** false positives	72	50%
Unknown	36	25%
Arrests	14	10%

IDENTIFY

REQUESTS & CHARGES	
	TOTAL
Number of requests	2136
Number of requests on S17	612
Number of requests on M20	1524
Number of charges (April 2018)	100+

OUTCOMES FROM S17 'OLD' ALGORITHM		
	Total	**% of total**
Matches generated by the system and confirmed by operators (AFR1)	107	17%
Matches generated by the system but disconfirmed by operators (AFR2)	95	16%
Rejected by the system (poor quality) but circulated to officers (AFR 3)	337	55%
Rejected by the system and too poor quality to be circulated to officers (AFR 4)	73	12%

OUTCOMES FROM M20 'NEW' ALGORITHM		
	Total	**% of total**
Matches generated by the system and confirmed by operators (AFR1)	333	22%
Matches generated by the system but disconfirmed by operators (AFR2)	136	9%
Rejected by the system (poor quality) but circulated to officers (AFR 3)	916	60%
Rejected by the system and too poor quality to be circulated to officers (AFR 4)	128	8%
% of AFR 1 results in rank 1	60%	
% of AFR 1 results in top 10	90%	

Key findings:

- Overall, the proportion of true positives increased from 3% at the initial Champions League deployment to 46% at the Six Nations deployments. This was due to both the new algorithm and increased operator familiarity with the system settings.
- Across events (including all data on 'blue' police matches) there were: a total of 2,900 possible matches generated by the AFR system; a total of 144 confirmed true positives by operators; and a total of 2,755 categorised as 'false positives'.
- Data suggest that there is a relationship between similarity scores and operator decisions - as scores increase, operators are more likely to deem a match a true positive. The strength of this relationship seems to have increased with improvements to the system.
- Of the images that were good enough for AFR Identify analysis, 73% generated possible suspect matches as reviewed by the operators when using the new M20 algorithm. This was an increase of over 20% compared to the old S17 algorithm.
- Figures from February onwards show that 90% of AFR Identify 'possible matches' appear in the first ten (ranked) results.
- Across all deployments (excluding the Campions League), 107 females and 17 individuals from a black or minority ethnic background were matched to persons of interest by the AFR Locate system.
- The AFR Locate system was observed struggling with large crowds as it froze, lagged, and crashed when frames were full of people. This was disappointing, as deployment requirements had not changed since NEC were awarded the contract.
- The current camera equipment does not operate well in low light: as daylight fades, the cameras compensate by increasing ISO levels, this then increases the 'noise', and in turn impacts the ability of the system to detect and analyse faces.
- A small-scale field trial designed to test the technology indicated the system has the potential to achieve a 76%+ level of facial recognition for enrolled individuals (this is a very rudimentary indication of accuracy, due to the limited scope of the trial). Some forms of clothing and accessories, which obscured parts of the face when worn, were found to impact accuracy.
- Operator views of the technology were optimistic on the whole, and confidence in the system increased when the new algorithm was implemented.

This research also highlights that multiple organisational reforms and innovations were required to actually making this technology 'work' in a practical policing context:

- Good quality images are key. South Wales Police quickly realised this and acted to improve how custody images are taken force-wide.
- The introduction of the new M20 algorithm was also an integral part of making the system 'work'. The improvements between the two algorithms were quite remarkable: the accuracy of matches was much higher, and the system responded quicker and was more stable.
- Hardware configuration is an important element of overall performance system, as the AFR system can be quite demanding in terms of its processing power requirements.

Subsequent sections of this report aim to elaborate and explain aspects of these summary findings at a detailed level, covering both aspects of the technology and the human behaviours and decision-making processes that accompany them. This is necessary to capture some of the complexities involved in rendering AFR an operationally viable tool for policing. It is also intended that these details should engage in an element of 'myth-busting' owing to how some misinformation and disinformation has emerged in public and policy discussions of what AFR can and cannot do. In part, these myths are an artefact of some highly stylized but influential fictional portrayals of facial recognition technologies in film and television. Set against this backdrop, the intent underpinning this report is to provide an independent, balanced and evidence-led assessment of AFR's application in policing. This is vital in terms of clarifying both what benefits might be accrued if this technology were to be adopted more widely and routinely, but also the levels of investment and effort required to deliver any such return. Such an account is also important in terms of helping to clarify, based upon evidence, where there are areas for legitimate public concern, and hence where regulatory effort and ethical interests need to be directed.

SECTION 1: **INTRODUCTION**

This document reports key findings from an evaluation conducted by the Universities' Police Science Institute at Cardiff University, of South Wales Police's operational deployment of Automated Facial Recognition (AFR) technology. The evaluation is intended to generate systematic and detailed evidence and insight about:

- The outcomes achieved as a result of deploying this new technology for the identification of persons of interest to the police;
- The issues and challenges that were encountered in implementing it, and;
- How these were managed in terms of organisational reforms and innovations.

In describing how processes of operational implementation translated into the delivery of policing and project outcomes, the multi-method evaluation strategy was designed to provide a 'holistic' and multi-dimensional understanding. This is on the grounds that previous evaluations of AFR technologies have tended to be based in relatively 'controlled' environments, and / or utilised relatively 'narrow' performance indicators focused upon aspects of the technology, rather than how and why it is (or is not) able to make defined contributions to police work.

This evaluation has therefore adopted a 'realistic' approach, taking into consideration the complex and multifaceted nature of implementing a technology such as AFR in a policing context. One of the strengths of this approach is that it allows a range of influences and factors to be examined,[1] including, for example: the social context of the 'thing' being implemented; the factors working to facilitate or impede its implementation; and the human agents involved. Informed by the empirical evidence collected as part of the evaluation effort, it is argued that the overall contribution made by AFR to policing is the product of interactions between several core inter-dependencies: the affordances of the technology in terms of its hardware and software; the role played by human operator interpretations, judgements and decision-making; and, organisational policies and procedures. To understand how and why AFR delivers certain policing and project outcomes, it is necessary to understand each of these domains and the mutual and recursive impacts that they have upon each other.

Developing this theme of complexity and nuance, in interpreting the empirical evidence, we avoid reaching any reductive and simplistic conclusion as to whether the technology 'works' or does not (as Pawson and Tilley (1997) argue is often the case in experimental evaluative approaches). The evidence collected via this evaluation suggests that:

- Integrating AFR technologies into policing enables the more efficient identification of some suspects, and it also provides for the identification of individuals who probably would not have become known to the police without the technology being available;
- Equally however, the empirical data document how considerable investment and effort was required to generate these effectiveness and efficiency gains.

Starting with the UEFA Champions League Final in June 2017, South Wales Police deployed AFR technologies in support of a number of policing operations, across a range of contexts and settings, over a ten-month period. This is consistent with the purposes set out in their original bid for Home Office funding support to purchase and trial an AFR system, where six areas were listed for a proof of concept test of AFR:

- counter-terrorism;
- major events;
- body worn video;
- mobile phone app;
- Automated Number Plate Recognition;
- child sexual exploitation.

As part of the original bid to the Home Office Police Transformation Fund, a number of benefits of implementing AFR were anticipated, which can be defined as both policing and project outcomes.

In the context of this report, 'project outcomes' relate to the goals outlined in the original bid. For example: applying the technology across multiple events; the ability to analyse moving images; the ability to enhance images for recognition; and the use of the AFR app to process bodycam and mobile phone images from officers. Policing outcomes on the other hand were defined in terms of improving the efficiency of key policing processes in relation to both 'real time' and protracted investigations, on the grounds that "This will provide a measurable increase in potential detections and a reduction in the quantity of current occurrences and officer time spent in further investigation" (South Wales Police, 2016: 11). In addition to assisting with the prevention and detection of crime, it was asserted that the integration of the technology would produce cashable savings through the reduction in investigation and prosecution time. It was also asserted in the bid that reductions in repeat offending and increased community cohesion could be achieved through the deployment of AFR technology. The funding application was for a total of £1,950,000, with the force committing £600,000 of their own funds to the project over the next two years.

As a condition of the funding secured from the Home Office, Cardiff University were commissioned to undertake an independent evaluation. The purpose being to provide an evidence-led assessment of this technology, and what procedures and processes were utilized to integrate it into South Wales Police's working practices, as well as some of the challenges associated with its use. Accordingly, the reporting of the evaluation findings is organized around three principal themes:

- Process of implementation: this element focuses upon how the technology was implemented and integrated into policing processes. As a new technology application for UK policing, a number of innovations were introduced,

1 Source: Pawson, R. and Tilley, N. (1997) Realistic Evaluation. London: SAGE.

and a range of challenges encountered as part of efforts to make it operationally effective and efficient.

- Outcomes achieved: the second component of the evaluation is concerned with setting out the empirical evidence about what the application of AFR delivered in terms of relevant policing and project outcomes. The purpose being to bring some clarity about what AFR can and cannot do.
- Legal, ethical and regulatory implications: reflecting the innovative nature of the use of AFR in support of a number of policing purposes, there are a range of concerns and issues that warrant public and political consideration as these kinds of technological applications become more widely used.

To support these different aspects of the analysis a multi-method research design was implemented. More detail on this is contained in the Appendix to this report, but in overview it included:

- Extensive observations of police operators using AFR in both its 'Locate' and 'Identify' modes, across multiple events and situations;
- Analysis of CSV file data downloaded from the computer system, which provides a record of all searches and results conducted using the system;
- Documentary analysis of 'Operator Logs' completed by system users, intended to establish an audit and record of all of their decisions in respect of possible 'matches' generated by the system;
- Image analysis of key visual materials where these seemed to generate unusual or unexpected results from the system;
- Questionnaire – a structured survey was delivered to police officers and staff participating in AFR training run by South Wales Police, designed to ascertain their views on the quality of training they had been provided with prior to using the system;
- A field trial to systematically test the accuracy of the AFR algorithm in terms of its ability identify individuals under a number of defined conditions.

This is one of, if not the first, independent academic evaluations of how policing applications of AFR perform in real world situations and settings, with all their complexities and challenges, as opposed to in more controlled environments. Accordingly, it is important that the analysis documents and describes all of the interacting social and technical components that together influence the effectiveness and efficiency of how AFR practically supports the delivery of policing functions. As will become apparent in subsequent sections, the success or otherwise of AFR depends upon far more than just the technical components of the system and its algorithms, in terms of it being able to detect a face and link it to a similar image in a large database.

Reflecting this imperative, the analysis was structured around three cross-cutting concepts:

- System performance – focuses upon the technical accuracy and precision of the technology itself, and factors directly impacting upon this.
- Organizational performance – looks at the policies, processes and procedures that are wrapped around the technology to make it useful to the police.
- Operator performance – includes the tactical decision-making and interactions with the technology of the direct users, and what is involved in making the system work.

Taken together, these concepts have an important role in helping to capture all the elements that have to be taken into consideration in terms of understanding how AFR works in a policing context. This is important because, in keeping with the research evidence base derived from other studies of police technologies, in terms of getting it to work, 'AFR' is not a 'plug and play' option. Arguably the key finding of this report is that, in terms of its implementation and integration into policing practice, there are processes of adoption and adaptation that have to be negotiated and navigated. By adoption, we mean that officers have to become 'socialized' to the technology and its requirements, wherein they tend to find some functions useful and practically relevant, but other elements less so. In operational terms, they gravitate towards the former components. At the same time, integrating any new technology induces adaptations to established organizational routines and rhythms. These can be more or less significant, but are important to chart.

These inter-linked processes of adoption and adaptation are guided by the extent to which the technological innovation can assist police officers in performing their core functions. The drivers for AFR relate to the challenge for police associated with the identification of suspects. As an issue this has become more acute recently, reflecting the increasing general mobility of populations, inasmuch as people are more likely to live, work and take their recreation in a wider and more diverse set of geographic locations than they were fifty years ago. Likewise, police officers tend to shift roles and responsibilities with a fair degree of regularity. Overlain upon which, has been the overall growth in population size, accentuated by increased tendencies for urban living. These are all trajectories of development that weaken the ability of individual police officers to know 'on sight' and recall 'the histories' of citizens who they might be interacting with.

Automated or Assisted Facial Recognition

Automated facial recognition systems are starting to become more commonplace across a number of everyday situations. Over the past few years they have been utilized at airports for border checks, and are incorporated in the latest version of the iPhone. They have also been the subject of film treatments (such as in Minority Report and Bladerunner) which have played an important role in capturing the public imagination in terms of what might be possible from a public safety and security perspective.

AFR works by scanning an image of a face and measuring the distances between key features, which it uses to construct what is akin to a 'facial signature'. The algorithms then compare this signature against a database of stored images, bringing back the records of those images that it assesses as being most similar to the scan image. This process is most accurate where there is a clear view of the face, which is presented straight on to the camera – as happens in border check processes. However, in terms of its deployment in support of operational policing, South Wales Police's use of AFR technology was in outdoor locations

where people would be walking past the cameras from different directions, in a range of environmental and climatic conditions.

In their promotional material for the NeoFace Watch system, NEC (South Wales Police's supplier) assert that it "has been proven to work in the real world, not just in the laboratory. A robust algorithm tested and improved over years in actual deployments NeoFace Watch overcomes challenges such as crowded environments, poor lighting, moving subjects and multiple variables as small yet significant as spectacles, hats and scarves."[2] These are all claims that the evaluation has tested, to some degree.

Reflecting the complexities of the operating environment, and for reasons that are elaborated in subsequent sections of this report, rather than casting the technology as 'Automated' Facial Recognition, we prefer to conceptualise it as 'Assisted Facial Recognition'. The latter provides a more accurate description of how it was used by police. The term 'automated' implies the identification process, in terms of linking a particular probe image to a specific database record, is conducted solely by an algorithm. This was not the case. In fact, the process observed was that the system assisted human operators to make identifications. Ultimately, decisions about whether a person of interest and an image (and its associated database record) matched were dependent upon the interpretations and decisions made by the police operators. As such, the facial recognition system was functioning as a decision-support tool, rather than an autonomous machine-based process devoid of user input. Accordingly, for the rest of this report when we refer to AFR we mean 'algorithm assisted facial recognition.'

The AFR system was deployed in one of two modes:

- Locate – is a 'real-time' application that uses a series of cameras to scan the faces of people in an area, looking for possible 'matches' against a pre-selected database of facial images of individuals deemed 'persons of interest' by police.
- Identify – takes images of unidentified suspects from a past crime scene (usually captured via CCTV or mobile phone camera) and compares these against a large database of police custody images in an effort to generate investigative leads.

These two applications involve different technical challenges and policing processes.

An additional layer of complexity involved in interpreting and making sense of the findings reported herein relates to an algorithm upgrade made to the AFR system during the evaluation period. Following an Interim Evaluation Report delivered by the Cardiff University team following the Champions League pilot study, the technology supplier, NEC, offered to install a new algorithm (M20) across the South Wales Police platforms. As documented, in subsequent sections of this report, this significantly improved the performance of the AFR system in both Locate and Identify modes. These were of a magnitude such that, in terms of detailing outcomes, a decision has been taken to report them separately. This is to enable readers to gain a sense of the performance levels of the current 'state-of-the-art' in terms of AFR technologies.

The next section provides an introductory overview of the AFR technology and how it was implemented by South Wales Police between May 2017 and March 2018. Sections 3-6 contain discussion of empirical evidence and findings in relation to both the process evaluation and outcome evaluation aspects. Sections 3 and 4 are dedicated to AFR Locate, and sections 5 and 6 are dedicated to AFR Identify. These sections each capture the range of operator, organizational and system level factors that shaped the support the AFR system provided across a range of policing functions. Section 7 of the report focuses upon discusses findings from a small-scale field trial conducted to provide a systematic test of the algorithm's accuracy in identifying faces in 'Locate' mode. The penultimate section attends to a number of ethical, regulatory and legal issues raised by the innovative nature of the AFR technology and its application for policing. The Conclusion section reprises some of the principal findings and considers their implications. A description of the multi-method research design, and how data were collected and analysed is recorded in the Appendix.

2 Source: https://www.nec.com/en/global/solutions/safety/face_recognition/PDF/Face_Recognition_NeoFace_Watch_Brochure.pdf [Accessed 21 May 2018]

SECTION 2: **BACKGROUND AND CONTEXT**

South Wales Police's original Home Office funding application was set against the backdrop of the upcoming UEFA Champions League Final (UCL), a week-long event taking place in May-June 2017. The event presented a spectrum of policing challenges, including large crowds and associated public order issues and terrorism risks. Pre-event estimates indicated that 170,000 football fans would likely be in the city[3] (in fact, it later transpired that there were over 310,000 people in Cardiff on the 3rd of June).[4] Given the likelihood that the final would involve foreign teams, there was a particularly acute challenge associated with the fact that the identities of potential troublemakers would not be well known to British police. This is the sort of environment in which 'AFR Locate', the real-time mode of the technology, was anticipated to be of particular value for policing. This deployment was the first use of facial recognition by a UK police force at a large sporting event.

Following the award of Home Office funding in January 2017, South Wales Police commissioned a procurement process to select a technology partner. Several companies responded to the Invitation to Tender and NEC and its 'NeoFace' application was selected through a process including both product quality and value for money criterion. The company originally supplied their 'S17' algorithm (and later their updated 'M20' algorithm). It is worth stating that the precise functioning of these algorithms is 'black boxed', in that it has not been revealed by NEC to the police or the evaluation team precisely how matches are being calculated by the system.

The principal functions of these algorithms can, however, be broken down as follows:

- detects visual images that have the features associated with a human face;
- analyses and measures distances between specific facial features;
- generates a mathematical representation (a 'facial signature');
- compares the signature against a database of stored images;
- provides a possible match (or matches), together with a similarity score based upon the comparison made.

Locate and Identify differ slightly in how the scores generated by the system are used:

- AFR Locate provides one possible match at a time, displaying the possible matched image retrieved from the 'watchlist' to the operator.
- AFR Identify works rather differently, receiving a list of up to 200 possible matched images presented in order according to the similarity score.

During training sessions provided to South Wales Police by employees of NEC, the scoring function was described as a 'gradient score' that indicates the probability of two people being the same. Possible matches are provided to the operator if they score over a certain threshold. Operators are able to set this threshold score themselves, though NEC initially recommended a score of '.55' (where 1.0 is the highest possible score attainable). Operators were informed that in live-time deployments, scores can increase as people get closer to the camera, but that they would only see the first score given when the person initially enters the frame. They were also advised that they should use their own judgement, not the score, as the basis of any decision to intervene.

NEC were responsible for advising South Wales Police on the computer hardware they would need to purchase ready for installation. The idea was that by the time of the Champions League Final, the force would be ready with the equipment, technology and trained staff needed in order to pilot the AFR Locate function. The force acquired: 12 Alienware laptops; 4 mobile camera vans, each fitted with two 'pan, tilt, zoom' cameras; and 14 additional cameras. The vans were then fitted-out to accommodate the technology. A dedicated team of officers were assigned to work on the project, and NEC assigned several members of their own staff – including engineers/technicians – to support the police team.

AFR Locate

Familiarisation and training of police staff to operate the new system commenced in May 2017. This involved a series of small groups being shown what the software looked like and how to use it by one of the engineers. There were two test deployments ahead of the Champions League Final week (one in Cardiff and one in Swansea). On the 22nd of May 2017, South Wales Police gave a press release detailing their new project and how and when it would be rolled out. They said they would be piloting the new technology during the Champions League Final week in Cardiff, working in partnership with NEC.

For the purposes of the Champions League pilot four 'watchlists' were prepared. These contained images of: wanted/possible suspects, persons of interest and missing persons. The specific rationale for including specific persons on the watchlists (e.g. offence severity, intelligence that they would be present) was not provided to the evaluators. The four watchlists were colour coded according to the level of perceived risk and threat associated with the individuals concerned:

- "Red" watchlist contained only a small number of individuals, who were perceived to pose a serious risk to public safety;
- "Amber" watchlist contained more individuals with previous convictions for more serious offence types;
- "Green" watchlist inclusion related to individuals of possible interest to police, whose presence did not pose any immediate risk or threat to public safety;
- In addition, a blue 'friendlies' watchlist was also incorporated into the system. This contained images of police officers and was designed to provide a sort of baseline test of the system's effectiveness to detect faces.

3 Source: https://www.bbc.co.uk/news/uk-wales-south-east-wales-40122286 [Accessed 6th March 2018]

4 Source: http://www.itv.com/news/wales/2017-06-13/record-number-of-people-visited-cardiff-for-champions-league/ [Accessed 16th August 2018]

The latter was used on the basis that it could be known when police officers were in the vicinity of the cameras and so whether the system should have spotted them. Whereas, for suspects it was less knowable whether or not they were present in the crowd. In total, around 1200 images were enrolled onto watchlists for this event. These were a mixture of custody and 'non-custody' images (e.g. some were traditional 'mugshots' while others were taken outside). Protocols for appropriate policing interventions were drawn up should individuals on the red, amber or green lists be detected.

In outlining the basic configuration of the AFR system, it is important to note that some confusion and misconceptions have been evident in aspects of the public commentary on AFR about watchlist sizes:

- 'AFR Identify' compares a 'probe image' (such as that taken from a crime scene) against a database of all custody images (circa 450,000 in number);
- 'AFR Locate' uses much smaller databases of images. For the latter mode, images are pre-selected based upon a range of possible criterion and compiled as a watchlist, which the system then checks against. Across the 10-month period of the evaluation, the Locate watchlist ranged between 400 and 1200 individuals.

Figure 1 below provides a simplified visual representation of the socio-technical process associated with AFR Locate. When the AFR system scans camera feeds, it is analysing multiple faces per frame and multiple frames per second. Limits on the number of frames and faces are configured manually, and they impact the processing 'load' placed on the system's computer hardware. Pixels between the eyes is another manually configurable setting, and it relates to the quality of the live feed.

When an alert is triggered because a possible match has been detected by the algorithm (often colloquially referred to as a 'hit' by operators), the operator has to decide whether they think this is a 'true positive' or 'false positive' and respond accordingly. This 'true / false' terminology originates in computer science and has been adopted by South Wales Police and its partners, in respect of human judgements and decisions in respect of AFR images. Given this, we reproduce these concepts throughout the evaluation:

- 'True positive' is defined as a match between a scanned visual and an image stored in a database generated by the system, which is judged to be correct by human operators;
- 'False positive' is a 'match' generated by the system assessed as incorrect by the operators.

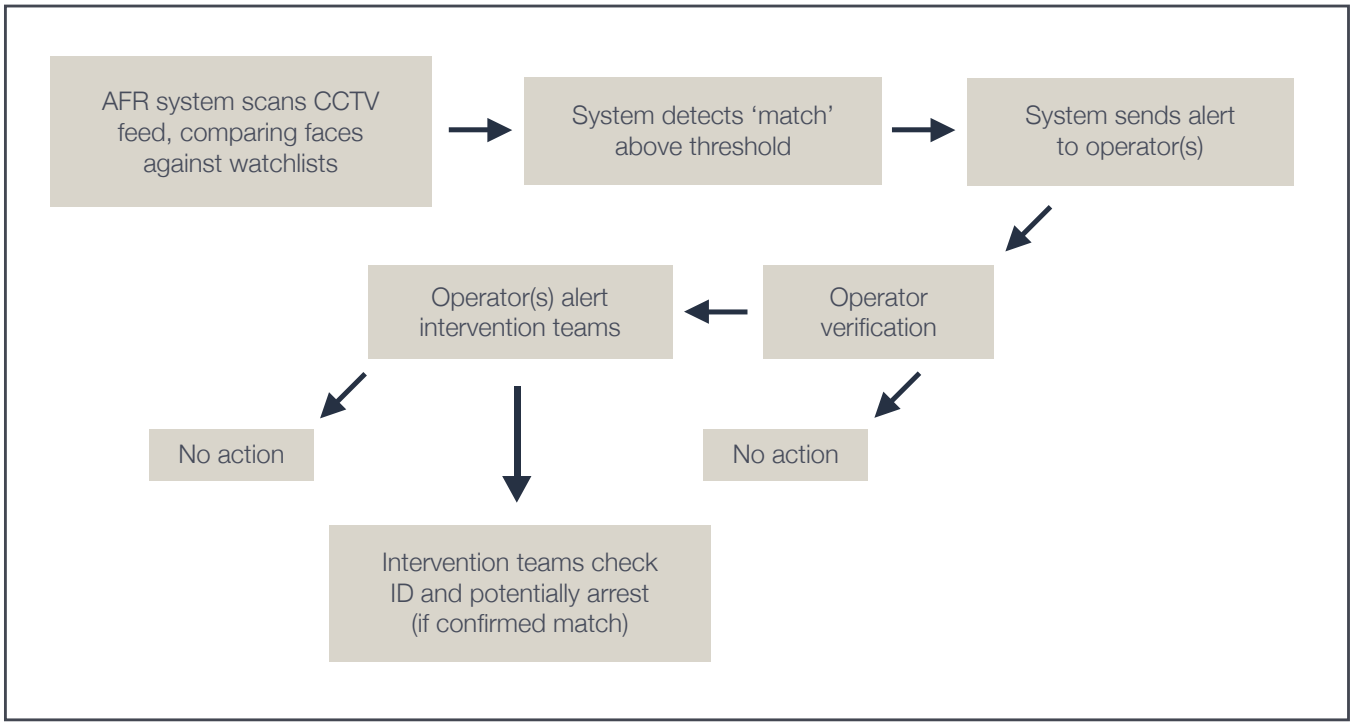


Figure 1 – process diagram of AFR Locate

In respect of the Champions League Final, AFR Locate was deployed in locations in and around Cardiff city centre. In addition to the four mobile police vans (which were clearly marked 'Facial Recognition'), four 'static' camera locations were established in: Cardiff Airport, Cardiff Central Train Station, Churchill Way and Callaghan Square. The latter two locations were to be used as dedicated 'fanzones' and would be key busy locations, and Cardiff Airport and Central train station would be two of the main transport hubs during the event. Unlike with the vans, the static cameras could not be moved by the operators. A van was also set up at the Principality Stadium, which remained in place throughout the event.

The plan was to have two operators in each location (van and static), with key members of the project team and someone from NEC in and around the city to assist if needed. Members of the UPSI evaluation team were also moving between locations, observing throughout the week of the event. An Operator Guide, detailing how to use the system and how to log matches correctly, was produced and available in all locations. Given that this real-time application of AFR technology is designed to locate persons of interest as they pass a camera, attempts were also made to station intervention teams nearby. In the event of a positive match (a 'match' identified by the system and confirmed by the operator(s)), members of the intervention team would be tasked to stop the person in question, check their ID, and take appropriate action. Evaluative aspects of this and all other deployments are discussed in later sections of this report.

Following the Champions League, South Wales Police deployed the AFR Locate function at a series of events (as set out in the original project plan), including:

Elvisfest, Porthcawl The annual Elvis Festival in Porthcawl was held on 22nd and 23rd of September 2017, and the force were again using the S17 algorithm supplied by NEC. This time, the system was loaded with 472 images of persons of interest from the South Wales Police area. Though on a smaller scale, the deployment followed the same basic principles as those established for the Champions League with two vans and two operators per van.

Operation Fulcrum, Cardiff Operation Fulcrum was deployed on the 18th of October in Cardiff City Centre. This is a recurring operation where divisional police proactively concentrate on warrants and outstanding suspects. The system was loaded with the images of persons of interest, in the hope of quickly identifying them if they walked past the van. The deployment lasted all day and was based out of one van in the city centre, again using the S17 algorithm. This same operation has been repeated on other occasions, including 22nd December, using the new M20 algorithm.

Anthony Joshua Boxing Match, Cardiff The next major deployment of AFR was for the Anthony Joshua fight on the 28th of October 2017, held in the Principality Stadium. For this event there was a van stationed at the stadium and two others in the city centre, and they were generally manned by two operators each. Watchlists were again made up persons of interest, with 608 individuals included in total. The severity threshold for offences by persons on watchlists is not known. The old S17 algorithm was still being widely used at this event, but the force had been given limited access to the new M20 algorithm on one laptop. This laptop was moved around the various locations during the day, running in tandem with the S17 algorithm. Though the majority of the results from this deployment were therefore based on the old algorithm, operators could begin to see performance differences between the two versions of the system.

Autumn Rugby Internationals, Cardiff AFR was deployed during November and December at four international rugby games in Cardiff: Wales vs. Australia (11/11/2017); Georgia (18/11/2017); New Zealand (25/11/2017) and South Africa (02/12/2017). Watchlists contained persons of interest (502 for Australia; 1,261 for Georgia; 892 for New Zealand; 911 for South Africa). By the time these events came around, the new M20 algorithm was available on all laptops. In terms of resource, there were two mobile vans and six operators for the Australia match, and two mobile vans, a stadium van and six operators for the Georgia, New Zealand and South Africa matches. Issues with the computer network were experienced at the Georgia match, resulting in the system failing and thus very little data being produced.

An important social innovation was introduced at this point of the evaluation process. There was growing interest in the number of 'transactions' being conducted by the system - that is, the number of times the system recognises a face. If the system calculated that it could be a match, an alert was given to the operators. But there was no auditable record about the number of times the system was concluding 'no match'. A system called 'Flow' was therefore developed to capture these numbers, and was deployed in one van at the Australia and New Zealand games. These figures begin to contextualise the ratios of true / false positives.

Music Concerts at Motorpoint Arena, Cardiff In December 2017, South Wales Police publicly announced a new approach to pickpocketing and mobile phone theft using the facial recognition system. This would be in partnership with Cardiff's Motorpoint Arena, who also made the announcement on their own website. This would entail fixed cameras inside the foyer of the venue, and officers on duty at events to operate the system and intervene with those identified as possible offenders. Perpetrators of mobile phone theft often travel to the target venue, as a consequence information on suspects to include in watchlists was gathered from other UK forces through 'Operation Gothic'. The first event to receive this intervention was the Kasabian concert on the 4th of December: watchlist numbers here totalled 442, and the outcomes were communicated to the public the following day. Later in December, the system was used again for the Liam Gallagher concert. Subsequent to these deployments AFR cameras are now fixed at the arena.

Operation Malecite, Swansea Operation Malecite was deployed in Swansea on the 23rd December, and was similar in nature to the previous Op. Fulcrum deployments in Cardiff.

Six Nations Rugby Internationals, Cardiff AFR was deployed at three Six Nations matches in Cardiff (all of the home games): Scotland (03/02/2018), Italy (11/03/2018) and France (17/03/2017). Watchlists were again made up of persons of interest. In terms of resource, there were generally 2 mobile vans, 1 stadium van, and 2 operators per van. The UPSI evaluation team also carried out a small-scale field trial during the Italy deployment.

AFR Identify

Although it uses the same technology, 'Identify' is a rather different application of AFR than described above. It involves taking still images of unidentified suspects from crime scenes (from CCTV and mobile phone footage etc.) and uploading them to the AFR system, which then compares them against a database of around 450,000 police custody images (a number that increases every day).

For each 'probe image' submitted, the AFR system provides 200 results ranked in order of similarity based upon an assessment made by the algorithm. The operator looks through them to determine whether there is a positive match which can be actioned (see figure 2). Within South Wales Police, results are categorised as AFR 1, 2, 3 or 4. These codes are noted below alongside the corresponding definitions used in this report:

- Match generated by the system and confirmed by operator (AFR 1);
- Match generated by the system but disconfirmed by operator (AFR 2);
- Image rejected by the system (poor quality) but circulated to officers (AFR 3);
- Image both rejected by the system and of too poor quality to be circulated to officers (AFR 4).

Responsibility for operating AFR Identify was assigned to the existing PROMAT / Identification team in South Wales Police. These individuals were not the same as those delivering 'Locate'. The Identify function was an additional task given to the unit involved in organising witness viewings and ID parades for identifying suspects. In July 2017 an organisational process was established enabling officers to submit requests for AFR work on cases where they have images of unidentified suspects. On three occasions during the evaluation period, UPSI researchers visited the Identify team to observe AFR Identify in action. A detective who used AFR to identify an unknown suspect has also given his feedback on how this technology impacted his investigation.

During observations in 2017, the Identify software was based on a single Alienware laptop, kept in the office and used by all members of the team (the software has to be used on this laptop, but normal workstations are used for communicating with officers etc.). As with Locate, AFR Identify also used the old S17 algorithm to begin with. At the end of October 2017, the Identify team were given access to the new M20 algorithm. Some cases which had previously returned no matches on the S17 algorithm were re-run on the new algorithm, and some differences in system performance were seen. Towards the end of the evaluation period, a new server-based version of AFR Identify was being developed. The idea being that officers would be able to run their own AFR search remotely, and the AFR capability could also be given to police staff in the Public Service Centre (PSC) for any urgent, 'out of hours' requirements.

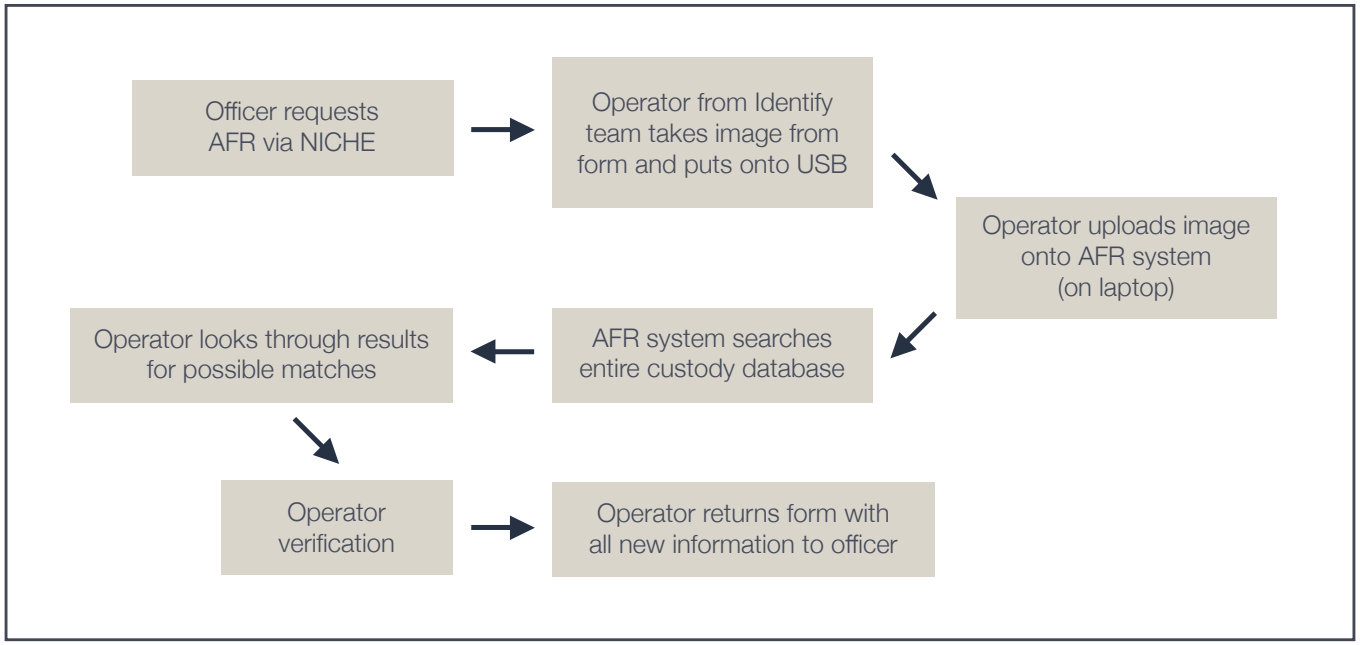


Figure 2 – process diagram of AFR Identify, correct as of October 2017 (after January 2018, USBs were not required to transfer information to the laptop)

SECTION 3: **AFR LOCATE PROCESS EVALUATION**

This section discusses the process of implementing AFR Locate in South Wales Police across the whole evaluation period. It describes the work involved in setting the system up, and the challenges and issues encountered in getting it to integrate with existing policing processes and systems. In essence, the focus is upon how the technology was adopted, and the amendments and adaptations that had to be introduced to both the system configuration and the policing processes wrapped around it, to render it a useful policing tool. Within this analysis of process implementation, the organisational performance, system performance and operator performance themes are used to orient the discussion.

AFR Locate and Organisational Performance

In preparing for the roll out of AFR Locate, South Wales Police held training and familiarisation sessions for Champions League operators, and also selected people for inclusion in watchlists. The scale of the event meant that a lot of operators were needed. Multiple training sessions were organised to ensure all operators were trained sufficiently (38 operators in total). These were held at Bridgend Headquarters during May 2017. One session was attended and observed by two researchers from the UPSI evaluation team. Around nine trainees (a mixture of officers and police staff), as well as a couple of officers from the AFR project team attended. The trainer had only been expecting five.

The training was delivered by one of NEC's software engineers, lasted around 3 hours, and was quite basic. It started with a 'slideshow' introducing the concept of facial recognition. The second part was a more 'hands on' demonstration: operators sat at laptops in pairs and were shown how to navigate the desktop and the application; how to create a new watchlist; and how to 'enrol' their own pictures. They were also shown what 'matches' look like on the system. The final part of the session covered additional features of the software. The operators were told they would have a 'crib sheet' on how to use the system in the vans with them during the deployment, and that officers and NEC staff would be around to assist if needed.

In addition, a larger familiarisation session was held on the 30th of May, again at Bridgend HQ. The purpose here was to brief operators on the more practical elements of the deployment and to explain how the technology would fit into the wider policing operation. Officers from the AFR project team led this session, and it was far more policing focused than the previous training sessions run by NEC. It covered, for example, how interventions would be made. Around 25-30 people attended this session, which included the would-be operators, as well as some senior officers from the AFR project team, and some NEC engineers. During the session, officers provided attendees with a detailed overview of the AFR project and additional information around the Champions League deployment itself. The engineers from NEC gave an overview of the hardware, explaining how it would be set up and what would be running on each laptop. An event handbook was also distributed at this session. This was a 14-page booklet covering a range of topics including: deployments; cameras; practicalities; watchlists; the 'score'; body worn video; passing on information; and intervention teams. During the session, operators raised questions and discussed issues such as the legitimacy and cost effectiveness of the new technology, and practicalities of liaising with the intervention teams.

Following the initial Champions League deployment, no other formal training was provided to operators. Whenever someone new worked on a deployment, they were simply given brief instructions on how to operate the system by an operator who had used it before. This means that many operators are currently using the system without formal training.

Another organisational performance issue for AFR Locate concerned the creation of watchlists. The evaluation team were given brief overviews of who was included in these at each event, but the specific inclusion criteria / offence severity thresholds were not specified. Subsequently, there were changes in watchlist composition and size for each event. It has not been possible to determine how manipulation of these variables influence the overall performance and contribution of AFR.

At each deployment, multiple different colour watchlists were used to denote the type of individuals within them. The colour codes for the Champions League have already been listed previously. At subsequent events, the watchlist colours were 'blue', 'amber/yellow' and 'purple/pink'. People wanted for offences were included in the latter category, and individuals suspected of offences were included in the 'amber/yellow' watchlist. For concert deployments, such as Kasabian and Liam Gallagher, watchlists included individuals 'known' to police for mobile phone theft and pickpocketing across the UK. During the evaluation, it was unclear precisely if and when missing and vulnerable people were included in watchlists for public safety reasons. It is our assessment that greater clarity about the criterion for inclusion (and exclusion) of individual images on watchlists should be established through creating a pre-defined policy framework.

AFR Locate and System Performance

For both the initial roll-out of AFR and all other deployments of the technology, there were system-related process issues that impacted upon its performance. These related to the ways the system is not 'plug and play' and must be prepared and loaded prior to outputting its results. This section discusses: image similarity scores; pixels; the setting of 'faces per frame' rates; the cameras; and the quality and type of images included within the watchlists.

As outlined previously, as part of its operations the algorithm generates a score pertaining to the similarity (at least as measured in computational terms) between the images stored in the

database and the image submitted for enquiry. This score was set to .55 for the Champions League event, as recommended by NEC. This meant that when the system compared passers-by to the 1200-strong watchlists, if the score of their similarity to someone on the watchlist was deemed .55 or above, then operators would be sent an alert for review.

However, due to the high number of false positives generated during the Champions League, this score threshold was revised and increased for later deployments. This meant that fewer alerts were sent to operators, as subjects had to achieve higher similarity scores in order for the system to flag it as a possible match:

- For Elvisfest the score was set at .60, but the number of matches made by the system was then very low.
- As a result, the score was lowered to .57 for the Anthony Joshua fight. This increased the number of matches made by the system, but did not bombard the operators with alerts. This score was then maintained for deployments including the Autumn Rugby Internationals.
- The score was later raised to .59 for the Six Nations. The precise reason for doing this is not known, but it did mean that matches made by the system were closer in similarity.
- For the Italy Six Nations game, there were scores of .55 and above in the spreadsheet provided to the evaluation team. It seems this was due to one van having the old settings on their laptop.

A second user-configurable item in the AFR NeoFace platform relates to the number of pixels needed between the two pupils of the eyes for the system to detect a face. This is affected by several factors including resolution of the image, distance from the camera to the subject and level of optical zoom. Following NEC's recommendation, South Wales Police set the number of pixels at 80 at the beginning of the Champions League deployment:

- On day 2 this was changed to 40 pixels. The precise reason for changing this is not known, but this setting was maintained for all subsequent deployments.
- If images are of a high enough resolution, they can be captured at a wider angle and still meet the minimum requirement of pixels between the eyes. However, the higher the resolution of the image, the longer the system will take to process it.
- Revising the number of pixels down from 80 to 40 suggests that on the first day, the operators needed to use a high level of optical zoom to detect faces. Reducing the pixels to 40 means the algorithm could analyse the faces being captured without having to set the camera so narrowly that most of the scene was lost.

'Frames per second' or 'frame rate' refers to how many frames are analysed by the software in a second. For the Champions League, this was initially set to 30 frames, meaning every second the software scanned 30 separate frames. 'Faces per frame' refers to the maximum number of faces the software can analyse in any given frame. On the first day of the Champions League, this was set at 10 faces per frame:

- With these settings – 10 faces per frame and 30 frames per second – this would mean that the software could theoretically analyse up to 300 faces per second (settings figures taken from presentation at South Wales Police AFR Board Meeting in December 2017).
- These settings were repeatedly revised during the evaluation, as South Wales Police tried to optimise the performance of the system hardware.
- On day two of the Champions League, the 'faces per frame' setting was lowered (and halved) to 5 in the hope of reducing the load put on the computer hardware (analysing up to 150 faces per second).
- In later deployments (Anthony Joshua and the first two Autumn International games), the 'frames per second' was also lowered and set at 25, giving a rate of (up to) 125 faces per second.
- This was again lowered to 10 frames per second (with 5 faces per frame, up to 50 faces per second) for the latter two Autumn games (New Zealand and South Africa). These settings were then kept for the Motorpoint and Six Nations deployments.
- These continuous revisions to the frame and faces rates highlight the severity of the system's hardware performance problems.
- Although the reduced load on the system improved performance (the results are discussed later), up to 50 faces per second is dramatically lower than the initial rate of up to 300 commissioned from NEC.
- Towards the end of the evaluation, NEC were recommending further lowering the settings to 2 faces per frame, 10 frames per second (up to 20 faces per second) to improve performance.
- Operators expressed disappointment with the need to lower the settings to such a level. Their use case had not changed, and it was not the level of system performance they had been expecting.
- With NEC recommending this lower number of faces but South Wales Police maintaining the higher number, it could be argued that they were changing the manufacturer's operating criteria. There could potentially be error being built into the system as a result of this, impacting the performance of the system (though this is not fully understood at present).

Operators were also able to change the position and zoom of the 'Pan, Tilt, Zoom' cameras located on top of the mobile vans. This was done using a 'Xbox' type gaming controller linked to the computer. When initially setting up the system at any deployment, the operators were responsible for deciding the best area to point the cameras at, and the appropriate level of zoom. These decisions are important – if the cameras are zoomed in too much or too little, or if there is glare from the sun, this will impact the pixel quality and lighting, and therefore the system's ability to read and analyse faces:

- Operators used the white boxes that the software displayed around faces to judge whether cameras were in the correct place: if the system was ' white boxing' faces, they felt

it was zoomed in appropriately. There was no specific guidance provided about this.

- As there were two cameras atop each van and two camera feeds displayed on the system, operators usually pointed the feed shown on the left-hand side of the screen towards the left side of the van, and the feed on the right-hand side of the screen towards the right side. This made the setup more intuitive for the operators. Remembering where the cameras were pointed became key to responding to alerts effectively.

As part of the contract and project plan, NEC were commissioned to deliver a 'mobile app' to South Wales Police. The plan had been to provide Intervention Teams with this app, so that they would also receive alerts even if they were not in the van. However, this tool was ineffective at all events. The reasons for this were:

- In order to work, the phone needs to be on the same Wi-Fi network as the laptop, which is a secure police network. However, the security required to connect to and operate on this network seemed to clash with the app.
- In practice, the app was observed randomly disconnecting from the network, unbeknown to the operators or the intervention teams. It was also observed being delayed in its delivery of alerts.
- South Wales Police are now exploring their own options for developing this sort of app, given the ongoing problems with NEC's version.

During the Autumn Internationals and later deployments, South Wales Police also deployed a system named 'Flow'. This was designed by NEC to count the number of 'transactions' carried out by the system (i.e. the number of faces scanned and compared against the watchlists that did not generate matches). This has also been referred to as 'the number of times the system said 'no''. This was a figure that the force had previously been unable to ascertain, and their understanding of the technology was limited to the number of true and false matches.

The quality of images inputted to the system has a substantial effect on the results generated by the AFR system, as South Wales Police quickly learned during the initial pilot. The images used for this first deployment varied greatly in quality. This is partly due to the international nature of the event, but also the force's lack of understanding around the importance of good quality images. Images came from British, Spanish and Italian sources as well as other European agencies. The majority were 'custody' (controlled environment) images, but they differed greatly in standard. In addition, images taken from social media and surveillance were also used for some individuals – these have no 'quality control' and many were taken outside. People were often smiling or squinting, or had their heads at an angle, which affected the shape of their faces and therefore the algorithm's analysis of them. This significantly impacted the results attained, as discussed in detail later.

Having realised the importance of using good quality images from the outcomes of the Champions League deployment, South Wales Police revised their choice of input images. At all subsequent deployments, only custody images of a high standard were entered into watchlists. Indeed, they undertook a specific force-wide programme to enhance the quality of all their new custody images. The requirement for and costs of this change in how custody images were taken, was not anticipated at the start of the roll-out of AFR.

As part of their equipment purchases for AFR, South Wales Police purchased 14 additional cameras at £4,800 each (totalling over £67,000) under the guidance of NEC. During the New Zealand game at the Autumn Internationals, one of these cameras was observed steaming up. As a result, the system was unable to detect any faces through it, rendering the camera inoperable. At the time, one of the operators explained he thought a seal had blown on it.

By February, all 14 cameras were experiencing the same issue, and it transpired that the problem was caused by a manufacturing fault. Fortunately, the cameras were still under their three-year warranty, and were replaced by the provider in time for the Six Nations deployments. However, if issues like this occur in future and the cameras are out of warranty, the cost of purchasing new ones will fall to the force. Camera equipment and computer hardware also has a certain 'lifespan' and will eventually need to be replaced, which will also be a cost to the force. Further to which, given the system's poor performance in low light, there has been talk of trialling new, more expensive cameras.

Although the costs of the hardware to support AFR deployment are not a principal focus of this evaluation, the above issues are worth considering in terms of what return on investment police forces should reasonably expect when purchasing an AFR system. The evidence generated by this research is that the overall performance of the system involves a number of overt (costs of purchasing and replacing equipment) and more 'hidden' costs (upgrading the quality of custody images and how they are taken).

AFR Locate and Operator Performance

This section examines how operators interacted with each other and the AFR system, and how they made their decisions when the system sent a match alert during Locate deployments. It also considers operators' perceptions of how they were trained. A structured questionnaire administered to Champions League operators provides insights into their personal views and experiences, and the observational data provide evidence about their interactions and decision-making. A key finding from this element of the evaluation is:

- In the early deployments in particular, officers with experience of using CCTV tended to try and use the AFR cameras in a similar way. This was inappropriate. CCTV cameras tend to be used relatively 'zoomed out' to track an individual and their behaviour, whereas AFR requires a close-up picture of a face.

A need to make quick decisions so as to maximize the utility of the platform was recognised by operators at all observed deployments. When an alert popped up, operators would immediately look at the two matched images on screen (sometimes double-clicking to get a wider image of the scene), deciding whether they thought it could be a match or not, and then looking outside of the van to try and locate the person of interest on the street. The alert noise was key in initially drawing the operators' attention. Sometimes, one operator would be looking for the person in the crowd while the other

was describing them aloud from the still image displayed on the screen:

- Operators reported using key facial features such as eyes, nose, mouth, jawline and hairline to inform their decisions. They also found the background information provided about persons of interest (e.g. offence type) to be useful.

There were some slight differences in decision-making processes across deployments, and so this element was not consistent. Sometimes, one operator became the primary decision-maker while others who were co-present contributed little (when police officers were present with police staff, officers were often defaulted to). Locate deployments involve looking at faces on laptop/desktop screens for long shifts in a confined space. One concern here is that operators might become 'face blind' and fatigued by this environment.

Particularly at the Champions League, operators were observed being quite bored and frustrated as they were responding to large numbers of 'false positives' with low similarity scores. This was improved by raising the score, as less false positives were generated. The new algorithm improved this further (though the inner workings of this algorithm are not understood). As a result of these changes, operators were not overwhelmed with alerts, and when they did receive them, they could see clear similarities between the subjects, and morale improved.

Following the Champions League deployment, operators were asked about their experience with AFR via a questionnaire (31 responses were received). Two thirds of the sample responded to an online version, and the rest via pen and paper. Key findings from this included:

- Although most operators had a fair amount or a lot of experience with computers, relatively few of them had any experience with CCTV or surveillance techniques (only 23%).
- On average, respondents: agreed that the amount of training they received was adequate; agreed that the training covered everything they needed to know for the event; agreed that the system was clearly explained to them; and strongly agreed that they understood why South Wales Police was using AFR.
- Although on average respondents agreed that the training covered everything, it is worth noting that 23% disagreed with this statement.

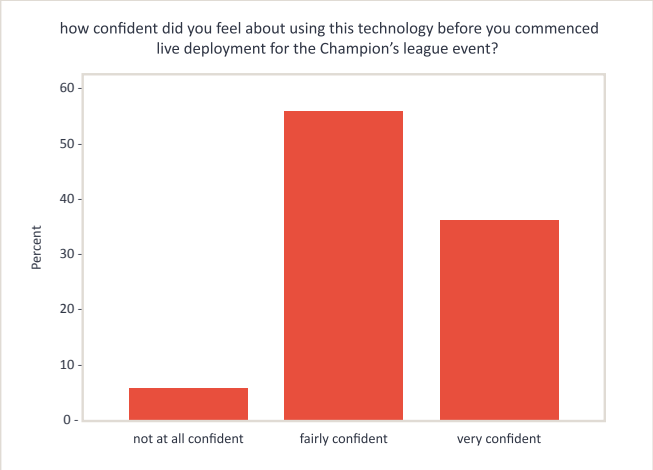


Figure 3 – confidence using AFR before the Champions League event

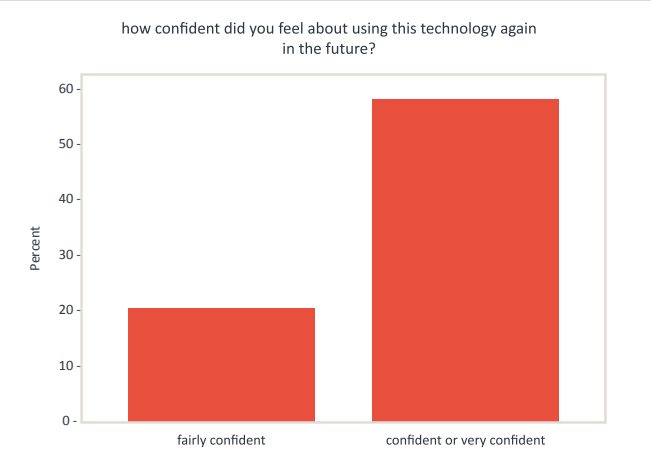


Figure 4 – confidence using AFR in future, after the Champions League event

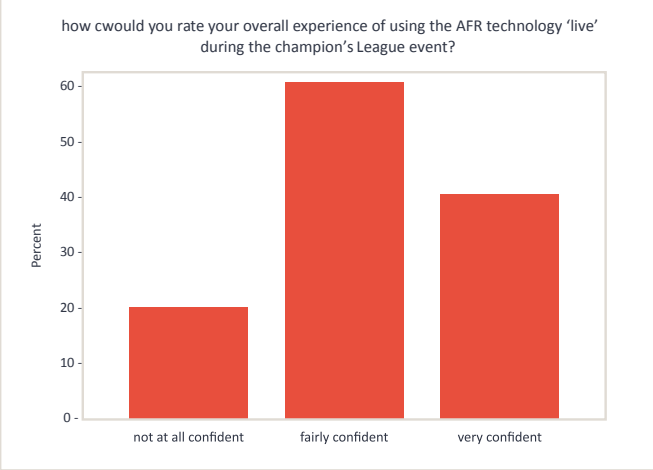


Figure 5 – overall rating of AFR experience at the Champions League event

- Most operators said that prior to starting their Champions League deployment, they felt a degree of confidence about using the system. However, post-deployment this changed, and all operators felt confident or very confident about using it again in the future (see figures 3 and 4).
- On a scale of 1 to 5, 81% of the respondents rated their experience of using AFR 'live' at 4 or above, indicating that the overall majority had a valuable highly rated experience (see figure 5).
- The System Usability Scale, a generic and validated methodology for assessing a technology's ease of use, was also incorporated into the questionnaire to gauge how 'usable' operators deemed the system. 'Average' on this scale is 68 - 15 scores were above average, 4 were average, and 11 were below average, suggesting the majority of AFR operators found it user-friendly.
- The problems most commonly cited by operators were environmental conditions, camera positioning and match logging.
- 84% of respondents said they would volunteer to be an operator again in the future, some said they were not sure, and one individual said they would not volunteer again.
- The latter person had no experience with CCTV, gave the training sessions a low score, and did not feel confident using the system. Although they made some positive comments about their experience in the free text, this person also stated "I didn't feel like the training for the software was tailored to what we were actually doing when we were deployed."

SECTION 4: **AFR LOCATE OUTCOME EVALUATION**

Th focus of this report now shifts to an assessment of the outcomes achieved by South Wales Police that are attributable to their deployment of AFR in the 'Locate' mode. As previously, the discussion is organised by separating out organisational, operator and system level influences.

S17 (Old) Algorithm Outcomes

Between May-October 2017 the S17 algorithm supplied to South Wales Police was deployed in relation to several major events including: the Champions League Final; Elvisfest; and Anthony Joshua boxing match (although for the latter, the new M20 algorithm was used at one location). Data from three such event deployments, along with key performance indicators, are set out in the Table below:

		EVENT SIZE	MATCHES	TRUE POSITIVES	ARRESTS
EVENT	**Champions League (4 days)**	extremely large	2632	78 (3%)	1
	Elvisfest (2 days)	small	18	11 (61%)	1
	Anthony Joshua	large	60*	5 (9%)	2

*South Wales Police figures show this as 51, as a police officer was included in the Amber watchlist for app testing purposes.

- Four arrests were made across the 'Locate' deployments.
- Operators were mostly optimistic about the system.
- The outcomes achieved were constrained by difficulties that the operators encountered in communicating with intervention teams (the latter were intended to conduct the stops of individuals identified).
- If operators thought a match was true, sent intervention teams, and then realised the match was false, people were invited into the vans to see their match and to see the technology in operation.
- AFR Locate system struggled in large crowds – observers saw lags of a few seconds and the system freezing completely.
- AFR Locate struggled in low light (low daylight and night-time).
- Poor 'custody' images have a negative impact on results. Following UCL, only good quality custody images were used.
- Identified 'frequent hitters' in the data, which appear to be caused by 'non-custody' images, bad angles, squinting.
- Initial score recommended by NEC was .55 – this was too low.
- The relationship between scores and operator decisions were explored: as the system score goes up, are operators more likely to deem it a 'true' match? We did not see a relationship between these two aspects at UCL because there were so many false positives, and there was too little data from the Anthony Joshua deployment to be able to draw clear conclusions.
- Lots of alerts at Champions League resulted in frustration, boredom and a lack of confidence in the system from operators.
- Generic logbooks meant operators recorded matches with differing levels of detail.

Locate Outcomes: Organisational Performance

A number of organisational performance issues impacted upon AFR Locate outcomes between May and October 2017. For both the Champions League and Elvisfest events, challenges in communicating with the intervention teams were encountered. These teams were intended to be tasked by the operators to stop any individuals highlighted by the AFR system and conduct identity checks with them. However, Intervention Teams often got called away from their planned stations with or near the mobile vans, to assist with other matters. In some cases, there were no intervention teams provided at all, or operators were not informed of their location. Observations also indicated that operators would have found feedback from intervention teams useful after stops, but this was not always possible in these scenarios.

Even where Intervention Teams were present (if the weather was miserable, the teams sometimes stayed inside the vans, either sitting in the front or standing near the doors in the back), difficulties were sometimes encountered. In particular when dealing with crowds, if the system generated a possible match, the operators in the van would sometimes have difficulty guiding their Intervention Team colleagues to where the subject of interest was. In part, this reflected how the cameras were zoomed in quite closely to facial features, thereby limiting the

availability of contextual cues needed to identify where the individual was located relative to others in the vicinity. This constrained the capacity of police to make physical contact with individuals on their watchlists identified as present in the locality. There is potential for this effect to be exacerbated at large sporting events, when subjects are often wearing similar clothing (e.g. red rugby jerseys).

Things improved somewhat during later deployments, with more intervention teams available at the Anthony Joshua event in October 2017. In one van, when the operators decided that a match they received looked like a positive match, the uniformed officers in the van were quickly able to get out and stop the person. Having confirmed the match, an arrest was made. However, at a second location, an operator who said he would have sent officers to stop one woman (and check her ID) was unable to do so, because the Intervention Team was not available.

When stops or arrests were made, Intervention Team officers would ask for a person's name and check their ID, then explain why they had been stopped and for what offence. This included telling them how they had been identified by the AFR system and showing them the screens in the van. Following the arrests made during the deployments, the force also provided updates to the public. These updates were often shared via social media, where short posts / tweets told users where the vans were being deployed and what had happened during the day in question.

Locate Outcomes: System Performance

To evaluate the performance of the technological system, the research team collated and analysed data from observations, the outputs generated by the NeoFace system and the logbooks used by the operators. Results from the deployments using the old algorithm vary because the system settings and preparations were changed multiple times, as the AFR team sought to optimise the technology. The events were also vastly different in size (the Champions League attracted an unprecedented number of people to Cardiff, whereas Elvisfest was a relatively small event). Accordingly, results are discussed in relation to each key event, rather than being directly compared. During the Champions League, there was an emphasis on logging 'blue' (police) matches generated by AFR. However, as the implementation continued, this process became increasingly flawed. Therefore, for deployments after the Champions League, the data on 'blues' have been removed from the reporting.

CHAMPIONS LEAGUE FINAL

Empirical data from the very first deployment of AFR indicated the system did not always function as expected. Bright and sunny weather sometimes 'washed out' peoples' faces or led to them squinting. This directly impacted system performance in terms of it making identifications.

On many occasions the system hardware configuration appeared to be overloaded, running slow, stuttering and even crashing on some occasions. For example, one researcher described a 90 second "lag" between the system detecting a match and the actual alert being issued to the operator, and sometimes the alert sound was 'missing'. One operator commented that this made the tool "operationally useless". A second researcher described a recurring lag of 1-2 seconds, characterised by stuttering and 'jumping' in the camera feeds. This seemed to happen as soon as the system began analysing faces that had entered the frame and was a significant problem in dense crowds. There was also one instance where the system completely froze: it stopped responding and it was impossible for the operators to close the dialogue box. They could only sit and wait for it to respond again while other matches were still coming through in the background, which they were unable to action (this also created a backlog for the logging when the system did begin to function again).

A summary of the system's performance is provided by the following indicators:

- Across the whole event and all watchlists (including blue), there were 2632 matches (alerts) made by the system.
- Operator logs were then used to clarify whether or not these matches were accurate, and indicated that 3% (N=78) were 'true positives' (confirmed by the operators).

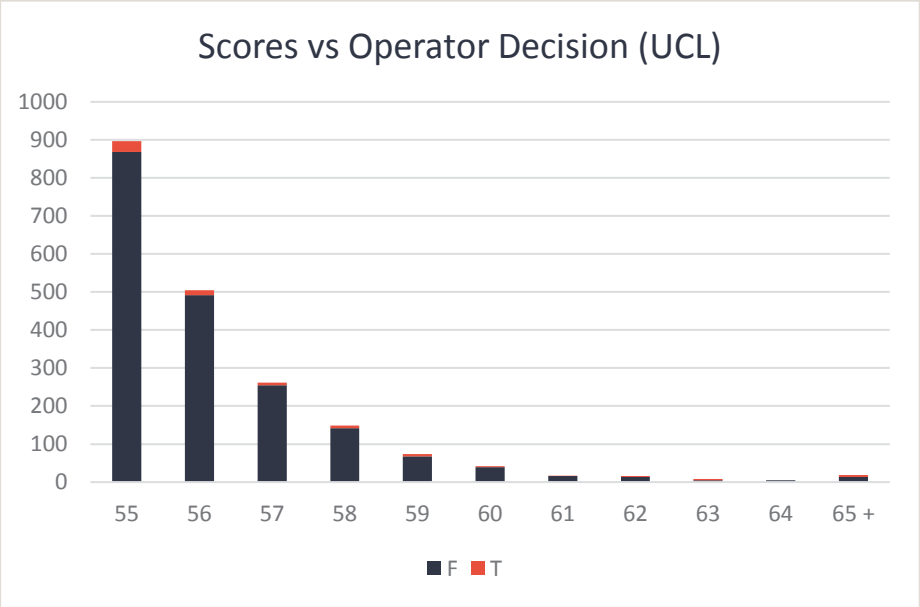


Figure 6 – bar chart showing relationship between match scores and operators decisions at UCL

- The majority of blue matches were confirmed as false positives (N=196 of 367) by operators, and for almost half (N=157) a decision was not recorded in the log books.
- There was a strong sense from observations that the threshold score for generating a match by the system had been set too low. This caused the high number of 'false positives' creating a very large workload for operators.
- Analysis also included a comparison of scores and operator decisions. The aim being to explore whether there was any link between the two: i.e. the higher the system score, the more likely the match would be classed as a 'true' match. No clear positive relationship between score and operator decision was found, because there were so few 'trues' overall (see figure 6).
- Also, as Figure 6 opposite shows, the vast majority of matches (more than 1,600 of the 2,632) scored .57 and below.

A key influence upon these figures was the poor quality of many of the images uploaded onto the system. As a result of these poor quality watchlist images, a small minority of people became frequent 'hitters':

- The top 15 individuals by match numbers (not from the blue watchlist) accounted for more than 25% of the total matches.
- Analysis of these 'power few' images gave some indication as to what may have caused them to appear in the data so frequently: 13 of these top 15 individuals had 'non-custody' images in the watchlists, meaning their photographs were not taken in controlled environments.
- The apparent disproportionate likelihood of these images triggering a match by the system may have been increased by their facial expressions or angles that their head was captured. Squinting, frowning, smiling and 'yaw' were all common features in the high match rate images.
- There were also very frequent hitters on the 'blue watchlist'. Five individuals accounted for 44% of the 'blue' matches, none of which were recorded as true positives.
- Interestingly, all of these blue images were classified as 'custody' following analysis, but facial features again appeared to be a key factor in these matches, as many had glasses, facial hair, a head tilt, or were smiling.

ELVISFEST

No field observations were conducted at the Elvisfest deployment in Porthcawl. However, data returns on system performance were provided to the research team by South Wales Police:

- Fewer matches (n=18) were reported during this two-day deployment.
- 11 (61%) were recorded as 'true positives', which is a much higher proportion than previously (see Figure 7 below) and 7 (39%) were 'false'.
- One arrest resulted.
- The fewer overall matches and higher 'true positive' proportion is likely due to the increased score threshold (.60 for this deployment) and the implementation of lessons learnt about image quality following the Champions League.

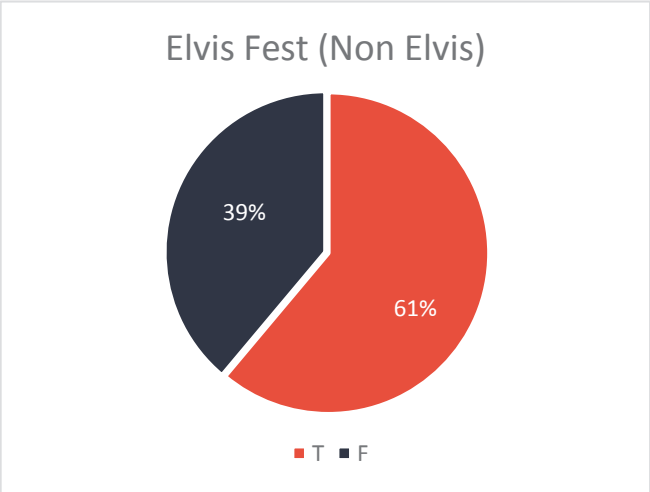


Figure 7: Proportion of true / false positives at Elvisfest

ANTHONY JOSHUA BOXING MATCH

Whilst the old S17 algorithm was still in use and was responsible for over 80% of the matches produced during this deployment, the new M20 algorithm was introduced for the first time on one laptop. This laptop was moved between sites at various points during the day, and so was not used consistently. This allows for a degree of comparison in performance, although this cannot be done systematically. Issues with the old S17 algorithm continued, as frequent instances of images 'freezing' and 'stuttering' were observed, when there were large crowds. At one stage (in one of the vans), a laptop stopped performing its alert function – though new matches were appearing in the list, there were no sounds or pop ups to alert the operators of possible matches.

In comparison, the M20 algorithm delivered superior performance. On numerous occasions, alerts came through on the M20 laptop when they had been missed by the old algorithm on the other laptop, despite them both using the same live camera feeds. The improved performance was particularly clear as daylight began to fade. At the stadium, the M20 algorithm was detecting the majority of faces in the frame (putting white boxes around them) when the S17 was not picking up any. The effectiveness of the M20 algorithm was greatly reduced as daylight continued to fade (and when it eventually completely faded) however, and this is a significant problem for the system. The live feeds became very grainy and the system could no longer recognise faces in the dark.

Data obtained and analysed indicated:

- 67 matches (alerts) were generated at this event. Removing 'blue watchlist' matches to focus on non-police personnel leaves a total of 60 (note: all blue matches were confirmed as 'true positives'). South Wales Police figures indicate that rather than 60, there were 51 (non-police) matches generated by the system, due to a police officer being added to the Amber watchlist for the purposes of app testing.
- The operator logs indicate that, of these 60, 9% (N=5) were confirmed as 'true positive' matches by operators.
- Of the remaining: 65% (N=39) were false positives, 15% (N=9) had no result recorded against them, 3% (N=2) had an unclear outcome and 8% (N=5) were missing from the logs (see Figure 8).
- Analysis again included comparison of system scores and operator decisions. For the Anthony Joshua event, the limited data does suggest a relationship between increasing scores and 'true positive' confirmation decisions by operators (see Figure 9).
- Figure 9 also shows that of the matches for which operator decisions were recorded, the majority (around 43%) scored .58.

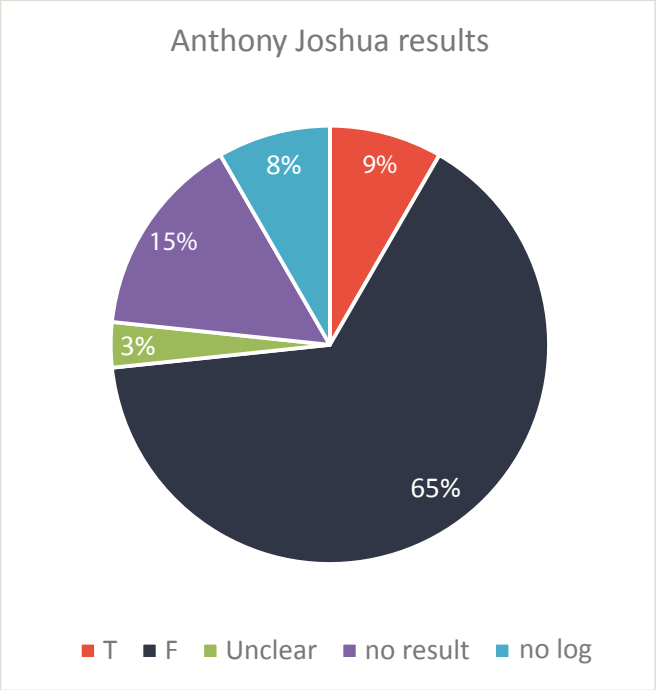


Figure 8 – pie chart of the true / false positives at Anthony Joshua fight (Oct 2017)

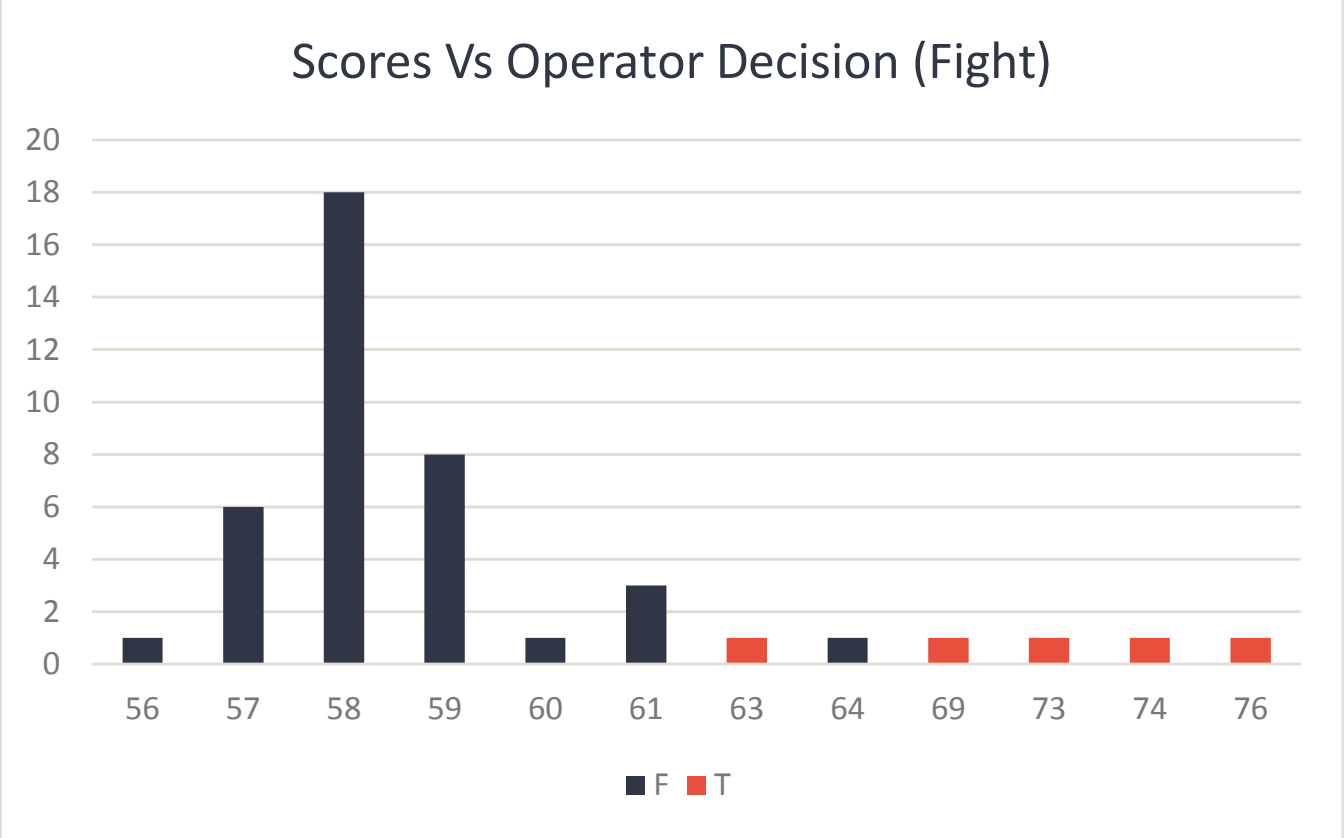


Figure 9 – bar chart showing relationship between match scores and operator 'true positive' confirmation decisions at Anthony Joshua fight (Oct 2017)

Consistent with the previous discussion, enrolment image quality improved significantly following the Champions League event. However, during the Anthony Joshua deployment, there was again one very frequent 'hitter':

- This person was female, and in total, she matched against nine different people (all false positive matches). This was in various camera positions and/or locations, and she also matched to both female and male images.
- Her enrolment image looked quite old and was not great quality, which may go some way to explaining the repeated matches.

Locate Outcomes: Operator Performance

The concept of operator performance concerns how operator decisions and behaviours impacted upon the outcomes achieved:

- After operators had made a decision on whether a match was 'true' or 'false', they were instructed to record it in logbooks provided. This was to allow for analysis of true and false positive rates. Logbooks were generic, not specifically designed for AFR, but operators were advised which information they should record (in-van document).
- Some operators were very thorough in their recording practices, noting all information about persons of interest – including date of birth, watchlist colour, name, offence, system score – and giving detailed information about how they had reached their decision.
- However, other operators provided much less detail, missing the decision taken, or not recording any information in the logs at all. This is problematic for both audits of deployments and the evaluation research.
- Observational data indicates that improvements in system performance and results by the Anthony Joshua deployment had a positive impact on operator confidence and attitudes towards the system.

Locate: M20 (New) Algorithm Outcomes

This section discusses outcomes associated with the second facial recognition algorithm used by South Wales Police, M20, also referred to as the 'Beta' algorithm and/or the 'new' algorithm. It thus covers the period between November 2017 until March 2018, the end of the evaluation period. During this time AFR was deployed in Cardiff at the Autumn rugby internationals, concerts at the Motorpoint Arena, and the Six Nations rugby matches. The headline indicators are provided in the Table below and listed in the following bullet points.

- Interactions between intervention teams and members of the public were observed as almost always amicable.
- The new algorithm performed better and more accurately than the original one.
- The AFR system continued to struggle in large crowds during Locate deployments, occasionally hitting 100% processor utilisation. The system does not take advantage of the four processors available on the Alienware laptops. Programming in the ability to use multiple processors ('multi-threading') would likely improve performance.
- When tested, the NEC mobile app frequently failed or was slow getting alerts to intervention teams, slowing down the intervention process. As a consequence, it was not implemented by South Wales Police.
- AFR Locate continued to struggle in low light. Once daylight was completely gone, the system failed (even with the improved algorithm).
- New AFR-specific logbooks were introduced for the Six Nations events. These improved the auditability of operator decision-making, but there was still a lot of missing data.
- Improved image quality on watchlists enhanced overall AFR performance, but an effect remains where some individuals appear as 'frequent hitters', suggesting they are 'lambs'.

	EVENT SIZE	MATCHES	TRUE POSITIVES	ARRESTS
EVENT				
Autumn rugby internationals (4 events)	large	91*	13 (14%)****	4
Kasabian Concert	small	7**	3 (43%)	1
Six Nations rugby (3 events)	large	48***	22 (46%)	2

*South Wales Police figures show this as 79, as a police officer was included in the Amber watchlist for app testing purposes.
**South Wales Police figures show this as 6, as a police officer was included in the Red watchlist for app testing purposes.
***South Wales Police figures show this as 49. It appears that in the data sharing process, one match was deleted.
****South Wales Police data shows 16 true positives at this event.

Locate: New Algorithm Outcomes

As a result of AFR Locate deployments between November 2017 and March 2018, South Wales Police made a number of arrests and other interventions. Updates were provided to the public via social media about these outcomes. Arrests per event were as follows:

- Autumn Internationals – 4 arrests
- Kasabian concert – 2 ordered to leave venue, 1 arrest
- Op Fulcrum (22/12/2017) – 2 arrests
- Op Malachite (23/12/2017) – 2 arrests
- Six Nations – 5 arrests

Across all deployments (excluding the Champions League), 107 females and 17 individuals from a black or minority ethnic background were matched to persons of interest by the AFR system.

Observational data suggests that South Wales Police had learnt from their previous experiences, and the use of Intervention Teams was becoming more effective and efficient. One of the Six Nations arrests involved an individual who had been initially identified via AFR Identify. Following a violent offence at a nightclub in late 2017, CCTV was sent to the AFR Identify team, who identified a suspect from their results. A warrant was then issued for his arrest, and when he walked down Queen Street in the city centre in February, AFR Locate picked him up. This was a sort of 'end-to-end' AFR case.

During the Kasabian concert deployment, AFR Locate intelligence resulted in two suspects being ordered to leave the venue, and one man was arrested for going equipped to steal (when searched, officers found he was wearing a women's swimsuit, often used by pickpockets to stash stolen mobile phones). By the next day, no mobile phones had been reported lost or stolen to the venue or the police, compared with previous concerts, where in excess of 20 missing phones were reported.[5]

As part of the adoption of the new algorithm system threshold scores were raised and the faces per frame / frames per second rate had been lowered. Both of these changes were designed to reduce the data processing 'load' on the system. This might have contributed to its improved performance.

Despite these changes, there were instances where the system struggled with crowds, with CCTV streams jumping and lags in alerts. When crowds were especially dense for around 15-20 minutes during the Scotland Six Nations game (as people exited the stadium and headed down Queen Street), though the system was technically still running, it had stopped 'white boxing' faces and was no longer detecting them. The likely cause of this, as had been the case earlier on in the day, was that one of the laptop's processors was at 100% utilisation. The system does not take advantage of the laptop's four processors (i.e. the system does not support 'multi-threading'). It therefore froze and was unable to process the live camera streams or display alerts. Reconfiguring the system to use multiple processors could potentially address this system performance issue.

Darkness continued to be a general problem during deployments with the new algorithm. During the Six Nations, it was assessed the new algorithm worked better in low light conditions, but stopped working in the dark.

5 https://motorpointarenacardiff.co.uk/news-and-alerts/successful-start-new-mobile-phone-theft-prevention-approach

Looking in a bit more detail at the results obtained by AFR Locate between November 2017 and end of March 2018 reveals:

- At the Autumn Internationals, there were a total of 104 matches suggested by the AFR system. Editing out those from the 'blue watchlist', this left 91 matches. Note that South Wales Police figures indicate that rather than 91, there were 79 (non-police) matches generated by the system, due to a police officer being added to the Amber watchlist for the purposes of app testing. The operator logs indicate that 14% (N=13) of these 91 were confirmed true positives by the operators.
- Of the remaining: 53% (N=48) were decided to be 'false positives', 8% (N=7) had no result recorded, 22% (N=20) were missing (no log) and 3% (N=3) had an unclear outcome (see figure 11).

These figures are summarised in Figure 11 below.

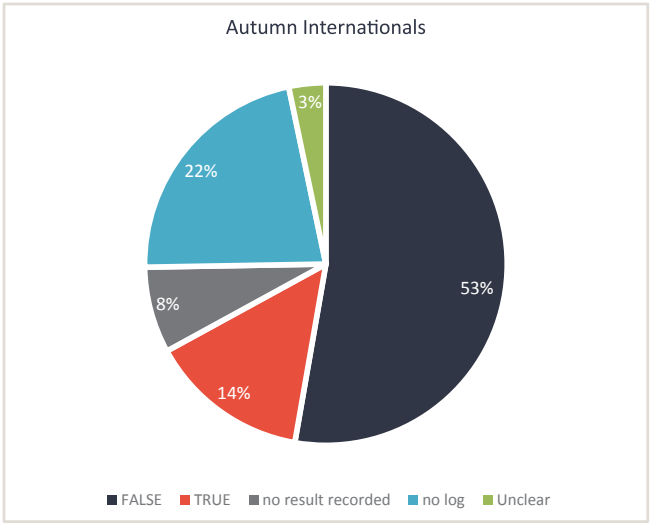


Figure 11 – pie chart of the true / false positives from all of the Autumn Internationals

- Across the three Six Nations deployments, there were a total of 56 matches suggested by AFR Locate. Again, removing 'blues' to focus only upon non-police matches, 48 were left. Note that South Wales Police figures indicate that there were 49 (non-police) matches generated by the system – it appears the lower number reported here is due to a match being lost in the data sharing process. 46% of these were judged 'true positives' by operators (n=22).
- Of the remaining matches: 46% (n=22) were false positives, and 8% (n=4) were missing data (see figure 12).
- The adoption of the new logbooks at this time prompted operators to elaborate on the initial 'yes' results and to note the outcome. These outcomes were recorded in a third column within logbooks, and figure 13 depicts these results.
- Of the initial 22 'true positives' (those recorded as 'Y' in the first column): 23% (n=5) were stopped, confirmed as 'true' and arrested; 32% (n=7) were stopped but turned out to be false positives according to Intervention Teams; 41% (n=8) had unclear/unsure outcomes; and 4% (n=1) no outcome was recorded by the operators.

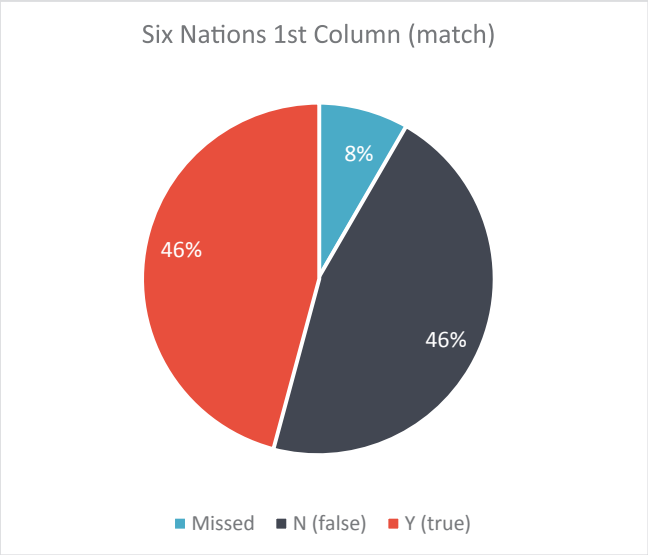


Figure 12 – pie chart of the true / false positives from all of the Six Nations deployments

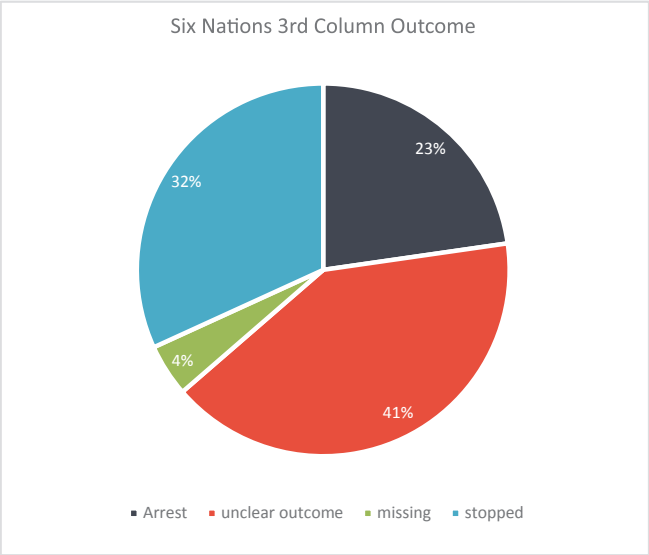


Figure 13 – pie chart of the outcomes of those noted as true or 'Y' in first column in operator log book (from all of the Six Nations)

As previously, a comparative analysis of system scores and operator decisions was conducted for when AFR Locate suggested a possible match:

- For the Autumn Internationals, as system scores increased, operators' decisions that matches were 'true' also increased (see figure 14). This figure also shows that of the matches for which there was an operator decision, the largest proportion (50%) were scored at .59.
- The Six Nations results again suggest a relationship between scores and operators decisions, as 'Yes' ('true') decisions increase with higher scores (see figure 15). This also shows that of the matches for which there was an operator decision, the largest proportion (48%) scored .59 or .60.

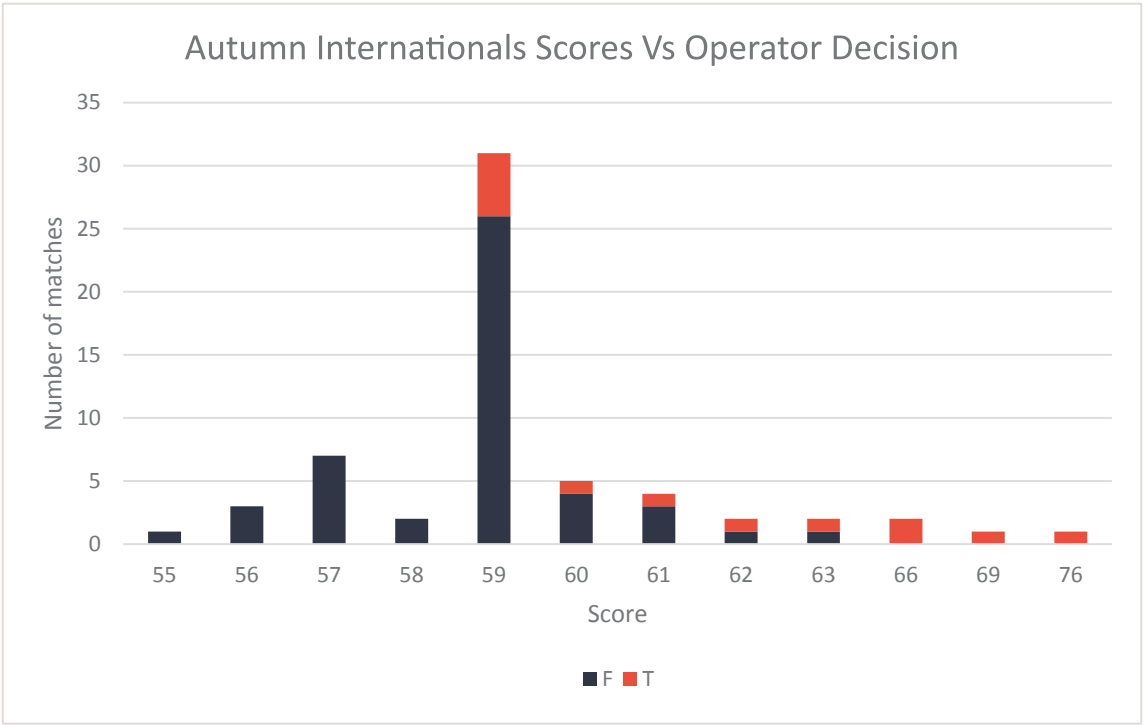


Figure 14 – bar chart showing relationship between match scores and operators decisions across the Autumn Internationals

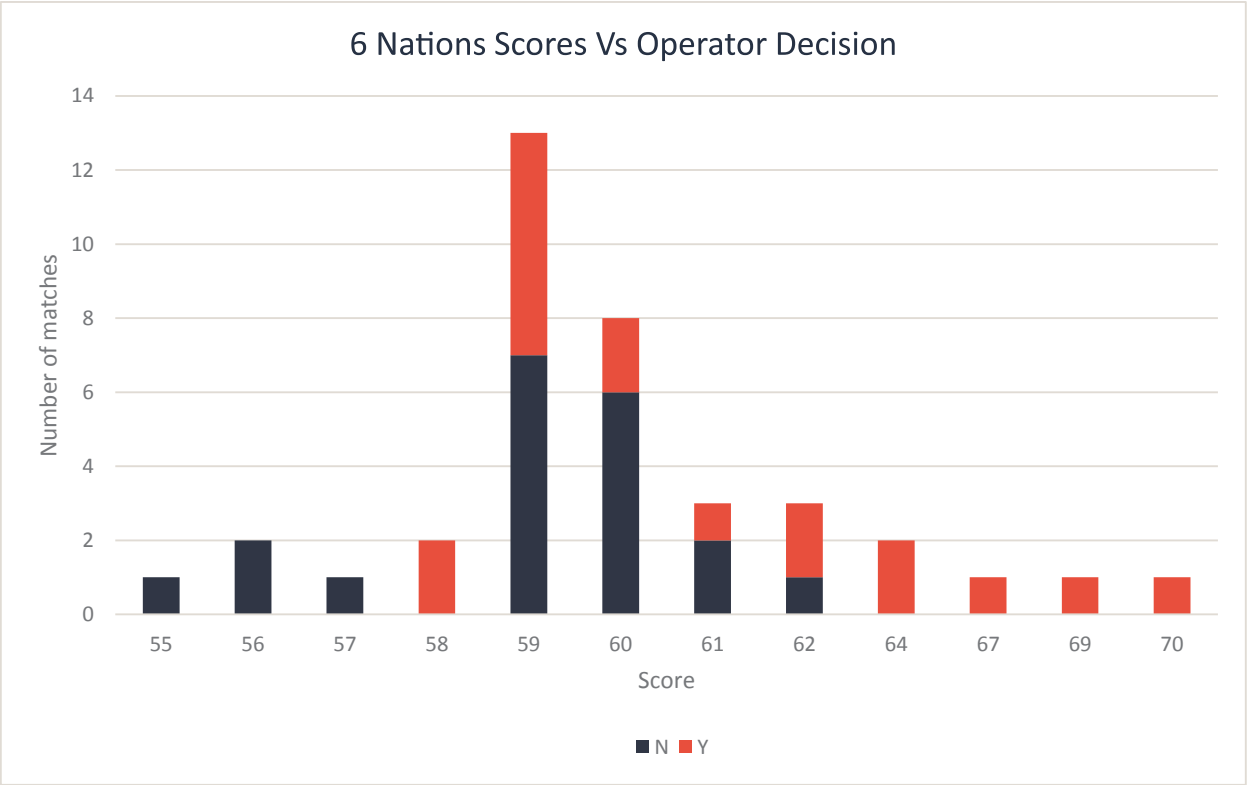


Figure 15 – bar chart showing relationship between match scores and operators decisions across the Six Nations

Although the new algorithm delivered superior performance compared with its predecessor, aided by improvements in the quality of watchlist images being used, there remained an issue with a small number of individuals generating multiple 'false positive' matches. Across the Autumn International deployments:

- Five people were wrongly 'matched' three or more times, with one woman matching 6 times (see figure 16).

The watchlist images responsible for generating the most 'false positive matches, were investigated across all the Six Nations and Autumn International rugby deployments. This revealed that:

- The most frequently matched image of Female 1, generated a total of 10 'false positive' matches.
- The images of these individuals do not appear to be particularly problematic in terms of standards or quality, and it may be that they are 'lambs'.

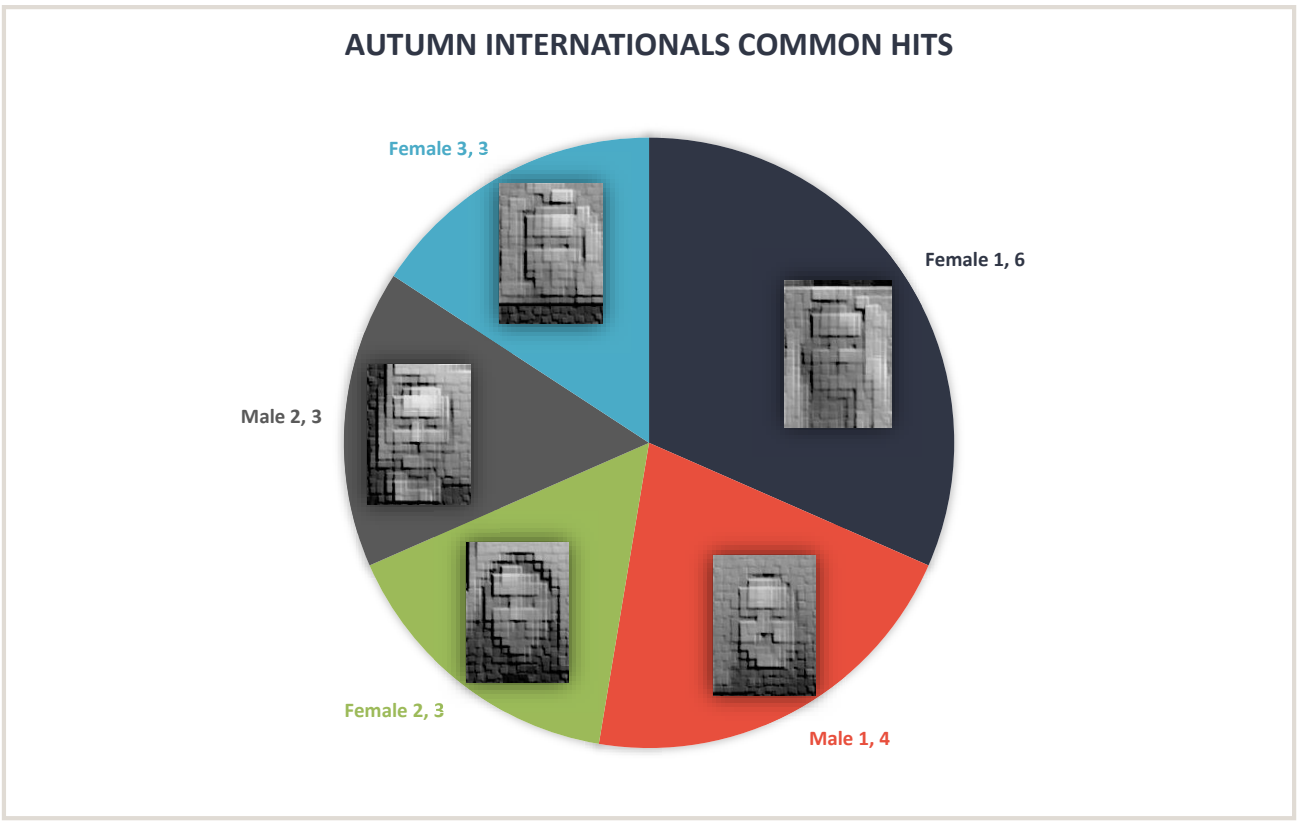


Figure 16 – pie chart showing the images of the most frequent 'hitters' across the Autumn Internationals

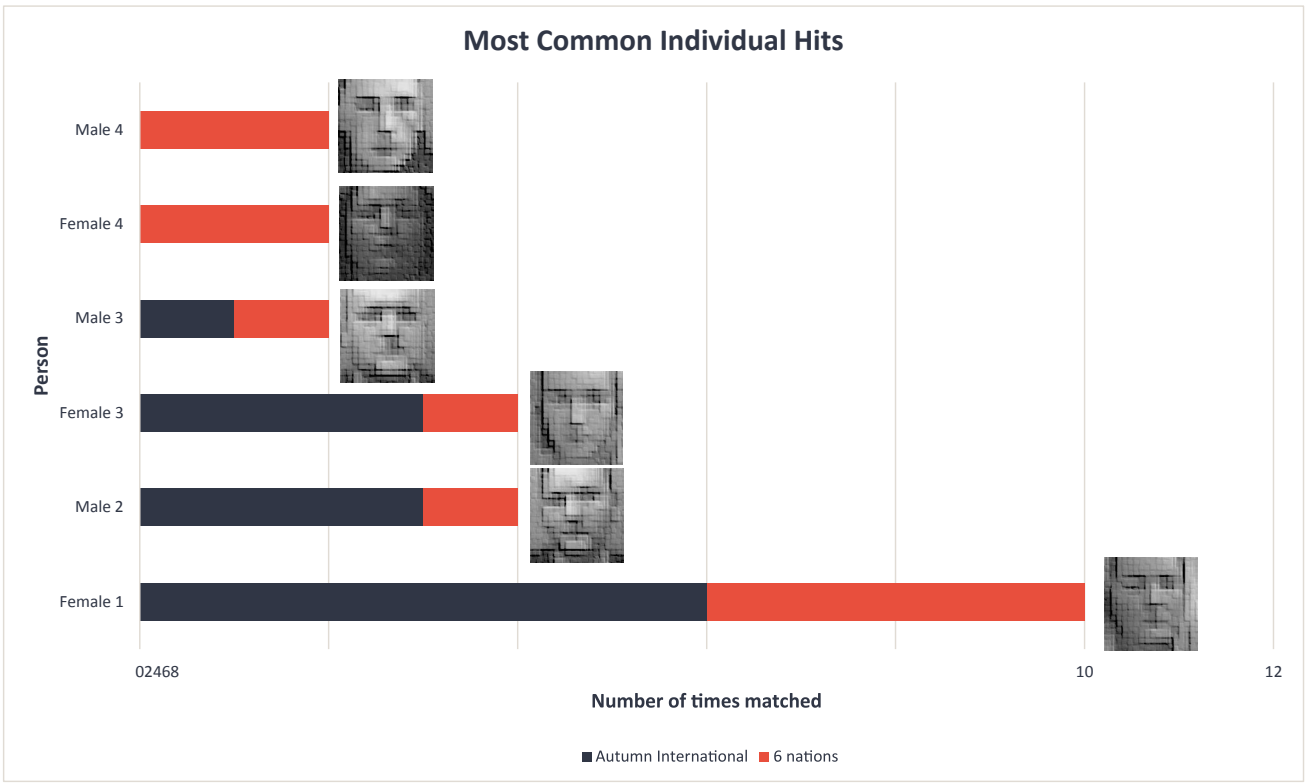


Figure 17 – bar chart showing the images of the most frequent 'hitters' who appeared across both the Six Nations (red) and the Autumn Internationals (blue)

SECTION 5: **AFR IDENTIFY PROCESS EVALUATION**

This section of the report discusses the process of implementing AFR Identify in South Wales Police across the whole evaluation period. It describes the work involved in operationalising the system, and the challenges and issues encountered in getting it to integrate with existing policing processes. The focus here is upon how the technology was adopted, how operators interacted with it, and the amendments and adaptations that had to be introduced (predominantly to the policing processes wrapped around it) in order to render it a useful policing tool. Within this analysis of process implementation, the organisational performance, system performance and operator performance themes are used to orient the discussion.

AFR Identify And Organisational Performance

In order to prepare for the roll out of AFR Identify, South Wales Police held training sessions for the Identify Team who would be using the system. These were different individuals from those involved in using Locate. This section describes this training, as well as findings from observations following roll-out.

A member of the UPSI evaluation team attended and observed the initial training for AFR Identify operators. The content was very similar to that provided for Locate. All participants were police staff and were existing members of the Identify Team. An NEC software engineer led the session, which lasted around 2 hours. The features of the system were explained, and operators were shown how to enrol images ready to search, how to know if the software detected a face, and what the results page would look like. They were told that 100 possible matches would be returned by the system (in reality, this was 200). Operators were also told they would be receiving a 'how to' document. Key terminology was introduced, including 'probe image' and 'enrolment'. Other potentially useful features of the system (such as using 'maybe' markers or 'flipping' images to view them from different angles) were demonstrated for reference during the session.

When AFR Identify went 'live', it was added onto the team's existing workload. In terms of the submissions received from operational officers, theft was the most common offence type with 71% (N=1519) of total submissions. Other offences include:

- Assaults (including common assault, serious assault): N=80
- Sexual offences (including rape, sexual assault, sexual communications): N=37
- Robbery (including armed): N=31
- Hate crime/racially aggravated: N=7
- Murder: N=1
- Kidnap: N=1

Within a couple of months of AFR Identify going live, the Identify Team were training as many South Wales Police officers as possible on how the new technology could assist during their investigations. This included inputs about how they could send requests to the Identify team, and what 'good' and 'bad' probe images looked like.

It rapidly became apparent, however, that a significant proportion of the images being submitted by officers were not of sufficient quality. The team were finding that many probe images submitted to them from across the force, were stills from CCTV displayed on screens and then photographed using the officer's mobile phone. The latter being submitted as the probe image. Operators established how the quality and resolution of the probe image had a significant impact upon the system's performance. As a consequence, by November 2017, operators had been talking directly to Police and Community Safety Officers about the issue (as they were often the ones dealing with those aspects). Operators tried to emphasise the importance of them obtaining good quality images, preferably copies of original CCTV footage, rather than taking pictures of CCTV screens on mobile phones.

During a later visit in December 2017, members of the Identify team also discussed their input into custody detention officers' (CDO) training. The focus of this training was on taking good quality custody images, and was informed by learning from 'what works' with AFR. The Identify team had pushed to be involved in this aspect and were hoping that investing their time in this way to improve custody images would yield 'downstream' benefits for them going forward.

AFR Identify And System Performance

The process for uploading probe images and setting up AFR Identify searches was relatively straightforward. It sometimes involved changing the number of pixels between the eyes on an image to ensure a face could be detected (depending on the probe image), but it was relatively easy to make such adjustments.

There was no ability to filter on gender in the old S17 algorithm. Operators said they pushed NEC to add this feature, and it was available in the M20 algorithm. This was deemed to be important as when the system returned database images that were clearly not the same gender as that of the individual depicted in the probe image, this negatively impacted operators' confidence in the system. Further discussions revealed that the new algorithm seemed more sensitive to image quality. Operators said that this could be unpredictable, and sometimes images they thought were good quality were not accepted, and images they thought would be too poor were accepted. When this happened, it was frustrating and confusing for them, but is a common problem when using 'black box' products.

In framing this evaluation considerable emphasis was placed upon understanding AFR as a socio-technical system, rather than just as a technological solution. This is exemplified in the evidence collected from observing the work of the Identify Team. While using the old algorithm, operators tended to set the similarity score lower than the usual .55 minimum that had been used by AFR Locate. This was because in their experience, matches could sometimes score very low (e.g. one was 197th out of 200 results). When the new algorithm was introduced, operators found it was much better and tended to start with a score of around .70 in anticipation of finding strong results. They understood they could lower the score if they needed to expand their search parameters, in order to have additional images returned to them by the system.

AFR Identify And Operator Performance

In terms of organising the AFR Identify Team's work, one member per day was tasked with performing AFR related functions. This typically included checking the inbox for new submissions and taking action on any requests that had been submitted. The process was as follows:

- Operator looks at request form on their workstation and saves probe image onto USB;
- Operator takes USB to Alienware laptop and uploads probe image onto AFR system;
- If the probe image is of sufficient quality for a face to be detected, the system 'enrols' the image (if not, the image is rejected);
- Results are checked for possible matches, comparing images side-by-side and looking in detail;
- If unsure, the operator notes the occurrence numbers from any potential matches, and returns to own workstation to check police systems for location, previous offences and so forth, to help inform decisions.

During observation in October 2017, operators said that typically, requests take around an hour to complete (start to finish). In assessing whether a returned custody image was a 'true positive' compared with the probe image, the following process was enacted by Identify operators:

- They focus on the facial structure of the people, and look at distinguishing features like hairline, nose, eyes, ears, jawline, facial hair shape, scars, tattoos etc.
- Operators also used their own experience and background knowledge to recognise faces and inform their decisions.
- When using the old algorithm, operators did not have a high level of confidence in the system. They reported using the score as a sort of tool to help make a judgement when they found a possible match. They did this by starting with a low score on the initial search, and when they found a match they thought was 'possible', they put the score up and re-ran the search to see if the same image was returned again. Several operators reported conducting several iterations of this process, increasing their confidence as the target image re-appeared in the results.
- When operators began using the new algorithm however, this informal practice changed. They would start with a high score in the hope of finding strong possible matches straight away. If they were unable to find any matches with a high score, they would then gradually lower the score to expand the number of images returned.
- This expectation that strong possible matches would probably be found straight away was linked to their increased confidence in the system.

SECTION 6: **AFR IDENTIFY OUTCOME EVALUATION**

In this section of the report, the discussion is centred around an assessment of the outcomes achieved by South Wales Police through the deployment of AFR 'Identify'. As previously, the discussion is organised by separating out organisational, operator and system level influences.

S17 (Old) Algorithm Outcomes

Between July and October 2017, the S17 algorithm supplied to South Wales Police was used for AFR Identify. Summary data from this period are set out in the Table and bullet points below:

S17 'OLD' ALGORITHM		
	Total	**% of total**
Number of requests	612	100%
Matches generated by the system and confirmed by operators (AFR1)	107	17%
Matches generated by the system but disconfirmed by operators (AFR2)	95	16%
Rejected by the system (poor quality) but circulated to officers (AFR 3)	337	55%
Rejected by the system and too poor quality to be circulated to officers (AFR 4)	73	12%

- A small number of arrests resulted from AFR Identify, although the precise figures are unknown to the evaluation team.
- Operators for AFR Identify were generally optimistic about the system.
- AFR Identify searches: when probe images were of a good enough quality to be accepted by the AFR system, 51% resulted in 'possible matches'.
- Over time, Identify became more 'business as usual' and well-known.

Identify Outcomes: Organisational Performance

This section provides a brief summary of arrests and feedback on AFR Identify between July and October 2017.

Arrest and charge figures relating to the old algorithm were not obtained during this evaluation (as the Identify team do not always receive feedback on their cases, their records of requests and AFR results categorisation do not include outcomes). However, by mid-December 2017, there had been over 50 charges using AFR Identify, many of which came prior to the new algorithm being rolled out.

As part of the evaluation, the research team interviewed an officer (Detective Constable) who had used AFR Identify to assist in one of his investigations. In September 2017, there was a series of eight burglaries in three days in an area of Cardiff, and the same male was wanted for all of the offences. Having heard about the new AFR capabilities, the investigating officer sent some CCTV stills over to the Identify team, who were able to return a possible 'match' image to him on the same day. The man was originally from the Swansea area and so was less likely to have been 'known' to officers in Cardiff. The perpetrator did leave DNA evidence on a window at one of the scenes, but results from that would have also taken some time to come back (3-4 weeks). Using the intelligence from AFR Identify officers were able to make an arrest the following day. Potentially, this rapid intervention may have prevented further crimes from being committed as part of an ongoing series.

Identify Outcomes: System Performance

A number of indicators of system performance for AFR Identify were derived. These show:

- A total of 612 requests for AFR Identify were completed using the old algorithm between the 27th of July and the 22nd of October, 2017.
- Of these, 107 had matches suggested by the system and confirmed by operators, 95 had matches suggested by the system but disconfirmed by the operators, 337 were rejected by the system (poor quality) but circulated to officers and 73 were both rejected by the system and were too poor quality to be circulated to officers. This data is depicted in the line graph below (see Figure 10).
- Of the images that were clear enough for the facial recognition software to analyse them (combination of matches suggested by the system that were either confirmed or disconfirmed by operators), 51% generated possible suspect matches as reviewed by the operators (match suggested by system and confirmed by operators).
- Each time operators ran a search, the system returned 200 results, which were then reviewed by the operators.
- The system did not always handle images taken side-on or at an angle very well, which CCTV images often are (due to the height of cameras), and operators felt that the effects of this could sometimes be seen in the results.
- As has been noted, scores were sometimes very low (e.g. results from one search seen by the research team all scored under .50).
- Operators also mentioned a 'lamb' who appeared "all the time". This was a man who had some deformed features, and operators said it was as if the system was "unable to read his face properly", hence his frequent appearance in results.

Figure 18 below summarizes these data (note data from July have been excluded from this analysis as there were very few results (9 in total), they were not business as usual, and they significantly changed the trend of the results).

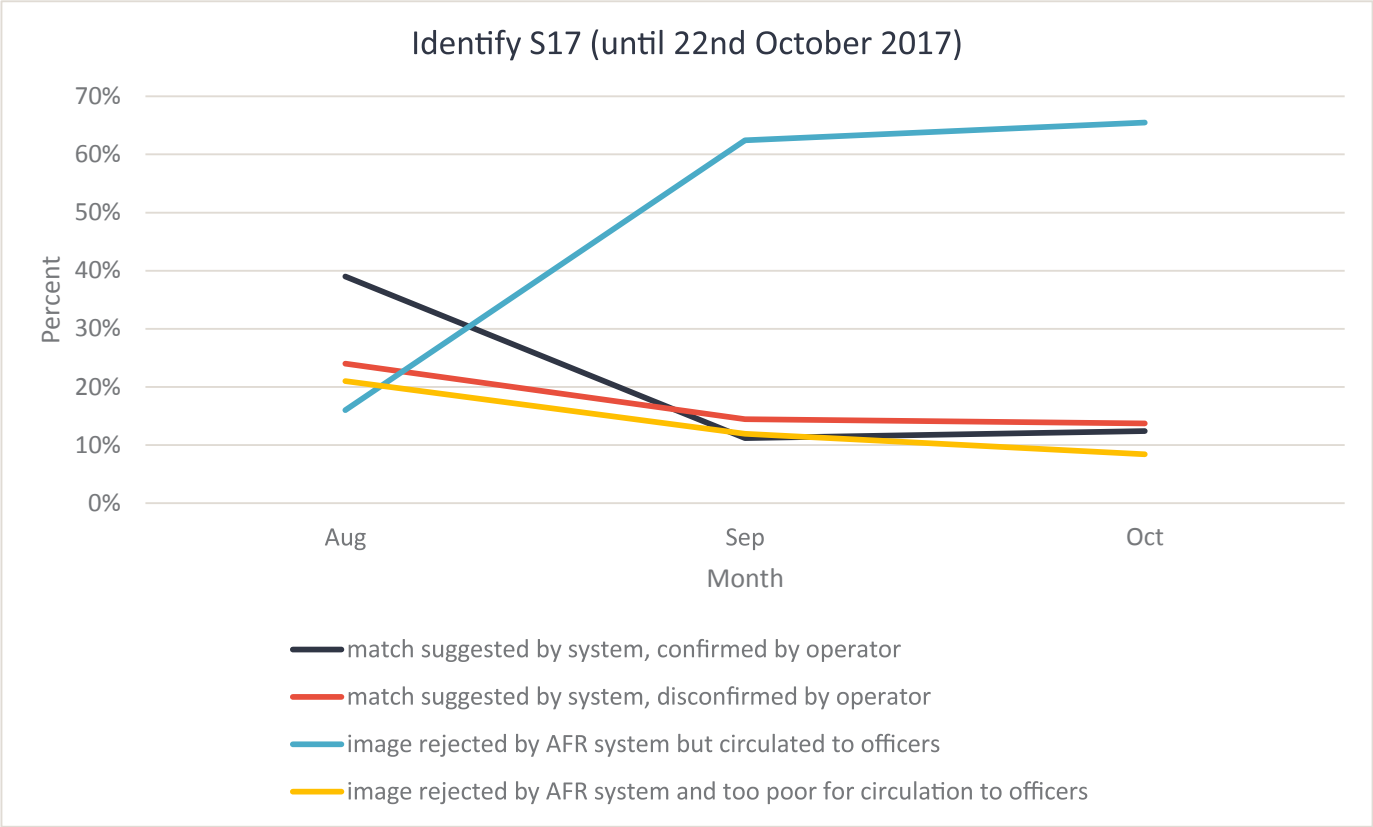


Figure 18 – line chart of AFR Identify results from August until 22nd October 2017

Identify Outcomes: Operator Performance

As indicated above, operator decision-making is a key determinant of overall outcomes achieved by AFR Identify. For each 'probe image' submitted to the system, AFR Identify returns 200 images of possible matches it has generated from the watchlists it is storing. The operators are required to select which of these images, if any, could be the individual depicted in the probe:

- In trying to confirm or refute possible matches, officers were observed conducting their own background research into the possible match(es).
- Operators did not always receive feedback on outcomes resulting from their work (such as arrests), which would help them gauge the usefulness of the intelligence they are providing to front-line officers.
- Operators' views on the system were optimistic, and AFR was seen as possessing a potential to become an important component of police suspect identification processes in the future.

Identify: M20 (New) Algorithm Outcomes

At the end of October 2017, South Wales Police were supplied with the M20 algorithm (also referred to as the 'Beta' algorithm and/or the 'new' algorithm) by NEC. Thus, the outcomes associated with the use of this algorithm for AFR Identify between the end of October 2017 and March 2018 (the end of the evaluation period) are discussed in this section. The headline indicators are provided in the Table below and listed in the following bullet points.

- Over 100 people charged following use of AFR Identify

IDENTIFY: M20 (NEW) ALGORITHM OUTCOMES

	Total	% of total
Number of requests	1524	100%
Matches generated by the system and confirmed by operators (AFR1)	333	22%
Matches generated by the system but disconfirmed by operators (AFR2)	136	9%
Rejected by the system (poor quality) but circulated to officers (AFR 3)	916	60%
Rejected by the system and too poor quality to be circulated to officers (AFR 4)	128	8%

between July 2017 and April 2018 (figure taken from board meeting presentation, 20th April 2018).

- For AFR Identify searches: when probe images were of a good enough quality to be accepted by the AFR system (combination of matches suggested by the system that were either confirmed or disconfirmed by operators), 73% generated possible suspect matches as reviewed by the operators (match suggested by system and confirmed by operators). This is an increase of 20% compared with the old algorithm.
- AFR Identify results are ranked on score, and 90% of 'possible matches' according to operator judgements are located within the top 10 results returned by the system.

Identify: New Algorithm Outcomes

AFR Identify outcomes to May 2018 indicate 675 'possible matches' to suspects have been generated by the system, with over 350 of these having led to charging decisions. In addition, there has been increasing use of the technology in innovative ways. This includes a suspicious death case where an unidentified body had been found in suspicious and unusual circumstances. Officers at the scene took a photo of the deceased person and sent it to the AFR Identify team. The AFR Unit were able to make a match and knowing the identity of the man, the investigating officers were able to look into his background, which helped make sense of the circumstances of his death. This case (minus the detail) was later reported by the mainstream media.[6]

In order to evaluate system performance, the research team again collated results from the Identify Team and observational data:

- In total, 1,524 submissions to AFR Identify were completed using the new algorithm between the 23rd October 2017 and the 19th March 2018.
- Of the requests during this time, 333 had matches suggested by the system and confirmed by operators, 136 had matches suggested by the system but disconfirmed by the operators, 916 were rejected by the system (poor quality) but circulated to officers and 128 were both rejected by the system and were too poor quality to be circulated to officers. This data is depicted in the line graph below (see figure 19).
- As the trend lines show, over time: the proportion of AFR matches confirmed by operators increased; the proportion of images rejected by the system but circulated to officers decreased; the proportion of matches suggested by the system but disconfirmed by operators slightly increased; and the proportion of images rejected by the system and of too poor quality for circulation to officers slightly decreased. This suggests organisational learning about how best to use the system
- Of the images that were good enough for the facial recognition software to analyse them (combination of matches suggested by the system that were either confirmed or disconfirmed by operators), 73% generated possible suspect matches as reviewed by the operators (match suggested by system and confirmed by operators).
- This was an increase of over 20% compared to the old S17 algorithm and is probably attributable to a combination of an improved algorithm interacting with enhanced probe image quality being submitted.

From mid-February 2018 onwards, operators documented another layer of data, recording the rank position of the matches

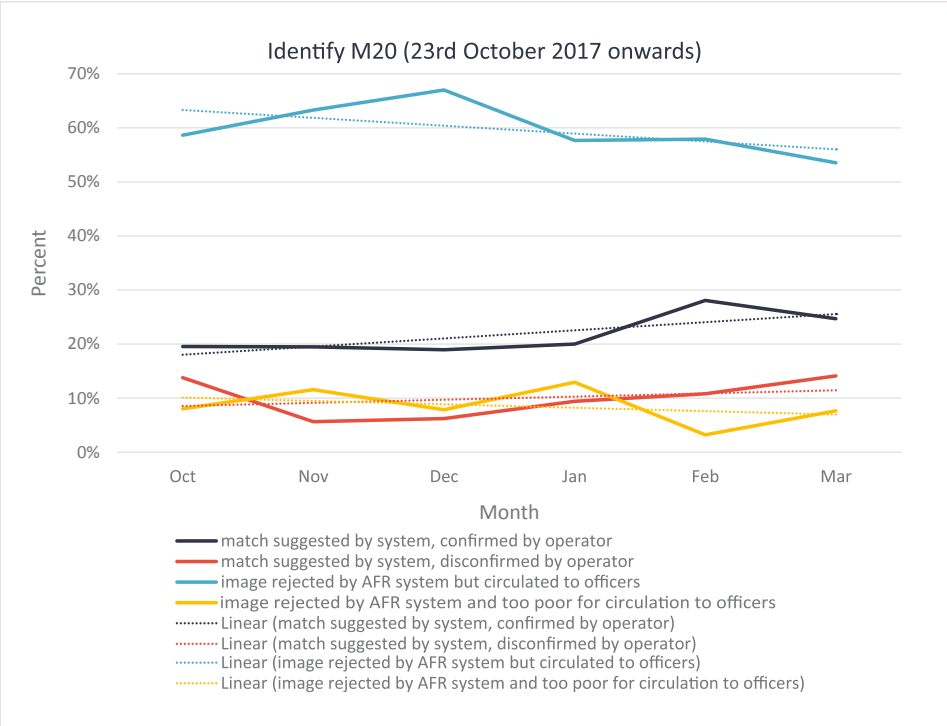


Figure 19 – line chart of AFR Identify results from 23rd October 2017 to 19th March 2018

6 https://www.walesonline.co.uk/news/wales-news/facial-recognition-software-used-police-14054626

suggested by the system and confirmed by them (also termed 'possible matches'):

- As the waterfall chart below (Figure 20) shows, in 60% of cases the possible suspect match image was returned ranked number 1 out of 200 by the system.
- Overall, 90% of AFR matches confirmed by operators appear in the first ten images as ranked by the system.
- Concerns have been raised regarding operator fatigue as they search through 200 results. Rather promisingly, these results suggest that if operators were pressed for time, nine times out of ten they would be able to find a possible match (match suggested by the system and confirmed by operators) within the first ten results.

Of note is that the previous algorithm did not handle side-on or angled images very well, and it was hoped that the new algorithm would be more effective. However, operators did not feel that they had seen a change in this respect. They did suggest however, that they were seeing fewer 'lambs' or frequently matched individuals compared with the previous algorithm.

An important element of operator performance was in maintaining the quality of probe images. Although there was a decline in the number of inappropriate images being submitted, AFR Identify team members still expended considerable time and effort in reviewing and rejecting large numbers of submissions from front-line officers.

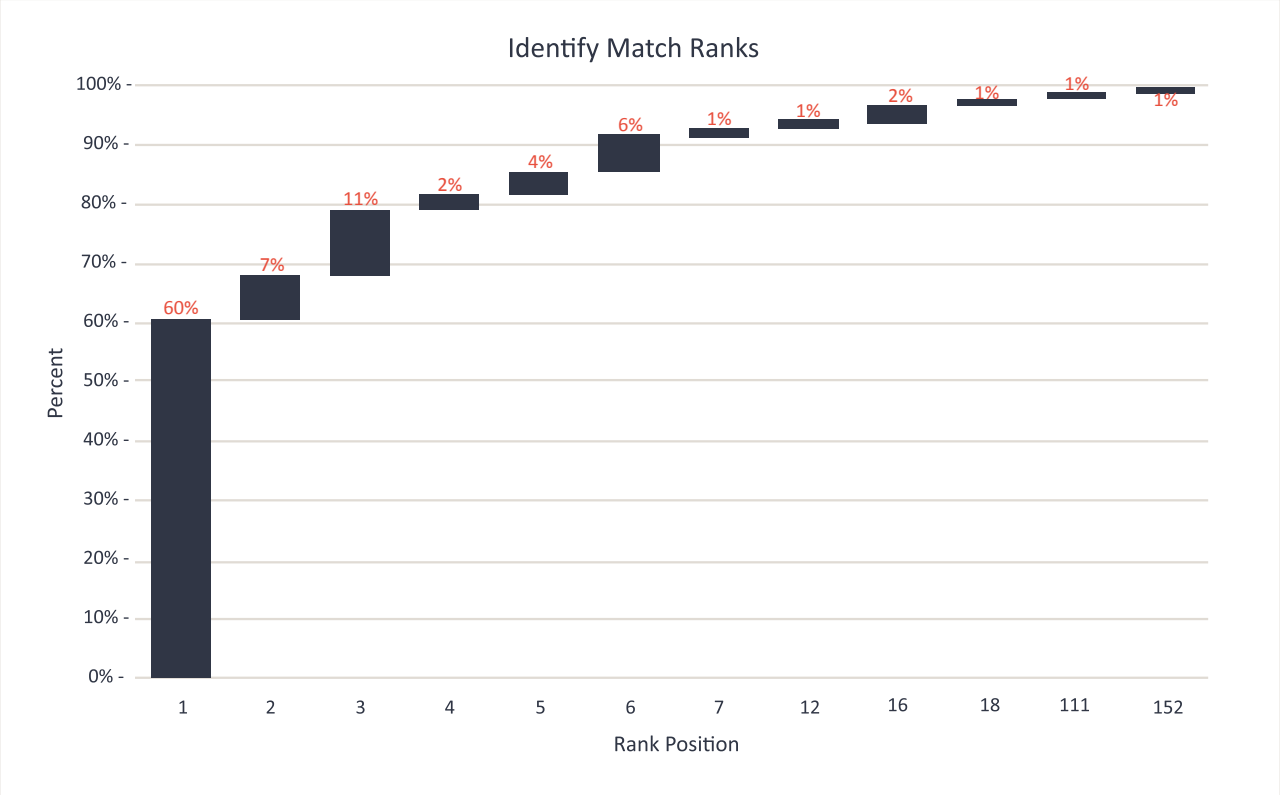


Figure 20 – waterfall chart of the rank position of the AFR Identify matches

SECTION 7: **UPSI FIELD TRIAL**

The previous sections have documented the kinds of outcome delivered by AFR Locate and Identify. It has been noted that there are a number of interacting dependencies that shape the performance of AFR in terms of generating policing outcomes. These complexities make it hard to establish just how accurate the system is in terms of its ability to identify individuals in real-time and match them to images stored on a watchlist. Accordingly, as part of the evaluation study, an attempt was made to test the system in a relatively controlled way.

This involved researchers enrolling images of themselves on a watchlist and then deliberately passing through the 'field of vision' of the AFR camera to see if it matched them. This experiment was conducted as part of the Wales v Italy Six Nations game deployment on the 11th Mach 2018. Albeit conducted on a limited budget, the idea was to determine what factors (e.g. walking speed/ direction, clothing and image quality), impact upon the AFR platform's ability to match an individual to their watchlist image. Due to the small scale of this trial, robust deductions regarding accuracy cannot be made.

Headline Field Trial Findings:

- Match rate of 76%+
- Custody images matter: one bad custody image included, and that volunteer was missed on many occasions by the system.
- Glasses and hats had some negative impact on system score and the ability of the system to detect faces. Specifically, problems with a baseball cap pulled down quite far (casting a shadow over the face) and glasses that obscured the eyes were observed.
- Seasonal variation factors may thus impact the effectiveness of the system. For example, people are likely to wear sunglasses and caps in the summer, while they might pull their scarves up high or put their hoods up during the winter.

Methods

The trial involved seven volunteers having their 'custody' photographs enrolled onto the AFR watchlist for the event. Each individual then carried out multiple 'walk-throughs' in the areas covered by the cameras under various 'treatment' conditions. As a result, there were over 90 'walk-throughs'. The 'treatment conditions' included:

- Different walking speeds (stroll, power walk, jog);
- Different walking directions (head on and diagonal to the camera);
- Group walk-through;
- 'On phone' stroll (head at an angle);
- Props: winter scarf; winter hat; glasses; baseball cap; headscarf (for women) / moustache (for men).

Each volunteer completed 'walk-throughs' under every condition, recording the time on a log. These logs were then compared to the AFR system CSV file post-event, in order to identify if people were picked up, or missed, and how they were scored by the system.

Results

The match rate was very high – between 76% or 81%. There is a degree of variance in the figures because, under the first test condition, 'head on stroll', there was an unusually high number of 'misses'. This was unexpected, as we anticipated this condition to produce some of the strongest results (given the clear facial view and slow speed), and so are unsure why the system missed individuals. If data from this 'head on stroll' condition is included in the analysis, the match rate is 76%, whereas if it is disregarded, it rises to 81% accuracy.

The 'group walk-through' (also referred to as the '5+ stroll') condition was also included in order to test how the system would respond to multiple watchlist individuals being in the frame at the same time. We were unsure if it would be able to pick up all seven individuals or if it would just lock onto the first five it detected. Under this condition, the system successfully identified all seven volunteers and did not 'lock' on to anyone in particular.

One of the volunteers had a very poor 'custody' image. Though the image was taken on a modern mobile phone and therefore had good pixel quality, the volunteer was blinking and trying not to laugh, meaning his facial features were distorted. As a result, this volunteer was picked up the least on his 'walk-throughs' and ended up with a 50% miss rate. Others who had much higher match rates also had much better 'custody' images.

Figure 21 depicts the impact of walking direction on system scores for 5 volunteers. We were expecting to see higher 'head on' scores (given that would mean a clearer view of the face), but the results are mixed. For some volunteers, the score is higher when walking head on towards the camera, while for others it is higher when walking diagonally. It is not clear why the results are patterned in this way, but does suggest that there is no direct relationship between walking direction and score.

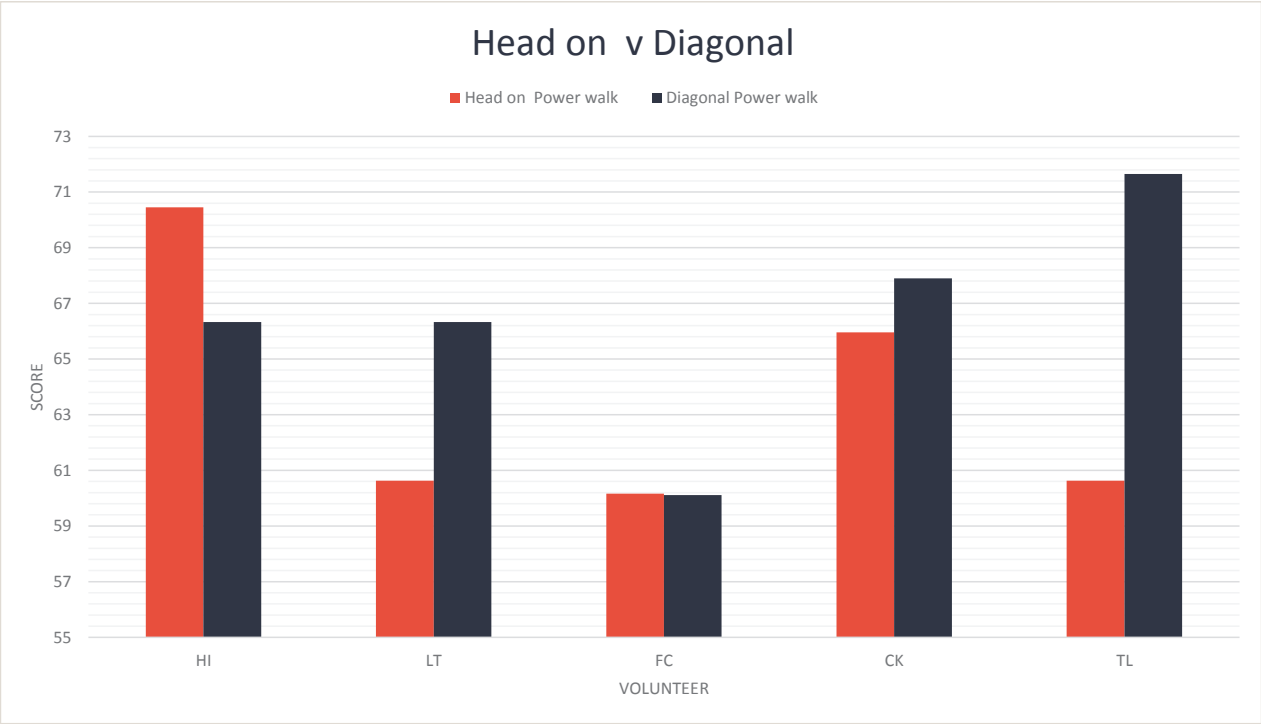


Figure 21 – bar chart showing the impact of walking direction on score

We were very interested in the impact of 'props' on the system. These were common accessories that people would likely wear at various times of the year. In particular, two different types of glasses were worn by different individuals: one 'normal' pair and one set that deliberately obscured the eyes and bridge of the nose. Additionally, a false moustache was used by male volunteers and a headscarf for the women, were also included. The moustache was used in order to explore the impact of a change in facial hair for men, which is likely to happen over time. The headscarf was included as this is now quite a common item of clothing, and volunteers wrapped this loosely around their head, covering their hair.

Figure 22 shows how the average scores generated by the system change by each prop type. So the cap and glasses appear most likely to 'suppress' the system score. The cap obscured the forehead casting a shadow on the rest of the face (if pulled down far enough, so the system was completely unable to 'read' the face). The glasses (especially the larger pair) obscured the eyes.

Each volunteer's average scores with and without props were compared. As might be predicted, for most, their average score was higher when picked up on camera without props (see Figure 23). There were a couple of outliers to this trend

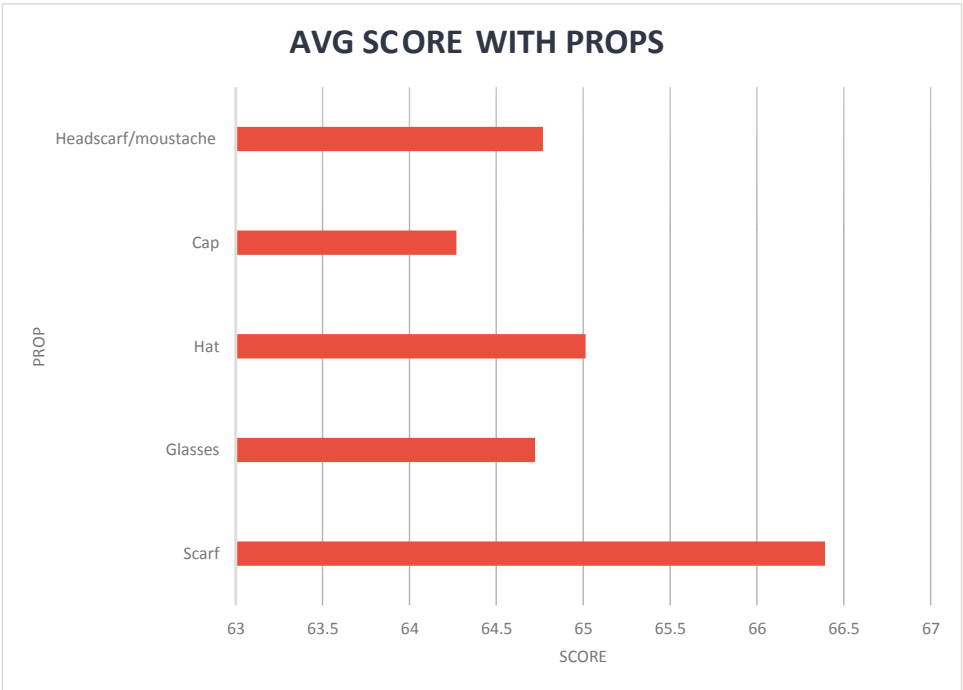


Figure 22 – bar chart showing the variance in score with each prop

however (AD and TL), whose scores were higher with the props. It is unclear why this happened. Overall, the data depicted in both Figures 22 and 23 clearly indicate that certain items of clothing can potentially impact system accuracy scores. Other data collected suggest that when people were jogging past the cameras their 'scores' tended to be lower also. The practical implication of this concerns setting the right threshold at which 'alerts' for possible matches are triggered by the system.

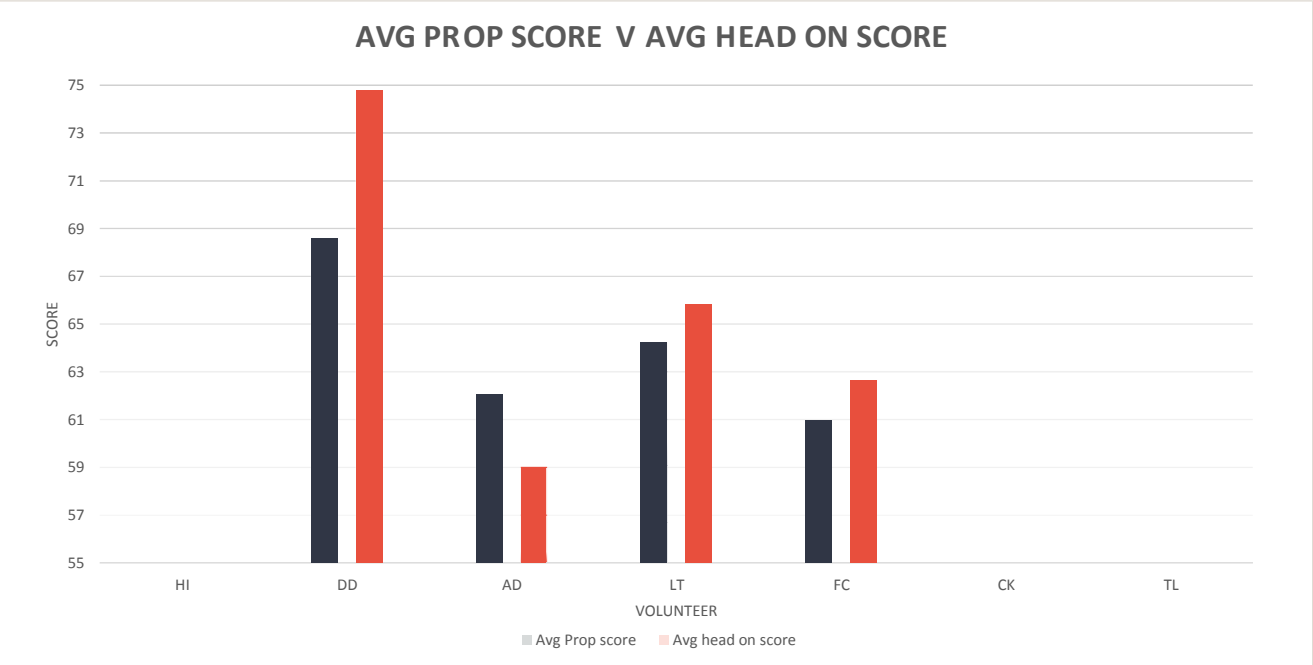


Figure 23 – bar chart showing the average scores with and without props for each volunteer

Interpretation And Limitations

Overall, the results from this small-scale embedded field trial indicate that the AFR platform is quite accurate in terms of its ability to detect faces of individuals in real-time and match them to images on a watchlist. A match rate in excess of 76% was achieved. Testing the system in this way allowed the evaluation team to offset some of the inconsistent logging issues by operators seen during the evaluation period.

The results also confirm the importance of using good quality custody images, in terms of image type and resolution. It appears that non-custody images (such as surveillance photographs or from social media) negatively impact the system's accuracy. The findings also indicate that clothing items can significantly impact scores and match rates where they obscure facial features.

Due to the limited funds for this trial, conditions such as ethnicity, low light, weather and controlled environmental conditions were not tested. Some of these have caused issues in the past, particularly low light and uncontrolled environments. We suspect that differences would be observed if the system were tested under these different conditions, and given the opportunity, would recommend exploring them in further detail.

As a result of this trial, we have also become curious about the impacts of 'watchlist size' (the number of people on watchlists) and composition (gender / ethnicity / age balances) on match rates. It would be interesting to explore what impacts changing the make-up of watchlists might have on results.

SECTION 8: **ETHICS, PUBLIC PERMISSION AND MEDIA COVERAGE**

As part of the evaluation, it was agreed that we would examine aspects of 'public permission' for this new technology in terms of citizens' reactions and responses to it. In part, this reflects South Wales Police's policy decision to proactively advertise their use of AFR technology to the public.

Based on observational data, it seems the public response to AFR was broadly positive, though slightly mixed. Most of the time, people passing-by the AFR vans appeared unfazed and unconcerned by their presence. On a number of occasions where operators had left the van door open, members of the public stopped to look inside, seemingly curious about the technology and engaging in brief conversations with the operators.

Furthermore, for the most part, interactions with those who were stopped by police intervention teams on the basis of the real-time 'intelligence' generated via the AFR system but turned out not to be persons of interest, were amicable. In these situations, operators / officers fully explained the exercise being carried out, and the individuals were invited into the vans to see the software for themselves and to see their own CCTV image alongside the 'match'. They were also thanked for their time. This approach helped alleviate any awkwardness. Some people also joked or commented that they agreed with the resemblance to their images.

There were however a small number of occasions where interactions with individuals who had been (wrongly) stopped were less positive. These interactions were not observed by researchers, but were described briefly by operators who had been present:

- Following a match, officers stopped a woman on the street. The woman on the watchlist was around 50, but the woman stopped turned out to be in her early 20s. She was upset that she'd been mistaken for the woman on the system because of the difference in age and looks.
- A man was stopped following a match. It turned out not to be him, and he went on his way. However, he later came back to the van worried and asking if it was likely that he would be stopped again repeatedly.

Ahead of the Champions League deployment, a major press release was made by the force. This was published on the 22nd May 2017 on South Wales Police's website and on their main Facebook page. It included quotes from Inspector Scott Lloyd, Assistant Chief Constable (now Deputy Chief Constable) Richard Lewis and Alun Michael (the Police and Crime Commissioner for South Wales). It described the use of NEC's 'Real-Time' solution and included pictures of the AFR vans, showing the cameras mounted on top and the AFR signage on the sides. The release also included some details about the industry partner, deployment dates, general locations, the benefits of the technology, and the privacy and legitimacy considerations being made. The response to this Facebook post was limited in scale and mixed in content. Some comments made reference to '1984' and the 'Orwellian' nature of these kinds of policing techniques, whist others were more positive. Questions were also raised about things like future uses and data storage.
Between April and June 2017, multiple media articles were published reporting the use of the technology at events. The sentiments of these articles was mixed, as was the response to them (though there were relatively few responses on the whole). Examples include:

- Wales Online, 22nd May 2017
- https://www.walesonline.co.uk/news/wales-news/police-say-facial-recognition-software-13070455#comments-section
- Engadget UK, 26th April 2017
- https://www.engadget.com/2017/04/26/british-police-to-scan-faces-at-champions-league-final/#comments
- Daily Mail, 23rd May 2017
- http://www.dailymail.co.uk/sciencetech/article-4533404/Champions-League-final-fans-faces-scanned.html#comments
- Wales Online, 2nd June 2017
- https://www.walesonline.co.uk/news/local-news/first-arrest-using-facial-recognition-13126934#comments-section

With the roll out of the new algorithm in November 2017, interest from the public, civil liberties organisations and the media continued to grow. There was also a large press release in partnership with the Motorpoint Arena in Cardiff in December 2017. This described the new facial recognition technology that had been installed at the arena as part of a partnership with South Wales Police, designed to tackle mobile phone theft and pickpocketing at concert events including Kasabian and Liam Gallagher. Outlets reporting deployment updates included BBC, ITV, The Guardian, Wired and WalesOnline. This coverage was generally mixed throughout. Mainstream media tended to report neutral updates on the technology, where other outlets (for example The Canary and The Register) were more negative about its use by police. See:

- The Register, 5th February 2018
- https://www.theregister.co.uk/2018/02/05/south_wales_police_facial_recognition_arrests/
- The Canary, 29th March 2018
- https://www.thecanary.co/uk/2018/03/29/south-wales-police-under-fire-for-using-facial-recognition-technology-against-protesters/

Social media updates issued by South Wales Police, particularly via Twitter, were a regular feature throughout the evaluation process. Typically, these generated very limited online reaction. Some people responded by saying that Locate is an excellent use of new technology, some criticised the focus of the police (referring to other current news stories), and others raised privacy and data retention concerns.

Over time, there were increases in both the number of responses and their 'sided' nature. Although some praised the policing effort, more responses were negative and strongly opposing AFR. Many raised issues of privacy, surveillance, legislation, consent, warrants, true/false positive rates and data retention, among other things. In addition, civil liberties groups showed growing interest in the use of AFR. During the evaluation period, Big Brother Watch used their own social media to directly question police use of facial recognition and produced briefings on the subject. In a number of these messages it was asserted that South Wales Police were using 450,000 custody images during their Locate deployments, which, as has been detailed above, is not correct.

More recently, while this evaluation report was being finalised in May 2018, there was a shift in the volume and tone of media coverage of AFR. Details from a Freedom of Information request responded to by the force and posted on their official website were reported widely. Among other things, the request, which came from a journalist specialising in privacy and security, asked for 'true' and 'false match' figures. These figures were supplied, including the high false positive rate from the Champions League (2,470 total matches, 173 'true'). South Wales Police offered some explanations for this, including that it was their first live deployment of the technology, that they had used poor quality images and that results on the old algorithm were poor compared to the new. However, the bad figures received the most media attention – both online and on TV:

The Times, 18th April 2018

https://www.thetimes.co.uk/edition/news/no-easy-way-to-stop-facebook-s-new-facial-recognition-technology-nr3m3gn5j

Wired, 3rd May 2018

http://www.wired.co.uk/article/face-recognition-police-uk-south-wales-met-notting-hill-carnival

BBC, 4th May 2018

http://www.bbc.co.uk/news/uk-wales-south-west-wales-44007872

Telegraph, 5th May 2018

https://www.telegraph.co.uk/news/2018/05/05/police-defend-facial-recognition-technology-wrongly-identified/

Privacy and Legitimacy

As outlined in the previous section, the deployment of AFR technology raises a number of complex questions in terms of privacy, legitimacy and ethics. A key finding of this evaluation has been that the contribution that AFR makes to how police identify possible suspects is quite complex and depends upon interactions between the technological capabilities and capacities of the system, and interpretations and decisions of human operators. It is for this reason that we have introduced the concept of 'Assisted Facial Recognition'.

Fundamentally, the technology is directing officers' attention to a set of people, from whom they select between in terms of deciding whether to conduct some form of intervention. Because of the 'black boxed' quality of the technology it is not possible to ascertain precisely how the technology is framing the attention of officers. In terms of its practical application, South Wales Police were treating any information generated via the AFR system as 'intelligence' rather than evidence. That is, it was directing their attention towards certain individuals on the grounds that these people might warrant further consideration and investigation.

Empirical evidence from this evaluation has shown that in terms of the ability of the AFR platform to provision useful intelligence to officers, its accuracy is impacted by image quality. As such, non-custodial images probably should not be utilised. However, potentially this does raise concerns about the appropriate length of time that police should be able to retain custody images and utilise them on databases of this sort. Related to which, there is a case for saying that police should be explicit about what criterion they are using to select images for inclusion on a watchlist. This might help to secure a degree of public legitimacy by ensuring that deployment and application are proportionate to the crime prevention or crime-detection benefits that might result. Criteria for inclusion should be clear, defined and appropriate to the event in order to ensure auditability and avoid 'fishing expeditions'.

At the time of writing, it is not clear how AFR image use and retention fits with the new General Data Protection Regulation, which gives greater protection and rights to individuals. Although law enforcement agencies are subject to some exceptions from the GDPR, AFR is already being used under non-specific (and sometimes ambiguous) pieces of legislation, and the new laws layer on additional complexity.

The retention of custody images by police is a contentious issue. Rules do exist to regulate this, but campaigns calling for the introduction of clear and specific regulations are ongoing. Recently, the Liberal Democrat MP Norman Lamb alleged that millions of photos were being stored illegally, and that people who had been released without charge were unaware of their right to request their custody image be removed from police databases (BBC, 2018).[7] It is likely that custody images of people who were subsequently released without charge exist on SWP systems. It is unclear whether these might make their way onto Locate watchlists, but they are likely being searched when AFR Identify mode is being utilised. If AFR is used to identify someone from a custody image that should have been removed, does this invalidate the 'match'? It is very likely that this would at least be challenged in a court of law.

If AFR is to be adopted more widely by the police service, it will be important to achieve greater legal clarity and transparency around the range of police powers implemented in respect of the system's application. During the observations conducted, no member of the public refused to cooperate when asked for their name and to show their ID following an AFR match. However, it does seem likely that the operationalisation of police powers in this regard will be tested at some point soon.

Multiple research studies have reported algorithmic biases regarding ethnicity in facial recognition systems. This was not an aspect empirically tested by the current study. It is an area of concern though. One recent study (Buolamwini and Gebru, 2018) tested three commercial algorithms for classifying gender.[8] It found that they performed best for 'lighter skinned' males (and almost as well for lighter skinned females), and worst for 'darker skinned' females. One of the suggestions from this study was that individuals with darker skin may have been less represented in the training data of the algorithms, resulting in the poor performance. Others have also noted the training data issue (e.g. O'Toole and Phillips, 2015).[9]

7 Source: http://www.bbc.co.uk/news/uk-politics-42961025 [Accessed 12 May 2018]

8 Source: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf [Accessed 18 May 2018]

9 Source: O'Toole, A and Phillips, P J (2015) Evaluating Automatic Face Recognition Systems with Human Benchmarks. In: T Valentine and J P Davies (eds) (2015) Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV. Chichester: John Wiley & Sons, Ltd.

As intimated above, although we have been unable to systematically test it, the indications from the evidence accumulated by this evaluation are that watchlist composition and configuration plays an important role in shaping AFR accuracy and precision in terms of generating 'true positive' and 'false positive' results. During the evaluation period, no overt racial discrimination effects were observed. But this could potentially be an artefact of the demographic make-up of the watchlists utilised.

In summary, this discussion simply seeks to surface a number of 'unsettled' issues and potential areas of concern surrounding the use of AFR by police. Currently, officers are drawing upon provisions under the Police and Criminal Evidence Act (1984), the Human Rights Act (2000) and the National Decision Model from the College of Policing, to guide and account for their decisions. However, all of these instruments and the ethical principles they expound pre-date the kinds of affordances AFR technologies provide. As such, new regulatory frameworks will probably be necessary as this technology becomes more widely adopted.

SECTION 9: **CONCLUSIONS**

By way of conclusion, this report has sought to provide an assessment of the contribution made by innovative facial recognition systems to the delivery of policing. In so doing, the evaluation conducted has analysed a range of different use cases and technological configurations. Based upon the evidence collated three principal applications for policing can be discerned:

- Suspect identification – this was the principal focus of South Wales Police's implementation of AFR in both 'Locate' and 'Identify' modes, and hence the main focus for the evaluative work conducted.
- Victim identification – more innovatively, the police also made use of the system in a small number of cases to help them identify possible victims.
- Vulnerable persons protection – at the time of writing, the police are starting to think about how the technology might help them with managing vulnerable people, for example individuals with dementia or a proclivity to go missing.

A key imperative of the evaluation has been to try to capture and convey some of the complexities associated with implementing and using an innovative system such as AFR. With new technologies in policing there is sometimes a tendency to portray them as a 'silver bullet'. That is, as providing a straightforward solution for an identified problem or issue. Whereas, in reality, things are frequently more complex and nuanced. This is our assessment in respect of AFR for policing.

In the original bid submitted to the Home Office, South Wales Police outlined a number of outcomes they were looking to achieve through the adoption and deployment of the AFR technology. These outcomes are mentioned throughout the bid and they can be divided into policing and project outcomes, though the majority were more policing-related. The outcomes that have neen achieved can be summarised as follows:

- AFR technology tested at multiple events (project);
- Operators able to enhance images to increase the possibility of recognition (project);
- Ability to process moving images (e.g. from mobile phone footage or live video streams) enhanced (project);
- Multiple images, such as from a large event, can be separated and enhanced for recognition using the technology (project);
- Reduce investigative and prosecution time; and cost savings by reducing requirement to manually cross reference images therefore freeing officers' time (policing). Some limited evidence of this outcome was found in that AFR Identify matches were returned to operators (and therefore officers) rapidly, but evidence relating to precise cost savings and prosecution timescales was not acquired.

Outcomes for which conclusive evidence of achievement was not found in the evaluation include:

- A mobile phone app, which allows officers to run searches against images captured on their mobile phones and bodycams (project). This outcome was not successfully achieved as the app failed during multiple deployments;
- Reduce repeat offending (policing);
- Measurably increase detections with the onus on increasing public satisfaction (policing);
- Increase community cohesion, for example regarding problems such as hate crime, which has traditionally been affected by poor quality suspect images and the inability to enhance them (policing).

The evidence clearly supports the conclusion that AFR processes and systems can contribute to police identifying persons of interest that they would not otherwise have been able to do so. However, the effort and investment required to get AFR to regularly contribute to delivering tangible policing outcomes was probably more than was anticipated at the outset of the project. That this is so, is reflected in our suggestion that, in terms of its application to policing, the technology is more accurately described as affording 'Assisted Facial Recognition', than 'Automated Facial Recognition'.

In other use cases, such as border entry and passport control functions, that are situated in more controlled environments, the algorithm functions relatively autonomously and devoid of human input. However, this was not how it was observed being used by the police. The police are asking the system to function 'in the wild' with far greater degree of variation in terms of the images that are being analysed, and the surrounding environmental and climactic conditions. Notably, the 'probe images' routinely contain faces where a person of potential interest is not looking 'face-on' to the camera.

Given this level of complexity and variation, some of the results delivered by the technology were impressive. Especially with the introduction of the newer algorithm, there was a step-change in terms of what it could accomplish. At the current time, the Identify mode of deployment seems to be generating more positive outcomes in terms of helping police to identify and instigate proceedings against otherwise unidentified suspects.

That said, it is important to retain an understanding of the importance of the 'human in the loop' and the role played by operators in confirming or refuting possible 'matches' suggested by the algorithm. Through attending to the interactions between system performance, organisational performance and operator performance issues, we have sought to develop an understanding of the range of factors and influences that shape what outcomes AFR delivers. Adopting this 'whole system' view is important in terms of recognising some of the wider impacts that arise. For example, an unanticipated consequence for South Wales of adopting AFR was the identification, mid-way through the project, of a need to improve the quality of all of the custody suite images they took of suspects.

At the same time as acknowledging the potential benefits of AFR to policing and public safety, the evaluation has sought to catalogue some of the profound challenges that accompany the adoption of this approach. Briefly these include:

- The fact that the functioning of the algorithm is 'black boxed' makes it very difficult for police to explain and account for the results that they are having returned to them.
- A critical factor in terms of overall system performance seems to be the size and composition of a watchlist. It has not been possible within the context of this evaluation to evidence precisely how and why this matters, but it does seem to make a difference. Going forward, we would recommend that decisions around watchlist size and composition are made public for the purposes of public accountability.
- It does seem that there are a small number of individuals whose 'face type' means that the system has an increased likelihood of triggering a 'false positive' match.
- If a decision is taken to expand the roll-out of AFR systems, it will be important for operators to maintain an awareness that the machine is not 100% reliable and can get things wrong.
- Currently, police are using the results generated by AFR as a form of 'intelligence', that supplements and augments their investigative actions and case building functions. Presumably however, at some point the principles of AFR will be subject to an evidential test in law. This might be very interesting, especially in cases where a suspect is identified and prosecuted, even though they do not appear especially high in the rankings returned by the AFR Identify algorithm.
- There are a number of ethical concerns about the use of this kind of surveillance technology for policing. It is worth thinking about how these could be managed. For example, it may be worth thinking about whether it is appropriate to utilise it for preventing and solving only more serious kinds of crime, rather than all offences.
- Associated with the previous point, if other forces move towards adoption, some new legislation may be helpful in terms of clarifying police powers to use these kinds of new surveillance. For example, South Wales Police are using custody suite images for AFR Locate and Identify, but should the threshold for inclusion be limited to individuals, arrested, charged and convicted of an offence?

In addition, to these issues that are directly relevant to the implementation of AFR, there are some wider considerations that gravitate around it. For example, writing in the late 1960s, the sociologist David Matza differentiated between the 'incidental' and 'bureaucratic' modes of suspicion, in terms of how police identify suspects.[10] The former involves starting from a crime and then pursuing lines of enquiry to try and discern individuals who might have had a means, motive and opportunity to commit the offence in question. In contrast to which, 'bureaucratic suspicion' relies upon 'rounding up the usual suspects'. In many ways, AFR enhances police reliance upon 'bureaucratic suspicion' in that it involves testing probe images against visuals of people who are already known to police. Whilst we know that the majority of crime is committed by a relatively small number of individuals, it is important to consider the limitations involved in such an approach to crime management.

Another important issue concerns the difficulties in gauging the crime prevention effects attributable to AFR. South Wales Police utilised a proactive strategic communications approach to accompany their use of AFR, in the hope that this might have a suppressant effect on crime by impacting potential offenders' decision-making. We have been unable to quantify what if any impact this had.

In keeping with increasing wider political and public debates about the import of algorithmic transparency and accountability, there have to be wider concerns about the fact that the AFR algorithms are black boxed. Neither the police utilising them, nor the evaluation team, really understand how they function. Consequently, we are not in a position to be able to evidence whether or not decisions about watchlist size and composition has a significant impact or not upon the results that an AFR query might return.

The detail provided in this report is necessary to capture some of the complexities involved in rendering AFR an operationally viable tool for policing. There is a need for some 'myth-busting' owing to how an element of misinformation and disinformation has emerged in public and policy discussions of what AFR can and cannot do. In part, these myths are an artefact of some highly stylized, but influential, fictional portrayals of facial recognition technologies in films and television. Set against this backdrop, the intent underpinning this report has been to provide an independent, balanced and evidence-led assessment of AFR's application in policing. This 'realistic' approach to evaluation, is vital in terms of clarifying both what benefits might be accrued, but also the levels of investment and effort required to deliver any such return. Such an account is furthermore important in terms of helping to clarify where there are areas for legitimate public concern, and hence where regulatory effort and ethical interests need to be directed as part of future policy and practice development.

10 Matza, D. (1969) Becoming Deviant. Prentice-Hall.

APPENDIX: **RESEARCH DESIGN AND METHODOLOGY**

Empirical data to inform the evaluation were collected using a multiple-method strategy. The intent being that different kinds of data would afford multiple perspectives on the technology and its policing application, thereby providing for a richer picture understanding of it, and its nuances and complexities.

Ethnographic Observation

Forty five hours and thirty minutes worth of observations of operators using the software during deployments between May 2017 and March 2018 were conducted as part of the evaluation. These observations of the software in use by operators in practice and in situ were conducted by multiple members of the UPSI team. The purpose of this element of the research was to understand how the software and hardware itself worked in practice, how the operators interacted with the software, and how the public reacted to the presence of the technology. More specifically, we wanted to understand:

- How well the system functioned in practice across different field situations, environments and climatic conditions, in terms of what information it delivered and its usefulness.
- How operators configured the system, to make it practically understandable and suitable for use, and any 'field innovations' they introduced in terms of practices to 'make the system work'.
- How operators made decisions about what to do and when, and how these decision-making processes were structured by the prioritisation framework embedded within the watchlists (blue/green/amber/red).
- How operators interacted with one another, and in communicating information from the system to intervention team members, or other police personnel.

The periods of observation also provided the opportunity to speak to the operators and interview them about their understandings of the technology, their decision processes, and the challenges – if any – they may have encountered. Observations (11 hours) were also conducted at two training sessions and one familiarisation.

The operators were informed during their training and familiarisation sessions that members of the UPSI team would be carrying out observations throughout the Champions League period. They were told to expect the researchers at any time during this period, in all locations (vans, the fixed locations and the airport). At all later deployments, direct contact was made with officers on the AFR project team to organise observation sessions.

Observations were carried out across all of the multiple AFR deployment sites in order to capture the broadest possible picture. This was important in clarifying how different operators, with their different backgrounds, oriented to the technology in different ways and sought to manipulate it differently also. Researchers also observed at various times of the day. This showed how changes in footfall in the public spaces in Cardiff could impact the software, such as when huge numbers of people were moving around the city centre at 18-19:00 on UCL final day. Observing at various times further enabled us to check whether cameras could be affected by conditions such as weather and daylight (or lack of), and the extent to which overcast vs. sunny weather conditions, or daytime vs. evening light conditions, could impact system performance.

For the Champions League event, 6 members of the UPSI team took part in observations, every day between 31st May 2017 and 3rd June 2017. In total, researchers spent 20 hours and 32 minutes observing, with observation periods ranging from 4.5 hours to 30 minutes. Subsequent Locate deployments were observed by one researcher on four occasions, with observation time totalling 17 hours. This researcher also visited the Identify team in Talbot Green on three occasions, with observations here totalling 8 hours. Researchers made handwritten fieldnotes and drew diagrams while they were observing. These notes were as detailed as possible and were typed up at the earliest opportunity. The notes successfully captured the range of issues of interest, with a high level of detail, and provide useful insights into how operators interact with the software, how the software works in practice, and how interventions are made.

Operator Logs, CSV Files and Analysis

When using the AFR Locate system, operators were required to keep a log of all matches generated by the algorithm. This involved logging details (such as the date, time, match details, decision and any action taken) for each individual match in a logbook. The logs were obtained (scanned into PDF format) from South Wales Police at the earliest opportunity following each deployment. These logs have been used to provide a record of the decisions made by operators, as they indicate whether the match identified by the system was deemed a 'true positive' match, or a 'false positive' match. Once the logs were received in PDF format, members of the UPSI team worked on transferring them into an electronic format, namely an Excel spreadsheet. The details transferred into the spreadsheet were (where applicable): date; time; watch list colour; name of the individual identified; decision made in respect of any intervention taken; and any other relevant comments. New logbooks were introduced in 2018, with several additional fields accommodated in the UPSI analysis.

The purpose of the log books was that they could be compared with the CSV file from the system in order to determine how many of the matches recorded by the system were true matches. Thus allowing researchers to determine the true / false positive rate, and to explore the relationship between system scores and operator decisions. Unfortunately, over time operator compliance with the instruction to complete the written log books, especially in relation to matches from the 'blue watchlist', was variable. This limited the ability of the research team to adduce reliable evidence about the overall accuracy of the AFR system.

Image Analysis

During the course of the evaluation (as described in more detail below) evidence emerged of a small number of database images demonstrating an increased propensity to be repeatedly matched to various people in the crowd by the facial recognition software. These are referred to as 'lambs'. This effect was evident across both the old S17 algorithm and the new M20 algorithm, though it was more pronounced in the older version. To investigate this effect, the images of the individuals who matched most frequently overall were subject to comparative analysis to determine any patterns. To do this, from the Champions League watchlists the top 20 most matched images were extracted and an assessment was made about whether they were from 'custody' or 'non-custody' sources. Non-custody images were images not taken in controlled environments (e.g. outside). This process was repeated for later deployments where relevant and when there was enough data.

Operator Questionnaire

Immediately following the Champions League deployment, a questionnaire was distributed to all AFR operators and 31 responses were received (82% return rate). This asked respondents about a range of issues, including: their usual policing roles; previous experience with technology and CCTV; views on the training session; rating of their experience, confidence in using the system pre- and post-event; system usability; and willingness to volunteer again. The questionnaire was available online and on paper, with most respondents completing it online. Once responses were received, the questionnaire data was quantitatively analysed and results discussed within Section 3 of this report.

Field Trial Methods

One aim of the evaluation was to try and assess the accuracy of the AFR platform in terms of identifying faces and matching them to databased images. A methodological problem in this regard, however, is that it would not be possible for police to know when all persons of interest on the watchlist were present in the area and thus there was a reasonable opportunity for the system to pick them up. Accordingly, in the initial design of the evaluation, this was the purpose of integrating a 'blue watchlist'. Because it would be possible to know when police staff were in the vicinity of the AFR cameras, this could be used to derive a measure of how frequently these individuals' faces were detected by the system. However, as reported in the preceding sections, this approach did not work owing to operational pressures and operators' non-compliance with the requirement for completing log-books for all matches.

In an effort to try and compensate for the implementation failure of this component of the evaluation, a decision was taken to establish an embedded small-scale field trial, designed to systematically test the accuracy of the AFR system to detect faces included on a watchlist. It should be emphasised that there were budgetary and timing constraints on what could be achieved, and as such, not all variables that it would have been desirable to test could be assessed (most importantly ethnicity was not tested). These limitations notwithstanding, the data derived from this trial provide a useful baseline for understanding how well AFR technologies perform under 'naturalistic' street-based conditions.

The trial involved seven volunteers having their 'custody' photographs enrolled onto the AFR watchlist for the event. Each individual then carried out multiple 'walk-throughs' in the areas covered by the cameras under various 'treatment' conditions. As a result, there were over 90 'walk-throughs'. The 'treatment conditions' included:

- Different walking speeds (stroll, power walk, jog);
- Different walking directions (head on and diagonal to the camera);
- Group walk-through;
- 'On phone' stroll (head at an angle);
- Props: winter scarf; winter hat; glasses; baseball cap; headscarf (for women) / moustache (for men).

Each volunteer completed 'walk-throughs' under every condition, recording the time on a log. These logs were then compared to the AFR system CSV file post-event, in order to identify if people were picked up, or missed, and how they were scored by the system.

Subsequent to the data collection phase, the CSV files were downloaded from the system and analysed to calculate the number of times the 7 volunteers were detected by the AFR system, and the variance according to different treatment conditions.

Get in touch.

Crime and Security Research Institute, Cardiff University
Level 2, Friary House, Greyfriars Rd,
Cardiff CF10 3AE

+44 (0) 2920 875440

crimeandsecurity@cardiff.ac.uk

@CrimeSecurityCU

www.crimeandsecurity.org