Research Protocol: Using CCTV in Police Investigations: A comparison of facial recognition technology-assisted reviews and manual reviews

Purpose

This document outlines a methodology to compare the outcomes of reviewing CCTV footage to find people when assisted by facial recognition technology, with manually reviewing the footage.

Background

Demand on policing is increasing. Complex and serious crime has increased at a disproportionate rate in comparison to overall crime in recent years[1]. At the same time, the number of qualified investigators has not risen in line with demand, leading to a detective crisis and very high workloads for investigators[2].

CCTV footage can be a useful source of evidence and intelligence in investigations[3]. However, it is resource intensive to review footage. Research comparing the ability of different groups of people to identify faces in CCTV footage measured the how long it took specialist 'police identifier' officers. It took police identifiers an average of 59 minutes to review 18 minutes' worth of footage[4]. Whilst the footage was of crowded scenes (so will likely take longer to review than footage containing less people) these data illustrated the resource burden of reviewing CCTV footage. Research has also demonstrated that recognising people in footage is susceptible to human errors[5], can increase stress and cognitive load[6], and result in health and safety issues.[4] Data measuring CCTV usage and usefulness in investigations is limited. However, interviews with police officers and staff in user research undertaken by the Home Office indicated a key barrier to using CCTV is the time it takes to review the footage.
A study focused on use of CCTV cameras on the UK rail network from 2011 to 2015 shown that the longer the time period in which an offence could have occurred the less likely it is that footage will be obtained or considered useful to the investigation[3].

Facial recognition is an emerging technology that may aid the police in reviewing CCTV footage, enabling them to find people in footage quicker. The Police Digital Service team in the Home Office has developed an Assisted Facial Recognition (AFR) platform to allow investigators to put CCTV footage through an algorithm to help with recognition. The tool has been developed in a user-centred way, guided by user research with police officers and staff across twenty-two police forces. To date, all testing of the tool has been conducted in a laboratory environment using non police data (e.g. actors and mock scenarios). The study outlined here seeks to understand how the technology might perform when used in an operational setting.

In this study CCTV footage used in historic investigations will be reviewed by police officers and staff using AFR to simulate some aspects of using the tool in a live investigation. Simulations will

[1] https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingjune2019

[2] https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/state-of-policing-2018-double-page.pdf

[3] Ashby, M.P.J. Eur J Crim Policy Res (2017) 23: 441. https://doi.org/10.1007/s10610-017-9341-6

[4] Davis, J. P., Forrest, C., Treml, F., & Jansari, A. (2018). Identification from CCTV: Assessing police super-recogniser ability to spot faces in a crowd and susceptibility to change blindness. *Applied cognitive psychology*, *32*(3), 337-353.

[5] Hillstrom, A., Hope, L., & Nee, C. (2008). Applying psychological science to the CCTV review process: a review of cognitive and ergonomic literature.

[6] Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. Human factors, 50(3), 433-441.

be done in a controlled environment under the observation of researchers. Footage will be drawn from investigations where the technology is most likely to have the greatest benefit – serious and complex investigations that required investigative teams to review large volumes of CCTV footage. By utilising material from serious cases, issues of proportionality and necessity are more likely to be addressed. Footage will be provided by Major Investigation Teams and Missing Persons Teams. The conclusions made by the reviewers using AFR will be compared to the conclusions made when the footage was originally reviewed by extracting data from original case file documentation. The time taken to review footage using the technology will be measured to build evidence of any potential time savings, compared to the original manual review. The findings from this study will inform future trials of the technology, further the evidence base on AFR in policing, and contribute to the wider public debate about the appropriate use and oversight of facial recognition.

Study aims

The study aims to:

1) Compare reviewers' conclusions when reviewing CCTV to find people using facial recognition technology with reviewing the footage manually
2) Compare the speed at which investigators can review CCTV footage to find people using facial recognition technology, with manually reviewing footage
3) Test the performance of the tool using ecologically valid footage captured from CCTV cameras used in real police investigations
4) Identify technological, organisational, usability issues that may hinder implementation of the technology

Research questions

The aims of the study will be met by answering three questions:

1) Do reviewers using facial recognition technology come to the same conclusions as those who reviewed footage manually in a live investigation?
2) How long does it take to review CCTV footage when finding people using facial recognition technology, compared to manually reviewing the footage?
3) Are there factors related to CCTV footage that might affect the performance of the facial recognition technology?

How the software works

The software works as follows:

1. The user uploads one or more static images of the target. The tool will create a face template based on the facial geometry.
2. The user uploads one or more videos.
3. The user will input their e-mail address to receive notifications of the results and start processing the footage.

4. During processing the software will identify faces in the video and create face templates of them. It will then do a probabilistic matching of the facial signatures in the static images against the facial signatures from the video.
5. The user will receive e-mail notifications of any possible matches.
6. The user will view the results and judge whether the possible matches suggested by the tool are accurate and record their decisions.

Face templates will be deleted immediately once a query returns a result.

Possible matches have differing levels of certainty indicated by a percentage figure from 0 to 100. The threshold for returning possible matches in the results will be initially set at 55%. This threshold is based on experimentation with test footage. Experiments have found that false positives tend to occur at an unacceptable rate when the threshold is set below around 55% but this measure is configurable. The ideal threshold will be dependent on the specific use case and the tolerance for false negatives and false positives in an operation.

Study Design

The study seeks to compare reviewing CCTV manually in historic major crime investigations with reviewing the same CCTV with reviewing using AFR. Data relating to the original manual reviews will be drawn from the investigation case files - new manual reviews will not be conducted as part of this study. Data will be collected from an exercise simulating how facial recognition could be used by investigators. Investigators will use the tool to review the footage and record their conclusions as to whether specific individuals were in the footage and what time they appeared.

We will be using real footage drawn from previous investigations, selected to be representative rather than strictly controlled for multiple variables. This will help us to identify areas for focus on in subsequent operational trials. We will therefore not be able to draw precise conclusions about accuracy with and without an AFR tool. These questions will be more suitably addressed in a planned lab trial.

As the footage used in the study is from real cases there is no definitive record of the identity of those who appear in the footage. Participants who review footage using the tool will also be asked to review the results of footage reviewed by other participants and indicate whether they think each match presented by the tool is a true match. This data will then be used to indicate whether possible matches made by the tool but not by manual reviewers are likely to be correct matches or not.

Participants

The participants in the simulation task will be drawn from a pool of investigators within a single police force who fit the below criteria:

- Have not had any involvement in the original case that the footage relates to
- Are unlikely to recognise the individuals in the footage (i.e. work in a different division)
- Have experience of reviewing CCTV for investigative purposes

Nine participants will review all the footage using the tool. Nine participants has been chosen as research suggests this is the optimum number of people to balance accuracy and resource[7]. Eight were suggested White et.al 2103, but we will use an odd number to get a consensus position.

The participants in the manual review group are pre-determined. They are the officers who did the CCTV review in the original investigation. Their success and time spent will be obtained from the original case viewing logs. We will not be contacting the original reviewers during this trial. Where possible we will match the experience levels and role of the participants to those of the original reviewers

CCTV Footage

Each participant will only be looking for one "target" individual in the footage. This person will not appear in all CCTV clips that the participant views. The participant will be shown an image of the person to be searched for and told that they may or may not appear any number of times in the footage they will review.

Requirements

- Comes with well-documented viewing logs
- Case is finalised, no appeals pending
- Still image of target available (possibly drawn from video footage)
- Has clear views of faces
- Three (TBD) pieces of footage, with the target present in at least one

In total we will use X (TBD) pieces of footage, each of a duration of approximately 20 minutes (TBD)

Footage will vary in the following ways

- Number of people in the footage at a single time
- Resolution/quality of footage
- Camera types and angles
- Lighting – inside/outside, day/night
- Race & sex of target
- Clothing type

We may use footage from multiple cases in the trial, but each participant will only see footage from one case.

We will note as much as possible about the footage used and the case it was taken from in order to see what if any factors may influence the comparison. Information we will try to collect includes:

- Job rank and experience level of the original reviewer
- What if any CCTV training the reviewer had
- What brief did the reviewer have - what specifically were they looking for

[7] (White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. Applied Cognitive Psychology, 27(6), 769-777.

- To what extent did the original reviewer view the footage together with other colleagues
- What time pressure was the reviewer under - how urgent and serious was the case
- Whether the reviewer worked full-time on the CCTV footage or did it interspersed with other responsibilities
- Where the review took place - what distractions may have been present
- Type of investigation
- Any information relating to the camera used for CCTV and still images
- CCTV location
- Age, sex, ethnicity of person sought
- number and length of clips, number of targets per clip/case, number of sightings per clip/per case

Before using any footage in the trial, we will need to test it with the tool to ensure that the file format is compatible or can be made so.

Setting

The trial will take place in a convenient room in the station where the participant normally works. Each session is expected to last <TBD> hours. In addition to the participant and researcher there may be 1-2 observers from the project team.

Sessions will be recorded with a small video camera placed on the desk, which will record the screen and the audio of the participant and researcher, but not the face of the participant.

Procedure

Baseline

Footage and case files will be analysed to identify each sighting investigators made when reviewing the footage manually, including the name of the individual and the time they were identified. This data will be used as the baseline.

Simulation

Participants will take part in the simulation exercise independently under the supervision of a researcher. They will use a computer that is typically used to review CCTV in their force. This is to capture a realistic estimate of processing time based on the police force's infrastructure.

Prior to participation, participants will be briefed and provide informed consent to participate. They will be given a standardised demonstration of the software and asked to complete a practice task using footage created for usability tests.

The participant will then be instructed to process a selected set of clips of closed case footage. Each participant will process different footage. The footage and associated images will be stored in a clearly marked electronic folder on the computer.

If the computer is attached to the force network, the participants will be instructed to request notifications of when results are ready to view using their work e-mail address when running searches. They will then access results via these notifications to simulate how the tool is intended to be used.

If the computer is not attached to the force network, the participant will be instructed to wait until the status screen indicates that the results are ready to view.

When presented with the results of the searches, participants will be instructed to review the results, and complete and rate how certain they are that the AFR results are a correct match, using a Likert scale

The researcher will use a stopwatch to time the process from the point at which the participant logs into the system to the point they have recorded their conclusions from the results. The exercise will be audio, video and screen recorded to allow researchers to review the recordings and identify potential usability issues that may arise. Data will also be extracted from system logs to measure the time taken to process the footage and Google Analytics data will be used to verify the time recorded by the researchers.

Participant Debrief

The researcher will conduct a brief semi-structured interview with the participant to identify any usability issues they encountered that may have impacted on task completion speed or their interpretation of the results. This will be done as a retrospective think-aloud exercise using the tool screen and the task list as prompts (van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2007). Evaluation of an informational Web site: Three variants of the think-aloud method compared. *Technical communication*, *54*(1), 58-71.)

Participants will also be asked how the trial experience differs from how they would normally view CCTV footage and establish matches in order to identify to what degree the trial is a realistic simulation of operational use.

Participants will be asked to complete a short questionnaire to measure their self-assessed digital skills and confidence. These data will be used to identify if digital skills and confidence may have impacted on the time taken to complete the task. They will also complete the single-question system usability score.

Review of other footage

After the simulation participants will be given a short break and then instructed to review the results from the footage presented to other participants. The footage will be processed in advance. These results will be presented within the tool and all participants will be able to view the probe image, matched image and the video clip showing the match. Participants will note the degree to which they agree with the AFR match using a Likert scale.

The order of the presentation of footage will be counterbalanced to reduce order effects. All possible matches generated by the tool will be reviewed. These data will then be used to estimate

the extent to which the tool may have generated true positives and false positives and assess the potential risk that participants may make misidentifications from the results (described further in the analysis section).

At the end of the session, each participant will take the Glasgow Face Matching Test – an online test of face matching ability.

Materials

Glasgow Face Matching Test (GFMT)
The GFMT is a test of facial recognition ability widely used in studies of facial recognition ability in humans. Participants are shown pairs of faces, photographed in full-face view but with different cameras, and are asked to make judgements as to whether it is the same or a different person. The short version (40 faces) will be used. Normative data shows that the average score is 81% (of faces that are correctly judged) and the only the top 10% of performers score 95% or above.

Digital assessment questionnaire
A short questionnaire will be used to assess participants' self-assessed digital confidence and skills. The questionnaire will be comprised of items drawn from the Cross government Digital Inclusion Evaluation Toolkit[8]

Single-question System Usability Score
This provides an indication of how usable the participants found the tool
I thought the system was easy to use. 1 Strongly Disagree ----5 Strongly Agree

Outcomes

To answer research question one (Do reviewers using facial recognition technology come to the same conclusions than those who reviewed footage manually in a live investigation?), two outcomes will be measured:

1) **Review time** - The time from starting to review CCTV to having made a decision based on the results of the review, including any wait time during processing of footage.

This will be calculated by a researcher observing the participant completing the task and timing the episode from logging on to the system to the participant having made a decision on all the results presented to them. The data will be verified by cross referencing timings with system data extracted from the tool. This outcome seeks to understand if the technology helps investigators reach a conclusion about the footage quicker.

2) **Time on task** - The time a participant spends interacting with the footage. This includes uploading footage and viewing footage and/or results provided by the tool. It excludes any wait time during processing.

[8]https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605093/DigitalInclusion_BankOfOutcomes.pdf

This will be calculated by a researcher observing the participant completing the task and measuring the time they interact with the system. The data will be verified by cross referencing timings with system data extracted from the tool. This outcome seeks to understand if the technology frees up investigator time to conduct other tasks.

The data will then be evaluated against a reasonable best-case scenario. As footage is taken from historic investigations and investigators do not routinely record how long it takes to review CCTV, direct comparisons cannot be made. Previous research suggests that reviewing behavior can vary considerably dependent on factors such as the quality of footage, level of activity in the footage, software functionality and severity of the crime[9]. However, it is reasonable to assume that it will take at least as long as the length of the footage to review footage in investigations that justify the review of large volumes of footage (e.g. serious crime or high-risk missing persons cases). Therefore, potential time saving benefits will be assessed by comparing time outcomes to the total duration of the footage.

To answer research question two, conclusions made by investigators during manual reviews will be extracted from case files and viewing logs. These data will include the name of the individual identified and the time of the identification. Names will be anonymised during extraction.

These data will then be used as a baseline to compare with the conclusions made during the simulated exercise. Participants reviewing footage with AFR will record their assessment of whether each match is correct or not and possible matches will be paired with case file data based on the timestamp of the match:

- What proportion of sightings made by investigators in the manual review were also made by participants when using AFR?
- How many sightings did participants using AFR make that were not documented in the original manual review?
- What proportion of the possible matches presented to participants by AFR that were not documented in the original manual review were deemed correct by the majority of reviewers?

Analysis

Time measures

Time outcome measures described above will be obtained for each video clip used in the study. These data will be transformed into a ratio of the length of the clip. For example, if a 20 minute clip took 10 minutes to review it would be it assigned a score of 0.5. Ratios will be calculated for time on task and total review time separately. These data will then be analysed descriptively to find the mean and interquartile ranges of ratios. Any outliers will be investigated to understand the cause of particularly quick or slow times by examining system data, characteristics of the footage (e.g. time of day) and participant feedback. If the distribution of the data is relatively normal, these data will then be used to estimate the average time it is likely to take to review one hour of footage using the tool as a ‘real world’ baseline measure of speed.

Investigator conclusions

Every sighting made in the manual review and every sighting generated by the software will be recorded in a database. These sightings will then be coded to indicate whether:

[9] Based on 26 interviews with a cross section of police officers and staff in 11 police forces who regularly review CCTV in their job

- it was recorded in the manual review
- AFR presented the sighting to participants during the simulated exercise
- each participant assessed the sighting as a correct match, incorrect match or were unable to decide (unclear)

The data will then be coded as per the table below and analysed to answer the three questions described in the Outcomes section.

Category	Definition
Missed by tool	A sighting recorded by investigators when manually reviewing the footage was not identified by AFR (proxy for false negative)
Not recorded by manual reviewers – likely correct	A sighting not recorded by investigators when manually reviewing the footage is presented to participants using AFR and over half of the participants assess the match as correct (proxy for true positive)
Not recorded by manual reviewers – likely incorrect	A sighting not recorded by investigators when manually reviewing the footage is presented to participants using AFR and less than half of the participants assess the match as correct (proxy for false positive)
Matched by both	A sighting recorded by investigators when manually reviewing the footage is presented to participants using the tool and over half of the participants assess the match as correct (proxy for true positive)

System data (including confidence level, file format and resolution), characteristics of the footage and probe image (as described by the researchers), and participant feedback will be used to identify possible reasons for inconsistencies in participants' conclusions for each match. All possible matches presented to participants by AFR that have been assessed differently by participants will be investigated to identify any common features that appear to affect participants' ability to assess the validity of the match. Possible matches with a high number of 'unclear' assessments will also be investigated. The footage and probe images of all 'Missed by tool' sightings will be reviewed to identify common causes and to identify which combination of variables may account for differing responses<This is the focus of the Cardiff analysis>

Limitations

The study will provide some evidence of whether facial recognition might help investigators to review CCTV and if it does, in what circumstances it is likely to have greatest benefits. However, there are some limitations to the study and what evidence can be drawn from it.

As the footage will be drawn from historic (closed) cases there is no 'ground truth' of what occurred in the footage and the true identity of individuals in the footage. Whilst proxy measures of true positive (correct matches), false positive (incorrect matches) and false negatives (misses) can be calculated, these are based on probable truth. A proxy measure for true negatives cannot be calculated without extensive effort (e.g. multiple people manually reviewing every individual that appears in the footage). A laboratory-based trial is being considered to complement this work that will provide reliable measures of accuracy.

By using historic cases there is less control over what investigators documented when reviewing footage manually. Taking footage from open cases (and therefore having more control over the manual review) was explored but due to the types of incidents that are likely to justify review of extensive CCTV footage, these incidents will take a long time to be resolved. There is also a risk

associated with this approach that investigators still do not comply with requests to document the manual review in more detail than usual, without significant effort in monitoring. Therefore, historic cases which have detailed viewing logs will be used. This may limit the scope and variety of possible cases that can be used.

Investigators do not routinely record how long tasks take and therefore the estimate of potential time saving will be based on a comparison with an assumed best-case scenario. This assumption is based on interviews with a wide range of officers/staff about how they view CCTV. Whilst precise measures of effect won't be obtained from this approach magnitude of impact will be better understood and the time taken to review CCTV using the tool with a range of different types of footage will still be of great value.

There will likely be large variability between cases in relation to variables that may impact on the effectiveness of the tool (e.g. image quality, target's appearance). A sufficiently large and diverse sample to test the effects of these variables in a statistical sense may not be collected. However, potential factors that affect performance will be identified on a case by case basis.

In summary, the study will not provide statistics that can be used to reliably measure effectiveness and generalise benefits more broadly. However, it will provide valuable evidence that is currently lacking - how facial recognition performs on real world cases, the circumstances in which facial recognition may best assist investigations, and why the technology is suited to some cases more than others. The results of the study will inform the design of more statistically robust work to follow.

Ethical Considerations

As the tool has only been tested on a relatively small range of imagery it has been decided to run the initial study on footage relating to historic cases. This will minimise the positive or negative impact the tool might have on the public's safety and access to justice while more data is being gathered to assess the potential benefits and harms of the tool. There is a chance that information that did not come to light during an original investigation may become available during the study as a result of processing footage through the tool. If this is believed to have occurred, the results of the trial would be sent to a senior officer in the force for them to determine if further action was required.

Informed consent to participate in the simulation will be obtained from all participants at the onset of the study in the form of a written consent form. They will also be provided with a participant information sheet explaining the purpose of the study, what participation involves and their rights as a participant. Participants will be asked to provide consent for sessions to be recorded but have the right to decline. All participants will have the right to decline participation at any point during the data collection period including after providing initial informed consent. They will do so by informing any of the researchers. Any data supplied up to this point will be securely deleted unless it is anonymised already. Participants will be reminded of this right at the start of every session.

The participant's data will only be accessed by the research team and no identifiable data will be shared more widely, for example with colleagues. Participants' responses to tasks will be anonymised as soon as practicable.

Annex 1 Materials checklist

- ❑ Video camera, stand, batteries and/or charging cable
- ❑ Information sheet
- ❑ Consent form
- ❑ Discussion guide
- ❑ Glasgow test link
- ❑ Digital assessment questionnaire
- ❑ Likert scale questions as paper printouts

Annex 2 – background literature

From the studies reviewed the following factors have been shown to affect human performance in facial recognition:

- Familiarity of the face, performance on unfamiliar faces is worse (Burton et al, 2010)
- Time on task, performance diminishes over time (Donald & Thatcher 2015, Macworth, 1948)
- Quality of imagery, pixelated images result in poorer performance (Bindemann et al 2013),
- Number of reviewers, performance increases when decisions are made in groups. Groups of eight were found to be the optimum size (White et al 2013)
- Distractions, interruptions and noise (Hillstrom et al 2008)
- Tiredness (Hillstrom et al 2008)

References

Bindemann, M., Attard, J., Leach, A., & Johnston, R. A. (2013). The effect of image pixelation on unfamiliar- face matching. Applied Cognitive Psychology, 27(6), 707-717.,

Blanton, A., Allen, K. C., Miller, T., Kalka, N. D., & Jain, A. K. (2016). A comparison of human and automated face verification accuracy on unconstrained image sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 161-168).

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. Behavior Research Methods, 42(1), 286-291.

Calic, D. (2013). From the laboratory to the real world: Evaluating the impact of impostors, expertise and individual differences on human face matching performance (Doctoral dissertation),

Davies, B. Innes, M. & Dawson, A. (2018) An Evaluation of South Wales Police's Use of Automated Facial Recognition. Cardiff University.

Davis, J. P., Forrest, C., Treml, F., & Jansari, A. (2018). Identification from CCTV: Assessing police super- recogniser ability to spot faces in a crowd and susceptibility to change blindness. Applied cognitive psychology, 32(3), 337-353.

Donald, F., Donald, C., & Thatcher, A. (2015). Work exposure and vigilance decrements in closed circuit television surveillance. Applied ergonomics, 47, 220-228.

Fussey, P. & Murray, D. (2019) Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology. Essex University.

Gates, K. (2010). The Tampa "smart CCTV" experiment. Culture Unbound: Journal of Current Cultural Research, 2(1), 67-89.

Grother, P. J., Ngan, M. L., & Quinn, G. W. (2017). Face in video evaluation (FIVE) Face recognition of non-cooperative subjects (No. NIST Interagency/Internal Report (NISTIR)-8173).

Hillstrom, A., Hope, L., & Nee, C. (2008). Applying psychological science to the CCTV review process: a review of cognitive and ergonomic literature.

Introna, L., & Nissenbaum, H. (2010). Facial recognition technology a survey of policy and implementation issues.

Labelled Faces in the Wild. University of Massachusetts [Accessed in August 2019, here http://vis-www.cs.umass.edu/lfw/results.html#note_human]

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. Quarterly Journal of Experimental Psychology, 1, 6-21.
Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... & Chen, J. C. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proceedings of the National Academy of Sciences, 115(24), 6171-6176.

Stacy, E. M. (2017). Human and algorithm facial recognition performance: face in a crowd (Doctoral dissertation).

White, D., Burton, A. M., Kemp, R. I., & Jenkins, R. (2013). Crowd effects in unfamiliar face matching. Applied Cognitive Psychology, 27(6), 769-777.