Guideline for Facial Recognition System End Users

European Network of Forensic Science Institutes

Digital Imaging Working Group

V1.0 June 2022

Project Team
Margreet Aerts-Bruintjes (Center for Biometrics, Netherlands Police)
Zsuzsanna Bartha (Hungarian Institute for Forensic Sciences)
Elisabet Leitet (National Forensic Centre, Swedish Police Authority)
Eszter Orsolya Lévai (Hungarian Institute for Forensic Sciences)
Sergio Castro Martinez (Comisaria General de Policía Científica. Spanish National Police)
Reuben Moreton (The Open University)
Johanna Morley (Interpol)
Elisabeth Pickersgill (Germany BKA)
Arnout Ruifrok (Netherlands Forensic Institute)

Table of Contents

1 INTRODUCTION

Facial Recognition (FR) systems can be used to search and match faces (extracted from images or videos) against a database of facial images. The accuracy of FR systems is nowadays high for a diverse range of image quality, mainly due to the introduction of Artificial Intelligence (AI) or convolutional neural networks. In both the public and private sector, the technology has many uses, such as one to many (1:N) searches for Identification of an unknown subject, one to one (1:1) comparison for the Verification of a claimed identity, or N:M for Clustering individuals together under a single identity.

The vast potential of FR, mainly in relation to the analysis of large volumes of data, combined with the speed of analysis and the recent improvement in accuracy rates, has led to the increased use of FR for law enforcement and investigative purposes, in Europe and around the globe.

2 AIMS

This document aims to provide a framework for end-users of FR systems and deliver:

- Basic knowledge and information on FR system set up
- Best practices with respect to search optimization, analysis and decision making
- Guidance with respect to the training and competency of users

3 SCOPE

This guideline addresses the processing and examination of facial images for FR searching, optimizing the search parameters, the understanding and interpretation of FR search results and reporting outcomes.

The scope is specific to retrospective 1:N identification searches for law enforcement, investigatory and forensic applications.

Other FR technology applications, such as live (real time) FR, border control and access to secure sites are out of scope.

Human 1:1 facial image comparison and other uses of facial images are also out of scope of these guidelines. For 1:1 comparison please refer to the document "Best practice Manual for Facial Image Comparison" published by the ENFSI [1].

4 DEFINITIONS AND TERMS

The terms used throughout this guide are described below:

Accuracy: A measure of how well the facial recognition process performs in terms of false positive and false negative errors. It should be noted that the FR process combines the automated FR system and the human review.

False positive: when an FR algorithm and/or human reviewer/examiner incorrectly associates (matches) two faces together but the ground truth is that the two faces belong to different individuals (referred to as non-mated search).

False negative: when an FR algorithm and/or human reviewer/examiner fails to associate (match) two faces of the same person, but the ground truth is that the two faces are the same individual (referred to as mated searches).

False Positive Identification Rate (FPIR): The proportion of the total number of non-mated searches where one or more potential candidates are returned as a mate.

False Negative Identification Rate (FNIR): The proportion of the total number of mated searches where the true mate is not returned as a potential candidate. This may be because it is found below a configured threshold, outside the top rank candidates or not selected as a potential candidate by the human reviewer.

True Negative Identification Rate (TNIR): the complement of the FPIR, giving a statement of how often non mated searches are not returned as a potential candidate.

True Positive Identification Rate (TPIR): the complement of FNIR, giving a positive statement of how often mated searches are successful.

Audit trail: A chronological record or set of records that provide documentary evidence of the sequence of activities that have affected a specific procedure, material, system, analysis or decision.

Biometric data: EU directive 2016/680, definitions: Biometric data means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data.

Candidate list: List of images returned by the FR system, ranked according to similarity score. The number of candidates shown will depend on system configuration settings such as threshold and/or a defined candidate length.

Cognitive bias: A broad term that includes a variety of processes that may lead to inaccurate judgments or interpretations; cognitive biases can affect memory, reasoning, and decision-making.

Confirmation bias: is a type of cognitive bias, whereby people test hypotheses by looking for confirming evidence rather than for potentially conflicting evidence.

Cumulative Match Characteristic (CMC): Summarizes the accuracy of mated-searches and plots the proportion of mated searches returning the mate at rank R or better.

Deep convolutional neural network: A class of machine learning based on artificial neural networks, most commonly applied to analyze imagery.

Enrollment: The process of localizing and aligning the face from an image or video and encoding the facial features to generate a template.

Facial Examiner: A trained facial comparison practitioner that conducts the task of facial examination (see **Facial image comparison; Examination)**.

Facial image comparison: Is a manual process undertaken by a human to identify similarities and differences between facial images. Facial image comparison is used in different applications, involves different levels of evaluation according to the purpose of the comparison:

> **Examination:** Detailed and methodological process to compare one image of a face to another (1:1) in accordance with scientific recommendations for the purpose of effecting a conclusion. It is often used in forensic applications.
>
> **Review:** Is a comparison of image-to-image often used in either investigative or operational lead generating applications. Review encompasses a broad range of purposes and levels of rigor involved in the analysis. An independent technical review or verification by at least one additional reviewer should be conducted.

Facial Recognition system (FR system): Software, which is able to detect, enroll and compare faces from digital images or a video frame against a database of enrolled reference facial images.

> **Identification (1:N):** The automated comparison of an unknown biometric sample against a database of (N) reference images in order to return a corresponding identity.
>
> **Verification (1:1):** The automated comparison of a biometric sample against the reference biometric sample corresponding to the claimed identity, resulting in a computer-evaluated similarity score.
>
> **Clustering (N:M):** The automated grouping of biometric samples, for example, represented within a collection of facial images, based on computer evaluated similarity.

Facial Reviewer: A trained facial comparison practitioner that conducts the task of facial review (see **Facial image comparison; Review)**.

Holistic comparison: The innate human ability of comparing faces by looking at the face as a whole without specifically considering the component parts in isolation.

Image enrollment quality: An algorithm or supplier proprietary metric of 'photo acceptance', that provides a measure of subject characteristics (for example pose, expression) and environmental/capture characteristics (for example illumination, resolution, focus).

Intra-variability: Differences in (biometric) samples taken from the same person.

Inter-variability: Differences in (biometric) samples taken from different people.

Metadata: Additional (non-biometric) alpha-numeric information associated with the facial images. Common examples of metadata include sex, age/date of birth, offence type, date of image capture.

Morphological comparison: The direct comparison of class and individual facial characteristics without explicit measurement. It is the method of facial image comparison in which the features and components of the face are compared.

Potential candidate: A person in the candidate list that is judged by the reviewer to show significant similarities to the person depicted in the probe image.

Probe image: Imagery in which the identity of the depicted subject is unknown. The probe image may be captured under either controlled or uncontrolled conditions. Other terms used as synonyms for probe image are questioned image and query image.

Rank: The position of a reference image in the candidate list as based on the similarity score, where rank 1 is deemed by the FR system to have the highest level of correspondence to the probe image.

Reference database: The combination of reference imagery and associated biographic and other relevant information (such as name, date of birth, crime reference etc.)

Reference image: Imagery in which the identity of the depicted subject is usually known and has been verified. Reference imagery is often captured under controlled conditions, for example mugshot images.

Similarity score: The degree of correspondence between two facial templates as judged by the FR system. The range of possible similarity scores are proprietary to the algorithm. This is sometimes just referred to as 'score'.

Template: A digital representation, created by the FR system, of features extracted from a biometric (face) sample. Unlike fingerprints, there are no face template standards. Templates are specific to the algorithm.

Threshold: Configurable setting of the minimum image enrollment quality metric or similarity score which must be reached in order for the FR system to return a candidate.

True mate: Two images that are taken from the same person. Another term used as a synonym for true mate is true source.

Verification: The review and/or independent analysis of the search by another FR system user.

> **Blind verification**: A type of verification in which the subsequent examiner(s) has no knowledge of the original examiner's decisions, conclusions or observed data used to support the conclusion.
>
> **Non-blind verification**: A type of verification in which the subsequent examiner has access to the original examiner's decisions, conclusions or observed data used to support the conclusion.

5 General introduction to FR systems

Facial recognition systems have been deployed since the 1990s. Early systems were limited to use with high quality passport style images with an evenly lit frontal pose face and neutral expression. However, the adoption of deep convolutional neural networks means that FR systems are increasingly tolerant of poorly illuminated or ill posed subjects and low quality images.

The generalized accuracy of FR systems on 'good quality' images has improved dramatically and even over just the last three years there has been a circa 10% reduction in false match rates, which are now well below 1% at a false non match rate of 0.0001% [2].

Automated facial recognition involves four steps as illustrated in Figure 1.

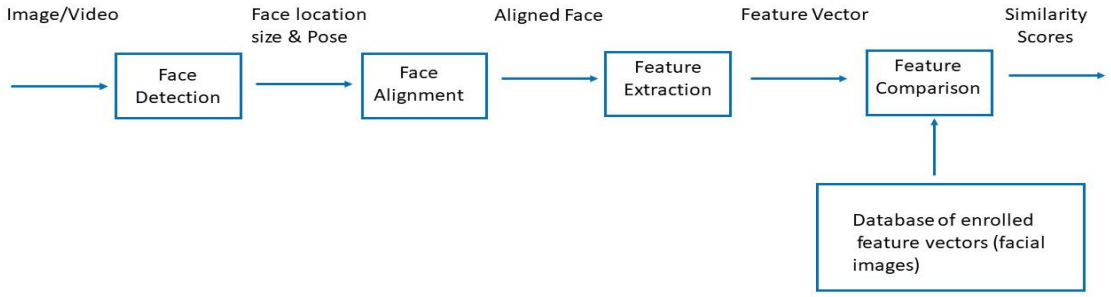

Figure 1: The steps involved in enrollment and search of a facial image using a FR system.

- **Face detection**. Locate one or more faces in the image and mark with a bounding box
- **Face alignment**. Normalize the face to a standard frontal pose
- **Feature extraction**. Extract features[1] from the face to create a feature vector, which can be thought of as a 'template' that is used in the recognition task
- **Face recognition/search**. Compare the face template against the collection of (known) face templates in the database to generate a similarity score[2] against each face

A given system may have a separate module for each step, which was traditionally the case, or may combine some or all of the steps into a single process.

[1] Feature extraction as part of a convolutional neural network (CNN) should not be confused with facial features as observed by humans. Early FR systems generated a template using geometric measurements of facial features. However CNNs use all the information available in a facial image and use machine learning to learn the face representation that correspond to different levels of abstraction, building the set of definition data, referred to as the feature vector.

[2] The majority of commercially available FR systems generate a similarity score, where a higher number indicates a greater level of correspondence between the two facial images. There are some FR systems that generate a difference score (a lower number indicates a greater level of correspondence). Throughout this document, the term 'similarity score' will be used.

It should be noted that in the European Union, the face template is considered as 'Biometric data', which is a special category of personal data. Searches using biometric data for identification purposes by law enforcement is only allowed when strictly necessary and requires certain safeguards [3].

5.1 Criminal investigative use case

The major use case for investigative applications is 1:N searching of an unknown subject against a reference database for the purposes of progressing an identification. Although this is referred to as 'Identification', it should be noted that, unlike fingerprint searches, the outcome of an FR search is not a positive identification (that can be, for example, produced as evidence) but a (list of) potential candidate(s) that can be proposed for further investigation.

An unknown facial image, referred to as the probe or query image, is searched against a database, generally containing reference images. The outcome by the FR system is in most cases a candidate list, ordered from highest to lowest similarity score according to the criteria of the algorithm. Similarity scores are proprietary to the FR system, generally with a higher similarity score indicating a greater degree of correspondence between the facial images. They are sometimes reported as a percentage. This **must not** be considered as a probability score that two images depict the same individual. A human review of the candidate list is required in order to determine if any potential candidate is present.

5.2 FR search output

A 1:N search using an FR system normally outputs a list of the highest ranked matches, where the highest similarity score is at rank 1. FR systems can be configured as threshold based, rank based, or a combination of both.

- For threshold based systems, all facial images with a similarity score above a fixed similarity score threshold (set by the system users) are returned in the candidate list
- For rank based systems, no threshold is set but candidates are returned in order of high to low similarity score with the number of candidates set by default or by the system user
- For combination systems, the similarity score threshold and the maximum number of candidates returned is fixed

For investigative purposes, the most common deployment configuration is rank based, which returns a candidate list to be reviewed by a (trained) human.

The output configuration may be adjusted to the use or objective of the system. For example, for high throughput systems such as real time FR, a combined threshold & rank configuration can be deployed and only the highest scoring candidate above the threshold returned. For these systems, the majority of transactions with the system will not result in any potential candidate.

5.3 Know your system

It is important to know the properties of the data within your FR system, as well as the performance of your algorithm.

5.3.1 The FR algorithm

The face representation of the algorithm is trained on large amounts of labelled data. The composition of the training data set in terms of the distribution of demographic variations can directly impact on the “fairness” of deep models, i.e. the models should be similar in the accuracy rates for different sexes or ethnicities. There are a number of different methods to mitigate demographic bias in FR systems. However, both the hierarchical architecture of the algorithm and composition of the training data are considered to be commercially sensitive information and are in general not shared with system users.

Therefore, it is incumbent upon system owners, administrators and users to ensure that they ‘know their algorithm’. Reference should be made to large scale, independent and transparent testing of FR algorithms such as those undertaken by the National Institute of Standards & Technology (NIST) e.g. [2] as well as undertaking due diligence testing with operationally representative data.

5.3.2 Image enrollment quality

Standardized facial image quality assessment is currently a topic of intensive research [4] and may become integrated into FR systems in the near future. Currently, quality scores, which incorporate factors such as resolution, sharpness and face localization, are proprietary to the algorithm. Depending on the system, quality thresholds can be applied, resulting in enrollment of the face only when the quality of the facial image meets the threshold.

It is recommended that initially, image enrolment quality thresholds are set according to specifications by the system providers. Agencies should undertake a quality assessment to profile their face data in order to ensure that any quality thresholds set are realistic based on the types of images that are received. Guidance on how to undertake such as assessment are provided in the Facial Identification Scientific Working Group (FISWG) document “Facial Recognition Systems Operation Assurance: Image Quality Assessment” [5].

5.3.3 Algorithm performance

The performance or accuracy of an FR system can be described using two key metrics; the False Positive Identification Rate (FPIR) and the False Negative Identification Rate (FNIR). The methodology for generating these metrics requires a large quantity of ground truth data and is outside the scope of this document. However end users should be

familiar with these terms and ask their system supplier to provide evidence of the supplied algorithm performance.

Further consideration should be given to the 'equitability' of the system such that similar algorithm performance should be observed across different demographics (such as ethnicity, sex and age). A detailed discussion of demographic effects can be found in the NIST publication 'Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects' [6].

5.3.4 Rank based systems

It is recommended that prior to operational implementation, testing with ground truth data of known mated pairs is undertaken. Ground truth data should contain images across a range of quality scores such as mugshot - mugshot, CCTV - mugshot etc. It is essential that test data should contain imagery that is representative of that expected in casework. The output from this testing can be used to plot a Cumulative Match Curve (CMC), which summarizes the accuracy of mated-searches and plots the proportion of mated searches returning the mate at rank R or better. This is not dependent on similarity scores (only the rank at which the mated pair is returned), so does not distinguish between strong (high similarity score) and weak mates.

An example CMC plot is provided in Figure 2. In this example, for mugshot probe images, 100% of mated pairs are returned within the top rank 1-3 positions. The number of candidate images that would need to be reviewed to ensure that every mated pair is returned for social media and high definition CCTV probe images is 12 and 14 respectively. However, it can be seen that the number of candidate images that need to be reviewed for low quality CCTV images is 38. Information like this can help provide guidance regarding resource implications and policy setting for use of the FR system.

It is important that agencies run similar tests with their specific algorithm, database and case relevant probe images. This testing will assist with similarity score threshold setting (where required) and determining the optimal candidate list size for review in order to provide reliable search result without putting an unnecessary amount of workload on reviewers[3].

Details on how to run system tests can be found in the FISWG document 'Understanding and Testing for Face Recognition Systems Operation Assurance' [7].

[3] The table under the "Investigation" tab at https://pages.nist.gov/frvt/html/frvt1N.html shows the error rates on whether the mated pair appears in the first 50 rank. Many algorithms with different image types were tested. This information may also provide guidance on optimum candidate list length.

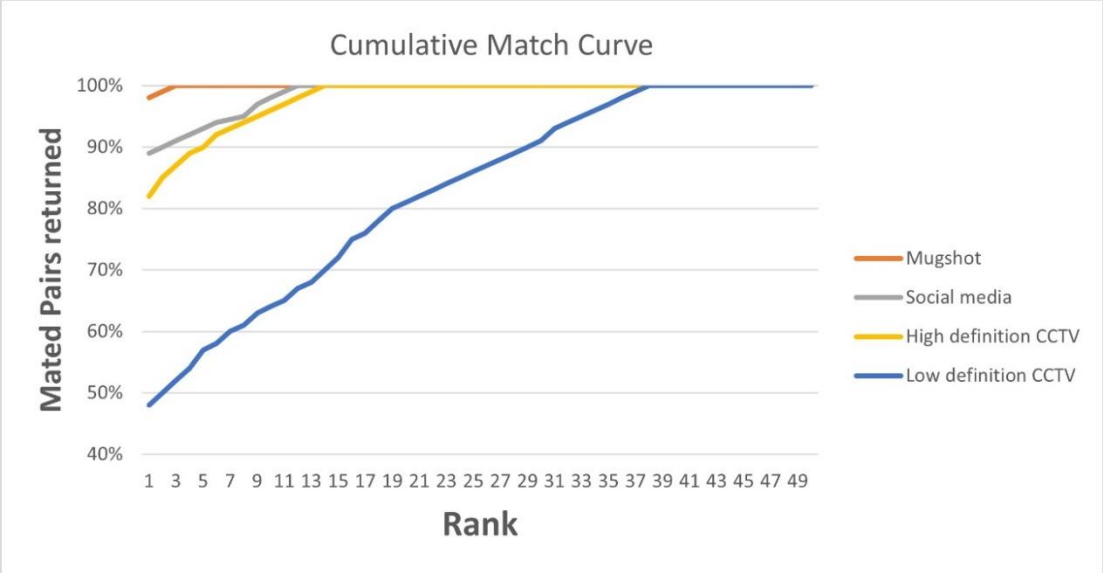

Figure 2: Example CMC for different quality probe images.

It should be noted that there is a relationship between the image quality (of the probe and gallery images) and the reviewer training requirements and time taken to review a candidate list. The following text is a summary from the FISWG document 'Facial Recognition System Methods and Techniques' [8], describing possible scenarios encountered when using FR systems.

Comparison of:

1) A high-quality probe against the high-quality portions of the facial gallery

Optimal images for facial comparison are high resolution and have sufficient focus to resolve features of interest, such as facial marks and facial lines, with minimal compression artifacts or distortion. The obvious advantage of comparing a high-quality probe against a high-quality gallery image is that the practitioner will be able to clearly view features, on each image, to support the morphological analysis of the face. The higher the quality of the probe image, the better the chance the system will return a potential candidate at a high ranking position in the list of reference images returned.

2) A low-quality probe against the high-quality portions of the facial gallery and *vice-versa*

Each agency and practitioner will have his/her own definition of what constitutes a low-quality probe image. These include, but are not limited to, distorted photos, low resolution face, and limited dynamic range, each of which may impede the practitioner's ability to clearly discern the subject's facial features. An FR system may accept a less-than-optimal probe image, but the lack of discernible facial features means that it will be more time consuming to review the list of returned images or that the examiner will be unable to validate a potential candidate.

3) A low-quality probe against the low-quality portions of the facial gallery

The most-challenging scenario, the submission of a low-quality probe image against a collection of low quality gallery images for search by an FR system may be disproportionately impacted by 'pose, expression, illumination' factors. Images returned in the candidate list may be influenced by similar imaging conditions or candidates may be returned on the basis of similar pose, rather than the face similarity. The time taken to review such images will be significant.

5.4 Database of reference facial images

Generally, FR systems for investigative purposes consist of a reference dataset of known individuals against which unknown probe images are searched. The reference dataset contains image(s) and metadata associated to the individual, according to each agency's policies, such as name, Date of Birth, sex, offence for which they were arrested etc.

It is recommended that reference images are captured under controlled conditions and that, **as a minimum**, they meet appropriate quality standards for automated facial recognition as depicted in Figure 3 [9]. Guidance on how to take images that meet this criteria can be found in 'Standard Guide for Capturing Facial Images for Use with Facial Recognition Systems' [10]. Further information can also be found in Police Standard for Still Digital Image Capture and Data Interchange of Facial/Mugshot and Scar, Mark & Tattoo Images [11].

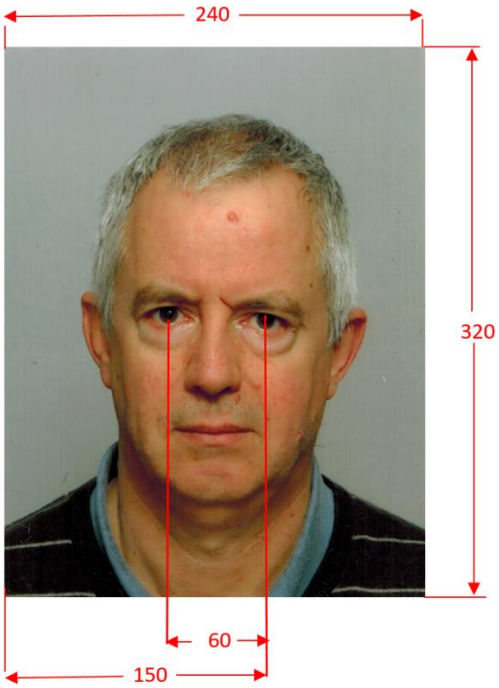

Figure 3: An example of ANSI/NIST-ITL or ISO/IEC 19794-5 compliant image with distances expressed in pixels.

Additional high resolution frontal and non-frontal images (for example 45° or 90° profile) will support post FR search human review of the candidate list.

Due to recidivism, many subjects in the reference database will have more than one image associated to them, where an image is taken every time they are arrested. Studies undertaken by the National Institute of Standards and Technology have demonstrated that FR accuracy can be improved *when* all reference (controlled) images associated to an individual are enrolled and available for searching [12]. Improvements may be a result of early 'template' level fusion or score level fusion. System suppliers should be consulted on the most appropriate strategy for consolidated multiple encounter enrollment. Multiple enrollments of a subject can also assist the human review process.

Where it is necessary to search a probe image against other collections of images, for example those taken under semi controlled conditions or collections of 'unresolved' crime images, it is recommended that these images are contained in a separate partitioned database. Based on agency policy, probe images can be searched against each database partition separately.

6 METHODOLOGY

The following sections describe the recommended methodology for 1:N FR searches as led out in the flowchart in Figure 4.

6.1 Flowchart

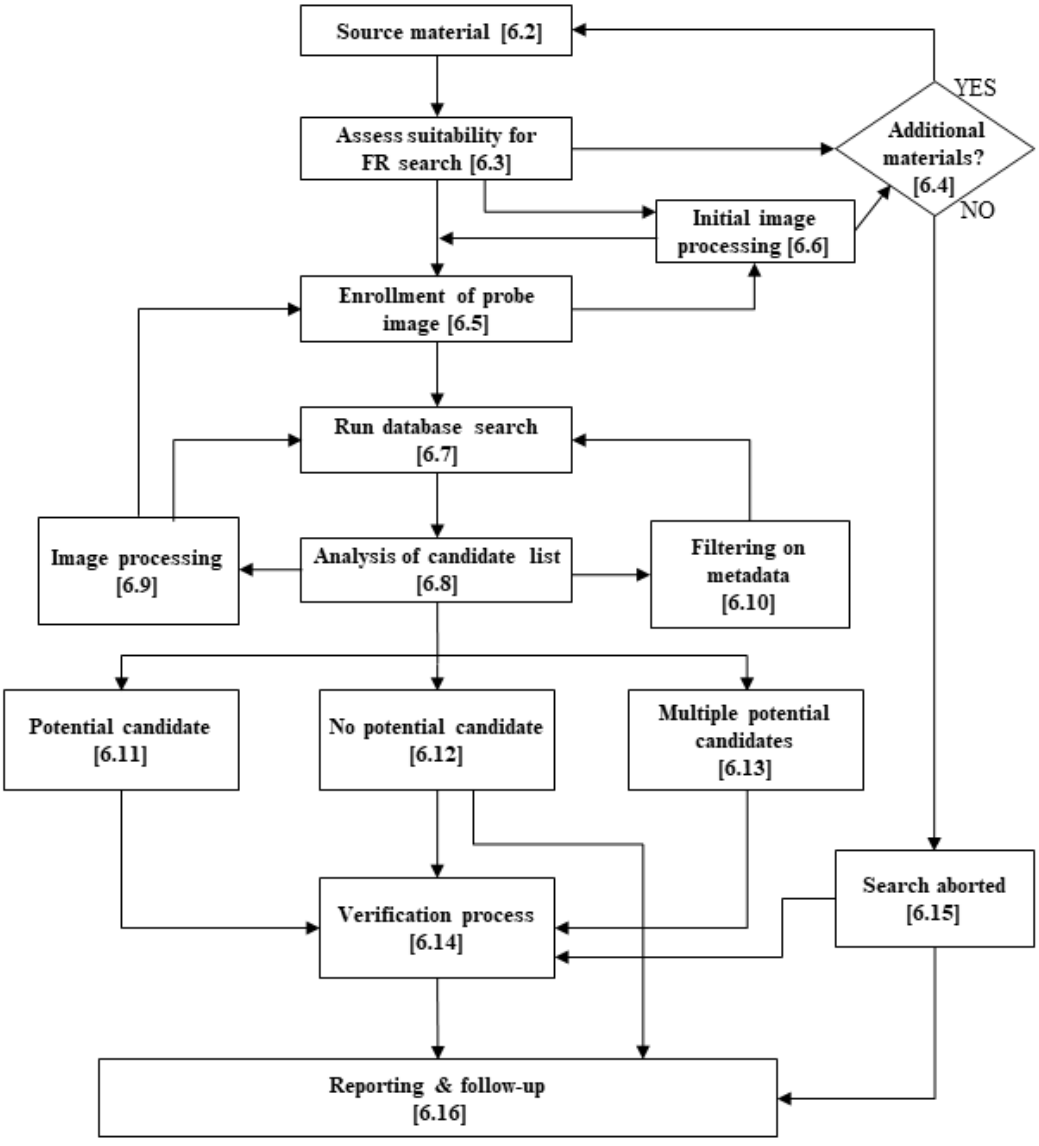

Figure 4: Flowchart showing the workflow.

6.2 Source material

Every FR search starts with the judgment of the material received. Image quality should be judged by the multiple criteria that influence the FR result. FR systems are designed to work best with ANSI/NIST-ITL or ISO/IEC 19794-5 compliant images (see Figure 3 above), and deviation from these requirements will degrade the performance of the FR

system. However, it is possible to achieve good results also from lower quality probe images.

It is recommended that, where available, still images from the original data are used. When the provided material is video, extraction of several frames from the original footage is advised. Frames should be extracted according to best practice digital imaging procedures [13]. Non-original material, for example screenshots, may add distortions, lower quality and quantity of details, resulting in an overall decrease of quality. Some FR systems can directly use video or multiple image search and have various mechanisms for optimizing the search.

The criteria and imaging acquisition factors described below can be used as a general guideline for assessment of the quality of facial images for use with FR systems.

For the remainder of this text it is assumed that one image is used as input to the FR system.

6.3 Assess suitability for FR search

6.3.1 Basic criteria

According to the ISO/IEC 19794-5 standard, some basic criteria should be met in a facial image. These include that the resolution of the image should be at least 60 pixels between the center of the eyes, and the eyes, nose and mouth should all be visible in a frontal image.

In reality, probe images are seldom depicted under these controlled conditions. Although FR systems work best with eyes, nose and mouth all visible in a frontal image, current systems may also work with off angle poses and significant parts of the face covered e.g. by a facemask. Many system vendors also claim that results with resolutions down to around 15 pixels between the eyes are feasible.

Apart from the basic criteria, the outcome of the FR system is influenced by other image properties. No strict criteria can be set for these properties, but the factors listed below should help in the judgement of the suitability of the image for FR. Some of these properties can be 'enhanced' using image processing, but it is important to have an understanding for how image processing might modify certain characteristics of the face. Image processing of the probe image is described in sections 6.6 and 6.9.

6.3.2 Imaging acquisition factors affecting FR

The following factors are known to negatively influence FR performance, including, but not limited to:

- Very low resolution images
- Compression: (Re-) compression should as much as possible be prevented because it will result in quality loss, unless lossless compression is used
- Lighting: Over or under exposure or hard shadows present in the image. Saturated or black areas over the face should preferably be avoided
- Low/high contrast

- Sharpness: The face should ideally be in focus. If blurred, details may get lost. Also, movement during acquisition resulting in motion blur may decrease sharpness
- Artifacts due to, for example, compression, motion, signal error or data corruption or re-acquisition artifacts like scanning of ID documents with security elements or photographing an image displayed on a screen
- Noise (e.g. images taken at low light). Some noise reduction using image processing may be performed, but will influence sharpness of the image.
- Distortion due to close distance to the camera, lens properties (e.g. fish-eye) or aspect ratio changes
- Occlusions of parts of the face due to e.g. dark glasses or mouth-covering
- Pose: High viewing angle and the degree of pitch and/or yaw. FR will work well on images with up to 45º yaw. Although many FR systems are now developing capability for recognition from profile images (up to 90º yaw), this is still nascent technology and caution should be exercised if these are the only probe images available for searching

Visualization of pitch/roll/yaw is depicted in Figure 5.

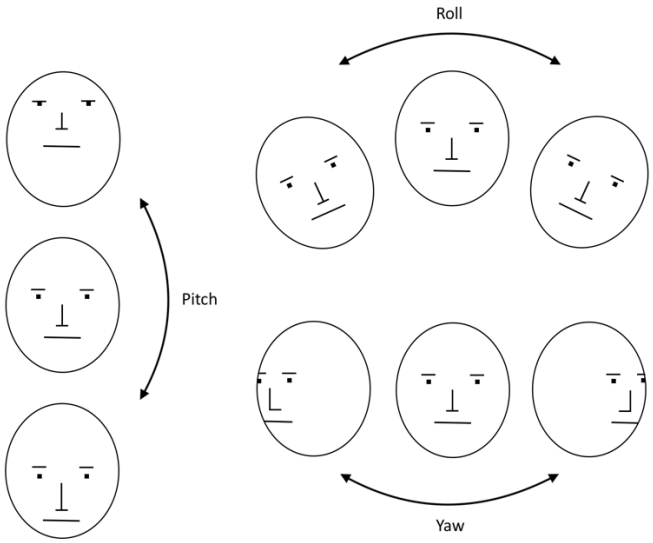

Figure 5: Visualization of Pitch, Roll & Yaw of the face

- Mirroring: Many selfie images secured from mobile phones are mirrored. Information in the background might in some cases help the assessment
- Image manipulation: Intentional manipulation (including morphing or “beauty filters”)
- Environmental factors: Weather (rain, etc.) windows/glass/reflections/insects between subject and camera

6.3.3 Subject (Human) Factors affecting FR

Changes in a person’s face may negatively affect FR performance, including, but not limited to:

- Expression
- Ageing
- Weight/ weight change
- Health/medical - Illness/hydration/drug use
- Post mortem changes
- Intentional alteration (makeup, surgery)
- Loss or change of features through self-mutilation or inflicted

While some facial features are more stable over time, others can change radically. The stability of facial features in adults is listed in [14].

FR system performance on images of children (under 16 years) is known to be more challenging because of the rapid changes in craniofacial morphology for infants through to adolescents [15]. There are two factors which need to be considered;

- The absolute age of the subject in the test images, where generally, young children (aged 0-13) are both harder to correctly recognize (high intra-variability) and harder to distinguish between (low inter-variability) [16]
- Age variation, the age difference between the probe and reference image(s) with a general trend of decreasing recognition rates with an increasing time gap between images, where the reference image was taken up to 16 years previously when the subject was a child [17]

6.4 Request for additional materials

If the submitted material fails to meet requirements in terms of quality, it might be relevant that the reviewer contact the requester and ask for supplementary materials. Such action should typically be balanced according to the importance of the case, the human resources available and agency policies. If no additional materials are available, the examination should be aborted according to section 6.15.

6.5 Enrollment of probe image

The probe image should preferably be enrolled in its original format along with relevant metadata. However, in some cases it may be warranted to start with initial image processing according to section 6.6, before the enrollment.

On enrollment, the face detection algorithm typically displays the automatically determined eye positions. Only if these positions obviously deviate from the correct positions, should the reviewer change these. The correct eye position should be positioned according to the technical specifications of the FR system.

Manual modification of the face quality thresholds (see section 5.3.2) on an image by image basis may be allowed according to agency policy in order to enroll the image.

If the probe image fails to enroll, the operator could proceed with adjustment of the eye positions or initial image enhancements according to section 6.6. If the image still fails to enroll, the reviewer should either ask additional materials, or the examination should be aborted.

If there is more than one image of the unknown person, for example multiple still images or frames from a video etc., it can be beneficial to submit more than one image for searching instead of only picking one, which appears to be the best. The reason is that a single image, which might seem the best for the human eye, is not always the best according to the algorithm criteria. Choosing the best image by the algorithm quality scores based on alignment and reliability of the face after detection are not necessarily the best predictors of success. There is a chance that the mate image in the database is captured in a less than ideal pose, corresponding better with an image of the unknown person that is not considered to be the 'best image'.

For video, some systems are able to select the most appropriate facial image(s) of the unknown person (based on internal quality metrics) for searching, or are capable of fusing several frames together. For the latter, this may not be the most effective strategy as lower quality images might be included in the search that adversely impact the outcome.

6.6 Initial image processing

If the face in the image can not be found, the enrollment will fail and the facial image search will not take place. To facilitate the enrollment, initial image processing might be performed.

It is recommended that intital image processing should not significantly alter the biometric data and should be kept to a minimum. Processing could include for example: cropping (to remove background and /or isolate the relevant face if there are multiple faces), marking the centre of the eyes, horizontal flip/mirroring (this should be utilized if the probe image might have been taken as a reflection, in case of a 'selfie' image, or if the image may have been flipped in transmission), enlarging using interpolation or aspect ratio corrections. No (additional) compression should take place and caution should be exercised when adjusting the aspect ratio as it may result in altering the geometry of the face.

Image processing is further discussed in section 6.9.

6.7 Run database search

After the face (on the probe image) has been correctly enrolled, a 1:N search can be run by the FR system against one or more selected reference database(s). The algorithm compares the template generated from the face on the probe image to the templates generated from the facial images in the database and returns a candidate list of reference images.

As was mentioned above under section 5.2, candidate lists can be rank-based, threshold-based or both.

6.7.1 Rank based approach

The rank of the true match may be affected by the quality of the probe image and reference images (all the factors described in sections 6.3.2 and 6.3.3) as well as the overall size of the reference database. Despite the possible effects of these factors, modern algorithms have significantly improved in the rate of returning a true mate at a high rank position, although its similarity score may be low.

In case of a rank-based approach, where the similarity score threshold is set to 0, the number of the candidates in the list (candidate list size) is configured before the search run. In most systems, there is a default candidate length, but in most cases this can be changed manually by the user. Law Enforcement Agencies in EU countries usually return searches with between 10-200 candidates [18]. For low quality images, the longer the candidate list, the greater the chance that a true mate is present. However, it is worth keeping in mind that the human review of the list needs time and long candidate lists will affect the workload. Additionally, research has demonstrated that long candidate lists of 100 or more images can result in increased false alarms, lower detection of true matches, lower decision confidence and increased response times [19]. These factors and the severity of the crime should be taken into consideration when determining candidate list size.

6.7.2 Threshold based approach

With threshold-based searches, the number of candidates returned in the candidate list depends on the similarity score threshold. This threshold can be default as recommended by the system supplier or manually set by the system administrator or the user. The threshold should be set according to the operational requirements and resources available. It should be noted that a high threshold setting minimizes the workload for human review (for example eliminating non-mated candidates being returned) but may result in true mates being excluded from the candidate list. This is depicted in Figure 6. The converse is also true; a low threshold increases the chance that a true mate is returned in the list but is resource intensive from a review perspective.

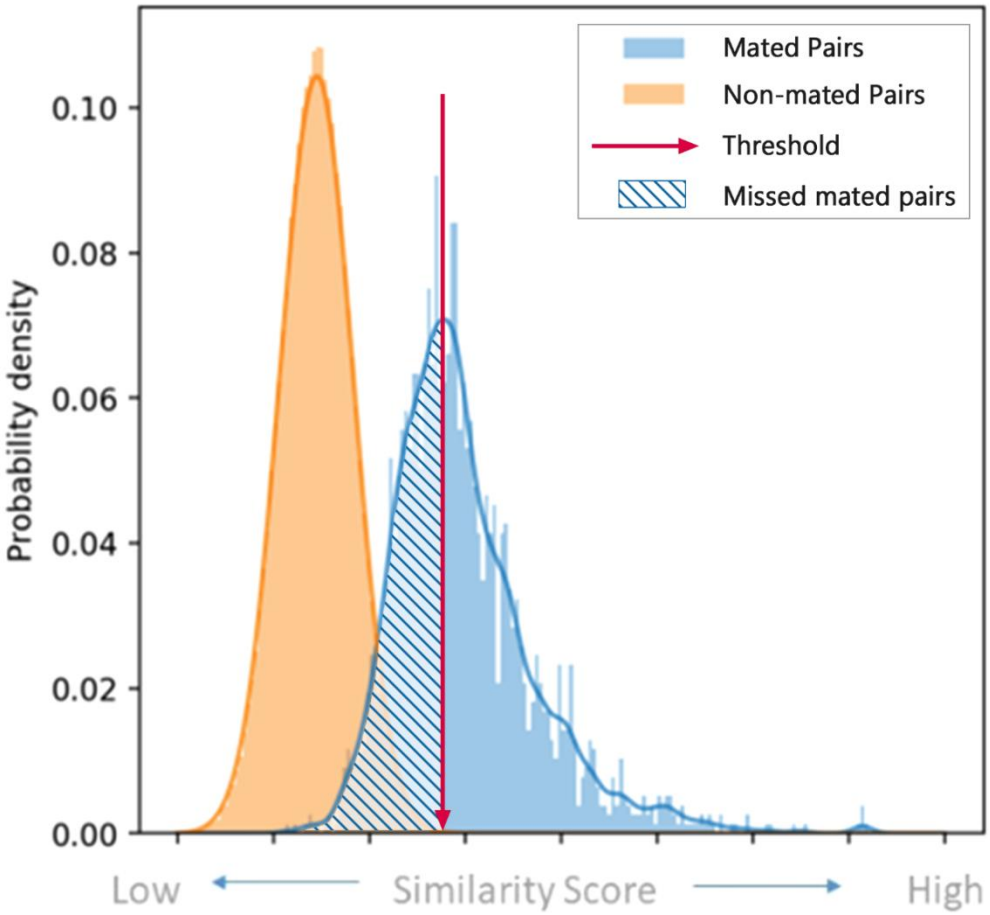

Figure 6: Similarity score distribution showing impact of setting a high threshold – everything (including a significant proportion of mated pairs) to the left of the threshold setting would be excluded from a candidate list.

The similarity score between two facial image templates is affected by both the quality of the probe image and the quality of the reference image. For example, when the probe image and one of the reference images has a similar low resolution or a similar type of noise, these may gain a high similarity score regardless of facial morphology. These reference images of candidates can outrank the true mate and may hinder the search results. Therefore, it is recommended to store different quality reference images in separate databases and run separate FR searches against each collection. Examples for quality separation are ICAO standard versus non-ICAO standard image database; mugshot gallery versus records of uncontrolled facial images; etc.

6.8 Analysis of candidate list

An image is merely a representation of reality, and a depiction of an individual is not perfectly identical to its original. Also, the appearance of a person will change over time.

This means that there will be a range of similarity scores returned by any FR system upon a comparison of two images of the same individual (intra-variability).

Conversely, there is also a risk that the system perceives different people as very similar to each other, even beyond obvious similarities such as close kinship, and returns a high similarity score. As can be seen in Figure 6, the similarity scores of the mated and the non-mated searches both show a range of values, which may be different depending on the data sets at hand (e.g. mugshots versus low-quality CCTV images). It is expected that there will be an overlap in the similarity score distribution for both same person (mated) and different person (non-mated) comparisons by the system.

When using an FR system with good quality images, the difference between images of the same person is small, resulting in high similarity scores, while the difference between images of different persons is generally larger, resulting in low similarity scores. Hence, the scores usually display a low intra-variability and high inter-variability when using good quality images, resulting in good discrimination between mated and non-mated score distributions. However, for low definition CCTV images the situation may be different. The poor quality (uncontrolled) images of the same person may show a large difference, e.g. due to differences in pose, resulting in a large variation in similarity scores (high intra-variability), while images from different persons may show a lot of similarity, e.g. due to similarity in pose, resulting in relatively high similarity scores for images of different persons.

The imaging or subject factors described in sections 6.3.2 and 6.3.3, may also hamper and affect the manual analysis of the candidate list and the reviewer should never presume that the correct identity will be at the top of the search list even if he or she exists in the database. Nor can it be presumed that the correct identity exists in the database at all.

Nevertheless, the rank and similarity score might give valuable information to the reviewer. Examples for such scenarios are (i) when the difference in the similarity scores between the persons at rank 1 and rank 2 is large, (ii) if different images of the same person appears more than once amongst the reference images in the candidate list or (iii) if the same person appears in different candidate lists from multiple database searches of the same probe image or (iv) if reference image(s) of the same person appears in different candidate lists from multiple searches with different probe images of the same person of interest.

Some agencies use the rank and score information, but other agencies prefer not to show the similarity score and/or present the candidate list in random order so as to not bias the reviewer by the FR system outcome. Whether information from the FR system is used or not in the evaluation is therefore highly agency specific.

The analysis of the candidate list normally includes a first quick assessment to exclude candidates, followed by a more in-depth comparison of the remaining prospective candidates. There might also be other relevant information that can be used in the evaluation of candidates.

6.8.1 Exclusion of candidates

The FR database search results in a list of reference images, generally ranked by similarity score. Depending on the quality of both the probe and the reference images, a first pass of exclusion of unlikely candidates can be performed using a holistic comparison method by a reviewer. In the holistic comparison, an exclusion of unsuitable candidates is made based on the basic features of the persons, such as sex, **obvious** skin tone differences[4] or other distinctive differences. Holistic comparison can therefore be regarded as a first and quick evaluation to eliminate candidates and identify candidates for further review using a more detailed morphological comparison.

6.8.2 Comparison of candidates

For a more in-depth morphological comparison (when the visible facial components and sub-components are compared), the remaining reference images in the candidate list should be viewed side by side with the probe image. Most FR systems will include a set of tools, such as linked zoom to facilitate analysis. However some of these tools such as superimposition (the placement of one image or video over another and adjusting the transparency) [20], wiping (to slide or fade the visible part of a superimposed image to illustrate similarities (slow wiping) and differences (fast wiping)) [21] or photo-anthropometry (the measurement of dimensions and angles of anthropological landmarks and other facial features) [22] are not reliable and should not be used [1].

When comparing facial features between the probe and the reference images, any factors that might have an impact on the appearance of a person need to be considered. These include technical and environmental conditions (e.g. resolution, lighting, reflections) and subject factors (e.g. ageing, face expression). Keeping these influencing factors in mind, facial features are analyzed that either support or oppose a possible match. If a person exists with corresponding features to the person in the probe image, this person might be regarded as a potential candidate depending on the level of details observed. More on the decision for potential candidate/no potential candidate/multiple potential candidates can be found in sections 6.11, 6.12 and 6.13. A potential candidate listed by the first reviewer during an FR run should be verified (according to agency SOPs) by a second FR reviewer, ideally in an independent review process according to section 6.14.

Depending on the analysis of the candidate list, image processing and/or metadata filtering according to sections 6.9 and 6.10 respectively might be applied and the search rerun.

6.8.3 Comparing images of children

Comparing or matching facial images of children is a challenging task for humans, just as for FR algorithms. People from novices to experts generally demonstrate significantly

[4] Care should be exercised when using skin tone to exclude an image as apparent differences may be due to image factors (capture, color tone etc.)

lower performance with images of children in face matching tasks compared to their performance with images of adults [23]. Both absolute (chronological) age and age variation of the children depicted on the images impacts the human performance [24]. Images of younger children, as well as greater age variation make these tasks more difficult, which is likely due to the rapid and great amount of facial changes happening throughout childhood and the less discriminating facial features of younger children [25]. Facial morphological development during childhood is not evenly paced over time, facial features have growth spurts and maturation occurs at different ages [25]. Studies on human performance have indicated that there is a false positive response bias for 1:1 comparisons of child images, and that this increases with wider age variation [24].

In summary, this means that reviewing searches with images of children should be done with extra care. According to research, methods recommended for the comparison of adults have their limitations with children's faces and to date there is no international standard methodology accepted for the latter [24]. Understanding the early facial development patterns and child-specific training of reviewers, who are expected to do work with images of children is recommended [24].

6.8.4 Other relevant information

When there is more than just the face visible of the person in the probe image, it is recommended to base the comparison on all visible features of the person. Some FR systems allow direct access to the police records where more reference images of the candidates might be stored. These images can include, for example, face profile (left/right), full body or scars/marks/tattoos.These images should be added to the comparison, especially when the reference facial image in the FR system does not provide all aspects of the person the probe image offers.

Other available information may also be taken into account, for instance additional probe images, capture date or written descriptions of the person in the database or police records.

6.9 Image processing

If the analysis of the candidate list does not yield a potential candidate, the reviewer may choose to process the image further so as to refine the search results. Such processing should be made in accordance with [26] by using either the system's own functions or a separate image editing tool (after which the image can be enrolled into the system again). It should be noted that the purpose of image processing is to optimize the image for searching by the FR system, not to create an aesthetically pleasing image.

Image procesing techniques that can be applied to influence the FR system performance and may include, but are not limited to:

1. Histogram equalization
2. Brightness or contrast adjustment
3. Color/tint corrections
4. Grayscale conversion

5. Noise reduction
6. Red eye reduction
7. De-blurring or sharpening

If no potential candidate is found after a search using an image processed in this way, the reviewer may choose to proceed with processing the image in a way that has a higher risk of altering its biometric/geometric contents, such as:

8. Aspect Ratio correction
9. Lens distortion correction. Some images, such as those from smart phones, automated teller machines (ATM's) and Body Worn Video cameras that use wide-angle lenses typically exhibit significant perspective ('barrel') distortion.
10. 3D pose correction

A log or audit trail should be kept of the image processing techniques and settings applied to the image.

Caution: Image processing should always be performed on a copy of the original image. The copy image should not be further compressed when image processing is applied.

The reviewer should primarily use methods that do not significantly alter the biometric/geometric data in the image.

The effect of image processing will vary with different FR systems and may in some cases even degrade performance rather than improve it.

The user should also use the original image (as well as the processed image) to aid review of the candidate list and decision making

6.10 Reducing the search space using metadata filtering

The reviewer should primarily search the probe image against the entire gallery. If such a search does not yield the desired results, refined searches can be performed by filtering the database on metadata such as sex or approximate age span.

By using filtering, the database size is reduced making it possible to generate a more case specific and relevant candidate list. The risks of applying such filters should be noted however, as metadata may have been filled out inaccurately during the booking procedure. Certain metadata may be used differently at different booking stations (for example the sex of transgender persons), while others may contain subjective assessments (for example age estimates).

Caution: It is important to be aware that in using metadata filtering, there is a risk that the correct identity will never be included in the candidate list despite existing in the database.

6.11 Potential candidate

When a candidate is considered, the reviewer must evaluate both the observed similarities and differences of the persons in the images and, given the different

imaging or subject factors, make a final conclusion. Also, other relevant information (see section 6.8.4) might be weighted into the decision.

A potential candidate means that the person in the reference image shows significant similarities to the person in the probe image. A potential candidate does not mean that the two persons must share the same identity.

In order for a potential candidate to be reported, no differences other than what is considered to be caused by imaging or subject factors (see section 6.3.2 and 6.3.3), should be allowed.

In some cases it might be appropriate to provide more than one candidate, see section 6.13.

6.12 No potential candidate

If none of the persons in the candidate list show significant similarities to the person in the probe image, or if all persons in the candidate list show dissimilarities that cannot be attributed to imaging or subject factors, the outcome of the search is *no potential candidate*.

That no potential candidate is found does not necessarily mean that the correct identity does not exist in the database. There are a number of reasons why an FR search might not result in a match against a person even if that person exists in the database, including imaging and subject factors as well as FR algorithm or manual review limitations.

6.12.1 Save probe to "unresolved cases database"

Depending on the agency specific procedures, the probe images resulting in *No candidate* might be enrolled into a database of unresolved cases. These images might also be searched regularly against new entries to the reference database, and give retrospective matches. Typically, when an unresolved cases probe image is matched to a known identity, the probe image is removed from the unresolved cases database.

Also, it is possible to match unresolved probe images against other unresolved probe images, in an attempt to detect links between different crimes. It should be noted that this is a challenging use case of FR technology, and that a low quality image matched against another low quality image may erroneously result in a high similarity score. The challenges of this scenario are discussed further in section 5.3.4.

The unresolved cases database use case is not further discussed in this guideline.

6.13 Multiple potential candidates

In some cases it might be appropriate to report more than one potential candidate. Examples when this might occur are when persons in the reference database are very similar to each other (such as identical twins), and the probe image quality does not provide the reviewer sufficient details to separate between them. Another example could

be when the reviewer suspects that the same person has been enlisted in the reference database under multiple identities (aliases).

Another scenario for reporting more than one candidate can be when no potential candidate was concluded, but when it was not possible to exclude all persons on the candidate list due to imaging or subject factors.

It is not recommended that the candidate list returned by the FR system is reported without having been subject to a human review process.

6.14 Verification process

To mitigate the risk of a false positive result, it is highly recommended that **all** potential candidates go through a verification process before the results are reported.

Some agencies also use a verification process for searches which have not resulted in a potential candidate, a strategy which is especially recommended in an operational setting where false negative results can lead to negative consequences.

The verification process can start at any point in the flow chart according to agency specific procedures (for example a second reviewer could run the search process independently) and may be conducted by;

- A second facial reviewer undertaking a blind verification process, or
- A second facial reviewer verifying the results of the first reviewer (non-blind), or
- One or more facial examiner(s) performing a 1:1 facial image comparison of the probe image and potential candidate(s). The process of such a comparison is described in detail in [1].

In blind verification, the result from the first reviewer is not known to the second reviewer, while in non-blind verification the result of the first reviewer is known. Non-blind verification will have a higher risk of confirmation bias.

If there is no consensus about the search result, the agency should have in place a policy for how disagreements will be handled. Each agency should consider the benefits and risks to the investigation when setting their policy, for example, reporting an incorrect potential candidate versus not reporting anything. A potential strategy is that one (or more) additional reviewer(s) provide their opinion(s).

6.15 Search aborted

The FR search may need to be aborted for a number of reasons, including:

- The probe image quality does not meet expected standard during the pre-assessment
- The probe image fails to be enrolled into the FR system
- The FR system fails to find the face in the probe image
- The FR system returns reference images that are matched on non-facial features such as image distortions
- The request is withdrawn by the requester

Whether any of these reasons to abort include a verification process or not will be determined by agency specific procedures.

6.16 Reporting and follow-up

Results after the FR search process can be either:

- A potential candidate
- No candidate
- Multiple potential candidates
- Search aborted
- Other agency specific answer

For all outcomes, the results must be communicated to the requester. Normally, the outcome of a FR search is reported as an investigative lead report.

The information to be included in an investigative lead report is normally agency or use case specific. FISWG recommend that reports should include any agency disclaimer, identifying the limitations of the method used and the recommended usage of the report [27].

6.16.1 Reporting a potential candidate

When a potential candidate is reported, it is recommended that the following information is provided:

- That a potential candidate does not indicate a positive identification of the person, and a summary of the reasons why.
- That important decisions such as fundamental rights limitations (for example arrests, crimes imputation, freedom of movement, etc.) should not be adopted based exclusively on the potential candidate reported.
- That the investigative lead report is not intended as evidence in court proceedings. The main purpose for the investigative lead report is to provide criminal investigation with intelligence.

6.16.2 Reporting no potential candidate

When no potential candidate is reported, it is recommended that the following information is provided:

- That a “No potential candidate” report is no guarantee that the correct identity is not in the database, and a summary of the reasons why.

6.16.3 Reporting multiple potential candidates

When multiple potential candidates are reported, it is recommended that the following information is provided:

- Clarification that the manual analysis could not conclude a single potential candidate.
- Whether the potential candidates provided are listed according to any specific ranking and if so which (for example the FR ranking or random).

6.16.4 Auditing trail

During the whole FR search process, from receiving the request, through the FR search, the human review of the candidate list and reporting, it is recommended that an audit trail of the case is recorded. Some topics that could be recorded, include;

- Administrative details
- FR search details
- Human review and verification details

Which information is recorded should follow agency specific protocols. The level of documentation should be proportionate to the task and balance the workload, resources and intended use of the report.

6.16.5 Follow-up

If the investigation against a reported potential candidate is pursued and the image materials are to be used as evidence, it is recommended that a forensic 1:1 facial image comparison is requested, to evaluate the imagery evidential support or not of the proposed candidate. The process of such a comparison is described in detail in [1].

Some agencies do not report a potential candidate when such is found, but instead require that a forensic 1:1 facial image comparison between the probe and reference images is performed and reported instead.

7 TRAINING AND COMPETENCY

For the workflow described in this document, a human intervention is required to review the results from the FR system.

Studies have consistently demonstrated that human performance in comparing unfamiliar faces is highly varied, and on average much poorer than our ability to recognize familiar faces [28].

Research has also shown that human operator performance can substantially impact on the reliability of results from an FR search [29], which is often overlooked when evaluating the performance of FR systems [30].

Therefore, the selection, training and testing of FR users requires careful consideration to mitigate the risk of error from human review. This section provides recommendations on the following aspects of FR users:

- Types of FR users
- Selection of individuals for FR user roles
- Formalized training
- Ongoing competency assessment

Untrained individuals, who have not received any specific training in FR review beyond basic vendor training for their FR system and have not been specially selected for the role, are not recommended as FR users. This is due to their lack of formalized training and absence of demonstrable competency.

Untrained individuals are unlikely to apply any specific processes or methodology when reviewing candidate lists, therefore the accuracy of the review will be largely dependent upon their innate ability at comparing unfamiliar faces. In the absence of formalized procedures basic users may be particularly susceptible to sources of cognitive bias, such as contextual information, potentially increasing the risk of error.

7.1 Types of FR users

FR users are responsible for operating the FR system, including enrolling unknown images, searching against a database, reviewing the candidate list to determine potential matches and verifying results.

FR users can be classified as follows, in accordance with the extent of the user's training, knowledge and demonstrable competency:

7.1.1 Facial reviewer

Facial reviewers are specialist FR users that have received formalized training and should be able to demonstrate ongoing competency and proficiency in FR review. 'Facial reviewers' typically do not receive as extensive training as facial examiners and work in high throughput environments with greater time constraints than examiners.

Due to the large number of comparisons that a reviewer must make when reviewing candidate lists it is expected that the processes used will be less rigorous and with

comparably less documentation of observations, than a Facial Image Comparison (FIC) undertaken for forensic or evidential purposes.

FISWG [31] defines a facial reviewer as a FR user who:

"Performs a comparison of image(s)-to-image(s) generally resulting from the adjudication of a candidate list generated by an FRS. The comparison results are often used in either investigative and operational leads or intelligence gathering applications."

Facial reviewers represent a diverse set of users and the results from studies of facial reviewer performance on facial comparison tests are similarly diverse, with some groups of reviewers demonstrating superior performance to lay persons and others not [32]. Approaches to training also vary from agency to agency, from one day courses to long-term training programmes that can last for a year or more. The content of training materials are also highly varied [33].

At the time of writing there is no published data to support a recommended approach to training for facial reviewers, however, when selecting and training reviewers the following general principles should be adhered to:

- Facial reviewers typically receive shorter durations of training compared to facial examiner [33], therefore reviewers should also be selected based on innate face-matching ability, using ecologically-valid tests that are representative of the operational work they will undertake [34]
- Facial reviewer training should be evidence-based and validated as suitable for the intended use case [35]
- Facial reviewers should undergo continuous professional development, including ongoing competency testing using operationally-representative images and following agency specific policies and procedures

7.1.2 Facial examiner

Highly trained specialists in forensic FIC, 'facial examiners' typically work in small teams and operate within a controlled quality management system that is sometimes accredited to an international standard (e.g. ISO 17025). Facial examiners are considered to be experts in FIC and can provide opinion-based evidence in a court of law.

FISWG [31] defines the role of a facial examiner as:
"*performs a comparison of image(s)-to-image(s) using a rigorous morphological analysis, comparison, and evaluation of images for the purpose of effecting a conclusion, often used in a forensic application."*

Facial examiners follow detailed procedures to perform FIC, generally based on the ACE-V framework (Analysis, Comparison, Evaluation and Verification) [36], according to international recommendations such as the *ENFSI BPM for Facial Image comparison* [1]. Studies have consistently demonstrated that facial examiners have superior performance in FIC when using their standard policies and procedures [32]. However, given that the procedures used for forensic FIC are overly time-consuming they are unlikely to be directly applicable to the task of FR review.

If working in the role of an FR user, facial examiners may apply a less rigorous approach to the review of the candidate list, followed by a more detailed 1:1 comparison of viable candidates. In some situations, the facial examiner may only conduct a 1:1 comparison of viable candidates, with the candidate list review being conducted by a facial reviewer.

Regardless of how the facial examiner operates as an FR user it is recommended that, as for facial reviewers, they have received validated training in the task of FR review and undergo continuous professional development and ongoing proficiency testing.

Given the different task demands between FR review and forensic FIC, it is not appropriate for a facial examiner to carry out FR review without relevant training and testing to demonstrate competency in the task, even if they have demonstrated competency in forensic FIC.

7.2 Selection of FR users

Human innate ability in unfamiliar facial comparison is highly varied and falls onto a wide distribution of performance. At the upper end of the distribution, a small number of individuals consistently perform exceptionally well at the task, often referred to as *super recognizers* in the academic literature [37]. At the bottom end of the distribution individuals perform exceptionally badly and may meet the diagnostic definition of *prosopagnosia* (or face blindness) [38]. The majority of individuals are somewhere in between.

Given this heterogeneity in facial comparison ability, agencies should consider testing personnel prior to selection as FR users and enrollment in formalized training, to identify higher performing individuals.

There are numerous laboratory-based tests for unfamiliar face comparison ability that are freely available, and many have normalized control data for comparison of performance [39]–[42].

Whilst such tests may provide an initial indication of face comparison ability, there is limited evidence that performance on laboratory-based tests directly correlates with improved operational performance in real-world settings [43]. Additionally, individual performance can vary across different, related face-processing tasks [44] and even within the same task due to extraneous factors such as fatigue and motivation.

Therefore, a single laboratory-based test may not provide an accurate picture of an individual's consistency of performance in facial comparison, and may not be directly applicable to applied tasks, like reviewing FR candidate lists in operational settings.

Given that many laboratory-based tests are freely available online there is also a risk that individuals may repeatedly take tests or employ some other means to fake a high score. So, whilst online laboratory-based tests can provide an initial indication of facial comparison ability, agencies should also use a variety of screening tests that are not available to the public [42].

Agency screening tests should be task-specific and representative of the types of images that an FR user will encounter operationally [34]. Potential FR users should ideally

undergo multiple screening tests, taken on different days to give a measure of consistency in performance.

FR user screening tests should be validated for their intended use-case and provide an accurate measure of an individual's performance by having a criterion score or cut-off for superior performance.

7.3 Training

Facial reviewers and facial examiners undergo formalized training to achieve competency in their role. There has been limited research into the efficacy of formalized training for FR review, however studies have shown that short, one-off training courses of three days or less are largely ineffective at improving facial comparison performance [35], [45].

Approaches to training FR users vary substantially between agencies. In some cases training lasts one day or less, whereas other agencies provide months of training that includes one to one mentoring [33]. There is some indication that extensive on the job training and mentoring is a source of expertise for facial examiners carrying out forensic FIC [46], however the benefits of such training has not, to date, been evaluated for FR users in the review of candidate lists.

In the absence of empirical studies of FR training efficacy agencies should validate their training to ensure that the approach is effective at improving operational performance and reducing the risk of error.

Training should also be evidence-based and incorporate exercises that have been demonstrated to improve facial comparison performance. Feed-back training is one such approach for improving facial comparison performance [47]. Feedback should be provided to trainees on their performance during training tasks, which allows learning from mistakes and indicates when and why trainees have performed well at a task. There is also some evidence that collaborative working on facial comparison exercises can provide a training benefit, particularly for lower performers [48].

Given the limited effectiveness of short, one-off training courses and the absence of data supporting the use of longer-term training and mentoring for FR users, agencies should supplement training with screening tests for selection prior to training (section 7.2 and task-specific competency testing for completion after training (section 7.4).

7.4 Competency Testing

In addition to screening tests and training, FR users should also participate in ongoing competency testing. Competency testing is used to demonstrate that an individual can reliably and accurately complete a particular task, in this case FR review and related sub-processes (e.g. processing of probe image, review of candidate list, and verification of results).

FR users should be competency tested after selection and training, and prior to undertaking FR reviews.

It is preferable for competency tests to be conducted using test items of known ground-truth, however, in some instances competency may be demonstrated through on-the-job training, such as during workplace mentoring.

Competency tests should encapsulate all of the processes an FR user is expected to undertake and should be conducted according to local policies and procedures. When designing competency tests, in addition to ensuring they are task-relevant, agencies should also ensure that the purpose of the test is clearly specified and is testable (i.e. can be assessed as pass or fail). Tests may also require multiple measures of accuracy for evaluation of results. The following measures of accuracy for human review may be useful:

- True and false positive identification rates
- True and false negative identification rates

The stimuli used in a test should be representative of the material FR users will encounter operationally with consideration given to the difficulty of the test, ensuring that the test is sufficiently challenging to provide a meaningful test of competency but not so difficult that a pass is unachievable.

For further advice on the design of human performance tests in forensic disciplines see [49].

Agencies should have a documented procedure for competency testing that defines at what intervals ongoing competency testing should occur and what action should be taken if an FR user´s competency has lapsed.

8 VALIDATION

The final output from an FR process is derived from multiple interactions between automated computer components and human users, which can all have an impact on the accuracy and reliability of the result [30].

To ensure that the FR process is fit for purpose agencies should consider carrying out an end-to-end validation of the entire FR process, including an evaluation of the performance of the algorithm, the competency of the users and the impact of any interactions between the automated components and the users. Such a validation study encompasses the entire FR method, including the testing processes discussed in section 5.3 (Know your system) and section 7.4 (Competency testing) of this guideline.

For agencies intending to gain accreditation for their FR process under ISO 17025:2017 standard, it is a requirement of the standard that all methods are formally validated prior to implementation [50].

ENFSI Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science [51] provides guidance on validation for both quantitative and qualitative methods.

In addition to the requirements of ISO 17025:2017 [50] and validation guidance provided by ENFSI [51], when designing validation studies for FR processes the following should also be considered:

- Validation of FR processes should only be conducted using ground truth material that is representative of the types of images that will be encountered in cases.
- A formal procedure for the planning, carrying out, reporting and approval of the validation exercise should be documented prior to starting the validation.
- When documenting the validation plan the following should be clearly established:
    - The requirements of the end-users of the method, which may include the FR user, the investigator and the wider criminal justice system.
    - The specifications for how the method meets the end-user's requirements. Specifications should be single testable statements that can be tested during the validation exercise.
    - Acceptance criteria that determine whether testing has satisfied the specifications of the end-user requirements.
- Validation tests should only be conducted by competent FR users.

Any major changes to agency procedures or updates to software, in particular the FR algorithm may require all or part of the validation exercise to be repeated.

9 REFERENCES

[1] European Network of Forensic Science Institutes, “ENFSI Best Practice Manual for Facial Image Comparison,” 2018.

[2] P. Grother, M. Ngan, and K. Hanaoka, “Face Recognition Vendor Test (FRVT) Part 2: Identification,” 2019.

[3] EU Directive, *2016/680 Article 10*. Official Journal, 2016.

[4] P. Grother, A. Hom, M. Ngan, and K. Hanaoka, “Ongoing Face Recognition Vendor Test (FRVT) Part 5 : Face Image Quality Asssessment,” 2021.

[5] Facial Identification Scientific Working Group, “Facial Recognition Systems Operation Assurance : Image Quality Assessment,” 2021.

[6] P. Grother, M. Ngan, and K. Hanaoka, “Face Recognition Vendor Test (FRVT) Part 3 : Demographic Effects,” 2019.

[7] Facial Identification Scientific Working Group, “Understanding and Testing for Face Recognition Systems Operation Assurance,” 2020.

[8] Facial Identification Scientific Working Group, “Facial Recognition Systems Methods and Techniques,” 2013.

[9] *Information technology - Biometric data interchange formats - Part 5: Face image data*. ISO/IEC 19794-5:2011, 2011.

[10] Facial Identification Scientific Working Group, “Standard guide for capturing facial images for use with facial recognition systems,” 2019.

[11] National Police Improvement Agency, “Police Standard for Still Digital Image Capture and Data Interchange of Facial / Mugshot and Scar , Mark & Tattoo Images,” no. May, 2007.

[12] P. Grother, G. Quinn, and J. Phillips, “Multiple-Biometric Evaluation (MBE) 2010: Report on the Evaluation of 2D Still-Image Face Recognition Algorithms,” *NIST Interag. Rep. 7709*, 2011.

[13] *Standard Guide for Forensic Digital Image Processing*. ASTM E2825-21, 2021.

[14] Facial Identification Scientific Working Group, “Physical Stability of Facial Features of Adults,” 2021.

[15] K. Ricanek, S. Bhardwaj, and &michael Sodomsky, “A Review of Face Recognition against Longitudinal Child Faces,” *BioSiG 2015*, pp. 15–26, 2015.

[16] P. Grother and M. Ngan, “Face Recognition Vendor Test (FRVT),” *Natl. Inst. Stand. Technol.*, 2014.

[17] P. Grother, M. Ngan, and K. Hanaoka, “NISTIR 8271 DRAFT SUPPLEMENT Face Recognition Vendor Test (FRVT) Part 2: Identification,” 2022.

[18] TELEFI Project, “Summary Report of the project ‘Towards the European Level of Exchange of Facial Images,’” 2021.

[19] R. Heyer, C. Semmler, and A. T. Hendrickson, “Humans and Algorithms for Facial Recognition: The Effects of Candidate List Length and Experience on Performance,” *J. Appl. Res. Mem. Cogn.*, vol. 7, no. 4, pp. 597–609, 2018.

[20] A. Strathie, A. Mcneill, and D. White, “In the Dock: Chimeric Image Composites

Reduce Identification Accuracy," *Appl. Cogn. Psychol.*, vol. 26, no. 1, pp. 140–148, 2012.

[21] A. Strathie and A. McNeill, "Facial Wipes don't Wash: Facial Image Comparison by Video Superimposition Reduces the Accuracy of Face Matching Decisions," *Appl. Cogn. Psychol.*, vol. 30, no. 4, pp. 504–513, 2016.

[22] R. Moreton and J. Morley, "Investigation into the use of photoanthropometry in facial image comparison," *Forensic Sci. Int.*, vol. 212, no. 1–3, pp. 231–237, 2011.

[23] R. S. S. Kramer, J. Mulgrew, and M. G. Reynolds, "Unfamiliar face matching with photographs of infants and children," *PeerJ*, vol. 2018, no. 6, pp. 1–25, 2018.

[24] D. Michalski, R. Heyer, and C. Semmler, "The performance of practitioners conducting facial comparisons on images of children across age," *PLoS One*, vol. 14, no. 11, pp. 1–17, 2019.

[25] C. Wilkinson, "Juvenile facial reconstruction," in *Craniofacial Identification*, C. Wilkinson and C. Rynn, Eds. Cambridge University Press, 2012, pp. 254–260.

[26] Facial Identification Scientific Working Group, "Standard Practice / Guide for Image Processing to Improve Automated Facial Recognition Search Performance," 2020.

[27] Facial Identification Scientific Working Group, "Minimum Guidelines for Facial Image Comparison Documentation," 2020.

[28] A. W. Young and A. M. Burton, "Are We Face Experts?," *Trends Cogn. Sci.*, vol. 22, no. 2, pp. 100–110, 2018.

[29] D. White, J. D. Dunn, A. C. Schmid, and R. I. Kemp, "Error rates in users of automatic face recognition software," *PLoS One*, vol. 10, no. 10, 2015.

[30] A. Towler, R. I. Kemp, and D. White, "Unfamiliar Face Matching Systems in Applied Settings," in *Face Processing: Systems, Disorders and Cultural Difficulties*, M. Bindemann and A. M. Megreya, Eds. Nova Science Publishers, 2017, pp. 21–40.

[31] Facial Identification Scientific Working Group, "Guide for Role-Based Training in Facial Comparison," 2020.

[32] D. White, A. Towler, and R. I. Kemp, "Understanding professional expertise in unfamiliar face matching," in *Forensic Face Matching: Research and practice*, M. Bindemann, Ed. Oxford University Press, 2021.

[33] R. Moreton, C. Havard, A. Strathie, and G. Pike, "An International Survey of Applied Face-Matching Training Courses," *Forensic Sci. Int.*, vol. 327, p. 110947, 2021.

[34] R. Moreton, G. Pike, and C. Havard, "A task- and role-based perspective on super-recognizers: Commentary on 'Super-recognizers: From the laboratory to the world and back again,'" *Br. J. Psychol.*, p. bjop.12394, Mar. 2019.

[35] A. Towler *et al.*, "Do professional facial image comparison training courses work?," *PLoS One*, vol. 14, no. 2, pp. 1–17, 2019.

[36] R. Moreton, "Forensic face matching: Procedures and application," in *Forensic Face Matching*, M. Bindemann, Ed. Oxford University Press, 2021.

[37] R. Russell, B. Duchaine, and K. Nakayama, "Super-recognizers: people with extraordinary face recognition ability.," *Psychon. Bull. Rev.*, vol. 16, no. 2, pp. 252–

257, 2009.

[38] S. Bate and J. J. Tree, “The definition and diagnosis of developmental prosopagnosia,” *Q. J. Exp. Psychol.*, vol. 70, no. 2, pp. 193–200, 2017.

[39] A. M. Burton, D. White, and A. McNeill, “The Glasgow Face Matching Test.,” *Behav. Res. Methods*, vol. 42, no. 1, pp. 286–291, 2010.

[40] M. C. Fysh and M. Bindemann, “The Kent Face Matching Test,” *Br. J. Psychol.*, pp. 1–13, 2017.

[41] D. White, D. Guilbert, V. P. L. Varela, R. Jenkins, and A. M. Burton, “GFMT2: A psychometric measure of face matching ability,” *Behav. Res. Methods*, 2021.

[42] J. D. Dunn, S. Summersby, A. Towler, J. Davis, and D. White, “UNSW Face Test: A screening tool for super-recognizers,” pp. 1–19, 2020.

[43] M. Ramon, A. K. Bobak, and D. White, “Super-recognizers: From the lab to the world and back again,” *Br. J. Psychol.*, vol. 110, no. 3, pp. 461–479, 2019.

[44] M. C. Fysh, L. Stacchi, and M. Ramon, “Differences between and within individuals, and subprocesses of face cognition: implications for theory, research and personnel selection,” *R. Soc. Open Sci.*, vol. 7, no. 9, p. 200233, 2020.

[45] R. Moreton, “Expertise in Applied Face Matching: Training, Forensic Examiners, Super Matchers and Algorithms,” The Open University, 2021.

[46] A. Towler, D. White, and R. Kemp, “Can face identification ability be trained? Evidence for two routes to expertise,” in *Forensic Face Matching*, M. Bindemann, Ed. Oxford University Press, 2021.

[47] D. White, R. I. Kemp, R. Jenkins, and A. M. Burton, “Feedback training for facial image comparison,” *Psychon. Bull. {&} Rev.*, vol. 21, no. 1, pp. 100–106, 2014.

[48] A. J. Dowsett and A. M. Burton, “Unfamiliar face matching: Pairs out-perform individuals and provide a route to training,” *Br. J. Psychol.*, vol. 106, no. 3, pp. 433–445, 2015.

[49] K. A. Martire and R. I. Kemp, “Considerations when designing human performance tests in the forensic sciences,” *Aust. J. Forensic Sci.*, pp. 1–17, Nov. 2016.

[50] *General requirements for the competence of testing and calibration laboraties*. ISO/IEC 17025:2017, 2017.

[51] European Network of Forensic Science Institutes, “Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science,” no. 001, pp. 1–31, 2014.

10 AMENDMENTS AGAINST PREVIOUS VERSION

None.