Forensic Facial Image Comparison: Examiners' Insights from an International Collaborative Exercise

Carolyn Dutot *, Stine Nordbjærg , Fredrik Stucki , Peter Cederholm

Posted Date: 16 January 2026

doi: 10.20944/preprints202601.1195.v2

Keywords: forensic facial identification; facial image comparison; morphological analysis; methodology; collaborative exercise; proficiency test; opinion scale

Article

Forensic Facial Image Comparison: Examiners' Insights from an International Collaborative Exercise

Running Head: Facial Image Comparison Collaborative Exercise

Carolyn Dutot [1,*], Stine Nordbjærg [2], Fredrik Stucki [3] and Peter Cederholm [3]

[1] Canada Border Services Agency, Forensic Facial Identification Section, Ottawa, Canada
[2] Danish National ID Centre, Biometric Team, Copenhagen, Denmark
[3] Swedish Migration Agency, Unit for Biometric Identification, Stockholm, Sweden
* Correspondence: Carolyn.Dutot@cbsa-asfc.gc.ca

Abstract: As the reliability and validity of forensic evidence, particularly in feature comparison disciplines, confront on-going scrutiny, forensic practitioners must ensure their processes, whether for investigative, intelligence or evidential purposes are robust, scientifically grounded, and validated. In forensic facial identification, morphological analysis is internationally recognized as the preferred method for facial image comparison, and is applied during the analysis and comparison steps of the Analysis, Comparison, Evaluation, Verification (ACE-V) process, commonly applied in feature comparison. While several international proficiency tests have assessed forensic facial examiners' accuracy in comparing mated and non-mated pairs (black box tests), fewer opportunities have focused on evaluating inter-laboratory procedures and methods. To address this gap, members of a small border and immigration focused expert working group participated in an inter-laboratory collaborative exercise designed to analyse and harmonize best practices across member laboratories. There are limited published validation studies of facial image comparison methods. This paper presents the results of a collaborative exercise that compares the methodologies of three different agencies, highlighting key similarities and differences in examiner process and decision making, and provides a foundation for the development of similar future initiatives.

Keywords: forensic facial identification; facial image comparison; morphological analysis; methodology; collaborative exercise; proficiency test; opinion scale

1. Introduction

Facial image comparison is a feature comparison discipline that focuses on analysing facial morphology to determine whether images represent the same individual or different individuals. It differs from automated facial recognition systems, which rely on algorithmic matching, in that human examiners evaluate facial feature observations based on recognized forensic methods. In forensic contexts, the comparison process is conducted under rigorous conditions, often following a structured workflow that includes the analysis, comparison, evaluation, and verification (ACE-V) process. Border and immigration agencies are increasingly applying these methods to operational decisions often for identity investigations, intelligence, or presentation of evidence for immigration tribunals/court, etc. Their casework frequently involves a mix of controlled and uncontrolled image sources, including identity and travel documents, intake/processing photographs, news media, law enforcement postings, and facial captures at ports of entry (e.g., counters or kiosks). However, border and immigration agencies are less mature in this space with respect to adhering to formal forensic protocols, and defending their facial image comparison opinions in court.

Quality management practices in forensic science emphasize participation in proficiency tests (PTs) and collaborative exercises (CEs) to ensure reliable and scientifically grounded processes. This

 doi:10.20944/preprints202601.1195.v2

is recognized by the International Organization for Standardization in ISO/IEC 17025 [1], and widely adopted by many traditional forensic science providers. PTs and CEs are key to standardising methods, validating techniques, and improving overall examiner expertise to ensure reliable results [2]. PTs are used for quality assurance and benchmarking the performance of forensic laboratories as an inter-laboratory comparison. PTs exist for facial image comparison (e.g. Collaborative Testing Services (CTS) [3], European Network of Forensic Science Institutes (ENFSI) [4]) but the results from these tests do not go into detail on the methods and procedures used by participants. CEs are a subset of proficiency tests designed for a particular purpose, e.g. method validation. The ENFSI also recognizes PTs and CEs as essential elements of laboratory quality management, offering guidance on conducting PTs and CEs, and promoting their use for developing best practices [5]. The CE presented in this paper was informally based on recommendations from the ENFSI.

While several international PTs exist for facial examiners—often evaluating accuracy in comparing mated and non-mated image pairs—there are few inter-laboratory CEs, particularly involving border and immigration-focused personnel. To address this gap, a small expert working group conducted a CE to examine and harmonize laboratory procedures, assess interpretation practices, and evaluate inter-examiner consistency in facial image comparison. While the agencies involved in this exercise are not accredited for their facial image comparison methods, they are committed to producing reliable forensic results, and continuously improving their processes. Recent research by Obertová et al. [6] highlights how accreditation can strengthen consistency and defensibility in this discipline, further underscoring the value of CEs such as the one presented here.

The aim of this study is to report the outcomes of this collaborative exercise, including the impact of examiner interpretation, image quality, and methodological variations on assessment results, and to provide insights for developing best practices and future inter-laboratory exercises.

2. Materials and methods

The expert working group members that facilitated the exercise meet regularly to discuss procedures and share information to improve processes, develop best practices, and work toward inter-lab consistency. Most meetings are held virtually, but to facilitate deeper discussions and information sharing the group also aims to hold in-person meetings every 12 to 18 months. For the in-person meeting held in 2023, the group explored using a CE to test their procedures and to stimulate dialogue regarding interpretation and evaluation of observations, method validation, and refinement of laboratory procedures based on the results of the exercise.

2.1. Study Design

The CE was conducted in 2023 with participation involving examiners from three different member countries performing facial image comparisons. The exercise was experimental, with no pre-test of case material prior to distribution. The study design focused on replicating realistic operational conditions encountered by border and immigration agencies. Although the sample size is small, the results provide a detailed qualitative analysis of facial image comparison methods, and was a first step towards validating the observations and evaluations among facial image examiners. Member participation was voluntary. The test case was prepared by an external resource. The test case images and case scenario were e-mailed to participating members by the CE coordinator within the group. Participating members were provided the test seven working days prior to the in-person meeting. The test case material was sent to three agencies representing three different countries (herein referred to as Agency 1, 2 and 3), all of which prepared results, and supporting material for post-examination group discussion.

The intent of the CE was to elicit an in-depth analysis and comparison of laboratory procedures and methods. For the post-examination discussion each participating Agency provided a presentation and illustration on their case approach, including case acceptance/strategy, methodology/ procedures (ACE-V), decision models, results, opinion scale, and outputs (i.e. reporting).

 doi:10.20944/preprints202601.1195.v2

2.2. Test case

The case scenario was modelled on case material that would be considered representative of typical operational casework for all participating members. The overall level of difficulty was designed to be moderate. As part of the test preparation, images were reduced through cropping to the subject's head/neck region. Image quality was intentionally reduced to reflect typical operational conditions, while high-resolution originals were maintained for the test case discussion. This adjustment simulated real-world scenarios, in which facial marks and fine details may not be fully visible. Examiners were instructed to assess the images as they would in standard practice.

The test case consisted of a total of five (5) images, including a reference image, and two sets of two questioned images. Members were instructed to intake the case and examine as in routine casework, following internal procedures. It is common practice to receive multiple images per subject. Grouped images for any given subject may be analysed and compared together, allowing examiners to consider all available visual information collectively.

2.3. Test-case Images

The images used in this CE were sourced from a freely licensed stock image repository. No identifiable information about any individual was available, and no images are reproduced in this publication to ensure compliance with license and usage rules. Images were used exclusively for training, assessment, and research purposes consistent with standard forensic practice.

Reference Image (Identity A.jpg[1])

The reference image was selected as an image suitable for a forensic comparison with some limitations due to overall image resolution, head angle, and lighting.

Questioned Images (Identity B1.jpg[2], Identity B2.jpg[3], Identity C1.jpg[4], Identity C2.jpg[5])

The questioned images were selected as images suitable for a forensic comparison. The limitations varied across images, and were due to overall image resolution, head and camera angle/position, lighting, obstruction of features, and expression.

As part of the post-exercise discussion, the original higher quality images were introduced and displayed to all of the examiner participants. An in-depth conversation ensued on the interpretation and weight of the observed facial features in the test case and whether the new information (facial feature detail) could have impacted or changed the original opinion provided.

Table 1. Test Case Image Properties.

	Original higher quality (jpeg)		**Reduced quality (jpeg)**	
	Resolution (pixels)	Size (KB)	Resolution (pixels)	Size (KB)

[1] https://images.pexels.com/photos/2412408/pexels-photo-2412408.jpeg?auto=compress&cs=tinysrgb&w=1260&h=750&dpr=2

[2] https://images.pexels.com/photos/3705262/pexels-photo-3705262.jpeg?auto=compress&cs=tinysrgb&w=1260&h=750&dpr=2

[3] https://images.pexels.com/photos/3705263/pexels-photo-3705263.jpeg?auto=compress&cs=tinysrgb&w=1260&h=750&dpr=2

[4] https://images.pexels.com/photos/2481371/pexels-photo-2481371.jpeg?auto=compress&cs=tinysrgb&w=1260&h=750&dpr=2

[5] https://images.pexels.com/photos/2481372/pexels-photo-2481372.jpeg?auto=compress&cs=tinysrgb&w=1260&h=750&dpr=2

 doi:10.20944/preprints202601.1195.v2

Reference Image				
Identity A	640x800	188	200x250	32.4
Questioned Images				
Identity B (Image 1)	950x1188	557	200x250	27.5
Identity B (Image 2)	801x1001	522	200x250	26.9
Identity C (Image 1)	1184x1481	468	300x375	31.6
Identity C (Image 2)	1154x1443	448	180x225	25.8

2.4. Participation

The three participating agencies completed the exercise as anticipated. There were no specific instructions requiring the exercise to be introduced as a blind test. The only specification provided was that each agency should follow its standard operating procedures for forensic facial examinations. Some variability in case assignment occurred due to differences in local procedures and available resources. Specifically, two agencies treated the exercise as known test case, while one agency conducted it as a blind test.

Agency 1: The test case was reviewed by a case manager (who also served as a participating examiner (A1E1)), and the test material was prepared following standard internal procedures. The material was then provided to one additional examiner (A1E2). Independent examinations were conducted, with both individuals aware that the submission was a test case.

Agency 2: The test case was reviewed by a case manager, and the test material was prepared following standard procedures. The material was provided to two independent examiners (A2E1 and A2E2). Unlike Agency 1 and 3, the examiners were not aware that the examination was part of a test case; the case was entered into the system as a blind test.

Agency 3: The test case was reviewed by a case manager (who also served as a participating examiner (A3E1)), and the test material was prepared following standard internal procedures. The material was provided to one additional examiner (A3E2). As with Agency 1, independent examinations were conducted, and both individuals were aware that the submission was a test case.

The examiners who participated in the exercise had a range of experience conducting forensic facial image comparisons, spanning approximately three to seventeen years.

2.5. Case processing

All agencies assigned a case manager role to handle the case intake. Case reception was very similar across agencies (i.e. case number assignment and tracking information, file structure for saving original and working copies, etc.). Case managers also conducted a preliminary suitability assessment of the images, to ensure the imagery met their acceptance criteria to pass on for further examination. The case manager role in each of the agencies is also a trained facial examiner, with at least two years of experience. Following the assessment and the acceptance of the case, the case manager set a case strategy for comparison of the images to address the questions asked by the submitter. Part of the process included the removal of contextual information to mitigate bias, and establishing the sequence for image analysis to support consistency across examinations. In Agency 2, a case manager was also assigned to conduct an administrative review of the results from the two independent examiners, in addition to setting the case strategy.

2.6. Comparison of Images

 doi:10.20944/preprints202601.1195.v2

All participating agencies apply the ACE-V process in their regular forensic facial examination casework. This process "gives the expert specific phases of examination that can be used to document the perception, information gathering, comparison, and decision-making that takes place during an examination" [7]. During the phases of examination, morphological analysis was the primary approach used, in which the features and components of the face were analysed and compared, following the standard guide for Facial Image Comparison Feature List for Morphological Analysis (ASTM E3149-18) [8]. The feature list includes the main facial features (e.g., eyes, ears, nose, mouth). It also describes details of the characteristics of the features (e.g., lobe of the ear, nostrils of the nose). It further lists the characteristic descriptors that describe what should be evaluated when comparing the facial features and the characteristics of the features (e.g., symmetry of the lips, shape of the jawline).

The test case was handled following standard internal examination procedures, including notetaking and documentation processes. The common strategy included two independent examiners completing two facial image comparisons, (i.e. Comparison 1: Questioned images (Identity B) to Reference image (Identity A) and Comparison 2: Questioned images (Identity C) to Reference image (Identity A)).

3. Results

3.1. Opinion Result versus Ground Truth

Table 2. Summary of Opinion Results by Examiner.

		Comparison 1 Opinion	Comparison 2 Opinion	Accuracy
Agency A	Examiner 1	*+2*	*-2*	*100*
	Examiner 2	*+2*	*-2*	*100*
Agency B	Examiner 1	*+2*	*-2*	*100*
	Examiner 2	*+2*	*-2*	*100*
Agency C	Examiner 1	*+2*	*-2*	*100*
	Examiner 2	*+2*	*-2*	*100*

All participating Agencies reached accurate results for both comparisons conducted.

3.2. Strength of Opinion

Currently there is no standardized opinion scale for facial identification. There are frameworks of opinion categories, (i.e. OSAC 2022-S-0001 Standard Guide for Image Comparison Opinions) [9] which support the development of opinion scales. No specific scale was provided to participants as part of the instructions for the test case, specifically to allow for discussion on this topic. Participants were asked to use their internal agency scale; all three participating agencies used a 7-point opinion scale. All scales were balanced, comprising an inconclusive result (0) and three levels of support for the same person, and three levels of support for a different person. Each scale aligns verbal descriptors (e.g. Strong support or Moderate Support) with numerical values ranging from +3 to -3.

The comparison of Questioned Identity B to Reference Identity A had **an opinion/result of +2** on the scale **for all participating agencies.**

The comparison of Questioned Identity C to Reference Identity A had **an opinion/result of -2** on the scale **for all participating agencies.**

 doi:10.20944/preprints202601.1195.v2

A summary chart is below illustrating individual agency opinion scales, translated into English. This allows for a direct comparison of how different agencies interpret and express confidence levels in their assessments, recognizing there may be some nuances when translated from the native language.

Table 3. Agency Opinion Scales.

Opinion Scale	Agency 1	Agency 2	Agency 3
+3	**Strong Support** for Same Source. The observed similar characteristics far outweigh the observed dissimilar characteristics.	The analysis conducted **strongly supports** that the submitted images depict the same person.	**Very strong support** for the same source. Other possibilities are considered to be practically ruled out.
+2	**Moderate Support** for Same Source. The observed similar characteristics outweigh the observed dissimilar characteristics.	The analysis conducted **moderately supports** that the submitted images depict the same person.	**Strong support** for same source. Other possibilities are considered to be very small.
+1	**Weak Support** for Same Source. The observed similar characteristics slightly outweigh the observed dissimilar characteristics.	The analysis conducted **supports** to a certain extent that the submitted images depict the same person.	**Support** for same source. Other possibilities are considered to be small.
0	**Inconclusive.** The findings do not differentiate the same source/different source propositions.	The analysis conducted **supports no conclusion** regarding whether the submitted images depict the same person or not.	**Inconclusive**. Sufficient support for a conclusion for or against same source cannot be found.
-1	**Weak Support** for Common Source. The observed dissimilar characteristics slightly outweigh the observed similar characteristics.	The analysis conducted **supports** to a certain extent that the submitted images depict different persons.	**Support** for different source. Other possibilities are considered to be small.
-2	**Moderate Support** for Common Source. The observed dissimilar	The analysis conducted **moderately supports** that the submitted images	**Strong support** for different source. Other possibilities are

 doi:10.20944/preprints202601.1195.v2

	characteristics outweigh the observed similar characteristics.	depict different persons.	considered to be very small.
-3	**Strong Support** for Different Source. The observed dissimilar characteristics far outweigh the observed similar characteristics.	The analysis conducted **strongly supports** that the submitted images depict different persons.	**Very strong support** for different source. Other possibilities are considered to be practically ruled out.

3.3. Qualitative Results

3.3.1. Case handling

The standard procedures and methodology were very similar across the groups, however there was some variation between agencies, most notably:

Agency 1 was not resourced to have a separate case manager handle the contextual information, and therefore one examiner completed the examination with knowledge of the full case information. As this was a known test case, the additional information may not have introduced particular biases, however it was noted that this does not reflect accepted forensic best practice. For Agency 3, the case manager was also an examiner for the test case, although not their standard procedure for regular case work.

Further, Agency 2 ran the test case through their unit as a blind test, and would have had the most similar scenario to actual casework. Despite having to maintain the "confidentiality" of the test from the examiners, and manage the scheduling so they could participate in the final post-test discussion, the exercise was successfully completed as a double-blind examination.

For comparison purposes, images of the same person are generally grouped based on known identities, allowing examiners to assess similarities and dissimilarities across images. The case manager is responsible to set out the comparison strategy, including the "grouping" of images. In this exercise, the strategy for comparison had some differences, specifically, Agency 1 conducted a cursory review of the associated images (Identity B images and Identity C images) before introducing them as "grouped images" representing each purported identity for the comparison. Agency 2 admitted the images associated to each questioned identity automatically by "grouping" as informed of their association by the submitter. Agency 3 conducted additional formal facial image comparisons to assess and confirm the association of images submitted for each questioned identity before accepting and comparing to the reference image.

During the examination phase, Agency 1 conducted two examinations separately and sequentially, and therefore the reference image was analysed in depth during the first examination, and then was re-introduced into the second examination. The other agencies analysed each of the five images submitted prior to conducting any comparison, which has been noted as best practice.

There was no finding that these variations had an impact on the outcome of the case. Further there was no difference between the agencies that knew it was a test versus the agency that ran it as a blind test, as all agencies reached the same level of final opinion. There is no evidence that Agency 1 or 3 were overly confident in their decisions, or Agency 2 more cautious in their opinion.

3.3.2. Interpretation and evaluation of observations

The collaborative exercise had the aim of assessing one form of forensic reliability, specifically inter-lab reproducibility (i.e. do different examiners working in different labs reach the same opinion when analysing the same materials?). Despite some variation in procedure, all agencies landed on

 doi:10.20944/preprints202601.1195.v2

the same numerical value on the scale for both comparisons, reflecting similar strength of opinion (not withstanding differences in verbal opinion scales). This speaks to the alignment of the participating agencies in their interpretation of the evidence. To further assess the consistency at the image analysis and feature comparison level, a deeper dive into examiner notes looking at individual observations and notes was performed.

Summary of Examiner Observations

Image Quality: Examiners largely documented that the images were suitable for comparison, while noting factors like lighting and pose as having potential impact. Some examiners also noted additional factors, including camera angle, age, or expression. The imaging conditions mentioned during the analysis and evaluation of the material suggests a general shared interpretation of imaging limitations.

Facial Features: For Comparison 1 the majority of facial features were marked as similar, and for Comparison 2 the features were mostly marked as dissimilar, indicating consistency among participants' observations of individual features.

A depiction of individual examiner observations made during the examinations for each facial feature from the standard list is presented below (Figure 1, Figure 2).[6] To simplify the illustration, terminology was standardised based on common understanding and interpretation of terms, (and taking into consideration translation to English), specifically:

1) Inconclusive was used to suggest uncertainty. Other terms originally recorded were unreliable, insufficiently resolved, competing observations.

2) Not visible or not able to compare was used for features that were either occluded or not visible because of head angle, etc., but would normally be expected on a face (i.e. ears, forehead, neck).

3) Not observed was used for features that were not present on either image, and may not always be on a face (i.e. scars, alterations, facial hair).

Trends between Examiners and Agencies

Comparison 1

Comparison Identity B vs Identity A

Facial Features	Agency 1: Examiner 1	Agency 1: Examiner 2	Agency 2: Examiner 1	Agency 2: Examiner 2	Agency 3: Examiner 1	Agency 3: Examiner 2
Skin	Similar	Similar	Similar	Similar	Similar	Similar
Face/Head Outline	Similar	Similar	Similar	Similar	Similar	Similar
Nose	Similar	Similar	Similar	Similar	Similar	Similar
Ears	Similar	Similar	Similar	Similar	Similar	Similar
Mouth	Similar	Similar	Similar	Similar	Similar	Similar
Chin	Similar	Similar	Similar	Similar	Similar	Similar
Jawline	Similar	Similar	Similar	Similar	Similar	Similar
Eyes	Similar	Dissimilar	Similar	Similar	Similar	Similar
Facial Lines	Inconclusive	Similar	Similar	Similar	Similar	Similar
Eyebrows	Similar	Similar	Similar	Inconclusive	Similar	Similar
Neck	Dissimilar	Dissimilar	Dissimilar	Similar	Similar	Similar
Forehead	Not visible	Not visible	Similar	Similar	Not visible	Not visible
Hair/Baldness pattern	Dissimilar	Dissimilar	Inconclusive	Inconclusive	Dissimilar	Dissimilar
Facial Marks	Inconclusive	Dissimilar	Inconclusive	Inconclusive	Dissimilar	Dissimilar
Facial Hair	Not observed	Not observed	Inconclusive	Inconclusive	Not observed	Not observed
Scars	Not observed	Not observed	Not observed	Not observed	Not observed	Not observed
Alterations	Not observed	Not observed	Not observed	Not observed	Not observed	Not observed

[6] The order of the facial feature list has been modified to better illustrate the observations.

 doi:10.20944/preprints202601.1195.v2

Figure 1. Examiner Feature Observations (Comparison 1).

Agreement on Similarities

For Comparison 1, there was broad consensus among all agencies regarding similarity for most facial features including skin, head outline, eyebrows, nose, ear, mouth, chin, and jawline. These features consistently showed similarity across all examiners, regardless of agency. All examiners also identified scars and alterations as not observed.

The ground truth response for Comparison 1 was that the images represented the same person, and it is expected that examiners would observe numerous similarities in features. The breakdown below prioritises observations where examiners showed more disagreement.

Variability in Feature Observation/Evaluation

- Forehead: There were differences in the observations of the forehead where Agency 2 provided an opinion of similar where Agency 1 and 3 determined the feature was not visible. In the examiner notes it showed that both A2E1 and A2E2 felt that there was enough visibility of the brow ridge to compare the feature.
- Eyes: A1E2 concluded that the eyes were dissimilar whereas all other examiners determined this feature to be similar. Review of the notes found that A1E2 observed both dissimilarities and similarities in the visible feature and sub-feature detail of the eyes, however the overall observation of eyes was determined to be dissimilar. During the evaluation phase, the feature was given little to no weight because of the imaging factors and expression.
- Neck: There was variation across examiners/agencies, specifically three examiners (A1E1, A1E2, A2E1) noted the feature as dissimilar, noting that the neck was broader/wider in the questioned images, but attributed it to age and assigned little weight to the feature. Three examiners (A2E2, A3E1, A3E2) noted the feature as similar, specifically that the overall neck and Adam's apple/muscular shape appeared similar, however little weight was attributed in the evaluation.
- Facial Hair: The observations were broadly the same, with a variation in the feature being identified as not observed versus inconclusive.
- Facial marks: Three examiners (A1E1, A2E1, A2E2) noted that the comparison of the facial marks was inconclusive, while three examiners (A1E2, A3E1, A3E2) found that the facial marks were dissimilar.

Comparison 2

 doi:10.20944/preprints202601.1195.v2

Comparison Identity C vs Identity A

Facial Features	Agency 1: Examiner 1	Agency 1: Examiner 2	Agency 2: Examiner 1	Agency 2: Examiner 2	Agency 3: Examiner 1	Agency 3: Examiner 2
Nose	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Ears	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Mouth	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Chin	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Eyebrows	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Jawline	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar	Dissimilar
Face/Head Outline	Dissimilar	Dissimilar	Dissimilar	Similar	Dissimilar	Dissimilar
Neck	Dissimilar	Dissimilar	Inconclusive	Dissimilar	Dissimilar	Dissimilar
Eyes	Dissimilar	Dissimilar	Dissimilar	Inconclusive	Dissimilar	Dissimilar
Skin	Similar	Similar	Similar	Similar	Similar	Inconclusive
Facial Lines	Dissimilar	Dissimilar	Dissimilar	Inconclusive	Dissimilar	Dissimilar
Facial Marks	Inconclusive	Dissimilar	Inconclusive	Inconclusive	Dissimilar	Dissimilar
Hair/Baldness pattern	Inconclusive	Similar	Not visible	Not visible	Dissimilar	Dissimilar
Forehead	Not visible	Not visible	Inconclusive	Similar	Not visible	Not visible
Facial Hair	Not observed	Not observed	Inconclusive	Inconclusive	Not observed	Not observed
Scars	Not observed	Not observed	Not observed	Not observed	Not observed	Not observed
Alterations	Not observed	Not observed	Not observed	Not observed	Not observed	Not observed

Figure 2. Examiner Feature Observations (Comparison 2).

Agreement on Dissimilarities:

For Comparison 2, there was broad consensus among all agencies regarding dissimilarity for many facial features including eyebrows, nose, ears, mouth, chin, and jawline. These features consistently showed dissimilarity across all examiners, regardless of Agency.

The ground truth response for Comparison 2 was that the images represented different persons, and it is expected that examiners would observe numerous dissimilarities in features. The breakdown below prioritises observations where examiners showed more disagreement.

Variability in Feature Observation/Evaluation

- Face/Head Outline: One examiner found the overall observation to be similar, however no weight was given to this observation in their evaluation.
- Hair/Baldness pattern: The observation for this feature varied significantly across examiners. Agency 2 noted that the hairline was not visible as a result of the hair length and style. A1E1 and A3E2 found the texture and shape of the hair to be different. A1E1found the hair length and the angle of the right-side hairline to be dissimilar. A1E2 had competing observations but overall labelled the feature as similar with high uncertainty due to pose and potential grooming. In general, this feature was given little weight in the overall evaluation by those that compared the feature.
- Forehead: Agency 1 and 3 determined the feature was not visible (occluded by hair), whereas Agency 2 examiners provided an opinion on the feature. A2E1had an inconclusive comparison result, and A2E2 found the general shape of the brow ridge area to be similar.
- Facial Hair: The observations were broadly the same, with a variation in the feature being identified as not observed versus inconclusive.
- Facial Marks: The observations for this feature were the same as in Comparison 1, three examiners (A1E1, A2E1, A2E2) noted that the comparison of the facial marks was inconclusive, while three examiners (A1E2, A3E1, A3E2) found that the facial marks were dissimilar.

 doi:10.20944/preprints202601.1195.v2

De-brief with Higher Quality Images

Although the higher quality images provided some improved level of detail, limitations to the imagery were still present, and all participants were satisfied that their opinion would have remained the same. There was common agreement that only well resolved discriminating features such as facial marks would have changed the strength of opinion to the highest level of support.

4. Discussion

Evaluating the strength of examination findings in facial image comparison is a complex process influenced by multiple interacting factors [2]. During the evaluation phase, examiners assess the clarity, quantity, specificity, reproducibility, persistence, and extent of similarities, dissimilarities, and expected variations in their observations [10]. When dissimilarities are identified, examiners consider whether these may result from image parameters identified during the analysis phase, such as differences in expression, pose, perspective, lighting, age, or weight. Additionally, they evaluate the likelihood that observed similarities are coincidental, as certain features—such as general shapes—can be shared by multiple individuals without indicating a common source.

All observations—whether indicating similarity or dissimilarity—are valuable for evaluative purposes because they contribute to the overall assessment. However, the weight assigned to a particular feature varies depending on the likelihood of the observation under specific propositions (e.g., same person vs. different persons), thresholds for marking features as inconclusive or similar/dissimilar, individual interpretation, and examiner experience. Since facial image comparison involves a degree of subjectivity, opinions are shaped by the examiner's training, knowledge, and experience. Consequently, the confidence and certainty associated with evaluations may vary to some extent.

During the analysis of examiner observations in this exercise, variations were noted that could be attributed to factors such as examiner experience, uncertainty due to imaging parameters, and differences in individual approaches. For example, when competing observations arose regarding components of a facial feature (i.e., eyes), examiner experience could have influenced which details were assigned more weight. One examiner (A1E2) identified both similarities and dissimilarities in eye feature details but ultimately determined the feature to be dissimilar, whereas all other examiners determine the feature to be similar. The examiner that was an outlier assigned little weight however to the observation due to having a high level of uncertainty.

Some variation among examiners was linked to internal thresholds for marking features as inconclusive versus similar/dissimilar. For instance, observations of facial marks varied due to individual interpretations of low-quality imagery. Examiners who marked features as inconclusive often reflected a more cautious approach to lower-quality images. Those who made determinations of similarity or dissimilarity noted high uncertainty and assigned negligible or no weight to these observations during evaluation.

Low resolution and poor image quality contributed to discrepancies when assessing subtle features. Fine details were often not well-resolved or discernible, leading to variability in observations. In contrast, larger and more distinct features were consistently marked as similar or dissimilar across agencies. This suggests that lower resolution has a less pronounced impact on broader structural features than on finer details.

Factors such as lighting differences, head pose variations, and camera angles further limited certainty in some observations. Despite these challenges and occasional divergence in observations between examiners, there was consensus on the limitations posed by the imagery quality; thus, negligible weight was given to uncertain features during evaluation. This finding underscores that while examiners may differ in their assessment of fine details due to resolution limitations or other factors, these differences do not significantly affect their final opinions, given the holistic evaluation of all features. Future research could explore this further by dividing participants into groups with access to higher versus lower-resolution images to determine whether results remain consistent.

 doi:10.20944/preprints202601.1195.v2

Recent studies, such as by Bacci et al. [11] have found clear relationships between image quality factors and forensic facial comparison outcomes, They suggest a triage approach using a standardized image quality score, where images below a threshold are likely to yield inconclusive or unreliable results, enabling examiners to prioritize cases with sufficient quality. Based on the results of the CE, the group supports further research aimed at establishing image quality thresholds or shared practices that could help reduce the risk of inconclusive or unreliable outcomes.

Research by Towler et al. [12] suggests that examiners rank the usefulness of certain features differently than non-experts (e.g., students). Features such as face shape and forehead—both of which showed variability in this exercise—are generally considered less useful unless they exhibit notable anomalies. As such, these features are typically given less weight during evaluations unless they provide significant distinguishing information.

Variations in how agencies handled certain observations also contributed to differences in results. For example, discrepancies in marking features like facial hair as "not observed" versus "inconclusive" were attributed to internal procedural differences when features were visible in one image but not another. A review of examiner notes also revealed that Agency 2 excluded hair colour and texture from evaluations due to their determination that these traits are largely unreliable.

Despite variations in feature comparison observations across examiners and agencies, there was strong alignment in their evaluation processes based on consensus regarding final opinions and strength of opinion on the scale. This suggests that while individual observations differed due to factors such as image quality or examiner experience, there was general consistency in how observed similarities and dissimilarities were interpreted and weighted during evaluation. The shared framework for assessing findings highlights a robust evaluative process that accommodates subjective elements while maintaining consistency across agencies.

As Obertová et al. [6] revealed, accredited facial image comparison units exhibited better awareness of standard operating procedures and more consistent use of validated methods compared to non-accredited units. This underscores the importance of harmonizing practices and validating methods to enhance consistency and support defensible outcomes.

4.1. Considerations

Consensus Strength: The clear consensus in the feature observations logically aligns with reaching the same level of support in the final opinion. However, despite less agreement on individual feature comparisons for Comparison 2, the same end result was reached. A thorough review of the distribution of steps and thresholds within the opinion scale is warranted to determine if more granularity would be useful. Additional calibration exercises could help align interpretation of the scale across examiners, but also for end users of the opinion.

4.2. Possible sources of error and suggestions for improvement of performance

4.2.1. Case Intake

A procedural issue that was discussed was related to the general acceptance of case images as provided by the submitter, and how agencies handle situations where there may be a doubt or a question related to the provenance or association of individuals within different images (e.g. grouping of questioned images). There was good discussion about the role of the case manager to communicate with the submitter to address questions/concerns about the submission, versus accepting and processing the images as provided. There was general consensus that there is value in discussing the evidence with the submitter to ensure that the request is clear, and/or to identify potential limitations, or sources of error, or to better determine the benefit in completing the examination.

Although this best practice has been addressed to some extent in other forensic disciplines, such as Forensic Document Examination, there is currently no specific standard practice for setting a case strategy and managing image groupings in facial identification. In Forensic Handwriting

 doi:10.20944/preprints202601.1195.v2

Examination and Human Factors: Improving the Practice Through a Systems Approach [13], a comparable scenario is outlined for addressing questions about handwriting submissions. In such cases, it is suggested that the examiner discuss the issue with the submitter. If the submitter cannot provide clarification, the examiner may not be able to proceed with the images as submitted, and should document the reasons. If the examination continues without clarification, the examiner may need to subdivide the images into groups based on features potentially belonging to different sources. It is recommended that the grouping and the rationale for continuing the examination be thoroughly documented. Discussions prompted by this exercise indicate agreement that adopting a similar standard practice could add value to the outputs of facial identification. This approach should be explored in greater detail by the broader community of facial practitioners.

Equally important there was consensus for the need to clearly state that the opinion or result is highly dependent on the information provided. Highlighting this dependency ensures stakeholders understand the limitations of the opinion and clarifies that any changes to the information could lead to a different result. An example caveat that is provided by the Forensic Science Regulator Codes of Practice and Conduct Development of Evaluative Opinions FSR-C-118. Issue 1 [14]:

> "My approach to the examination strategy and interpretation of the observations in this case is crucially dependent on the information made available to me. If any of this information is incorrect or if further information is made known to me, it will be necessary for me to reconsider my interpretation."

4.2.2. Case Strategy

The approach or case strategy for analysis and comparison of the images is relevant to mitigate potential biases. For Agency 1, the analysis of the second set of questioned images (QC1 and QC2) after having used the reference image (RA) in the first comparison could have introduced a potential source of error. More specifically this approach could introduce cognitive bias where information from the reference material could bias the subsequent analysis of the questioned image(s). An improved practice will be to properly sequence the information [15], specifically to conduct an analysis of all five images independently to assess the image quality, visibility of features, feature detail, etc., prior to conducting the two comparisons, so as not to introduce influences or familiarity with an image unnecessarily. In normal casework, images related to a case may be submitted at successive time frames. Resource permitting, best practice would be to assign different examiners to examine the new material.

5. Conclusions

Overall, the exercise was considered to be a success by all groups given that participating examiners reached similar and accurate opinions, which provides tangible support for the current procedures in place. Further, the groups were able to present and discuss their local procedures in detail, which both confirmed existing inter-agency alignment, and provided an opportunity to explore ways to improve processes. It is recognized that the members of the expert working group meet regularly to share information to improve processes and develop best practices within their respective organisations, which may be an underlying factor in the success of the exercise.

Final discussions amongst examiners took place on lessons learned, if the aims of the test were met, and learning points for the future design of similar tests or exercises. The ability to have examiners meet to have a fruitful discussion and in-depth analysis of the observations was recognized as critical for the general knowledge gained, and development of examiners. All participating members felt that the exercise was a very good tool to go beyond the accuracy of the decision (i.e. black box), and delve deeper into the overall lab process. In addition, examiners felt that the opportunity to review together the higher quality images from the test case was a key element of the exercise.

 doi:10.20944/preprints202601.1195.v2

It was agreed by case managers that the turnaround time expected to intake the case and complete the examination was too short. It is recommended that coordinators plan ahead to give participating agencies the opportunity to better organize the test case into their regular practice. For those agencies that decide to run the test case through as a blind test, it is easier to manage and introduce as regular priority casework (i.e. not urgent) with fewer suspicions raised by participating practitioners.

Forensic science processes, should exhibit valid, reliable, and robust processes for the variety of intelligence, investigative and evidentiary applications that seek forensic opinions. The actual scientific method used for examination of evidence should be tested in its entirety, with practitioners trained in the appropriate use of the method. Forensic opinions should be provided such that there is objective evidence that a method and process is fit for purpose, and that the information and results obtained can be relied upon. An aligned approach to interpretation and evaluation of observations is a critical part of establishing consistency and reliability of forensic opinion, and calibration between examiners. As the discipline of facial identification evolves and expands, it will be critical to reduce potential variability in approaches, and demonstrate that the recommended method produces accurate and explainable results. The dialogue that the exercise incited was considered to be very useful for all of the participants, and provided an avenue for learning and development. It allowed peers to explain the way they evaluate and determine the relevance of the observations. This type of review and feedback amongst examiners is invaluable to confirm or refute if our assumptions and interpretations as experts are aligned, and thereby facilitate greater consistency, a cornerstone of forensic evidence.

In addition, the discussion brought about a conversation between participating agencies on overall case handling, including the evaluation and management of contextual information, and supporting/supplemental case images. The results of this discussion effected actual changes to standard operating procedures within the agencies to align with the accepted group approach. These improvements strengthen comparison results, exceed current best practice criteria, and may promote further transparency and alignment across the discipline.

In conclusion, this study contributes to ongoing discussions in the field regarding the need for evidence-based protocols. It emphasizes the necessity for standardized procedures, method validation, and inter-laboratory collaboration, particularly in agencies maturing their forensic services in this space.

Acknowledgments: This work has been funded with the support of the individual agencies governing the participation in the expert working group. The content of this publication represents the views of the authors only and is their sole responsibility. The associated agencies bear no responsibility for any use, interpretation, or application of the information contained herein. The authors wish to thank all participants, and supporting agencies in the collaborative exercise discussed in this paper. We acknowledge RELI Ltd for contributing to the preparation of the test material. Individual agencies supported the activities around this exercise and there was no external funding required.

Authors' contributions: Carolyn Dutot: Conceptualization, Analysis, Writing - original draft, reviewing and editing. Stine Nordbjærg: Conceptualization, Analysis, Writing – reviewing and editing. Fredrik Stucki: Analysis, and Writing - review and editing. Peter Cederholm: Analysis, and Writing - review and editing. All authors have read and agreed to the published version of the manuscript.

Declaration of interest statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

 doi:10.20944/preprints202601.1195.v2

References

1. ISO. ISO/IEC 17025 Testing and calibration laboratories. International Organization for Standardization; 2005. Accessed on 2025 Feb 26. Available from: https://www.iso.org/ISO-IEC-17025-testing-and-calibration-laboratories.html
2. ENFSI. Best Practice Manual for Facial Image Comparison. ENFSI, European Network of Forensic Science Institutes; 2018. Available from: https://enfsi.eu/about-enfsi/structure/working-groups/documents-page/documents/best-practice-manuals/
3. Collaborative Testing Services Inc. Facial Identification Comparison report. Accessed on 2025 Feb 26. Available from: https://cts-forensics.com/program-9.php
4. ENFSI. ENFSI, European Network of Forensic Science Institutes Digital Imaging Working Group. Overview Available from: https://enfsi.eu/about-enfsi/structure/working-groups/digital-imaging/
5. ENFSI. Framework for the Conduct of Proficiency Tests and Collaborative Exercises within ENFSI. European Network of Forensic Science Institutes; 2023. Available from: https://enfsi.eu/wp-content/uploads/2023/08/CQQ-FWK-004-PT-CE.pdf
6. Obertová Z, Siebke I, Schüler G. Challenges of accreditation in forensic fields concerned with human identification: a survey of European facial examiners. Forensic Sci Res. 2024;9:owae047. https://academic.oup.com/fsr/article/9/3/owae047/7731316
7. Vanderkolk JR. Examination Process. In: Justice NI, editor. Fingerprint Sourcebook. National Institute of Justice; 2011. p. Chapter 9. Available from: https://nij.ojp.gov/library/publications/fingerprint-sourcebook
8. ASTM. Standard Guide for Facial Image Comparison Feature List for Morphological Analysis. ASTM International; 2022. Available from: https://www.astm.org/e3149-18.html
9. OSAC. 2022-S-0001 Standard Guide for Image Comparison Opinions. The Organization of Scientific Area Committees for Forensic Science, National Institute of Standards and Technology; 2022. Available from: https://www.nist.gov/document/osac-2022-s-0001-standard-guide-image-comparison-opinions-3
10. OSAC. Standard Guide for Developing Discipline Specific Methodology for ACE-V. The Organization of Scientific Area Committees for Forensic Science, National Institute of Standards and Technology; 2020. Available from: https://www.nist.gov/document/standard-guide-developing-discipline-specific-methodology-ace-v
11. Bacci, N., Briers, N. & Steyn, M. Prioritising quality: investigating the influence of image quality on forensic facial comparison. Int J Legal Med 138, 1713–1726 (2024). https://doi.org/10.1007/s00414-024-03190-7
12. Towler A, White D, Kemp RI. Evaluating the Feature Comparison Strategy for Forensic Face Identification. 23(1):47-58; 2017. doi:10.1037/xap0000108
13. Taylor M, Bird C, Bishop B, Burkes T, Caligiuri MP, Found B, et al. Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach. U.S. Department of Commerce, National Institute of Standards and Technology; 2021. doi: https://doi.org/10.6028/NIST.IR.8282r1
14. Forensic Science Regulator. Forensic Science Regulator Codes of Practice and Conduct Development of Evaluative Opinions. 2021. Available from: www.gov.uk/government/organisations/forensic-science-regulator
15. Quigley-McBride A, Dror IE, Roy T, Garrett BL, Kukucka J. A practical tool for information management in forensic decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in casework. Forensic Sci Int Synergy. 2022;4:100216. https://doi.org/10.1016/j.fsisyn.2022.100216.