Mind the Gap: A Practical Framework for Classifiers in a Forensic Context

Chris Zeinstra, Didier Meuwly, Raymond Veldhuis, Luuk Spreeuwers
University of Twente, Faculty of EEMCS, DMB Group
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{c.g.zeinstra,d.meuwly,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

Abstract

In this paper, we present a practical framework that addresses six, mostly forensic, aspects that can be considered during the design and evaluation of biometric classifiers for the purpose of forensic evidence evaluation. Forensic evidence evaluation is a central activity in forensic case work; it includes the assessment of strength of evidence of trace and reference specimens and its outcome may be used in a court of law. The addressed aspects consider the modality and features, the biometric score and its forensic use, and choice and evaluation of several performance characteristics and metrics. The aim of the framework is to make the design and evaluation choices more transparent. We also present two applications of the framework pertaining to forensic face recognition. Using the framework, we can demonstrate large and explainable variations in discriminating power between subjects.

1. Introduction

Given trace specimens from a crime scene (for example finger marks or face images extracted from surveillance camera footage) and reference specimens taken from a suspect (for example, finger prints or good quality frontal and profile facial images), one of the tasks of the forensic examiner is to determine the strength of evidence supporting the hypothesis that trace and reference specimens have a common donor versus the hypothesis that the trace originates from another donor. We refer to this process as forensic evidence evaluation.

Not every modality has the same maturity level of automation. For example, on one hand, the finger print modality has a straightforward three level representation (of which often only the first two are used) and mature extraction and comparison methods [22]. On the other hand, the face is a complex modality and especially forensic evidence evaluation using images of faces taken under realistic conditions is largely a manual process [17]. Therefore, there is still a general need for research on biometric classifiers that are capable of producing strength of evidence, to be used in a forensic evidence evaluation process in which the human examiner remains to have a pivotal role. Biometric and forensic science have much in common, notably due to their strong interest in connecting individuals to traces (in the forensic nomenclature) or probes (in the biometric nomenclature).

However, a number of small but important differences between biometric and forensic science are easily overlooked. First, not every biometric modality has the same potential in forensic science and vice versa. Second, scores produced by biometric classifiers in most cases cannot directly be used as strength of evidence in a court of law. Finally, certain performance characteristics are relevant from a forensic perspective [23], but are hardly taken into account by standard biometric research.

Additionally, in both biometric and forensic science, the importance of subject based performance evaluation in relation to general performance evaluation seems to be somewhat underrated, since insight into this type of performance might be especially important from a forensic point of view. Here, a subject based performance evaluation uses traces from a single subject, whereas a general performance evaluation uses traces from multiple subjects. Having an extreme eye fissure angle opening might discriminate a specific subject well, while in general this angle has average to poor biometric performance. In general, the range of possible performances and their link to phenotype enforces the scientific foundation of the reported strength of evidence based on phenotypes. This is a relevant issue in light of the very critical NRC [9] and PCAST [26] reports on forensic science in the USA.

The contributions of this paper are:

- A systematic presentation (framework) of aspects to be considered during the design and evaluation of biometric classifiers for forensic evidence evaluation, including forensic performance characteristics and metrics;
- An emphasis on general versus subject based performance evaluation;

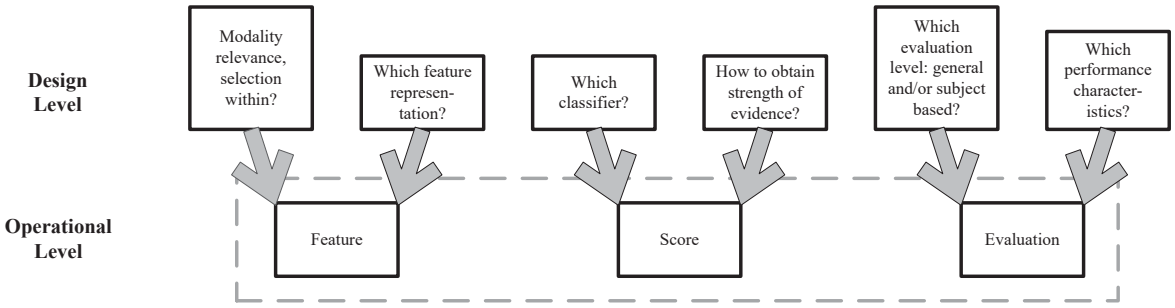


Figure 1. Framework that can be used during the design and evaluation of biometric classifiers for forensic evidence evaluation and relates to the forensic process published in [23]. Top row contains the aspects, bottom row shows essential components of an evaluated biometric system. Arrows are unidirectional "influences" relationships.

- A presentation of two relevant applications within the domain of forensic face recognition.

The paper is structured as follows. Section 2 presents the framework and discusses the six aspects. Section 3 demonstrates the first application of the framework. It studies nine facial measures that can be used when trace images depict a perpetrator wearing a balaclava. Section 4 contains the second, smaller, application of the framework that studies the location of facial marks. It considers the case when trace images originate from surveillance cameras. Both applications take the form of a small paper within a paper, as such they can be read independently from the description of the framework. Since related work is either confined to the framework or the two applications, we do not provide a separate related work section. Section 5 presents the conclusion.

2. Framework

As can be seen in Figure 1, the six aspects reside at the design level of biometric classifiers for forensic evidence evaluation. These aspects influence the biometric system and its evaluation at an operational level. The aspects can be grouped into three groups of two aspects; they influence the feature, score calculation, and the evaluation. Choices addressed in or related to the first five aspects can serve as a template for their influence on Aspect 6, the chosen performance characteristic(s). This framework is a generalisation of the six aspects discussed in [33] and includes forensic performance characteristics presented in a recently published forensic guideline [23].

2.1. Modality relevance and selection within

Modalities can be composed of several smaller entities which on their own might be modalities as well. Due to the forensic context, may be only some of these modalities can be used.

Jain et al. [15] identifies characteristics of biometric modalities. The distinctiveness property is used in forensic scenarios as identification (both closed and open), investigation, intelligence, and the evaluation of strength of evidence [10]. Distinctiveness does not need to be uniform amongst the subjects in order to have a forensic interest. Acceptability is a notable exception; robustness to forensic scenarios and the availability of biometric information as a trace are additional forensically important characteristics.

Another test for suitability is the forensic relevance of the modality, either at source (who is acting) or activity (what is the act) level inference. It involves the description of forensic use case(s) in which the modality could be used. A forensic use case describes an act that produces a particular type of trace material captured at a crime scene. An example is a robbery in which the robber wears a balaclava and trace material only shows a few facial parts like the eyes, eyebrows, mouth, possibly part of nose and part of the chin.

2.2. Feature Representation

The outcome of the modality selection steers the possible feature representation(s). An example is given by facial marks. Since the measurement of the location of a facial mark is subject to within variation, one might also consider the use of a facial grid to represent the location in terms of the grid cell it belongs to, as a method to compensate for this variation [33].

2.3. Classifier

The third aspect is whether we use any data to train a classifier and if so, whether that data is related to a general population or to a subject. Although we use the term classifier, we are mostly interested in the comparison score it produces, rather than a decision. For example, the probability of detecting facial marks depends on the considered region of the face [33]. Hence, using facial mark location

data in a classifier could enhance its ability to discriminate; a model based on a single subject could even be better than one based on general data, especially if the facial mark locations are very distinctive for that particular subject [33].

2.4. Strength of evidence

The desired outcome of a comparison process is either a comparison score in biometric science or strength of evidence in forensic science. The latter is commonly expressed as a likelihood ratio in modern forensic science[1]:

$$\mathrm{LR}(E) = \frac{\mathrm{p}(E|\mathcal{H}_s)}{\mathrm{p}(E|\mathcal{H}_d)}. \tag{1}$$

Here E denotes evidence, $\mathcal{H}_s$ is the same source hypothesis and $\mathcal{H}_d$ is the different source hypothesis. As described in [14], the forensic examiner is responsible for the calculation of $\mathrm{LR}(E)$, whereas a court of law determines the prior odds $\frac{\mathrm{p}(\mathcal{H}_s)}{\mathrm{p}(\mathcal{H}_d)}$ and ultimately the posterior odds $\frac{\mathrm{p}(\mathcal{H}_s|E)}{\mathrm{p}(\mathcal{H}_d|E)}$:

$$\frac{\mathrm{p}(\mathcal{H}_s|E)}{\mathrm{p}(\mathcal{H}_d|E)} = \mathrm{LR}(E) \times \frac{\mathrm{p}(\mathcal{H}_s)}{\mathrm{p}(\mathcal{H}_d)}. \tag{2}$$

Finally, often (1) is used in its $\log_{10}$ form as $\mathrm{LLR}(E)$. In the latter form, it emphasises the magnitude of the likelihood ratio rather than the exact value.

The evidence E in $\mathrm{LLR}(E)$ is either the occurrence of trace x and reference y or a biometric comparison score $s = s(x, y)$ computed on trace x and reference y.

In the first case, we obtain the *feature based log-likelihood ratio*:

$$\mathrm{LLR(x, y)} = \log_{10}\left(\frac{\mathrm{p}(x, y|\mathcal{H}_s)}{\mathrm{p}(x, y|\mathcal{H}_d)}\right). \tag{3}$$

Approaches to calculate (3) include the use of parametric models for $\mathrm{p}(x, y|\mathcal{H}_s)$ and $\mathrm{p}(x, y|\mathcal{H}_d)$ and copula models that relate joint probability distributions to their marginal distributions [31].

In the second case, $\mathrm{LLR}(E)$ reverts to the *score based log-likelihood ratio*:

$$\mathrm{LLR}(s) = \log_{10}\left(\frac{\mathrm{p}(s|\mathcal{H}_s)}{\mathrm{p}(s|\mathcal{H}_d)}\right). \tag{4}$$

Several techniques can be used to estimate the numerator and denominator of (4): parametric model (for example a normal distribution), non-parametric (for example Parzen windows [29]) or the Pool of Adjacent Violators (PAV) algorithm [12]. Given a training set of scores, the PAV algorithm estimates $\mathrm{p}(\mathcal{H}_s|s)$ from which the likelihood $\mathrm{LLR}(s)$ can be derived:

$$\mathrm{LLR}(s) = \mathrm{logit}(\mathrm{p}(\mathcal{H}_s|s)) - \mathrm{logit}(\mathrm{p}(\mathcal{H}_p)), \tag{5}$$

[1] Although Darboux, Appell, and Poincaré suggested its use already in 1906 for the appeal in the Dreyfus case [1], mostly during the last decade it has seen a mainstream acceptance.

with $\mathrm{logit}(x) = \log_{10}\left(\frac{x}{1-x}\right)$. Note that the prior $\mathrm{p}(\mathcal{H}_s)$ in (5) is the fraction of same source pairs in the training set and it is not the prior $\mathrm{p}(\mathcal{H}_s)$ set by a court of law. This process is an example of *score calibration* [5].

Both biometric comparison score functions and feature based log-likelihood ratio functions may include parameters that reflect general behaviour or subject based behaviour. In particular, the same source $\mathcal{H}_s$ and different source $\mathcal{H}_d$ hypotheses can be formulated in two, distinct, manners. The general formulation is

- $\mathcal{H}_s = \mathcal{H}_s^g$: the trace x and reference y originate from a common donor.
- $\mathcal{H}_d = \mathcal{H}_d^g$: the trace x and reference y do not have a common donor.

The subject based formulation is

- $\mathcal{H}_s = \mathcal{H}_s^s$: the trace x and reference y originate from the same specific donor.
- $\mathcal{H}_d = \mathcal{H}_d^s$: the trace x and reference y do not have the same specific donor.

Since the subject based formulation is tailored towards a specific subject (the suspect), one could argue that the subject based formulation should be favoured over the general formulation, although less data is available for a reliable estimate of parameters.

2.5. Evaluation level

Another consideration is at which level performance characteristics are evaluated. From a biometric point of view, often only the general discriminating power is of interest. We refer to this as a *general evaluation*. However, since some modalities are generally not very discriminative, they still might be for certain subjects. It suggests that it makes sense to also report at a subject based level, at least to get an idea of the range of attained performances. We refer to this as *subject based evaluation*. Observe that the use of subject based data is independent of subject based evaluation: it is indeed possible to perform a subject based evaluation on any classifier.

2.6. Performance Characteristics and Metrics

Biometric performance is often confined to ROC, AUC or EER. According to a recently proposed guideline by Meuwly et al. [23], used as a basis for an upcoming ISO standard, there are several other performance characteristics and corresponding metrics that are relevant in the context of our framework. In their work, they classify the performance characteristics into primary and secondary classes. The primary class encompasses

- Accuracy

- Discriminating power
- Calibration

Accuracy is the "closeness of agreement between computed likelihood ratio and the ground truth status" and is measured in Cllr. Given a set $\mathcal{S}$ of n_s and a set $\mathcal{D}$ of n_d scores under the same source hypothesis $\mathcal{H}_s$ and different source hypothesis $\mathcal{H}_d$) respectively, the cost of log-likelihood ratio [7] is defined by:

$$\mathrm{Cllr} = \frac{1}{2}\left(\frac{1}{n_s}\sum_{s\in\mathcal{S}}\log_2(1+\mathrm{e}^{-s}) + \frac{1}{n_d}\sum_{s\in\mathcal{D}}\log_2(1+\mathrm{e}^{s})\right). \tag{6}$$

Discriminating power is a "property representing the capability of a given method to distinguish amongst forensic comparisons where different propositions are true", and is either measured in EER or $\mathrm{Cllr}^{\mathrm{min}}$. If we apply the PAV algorithm to the set of scores and reapply (6), we obtain the minimal achievable cost of likelihood ratio $\mathrm{Cllr}^{\mathrm{min}}$. This quantity measures the discriminating power and can be used as an alternative to EER.

Calibration is a "property of a set of LRs (...)". Perfect calibration means that LRs can be interpreted as strength of evidence. Its performance metric is calibration loss:

$$\mathrm{Cllr}^{\mathrm{cal}} = \mathrm{Cllr} - \mathrm{Cllr}^{\mathrm{min}}. \tag{7}$$

Calibration loss essentially measures how well the computed likelihood ratio can be used as strength of evidence in a court of law.

The secondary performance characteristics are

- Robustness
- Coherence
- Generalisation

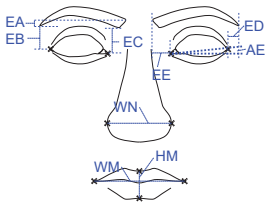
Robustness refers to "the ability of the method to maintain a performance metric when a measurable property in the data changes". Coherence is the "ability to yield likelihood ratio values with better performance with the increase of intrinsic quantity/quality (...)". Generalisation refers to the "ability to maintain performance under a dataset shift." The secondary performance characteristics are measured in Cllr or EER.

3. Application 1: Balaclava

3.1. Introduction

Figure 2a shows a representative balaclava that could be worn by a perpetrator. Although the shown example is of good quality, trace images are typically taken under challenging conditions that significantly impact the image quality. Shape information is typically lost in these low quality images, whereas it might be still possible to extract angles, positions, and distances [36]. Several forensic institutes participate in the Facial Identification Scientific Working Group [3]. It has published several recommendations regarding the forensic facial comparison process, including a detailed list of facial features (FISWG characteristic descriptors) [2] that might be considered during a comparison. Figure 2b shows nine simple balaclava related characteristic descriptors (angle fissure, five distinctive eyebrow measures A-E, the height and width of the mouth, and the width of the nose). We expect that these descriptors have limited general discriminating power, but have a potential to discriminate some subjects to a certain extent.

a) Balaclava



b) FISWG descriptors

Figure 2. a) Subject wearing a balaclava with three holes, b) Face showing the nine considered FISWG characteristic descriptors: angle fissure (AF), five distinctive eyebrow measures A-E (EA-EE), the height (HM) and width (WM) of the mouth, and the width of the nose (WN).

This leads to the following two research questions:

RQ1 What is the discriminating power of an untrained classifier and those trained on general or subject based data viewed at a general and a subject based evaluation level?
RQ2 Which feature phenotypes correspond to good and poor subject based discriminating power?

3.2. Related Work

One of the papers that initiated research in the periocular region is [28], exploring the use of a local (SIFT) and global (HOG, LBP) approach to describe the texture of the periocular region. Other studies investigated different variants of LBP [32, 21] and FISWG characteristic descriptors [35]. Also the eyebrow modality itself has been the topic of several studies [34, 18]. In the latter study, it was shown that the eyebrow region accounts for $\frac{1}{6}$ of the facial region while it retains $\frac{5}{6}$ of the performance of the facial region. The remaining modalities (nose and mouth) have almost never been studied. For example, the study of Moorhouse [25] considered the nose using photometric stereo images; lips as a biometric modality have been studied in Choraś [8]. In general, FISWG characteristic descriptors have been the

subject of several related studies, of which [36] systematically investigated them in various forensic use cases.

3.3. Framework applied

Modality relevance and selection within. The forensic relevance of the nine descriptors has already been explained. Moreover, these features are exemplary for a category of features with limited general discriminating power that can discriminate some subjects to a certain extent.

Feature Representation. All measures are one-dimensional real numbers; all but one (angle fissure) are either a distance or a relative position. The angle fissure is measured in degrees.

Classifier. We employ classifiers that are untrained and ones are trained on either general or subject based data.

Strength of evidence. For each of the FISWG characteristic descriptors, we choose three different score comparison functions. The Euclidean distance score is

$$s(x, y) = -|x - y| \tag{8}$$

and is PAV calibrated, from which the likelihood ratio (5) can be computed.

We also use the feature based log-likelihood ratio (3) and assume that the feature values are normally distributed. Using the general model, we assume that under the same source hypothesis we have

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathcal{H}_s^g = \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{9}$$

and under the different source hypothesis

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathcal{H}_d^g = \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right). \tag{10}$$

We also formulate a subject based model, for which under the same source hypothesis we have

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathcal{H}_s^s = \mathcal{N}\left(\begin{pmatrix} \mu_x^s \\ \mu_y^s \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right) \tag{11}$$

and under the different source hypothesis

$$\begin{pmatrix} x \\ y \end{pmatrix} | \mathcal{H}_d^s = \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \right). \tag{12}$$

The only difference between the general and subject based model are the subject specific means in the same source model.

Evaluation level. Given research question RQ1, we are interested in using both the general and subject based evaluation level.

Performance characteristics. We select discriminating power as the performance characteristic.

3.4. Experimental setup

For this experiment, we use a subset of the FRGCv2 dataset [4]. It consists of 12306 images taken under controlled conditions showing in total 568 subjects with a neutral expression. We adopt the following procedure on training and testing.

In total 376 subjects have less than 25 recordings and are used for training of the general model. For the remaining 192 subjects with 25 or more recordings, the first ten recordings are reserved as subject based training data; the remainder of the recordings constitute the test set.

The nine FISWG characteristic descriptors are automatically determined. We use the One Millisecond Deformable Shape Tracking Library (DEST) [19] for landmark location. We do not employ the default landmark model of DEST as it is too coarse for our purposes. We train DEST using all available (2330) images in the HELEN database [20] and the available ground truth annotation provided by STASM [24] of a model containing 199 landmarks. An affine transformation is then applied on the landmark positions such that the found pupil coordinates of each image are mapped to fixed locations. Finally, we extract the nine descriptors from the landmarks in this coordinate system.

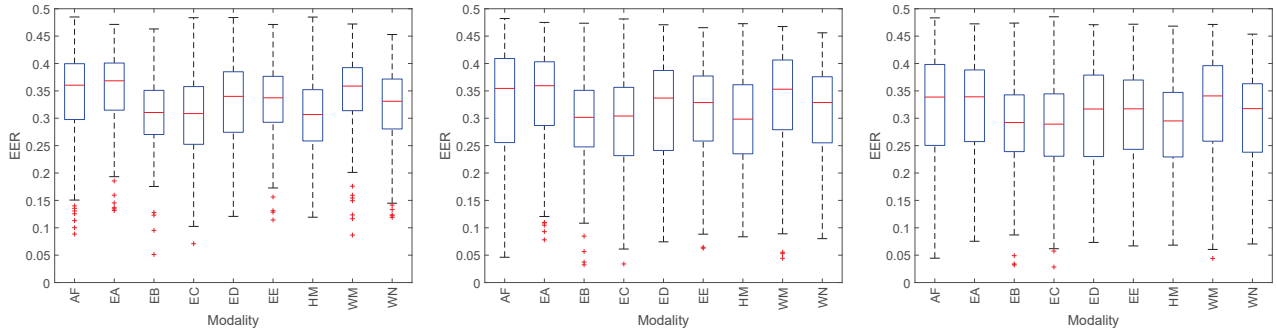
3.5. Results and discussion

Regarding RQ1, the discriminating power of the three classifier types at a general and subject based level, Figure 3 shows the box plots of the EER for comparison methods that do not require training, those trained on general data, and those trained on subject based data. We observe that all considered characteristic descriptors can be seen as soft biometric modalities as they have a very moderate median EER.

Although the box plots appear very similar, we can show, using a Wilcoxon signed rank test, that for each considered characteristic descriptor, the subject based method is better than the general method which in turn is better than the score based method ($p < 0.1\%$). This relationship is reinforced by their corresponding high correlation coefficients: $\rho \in [0.91, 0.99]$.

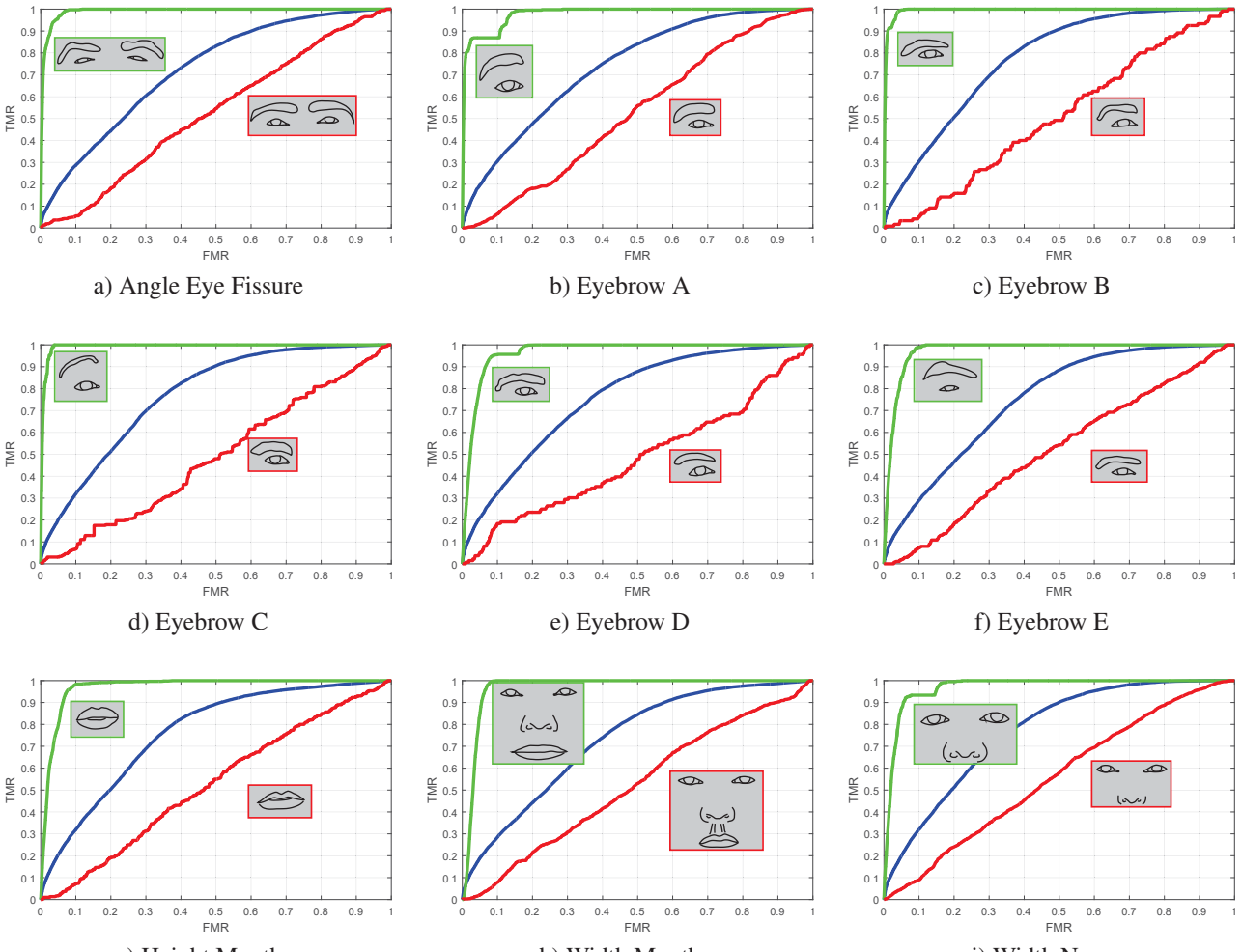
What makes the considered characteristics particularly interesting is the performance difference between some subjects. In Figure 4, for each of the descriptors, we show an outline[2] of the best (green) and worst (red) performing subjects with their performance, alongside with the general performance (blue) in a ROC curve. These examples indicate that the performance at a subject level can be explained in terms of the phenotype of the feature. For example, Figure 4a shows that having low outer eye corners in relation to the inner eye corners is discriminative, whereas they are more leveled, they are essentially random. These results address

[2] Showing the outline is clearer than showing the actual facial patch.



a) Score LLR b) Feature LLR, gen. c) Feature LLR, subj.

Figure 3. Box plots of (a) EER of score based likelihood ratio, (b) EER of feature based likelihood ratios using a general model, and (c) EER of feature based likelihood ratios using a subject based model. The following nine FISWG characteristic descriptors have been considered: angle eye fissure (AE), eyebrow A-E (EA-EE), height mouth (HM), width mouth (WM), and width nose (WN).



a) Angle Eye Fissure b) Eyebrow A c) Eyebrow B

d) Eyebrow C e) Eyebrow D f) Eyebrow E

g) Height Mouth h) Width Mouth i) Width Nose

Figure 4. The variation in performance of nine FISWG characteristic descriptors: (a) angle fissure, (b)-(f) eyebrow A-E, (g) height mouth, (h) width mouth, and (i) width nose. Green and red refers to the best and worst performing subject and performance respectively, blue is the general performance.

RQ2 on the connection between phenotype and discriminating power.

This is one of the key observations in relation to forensic evidence evaluation. In principle we are only interested in discriminating a particular subject from a group of subjects, rather than the much stronger property of discriminating everyone, including this particular subject. Although we do not claim this insight is new, but given the large variation between subjects, it seems warranted to emphasise this in the context of the validation of likelihood ratio methods for forensic evidence evaluation as described by [23]. This guideline does not specify the level of evidence evaluation.

4. Application 2: Grid Based Facial Mark Likelihood Ratio Classifiers

4.1. Introduction

The second application is more limited than the first application and is a small extension of [33] on facial marks. In that work, primary performance characteristics (discriminating power and calibration) of six classifiers operating on a facial grid representing facial mark locations were compared. Moreover, the influence of grid cell sizes ranging from 0.05 IPD (interpupillary distance) to 1.0 IPD on these characteristics was studied as well. This study did not consider secondary performance characteristics like generalisation of discriminating power. In particular, only a subset of images of FRGCv2 [4] taken under controlled conditions and showing subjects with a neutral expression was used. Therefore, we use another dataset, SCFace [13] to address the generalisation of discriminating power.

Due to the inherent poor image quality and low(er) resolution of surveillance camera stills, we expect a large reduction in the number of detected facial marks in trace images relative to the corresponding reference images, introducing a systematic difference and rendering even classifiers trained on general data useless. Therefore, we only consider the Hamming classifier that (a) only uses the presence of facial marks in grid cells and (b) does not use any general or subject based facial mark location data.

This leads to the following two research questions:
RQ1 Can we generalise the discriminating power of Hamming based classifiers that use a facial mark grid?
RQ2 How many subjects can still be discriminated, using their facial mark grid?

4.2. Related Work

Facial marks have forensic relevance [6, 2]. Several studies consider facial marks and the spatial patterns they form. They can complement face recognition systems [27, 11] or serve as a single biometric modality [30]. Applications include quering mugshot databases for matches to facial mark spatial patterns [16] and the calculation of strength of evidence [33].

4.3. Framework applied

Modality relevance and selection within As discussed in [33], not every facial mark type is suitable in a forensic context. In particular, in [33] and the present study only the mole, pockmark, raised skin, and scars are taken into account.
Feature Representation. We assume that the facial mark locations are given in a coordinate system for which the pupil coordinates are fixed. We superimpose a grid with square cells having sizes Δ ranging from 0.05 IPD to 1.0 IPD in steps of 0.05 IPD. The feature is a binary vector that indicates for each grid cell whether it contains no or at least one facial mark.
Classifier. As discussed before, we only consider the Hamming comparison score

$$H((b_1^{ij}),(b_2^{ij})) = -\sum_{i,j} |b_1^{ij} - b_2^{ij}|. \tag{13}$$

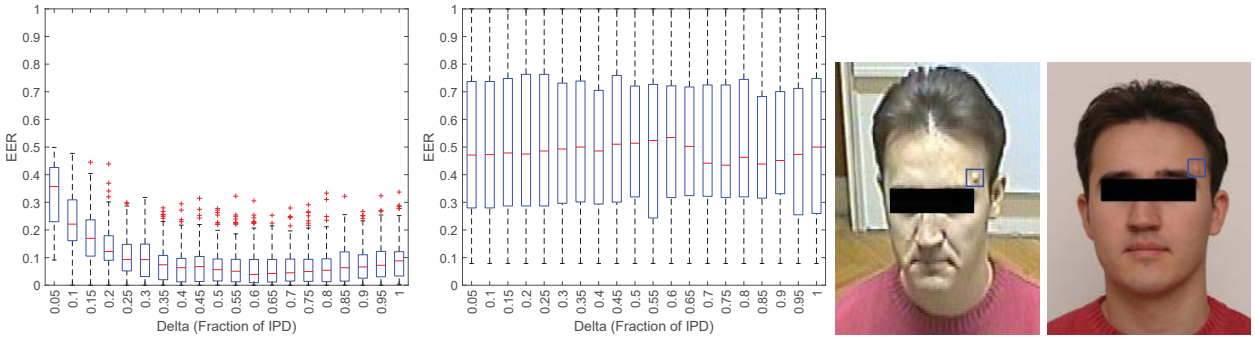
Strength of Evidence. Since the Hamming comparison function does not produce likelihood ratio values, we use PAV calibration and (5) to create a score based likelihood ratio.
Evaluation Level We evaluate both at a general and a subject based level, as we expect that facial marks on low quality images have poor discriminating power, but might discriminate certain subjects.
Performance Characteristics Generalisation is chosen to augment previous work [33].

4.4. Experimental Setup

The SCFace dataset contains surveillance footage of six cameras in seven different configurations (visible and IR), depicting a subject at three different distances (4.20m, 2.60m, and 1.00m), IR mugshot and high resolution images for 130 subjects. We manually locate facial marks in reference images and then annotate all trace images, in random order. The facial mark locations are subsequently mapped to the fixed coordinate system introduced before.

4.5. Results and Discussion

Regarding RQ1, we compare the EER on FRGCv2 (Figure 5a) with the EER on SCFace for Camera 1 and distance 3 (Figure 5b). Other cameras exhibit similar results and are therefore omitted. We observe that in the FRGCv2 case, the EER has some dependency on the grid cell size (notably with smaller grid cell sizes, its increase is explained by within variation) and some variation between subjects. On the other hand, in the SCFace case we observe a very poor EER that is even mostly independent of the grid cell



a) FRGCv2 b) SCFace (Cam 1, Dist 3) c) Trace d) Reference

Figure 5. EER of Hamming comparison score function on a) FRGCv2, b) SCFace Camera 1, Distance 3, a subject that can be perfectly discriminated based on the indicated facial mark location c) Trace, and d) Reference. Due to the SCFace license agreement we can only show anonymised images.

size. With respect to RQ1, we conclude that the discriminating power of Hamming based classifiers using facial mark grids cannot be generalised.

With respect to RQ2, we find that results between subjects vary to a large extent. There is a number of subjects that even have EER=1; this is caused by the large mismatch of observed facial marks in the trace and reference image belonging to that subject in relation to the differences in facial mark observations between its trace and reference images of other subjects. However, mostly for distance 3 (1.00m), we found up to ten different subjects that can be perfectly discriminated based on their facial mark grid. An example of such case is shown in Figures 5c and 5d. Subject 009 has a facial mark (raised skin) located at his left temple, clearly visible in both trace and reference images.

4.6. Acknowledgement

We would like to thank Prof. Grgic for his kind permission to let us use an anonymised version of a subject of the SCFace dataset in Figures 5c and 5d.

5. Conclusion

In this paper, we have presented a framework that considers six aspects during the design and evaluation of biometric classifiers for forensic evidence evaluation. We also presented two applications of this framework.

The first application deals with the situation that trace images depict a perpetrator wearing a balaclava. We explored the use of nine simple characteristic descriptors and found that the incorporation of either general or subject based data versus no training yields very similar classifiers. We observed a large variation in discriminating power between subjects that can be attributed to, in some cases the extreme, phenotype of the considered features.

The second application is an extension of existing work on classifiers that use facial marks as features on a large subset of the FRGCv2 dataset. The extension considered a secondary performance characteristic and used the SCFace dataset. The EER of Hamming based classifiers is in the case of the SCFace dataset very poor and we concluded that its good results on the FRGCv2 dataset cannot be generalised. Despite the lack of facial mark observations in the SCFace case, we did find subjects that could be discriminated based on their facial mark grid.

These applications show that this framework has an added value for the forensic biometric community. The framework makes design choices more transparant. Furthermore, both applications emphasize the importance of subject based evaluation. Especially the first application scientifically connects results to phenotypes and as such helps to enforce the scientific foundation of forensic science found lacking in the NRC and PCAST reports.

References

[1] Affaire Dreyfus, Rapport de Mr. les Experts Darboux, Appell, Poincaré. http://www.maths.ed.ac.uk/~aar/dreyfus/dreyfustyped.pdf. Accessed: 2016-12-12. 3

[2] FISWG Facial Image Comparison Feature List for Morphological Analysis. https://fiswg.org/FISWG_1to1_Checklist_v1.0_2013_11_22.pdf. Accessed: 2017-01-09. 4, 7

[3] FISWG website. https://fiswg.org. Accessed: 2014-04-22. 4

[4] FRGC website. http://www.nist.gov/itl/iad/ig/frgc.cfm. Accessed: 2014-04-22. 5, 7

[5] T. Ali. *Biometric Score Calibration for Forensic Face Recognition*. PhD thesis, University of Twente, Enschede, June 2014. 3

[6] A. Bertillon. *Identification anthropométrique: instructions signalétiques*. 1893. 7

[7] N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(23):230–275, 2006. 4

[8] M. Chorań. The lip as a biometric. *Pattern Analysis and Applications*, 13(1):105–112, 2010. 4

[9] N. R. Council. *Strengthening Forensic Science in the United States: A Path Forward*. 1

[10] D. Meuwly and R. Veldhuis. Forensic biometrics: From two communities to one discipline. In *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, pages 1–12, Sept 2012. 2

[11] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, 11(3):441–467, 2016. 7

[12] T. Fawcett and A. Niculescu-Mizil. PAV and the ROC convex hull. *Machine Learning*, 68(1):97–106, 2007. 3

[13] M. Grgic, K. Delac, and S. Grgic. SCFace - surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879, 2011. 7

[14] G. Jackson, S. Jones, G. Booth, C. Champod, and I. Evett. The nature of forensic science opinion - a possible framework to guide thinking and practice in investigation and in court proceedings. *Science & Justice*, 46(1):33–44, 2006. 3

[15] A. K. Jain, P. Flynn, and A. A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007. 2

[16] A. K. Jain, B. Klare, and U. Park. Face Matching and Retrieval in Forensics Applications. *IEEE MultiMedia*, 19(1):20–20, Jan 2012. 7

[17] Jason P. Prince. To examine emerging police use of facial recognition systems and facial image comparison procedures. www.churchilltrust.com.au/media/fellows/2012_Prince_Jason.pdf, 2012. Accessed: 2014-04-22. 1

[18] F. Juefei-Xu and M. Savvides. Can your eyebrows tell me who you are? In *Signal Processing and Communication Systems (ICSPCS), 2011 5th International Conference on*, pages 1–8, Dec 2011. 4

[19] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, June 2014. 5

[20] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive Facial Feature Localization. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision ECCV 2012*, volume 7574 of *Lecture Notes in Computer Science*, pages 679–692. Springer Berlin Heidelberg, 2012. 5

[21] G. Mahalingam and K. Ricanek. LBP-based periocular recognition on challenging face datasets. *EURASIP Journal on Image and Video Processing*, 2013(1):36, 2013. 4

[22] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer Science & Business Media, 2009. 1

[23] D. Meuwly, D. Ramos, and R. Haraksim. A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 2016. https://doi.org/10.1016/j.forsciint.2016.03.048. 1, 2, 3, 7

[24] S. Milborrow and F. Nicolls. Active Shape Models with SIFT Descriptors and MARS. *VISAPP*, 2014. 5

[25] A. Moorhouse, A. Evans, G. A. Atkinson, J. Sun, and M. L. Smith. The nose on your face may not be so plain: Using the nose as a biometric. In *3rd International Conference on Imaging for Crime Detection and Prevention, ICDP 2009*. Institution of Engineering and Technology, December 2009. 4

[26] P. C. of Advisors on Science and T. (US). *Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods*. Executive Office of the President of the United States, President's Council of Advisors on Science and Technolgy, 2016. 1

[27] U. Park and A. K. Jain. Face Matching and Retrieval Using Soft Biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, Sept 2010. 7

[28] U. Park, R. Jillela, A. Ross, and A. K. Jain. Periocular Biometrics in the Visible Spectrum. *Information Forensics and Security, IEEE Transactions on*, 6(1):96–106, March 2011. 4

[29] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962. 3

[30] N. Srinivas, P. J. Flynn, and R. W. Vorder Bruegge. Human Identification Using Automatic and Semi-Automatically Detected Facial Marks. *Journal of Forensic Sciences*, 61:117–130, 2016. 7

[31] N. Susyanto. *Semiparametric Copula Models for Biometric Score Level Fusion*. PhD thesis, University of Amsterdam, 2016. 3

[32] J. Xu, M. Cha, J. L. Heyman, S. Venugopalan, R. Abiantun, and M. Savvides. Robust local binary pattern feature sets for periocular biometric identification. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1–8, Sept 2010. 4

[33] C. Zeinstra, R. Veldhuis, and L. Spreeuwers. Grid-based likelihood ratio classifiers for the comparison of facial marks. *IEEE Transactions on Information Forensics and Security*, 13(1):253–264, Jan 2018. 2, 3, 7

[34] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Towards the automation of forensic facial individualisation: Comparing forensic to non-forensic eyebrow features. In *Proceedings 35th WIC Symposium, Eindhoven, Netherlands*, pages 73–80, Enschede, May 2014. Centre for Telematics and Information Technology, University of Twente. 4

[35] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Beyond the eye of the beholder: on a forensic descriptor of the eye region. In *23rd European Signal Processing Conference, EUSIPCO 2015, Nice*, pages 779–783. IEEE Signal Processing Society, September 2015. 4

[36] C. G. Zeinstra, R. N. J. Veldhuis, and L. J. Spreeuwers. Discriminating power of FISWG characteristic descriptors under different forensic use cases. In *BIOSIG 2016 - Proceedings of the 15th International Conference of the Biometrics Special Interest Group, 21.-23. September 2016, Darmstadt, Germany*, volume 260 of *LNI*, pages 171–182. GI, 2016. 4, 5