Manually annotated characteristic descriptors: measurability and variability

Chris Zeinstra, Raymond Veldhuis, Luuk Spreeuwers
University of Twente
Services, Cybersecurity and Safety Group, Faculty of EEMCS
P.O. Box 217, 7500 AE Enschede, The Netherlands
Email: {c.g.zeinstra, r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl
Arnout Ruifrok
Netherlands Forensic Institute, Image Analysis & Biometrics
P.O.Box 24044, 2490 AA The Hague, The Netherlands
a.ruifrok@nfi.minvenj.nl

***Abstract*—In this paper we study the measurability and variability of manually annotated characteristic descriptors on a forensic relevant face dataset. Characteristic descriptors are facial features (landmarks, shapes, etc.) that can be used during forensic case work. With respect to measurability, we observe that a significant proportion cannot be determined in images representative of forensic case work. Landmarks, closed and open shapes, and other forensic facial features show mostly that the variability depends on the image quality. Up to 50% of all considered evidential values are either positively or negatively influenced by annotator variability. However, when considering images with the lowest quality, we found that more than 70% of the evidential value intervals in principle could yield the wrong conclusion.**

***Index Terms*—Forensic facial features, FISWG, manual annotation, measurability, variability**

I. Introduction

When a person robs a shop, recordings of that person made by a CCTV camera might be usable as trace material. If a suspect is caught, a forensic facial examiner will compare the, often low quality, trace image(s) to high quality mugshot reference images taken from the suspect when he is in custody. There exist numereous studies ([1], [2], [3], [4], [5]) showing that in general anthropometric measurements are not suitable for forensic evaluation. Therefore, typically the general composition of the face, shape like features (for example the shape of the jaw), and when possible, highly discriminating features like facial marks are taken into account. The outcome of the process is evidential value.

Different forensic institutes use similar but not identical comparison procedures, see Spaun [6] for the operating procedures at the FBI and Prince [7] for other institutes as well. The Facial Identification Scientific Working Group [8] has published recommendations on the comparison process [5]. Their FISWG Facial Image Comparison Feature List for Morphological Analysis [9], FISWG Feature List for short, contains characteristic descriptors, that is, facial features, that can be used during forensic case work.

The comparison process itself is largely manual. We assume that the examiner endows the trace and reference images with manual annotation from which the characteristic descriptors are derived. Corresponding characteristic descriptors in trace and reference images are then compared. In this work we investigate the measurability and variability of manually annotated characteristic descriptors.

According to Jain et al. [10], measurability refers to *"how possible it is to capture the biometric feature using a suitable device (...). The raw data captured must also allow for (...) feature extraction."* Here the biometric feature is the characteristic descriptor, and the device is the annotator who creates annotation from which the characteristic descriptors are to be derived. The dynamic range of characteristic features is large, that is, they range from large scale features (for example the outline of the face) to small scale features (for example lip wrinkles). This suggests there is a relationship between the measurability of characteristic descriptors and for example the resolution of a trace image.

Since taking these measurements is is an inherently subjective process, they will exhibit variability. We will investigate this with respect to the placement of landmarks, open and closed shapes, and a selection of other characteristic descriptors. The variability of landmarks in a forensic context has been studied before by Tome et al. [11]. Since the ultimate output of a forensic facial expert is evidential value, we present four evidental value models, and study the variability of evidential value caused by annotation variability.

The remainder of this paper structured as follows. In Section 2 we introduce characteristic descriptors. In Section 3 we describe the experimental setup. In Section 4 we present and discuss the experimental results, and in Section 5 we formulate our conclusion.

II. FISWG characteristic descriptors

Since there are more than 250 FISWG characteristic descriptors, we only visually introduce a subset of them in Figures 1 and 2. We refer to [9] and [12] for a complete overview. Figure 1 shows the face from a holistic perspective that contains large scale structures like the outline of the face and landmarks that indicate the positioning of facial parts

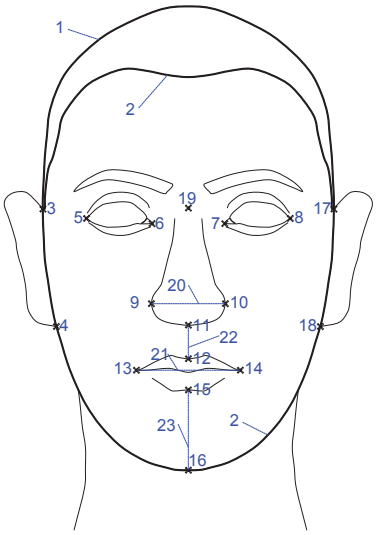

Fig. 1. Face from holistic perspective: (1) Cranial Vault shape/availability, (2) Facial shape, (3) location of 17 landmarks (upper/lower connection ears to face (3, 4, 17, 18), inner/outer corners eyes (5-8), nose (9-11), mouth (12-15), chin (16), and nasal root (19)), (20) width of nose, (21) width of mouth, (22) nose-mouth distance, and (23) mouth-chin distance.

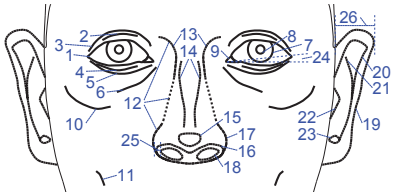

Fig. 2. Other Characteristic descriptors found in the middle part of the face: (1) Fissure shape/size/symmetry, (2) Upper Folds shape/availability/count, (3) Superior Palpebral Furrow shape/availability, (4) Lower Folds shape/availability/count, (5) Inferior Palpebral Furrow shape/availability, (6) Infraorbital Furrow shape/availability, (7) Iris shape, (8) Pupil shape, (9) Caruncle shape, (10) Cheekbone shape/availability, (11) Dimple Cheek shape/availability, (12) Nose shape/size/symmetry, (13) Nasal Root shape/size, (14) Nasal Body shape/size/symmetry, (15) Nasal Tip shape/symmetry, (16) Nasal Base size/deviation, (17) Alae shape, (18) Nostrils shape/size/symmetry, (19) Outer Helix shape/symmetry/size, (20) Inner Helix shape/size, (21) Anti-Helix shape/size, (22) Tragus shape/size, (23) Anti-Tragus shape/size, (24) Fissure angle, (25) Nostril thickness, and (26) Ear Protrusion.

within the face. Figure 2 shows the characteristic descriptors that reside in the middle part of the face.

Although the number of characteristic descriptors is large, each of them falls into one of four feature types. The first feature type is "low-dimensional" $\mathbb{R}^k$, that is either $k = 1$ (for example the eye fissure angle), or $k = 2$ (for example a landmark position). The second type is the visual occurrence of a facial feature (for example the cheekbone), expressed as a binary value. The third type is count (for example the number of upper eye folds). The final type are shapes. An example is the shape of the outer ear helix. The shape feature type is the most occurring type in the set of characteristic descriptors.

III. Experimental setup

A. Dataset

For this study we employ the ForenFace dataset [13]. This dataset contains a reference image and four different trace images for 87 subjects. The trace images are chosen such that they are representative of particular forensic use cases: (a) ID Card refers to use of a valid identity document of another person, (b) Debit Card refers to the use of a stolen Debit Card and (c) Robbery refers to a robbery on for example a bank or shop. Two images of different resolution and illumination represent the latter use case. Example images with their average interpupillary distance (IPD) are shown in Figure 3. ForenFace also contains annotation from which all characteristic descriptors can be derived.

B. Extraction of Characteristic Descriptors

The annotation of ForenFace consists of either landmark positions, see Figure 1, or points that collectively constitute a Hermite spline, representing a shape. A Hermite spline [14] is a piecewise third order polynomial that defines a smooth open or closed curve that can be subsampled into an arbitrary dense point cloud. Most point clouds are directly usable as a characteristic descriptor, for example the eye fissure shape (Figure 2, item 1). Other points clouds, possibly in conjunction with other point clouds, can be used to derive other characteristic descriptors. For example, the eye fissure angle (Figure 2, item 24) is derived from the eye fissure shape, whereas the nostril thickness (Figure 2, item 25) is derived from the alae and nostril shapes. The visual occurrence and count type features are extracted by counting the number of distinct shapes that constitute a characteristic descriptor, for example the number of upper eye fold shapes.

C. Measurability

For each characteristic descriptor and forensic use case we calculate the percentage of subjects for which we found a measurement in the ForenFace dataset. Due to the large amount of characteristic descriptors we average over each forensic use case and each of the 18 facial categories defined in the FISWG Feature List.

D. Variability

Annotation variability refers to the variability of multiple annotations of a single facial feature in a single image. The term variability is chosen instead of for example variation or standard deviation, as we use related but different measures for different characteristic descriptors.

We select five subjects (ids 1,4,19,82, and 101) from the ForenFace dataset for which five images are annotated three times by three trained annotators, yielding in total 25 annotated images with 9 annotations each. At least one week between every session is taken into account. Although we can identify three types of variability (the variability within an annotator, the variability between annotators, and the total variability), due to similar results, we only report total variability.

Landmarks. We investigate the variability of landmarks by measuring the standard deviation with respect to their

a) b) c) d) e)

Fig. 3. Available images with average IPD: a) Reference image (370px), b) ID Card (35px), c) Debit Card (65px), d) Robbery 1 (23px), and e) Robbery 2 (11px).

mean and report results relative to interpupillary distance. This enables comparison between different forensic use cases.

Open and closed shapes. There does not exist a point to point correspondence between two point clouds sampled from two Hermite splines. This implies that a mean shape cannot be defined. We therefore measure shape variability in terms of pairwise difference between two shapes. For closed shapes this is defined as the area constituted by all points that are inside of one shape and outside of the other shape; for open shapes it is defined as the area of the (possibly selfintersecting) closed shape resulting from concatenating corresponding begin and end points of the shapes. Both definitions are visualized in Figure 4.

In order to make comparisions between the variability of different closed and open shapes possible, we scale our results, as in the landmark case. The scaling factor is not interpupillary distance but its two dimensional analog: the interpupillary area (IPA) that we define as the area covered by a square with sides equal to the interpupillary distance. Note that with this approach small (resp. large) shapes will inherently have a small (resp. large) variability. An alternative is to scale to the relative size of the shape. This is straightforward to calculate and to interpret for closed shapes, but a similar approach does not exist for open shapes.

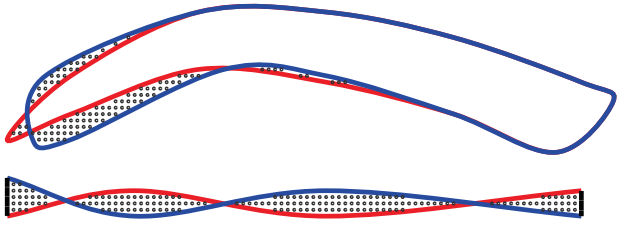

Fig. 4. Visualization of pairwise difference. Given blue and red closed (top) and open (bottom) shapes, the pairwise difference is indicated by dots.

Other characteristic descriptors. Other characteristic descriptors are derived from landmarks and/or shapes, and we will present the variability of a selection of them.

Evidential value. Evidential value is commonly expressed in a log likelihood ratio. In the next section we present four complementary models that are used to calculate evidential value. We report the variability of evidential value in terms of the standard deviation.

E. Models for Evidential Value

This section briefly introduces four different models to calculate evidential value. It follows the models presented in [12], in which more details on the derivation are given. In all cases x is a trace and y a reference. The prosecutor hypothesis $\mathcal{H}_p$ states that the trace and the reference are from the same source, whereas the defence hypothesis $\mathcal{H}_d$ states that the trace comes from a relevant population that does not include the suspect. The log is taken with respect to base 10.

Low dimensional features If we assume

$$p\left(\begin{pmatrix} x \\ y \end{pmatrix} |\mathcal{H}_p\right) \sim N(0, \Sigma_p) \text{ and } p\left(\begin{pmatrix} x \\ y \end{pmatrix} |\mathcal{H}_d\right) \sim N(0, \Sigma_d) \tag{1}$$

then (with $\Delta = \Sigma_d^{-1} - \Sigma_p^{-1}$)

$$l_N(x, y) = \frac{1}{2}\left(\log|\ \Sigma_d| - \log|\Sigma_p| + (x^T y^T)\Delta \begin{pmatrix} x \\ y \end{pmatrix}\right) \tag{2}$$

yields evidential value. We apply (2) only when $(x, y) \in \mathbb{R}^k \times \mathbb{R}^l$, $k, l \leq 2$.

Visual occurence features If we assume a bivariate Bernoulli distribution for $(x, y) \in \{0, 1\} \times \{0, 1\}$

$$\begin{pmatrix} x \\ y \end{pmatrix} |\mathcal{H}_p \sim p_{xy} \text{ and } \begin{pmatrix} x \\ y \end{pmatrix} |\mathcal{H}_d \sim \begin{pmatrix} \text{Bern}(q_x) \\ \text{Bern}(q_y) \end{pmatrix}, \tag{3}$$

then

$$l_B(x, y) = \log\left(\frac{p_{xy}}{q_x^x(1-q_x)^{(1-x)}q_y^y(1-q_y)^{(1-y)}}\right) \tag{4}$$

yields evidential value.

Count features The count comparison score function is applied on count descriptors and is given by

$$s_C(x, y) = -|x - y|. \tag{5}$$

Shape features We represent shapes in terms of point-clouds, so if $X = \{\mathbf{x}_i \in \mathbb{R}^2 | i = 1, \ldots, N_x\}$ and $Y = \{\mathbf{y}_i \in \mathbb{R}^2 | i = 1, \ldots, N_y\}$, then the shape comparison score function is defined by

$$s_{Shape}(X, Y) = -\frac{1}{N_x} \sum_{i=1}^{N_x} d_{pc}^2(\mathbf{x}_i, Y) - \frac{1}{N_y} \sum_{i=1}^{N_y} d_{pc}^2(\mathbf{y}_i, X), \tag{6}$$

where d_{pc} measures the minimal distance between a point $\mathbf{w} \in \mathbb{R}^2$ and a point cloud $Z = \{\mathbf{z}_i \in \mathbb{R}^2 | i = 1, \ldots, N\}$: $d_{pc}(\mathbf{w}, Z) = \min_{i=1,\ldots,N} \|\mathbf{w} - \mathbf{z}_i\|$.

The scores obtained by (5) and (6) are converted to log likelihood ratios by using the Pool of Adjacent Violators algorithm [15]. This algorithm outputs a monotonic transformation from which the mapping

$$s \mapsto \ell(s) = \log(\frac{p(s|\mathcal{H}_p)}{p(s|\mathcal{H}_d)}) \tag{7}$$

can be constructed.

IV. Experimental Results and Discussion

A. Measurability

In Table I we see that the measurability of the face/head outline and hairline/baldness pattern categories remain constant over the forensic use cases. These categories can be considered as large scale facial categories, explaining their constantness.

For most other lower scale facial categories the percentages decrease as the image quality decreases. Probably, either features are considered to be too bad (for example artefacts by image compression) or too small (feature size compared to pixel size) to be annotated. A special case is scars which were apparently not found in this dataset.

In some cases, the forehead brow structures are more visible in the Robbery 1 and 2 cases, compared to the Reference image. This is probably caused by differences in illumination. The reference images have frontal illumination, whereas the Robbery 1 and 2 images have an artificial illumination component from the ceiling, emphasing brow structures on the forehead.

When looking at the facial categories with the highest feature measurability we find that reference images favor facial marks, nose, and eyebrows. For ID Card the categories are eyebrows, face/head compisition, forehead, and mouth, and for the Debit Card nose, mouth, forehead and eyebrows. We see a shift towards larger scale structures when looking at Robbery 1 and 2 images: Hairline/Baldness pattern, Face Head/Outline, and Forehead.

Overall, with decreasing quality and image size the larger scale facial categories seem to be the relatively highest measurable features. However, the percentage of cases for which these (and most other) features are found decreases. This are two expected effects caused by alternating between higher and lower quality images.

We conclude that the FISWG Feature List contains a large number of features that are not measurable in realistic forensic use cases, notably the Robbery use cases.

TABLE I
Measurability of FISWG characteristic descriptors in percentage of subjects, averaged over characteristic descriptors within a facial category

	Ref	ID Card	Debit Card	Robb 1	Robb 2
Face Head Outline	70.9%	71.3%	71.3%	71.3%	70.9%
Face/Head Composition	94.4%	93.8%	91.0%	86.0%	74.1%
Hairline/Baldness pattern	77.9%	75.6%	75.9%	76.7%	79.3%
Forehead	72.0%	81.6%	83.9%	71.3%	80.1%
Eyebrows	90.9%	87.1%	78.6%	38.8%	9.9%
Eyes	78.2%	41.5%	35.6%	11.3%	7.0%
Cheeks	7.5%	11.2%	8.3%	4.0%	2.6%
Nose	93.6%	69.1%	87.5%	27.5%	8.1%
Ears	50.1%	16.9%	14.8%	5.7%	3.8%
Mouth	92.9%	80.3%	85.7%	7.1%	0.8%
Chin/Jawline	79.8%	76.6%	75.2%	52.4%	26.7%
Neck	41.5%	47.0%	23.6%	17.6%	6.2%
Facial Hair	20.7%	18.4%	19.5%	17.2%	8.0%
Facial Lines	42.0%	30.2%	27.9%	12.4%	4.0%
Scars	0.0%	0.0%	0.0%	0.0%	0.0%
Facial Marks	92.0%	55.2%	58.6%	37.9%	6.9%
Alterations	1.1%	0.0%	0.0%	0.0%	0.0%

B. Variability of landmarks

We report the total standard deviations of the landmarks in Figure 5. We notice that the ear upper landmark exhibits more variability than the ear lower landmark, this is probably caused by hair occlusion effecting the annotation of the ear top landmark. Also, we observe that the eye landmarks variability increases in the Robbery 1 and 2 use cases. As can be seen in Figures 3d) and 3e), the integrity of the eye region in those two use cases is significantly deteriorated. Moreover, in the same images we see that the nose and mouth region seems somewhat more robust to image degradation in the Robbery 1 and 2 cases. This fact can also be seen in the reduced variability of those landmarks. The chin landmark variability increases in Robbery 1, Robbery 2, and Debit Card use cases. The former two are caused by the blurring of the chin and

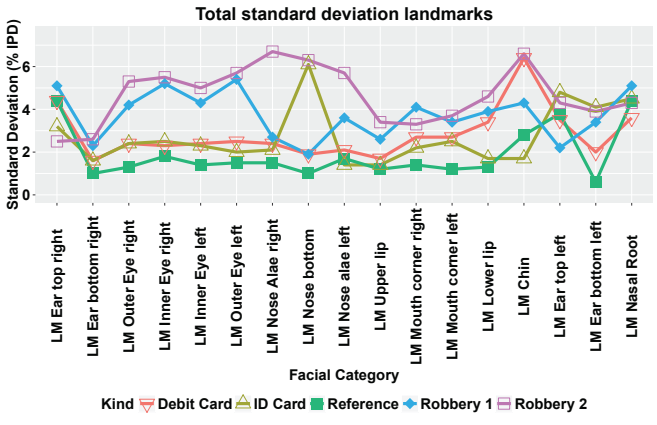

Fig. 5. Total standard deviation of landmarks, expressed as percentage of average IPD.

neck areas, causing an almost indiscernible chin landmark position. The Debit Card variability has an another reason. Since not all subjects look straight into the camera, some images are taken somewhat in an upright position, causing some annotator dubiety on the location of the chin landmark. Finally, the location of the nasal root exhibits variability, even in the reference images. We think that is caused by the inherent difficulty of locating this landmark in frontal view.

C. Variability of closed and open shapes

In Tables II and III we present the total pairwise differences in terms of IPA for some closed and open shapes. For most open and closed shapes we observe a trend that the variability increases with decreasing image quality. We notice that larger scale structures (for example the face outline) exhibit a larger variability than smaller scale structures like the eye fissure, as described in Section III-D. In some cases the variability increases in the Debit Card case and/or (partially) decreases in the Robbery case(s). We think there are two explanations for this observation. The former effect is due to a lack of contrast in the black and white images, causing annotator ambiguity; this effect is noticeable in the face outline. The latter effect is caused by the image quality degradation such that some facial features are so distorted such that are (a) still visible but (b) that there is almost no room for interpretation. In some cases the image degradation is so severe (for example the eye fissure), that there exist no measurements at all.

The cheekbone (location and its visual presence) is sometimes difficult to observe in well conditioned images. Also, its precise location is subject to interpretation variability. If we would compare the variability with respect to average feature area, we find that the eyebrow exhibits the largest pairwise difference. This might be caused by (a) interpretation issues regarding the eyebrow/skin boundary at the facial outline tip, and related (b) sensitivity of the eyebrow/skin boundary visibility to constrast and blur. One can argue that the eye fissure and mouth are less sensitive to these factors as they have better discernible boundaries in terms of color.

D. Variability of other characteristic descriptors

In Table IV we list a selection of characteristic descriptors derived from landmarks and shapes that represent the encoutered variability. In the first five rows we list four distinct distance measures and the fissure angle. As expected generally the standard deviation increases with the image quality degradation. In the last three rows we report three count descriptors. Notice the relationship between image quality and variability, caused by the lack of data in the Robbery cases. Also, we mention that facial marks exhibit the largest variability of all considered count like characteristic descriptors.

TABLE II
TOTAL PAIRWISE DIFFERENCE FOR SOME CLOSED SHAPES, EXPRESSED AS PERCENTAGE OF THE IPA.

	Ref	ID Card	Debit Card	Robb 1	Robb 2
Eyebrow	1.3%	1.4%	1.4%	1.9%	1.2%
Eye Fissure	0.3%	0.5%	0.5%	0.4%	N/A
Mouth	0.9%	1.1%	2.0%	2.9%	N/A

TABLE III
TOTAL PAIRWISE DIFFERENCE FOR SOME OPEN SHAPES, EXPRESSED AS PERCENTAGE OF THE IPA.

	Ref	ID Card	Debit Card	Robb 1	Robb 2
Face outline	2.8%	2.9%	9.1%	6.6%	11.2%
Hairline boundary	1.4%	1.1%	1.2%	2.2%	3.5%
Cheekbone	1.9%	8.0%	N/A	2.0%	2.0%
Nose outline	0.9%	1.4%	1.7%	2.5%	4.2%
Ear outline	0.3%	0.3%	1.3%	1.7%	3.7%
Chin outline	0.9%	0.7%	3.7%	2.6%	1.0%
Neck outline	0.8%	0.6%	1.1%	4.5%	3.9%

TABLE IV
STANDARD DEVIATIONS OF DISTANCES, THE FISSURE ANGLE AND SOME COUNTS. DISTANCES ARE REPORTED WITH RESPECT TO IPD, THE ANGLE IS IN DEGREES, AND COUNT IS DIMENSIONLESS.

	Ref	ID Card	Debit Card	Robb 1	Robb 2
Width nose	0.6%	2.2%	1.7%	2.9%	7.1%
Width mouth	2.1%	4.2%	4.3%	6.1%	6.0%
Nose-mouth distance	1.2%	2.1%	2.6%	1.9%	4.1%
Mouth-chin distance	1.9%	1.9%	8.4%	4.3%	6.8%
Eye fissure angle	1.5°	2.0°	2.4°	2.9°	4.4°
Eye lower folds count	3.3	0.1	0.2	0.0	0.0
Eye upper folds count	1.2	0.0	0.1	0.0	0.0
Facial marks count	5.4	1.2	2.8	0.4	0.0

E. Evidential value and its variability

For each forensic use case and characteristic descriptor, we construct the set of genuine (resp. impostor) evidential values, using the method explained in Section III-E. In Figure 6 we present histograms of the emperical means $\hat{\mu}$ of these two sets for each of the four forensic use cases. We observe that the evidential value of a single characteristic descriptor in general is limited, especially in the Robbery 2 use case. This is in line with the findings presented in [12]. However, in general (a) the combination of characteristic descriptors yields larger evidential value or (b) a single extreme feature value yields high discriminating power.

To each set of genuine (resp. impostor) evidential values having emperical mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$, we associate an interval $I = [\hat{\mu} - 2\hat{\sigma}, \hat{\mu} + 2\hat{\sigma}]$ that captures approximately 95% of the range of evidential values. In order to quantify the influence of annotation variability on the evidential value, we use I and $\hat{\mu}$ to define four distinct influences (Table V). In particular, "$\hat{\mu}$ in correct region" means that a genuine (resp. impostor) evidential value is positive (resp. negative).

TABLE V
TYPES OF ANNOTATION VARIABILITY INFLUENCE ON EVIDENTIAL VALUE

Influence	Description
$\mathcal{I}_+$	Interval I completely in correct region
$\mathcal{I}_\downarrow$	$\hat{\mu}$ in correct region, I partly in correct region
$\mathcal{I}_\uparrow$	$\hat{\mu}$ in incorrect region, I partly in correct region
$\mathcal{I}_-$	$\hat{\mu}$ in incorrect region, I completely in incorrect region

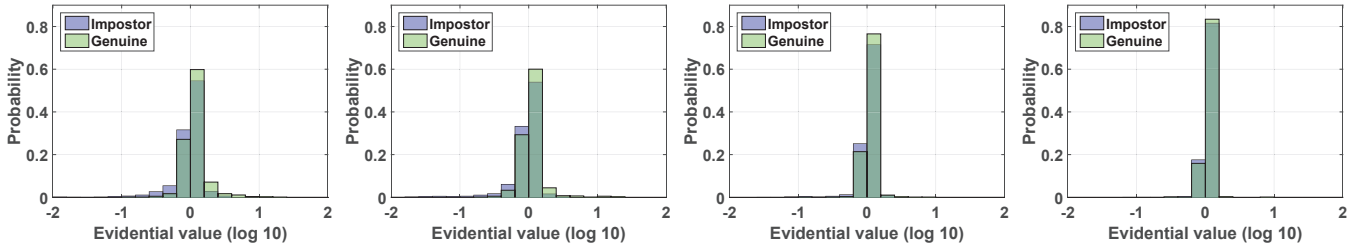

Fig. 6. Histogram of attained evidential values for genuine and imposter cases: (a) ID Card, (b) Debit Card, (c) Robbery 1, and (d) Robbery 2.

TABLE VI
INFLUENCE OF ANNOTATOR VARIABILITY ON EVIDENTIAL VALUE IN PERCENTAGE TOTAL NUMBER OF CHARACTERISTIC DESCRIPTORS FOR EACH FORENSIC USE CASE.

Type	ID Card	Debit Card	Robbery 1	Robbery 2
Genuine $\mathcal{I}_+$	17.5%	20.7%	14.3%	6.6%
Genuine $\mathcal{I}_\downarrow$	29.1%	31.1%	18.7%	12.1%
Genuine $\mathcal{I}_\uparrow$	21.4%	20.3%	15.6%	9.5%
Genuine $\mathcal{I}_-$	32.0%	27.9%	51.5%	71.9%
Genuine Total	100%	100%	100%	100%
Impostor $\mathcal{I}_+$	13.5%	14.5%	7.5%	7.4%
Impostor $\mathcal{I}_\downarrow$	30.5%	29.7%	20.2%	10.9%
Impostor $\mathcal{I}_\uparrow$	19.5%	22.5%	14.9%	9.9%
Impostor $\mathcal{I}_-$	36.5%	33.2%	57.3%	71.8%
Impostor Total	100%	100%	100%	100%

In Table VI we have compiled the annotator influence on evidential value. The influences are similar for genuine and impostor cases. When lowering the image quality, we observe that in general the neutral influence I_+ reduces from around 20% to 7%, wheras the negative (resp. positive) influence $I_\downarrow$ (resp. $I_\uparrow$) reduces from 30% to 10% (resp. 20% to 10%). The percentage of I_- rises over 70% in the Robbery 2 use case, or rephrased, 70% of the evidential value intervals yields the wrong conclusion. However, note that these evidential values are too low to be used in real forensic case work and are caused by using evidential value models that already were shown to have low discrimating power [12].

V. CONCLUSION

In this work we have presented the results of two related experiments using manual annotation from which characteristic descriptors can be derived. With respect to measurability we found that a large number of characteristic descriptors cannot be determined in images representative of forensic case work and therefore we question its detailed nature.

In general, the variability of landmarks, closed and open shapes (in terms of IPA), and other characteristic descriptors increases when the image quality decreases. In other cases we can explain slightly different variability dependence on image quality.

We found that the evidential value of single characteristic descriptors in general is very limited and this reiteres the results of a related study on discriminating power of FISWG characteristic descriptors. The annotation variability of the characteristic descriptors influences up to 50% of all considered evidential values. In the serverest case use case we found that more than 70% of the evidential value intervals in principle could yield the wrong conclusion.

REFERENCES

[1] Krista F. Kleinberg. *Facial anthropometry as an evidential tool in forensic image comparison.* PhD thesis, University of Glasgow, 2008.

[2] Martin Evison and Richard Vorder Bruegge. *Computer-aided forensic facial comparison*. Taylor and Francis Group, Boca Raton, Florida, USA, March 2010. Edited book. Evison and Vorder Bruegge also author the introduction (pp. 1-9) and 'Problems and prospects' (pp.157-168).

[3] Josh P. Davis, Tim Valentine, and Robert E. Davis. Computer assisted photo-anthropometric analyses of full-face and profile facial images. *Forensic Science International*, 200(13):165 – 176, 2010.

[4] M.M. Roelofse, M. Steyn, and P.J. Becker. Photo identification: Facial metrical and morphological features in south african males. *Forensic Science International*, 177(23):168 – 175, 2008.

[5] Fiswg guidelines for facial comparison methods. https://fiswg.org/FISWG_GuidelinesforFacialComparisonMethods_v1.0_2012_02_02.pdf. Accessed: 2017-01-09.

[6] N. A. Spaun. Forensic biometrics from images and video at the federal bureau of investigation. In *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*, pages 1–3, Sept 2007.

[7] 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.

[8] Fiswg website. https://fiswg.org. Accessed: 2014-04-22.

[9] 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.

[10] Anil K. Jain, Patrick Flynn, and Arun A. Ross. *Handbook of Biometrics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[11] R. Vera-Rodriguez, P. Tome, J. Fierrez, N. Expsito, and F. J. Vega. Analysis of the variability of facial landmarks in a forensic scenario. In *Biometrics and Forensics (IWBF), 2013 International Workshop on*, pages 1–4, April 2013.

[12] 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.

[13] Forenface website. http://scs.ewi.utwente.nl/downloads/show, ForenFace/. Accessed: 2016-06-08.

[14] Richard H. Bartels, John C. Beatty, and Brian A. Barsky. *An Introduction to Splines for Use in Computer Graphics And Geometric Modeling*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[15] Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68(1):97–106, 2007.