Discriminating power of FISWG characteristic descriptors under different forensic use cases

Chris Zeinstra, Raymond Veldhuis and 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

***Abstract*—FISWG characteristic descriptors are facial features that can be used for evidence evaluation during forensic case work. In this paper we investigate the discriminating power of a biometric system that uses these characteristic descriptors as features under different forensic use cases. We show that in every forensic use case we can find characteristic descriptors that exhibit moderate to low discriminating power. In all but one use cases, a commercial face recognition system outperforms the characteristic descriptors. However, in low resolution surveillance camera images, some (combination of) characteristic descriptors yield better results than commercial systems.**

***Index Terms*—Forensic facial features, FISWG, discriminating power**

I. Introduction

One of the tasks of a forensic facial expert is to compare trace images to reference images taken from a suspect in order to determine evidential value. This process is referred to as forensic face verification. Although there does not exist a de jure or de facto international standard for forensic face verification, a standardization effort is done by FISWG [3]. FISWG has published several recommendations on facial identification, including a one-to-one comparison list describing characteristic descriptors [2] that can be used during forensic case work.

We envision a forensic facial evaluation system that receives input from a forensic facial expert and computes evidential value. According to [11] and a recent proposed forensic guideline [15] discriminating power is one of six aspects that should be taken into account during the validation of such a forensic evaluation method. Therefore, the aim of this paper is not to present classifiers that have state-of-the-art results, but rather to investigate the discriminating power of biometric classifiers using FISWG characteristic descriptors as features under different forensic use cases.

Note that we present biometric results, that is, averaged results. In particular, genuine scores stem from same source comparisons of different subjects. In an ideal situation sufficient trace and reference material of one subject is available, making a specific subject based comparison possible. This will be investigated in future work.

Typically trace and reference images are wholly visible in casework, whereas in this work we extract FISWG characteristic descriptors prior to the comparison. Its advantage is that we can investigate the truly isolated FISWG characteristic descriptor, and we do not create a bias by seeing the features simultaneously in the trace and reference images. An overview of our system is given in Figure 1.

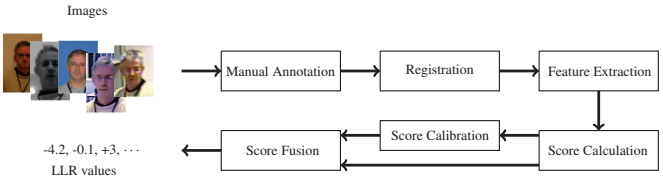


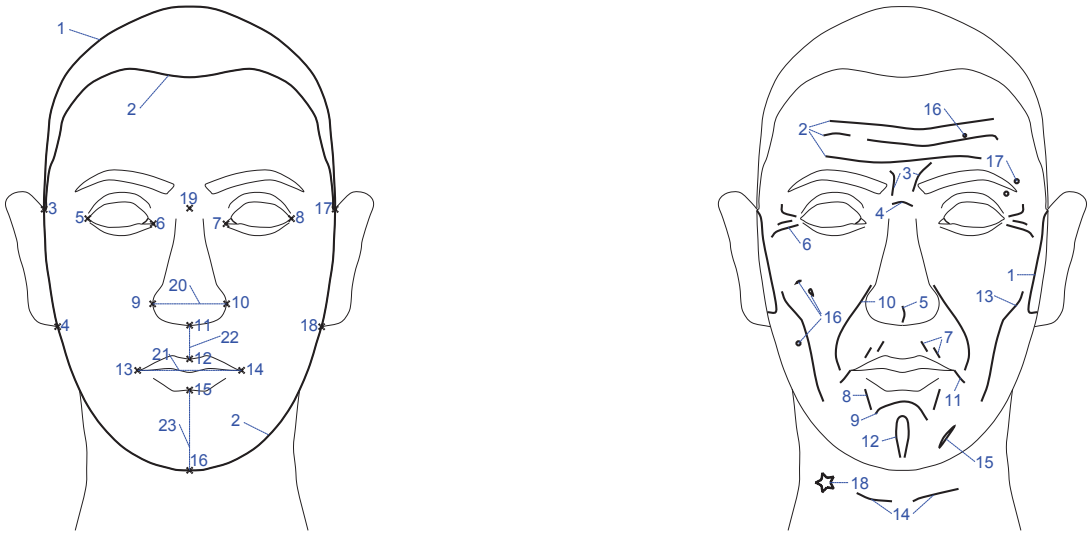
Fig. 1. Overview of our system. LLR refers to log likelihood ratio as a means to represent evidential value.

This paper is organized as follows. In Section 2 we discuss related work, in Section 3 we introduce the FISWG characteristic descriptors, and in Section 4 we present forensic use cases. In Section 5 the experimental setup is discussed. In Section 6 we discuss results of single and specific combinations of characteristic descriptors. We also compare our results to two commercial face recognition systems. Finally, in Section 7 we present the conclusion.

II. Related Work

There exist numerous studies [14], [8], [7] showing that in general anthropometric measurements are not suitable for evidential evaluation. Therefore, forensic face verification typically involves the examination of (dis)similarities of anthropomorphical facial features. It is remarkable that the FISWG one-to-one comparison list includes some anthropometric measures.

Two studies by Tome et al. are closely related to our work. In [17] the biometric performance of linear SVM classifiers on 15 forensic facial regions is investigated. Here the SCFace [10] and subset of Morph [16] are used. They conclude that "... depending on the acquisition distance, the discriminative power of regions change, having in some cases better performance than the full face". In [18] the performance of continuous and discrete soft biometric features are tested on the Morph [16] and ATVS Forensic DB [19] datasets. Experimental results show high discrimination power and good recognition performance for some specific cases. However, these cases correspond to relatively good quality images.



a) Face from a holistic perspective

b) Face from detailed perspective

Fig. 2. Face from a) holistic and b) detailed perspective. Prefix H (resp. D) refers to holistic (resp. detailed) perspective. (H1) Cranial Vault shape/availability, (H2) Facial shape, (H3) location of 17 landmarks (upper/lower connection ears to face (H3, H4, H17, H18), inner/outer corners eyes (H5-H8), nose (H9-H11), mouth (H12-H15), chin (H16), and nasal root (H19)), (H20) width of nose, (H21) width of mouth, (H22) nose-mouth distance, and (H23) mouth-chin distance. (D1) Facial Hair shape/symmetry/availability, (D2) Forehead Creases shape/size/availability/count, (D3) Vertical Glabellar shape/size/availability/count, (D4) Nasion Crease shape/availability/count, (D5) Bifid Nose Crease shape/availability/count, (D6) Periorbital Creases shape/size/availability/count, (D7) Upper Circumoral Striae shape/size/availability/count, (D8) Lower Circumoral Striae shape/size/availability/count, (D9) Mentolabial Sulcus shape/size/availability, (D10) Nasolabial Creases shape/size/availability, (D11) Marionette Lines shape/size/availability, (D12) Cleft Chin shape/size/availability, (D13) Buccal Creases shape/size/availability, (D14) Neck wrinkles shape/size/availability/count, (D15) Scars shape/availability/count, (D16) Facial Marks shape/availability/count, (D17) Piercing shape/availability/count, and (D18) Tattoo shape/availability/count.

There exist some smaller scale studies by Zeinstra et al. that also consider FISWG from an extract feature first based approach, on eyebrows [20], [22], and the periocular region [21].

Other research efforts focus on somewhat different aspects of forensic face recognition: facial aging, forensic sketch recognition, and facial mark based matching and retrieval [12], [13].

III. FISWG Characteristic Descriptors

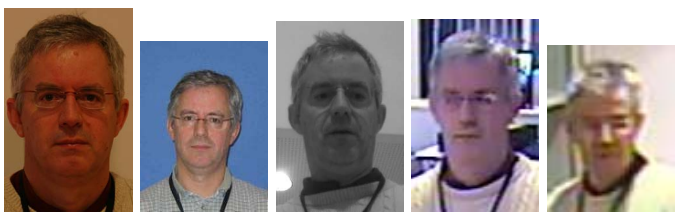
Characteristic descriptors capture information that is considered important during forensic casework. In most cases multiple characteristic descriptors are extracted from one facial trait. For example, from the eyebrow the shape, size, hair density, symmetry, and specific relative positions can be extracted. We present the characteristic descriptors only visually in Figures 2 and 3 due to the sheer number of them (250). In general, most characteristic descriptors fall into classes as landmark, shape, width, size, etc. Also, some very specific characteristic descriptors are defined. For example, the B position of the eyebrow is defined as the vertical position of the outer tip of the eyebrow with respect to the outer eye corner (Figure 3 U8).

IV. Forensic use cases and the ForenFace database

A forensic use case refers to a criminal act whose traces consist of distinct facial image types. In our work we use the ForenFace database [4]. This database contains manually annotated images of 87 subjects that are representative of three forensic use cases. Moreover, from the annotation the FISWG characteristic descriptors can automatically be derived.

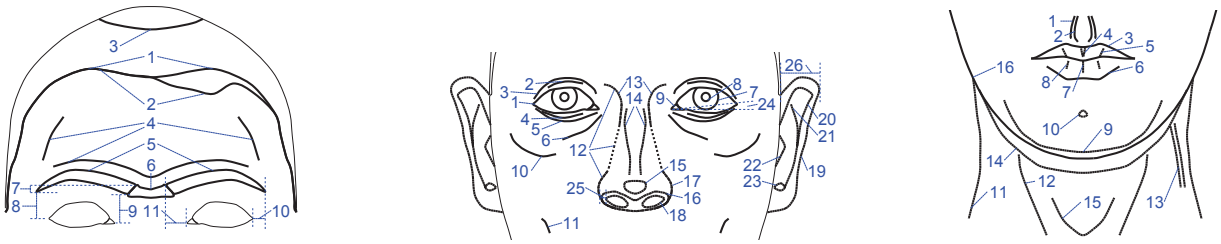
The ID Card use case is when a customs or immigration officer suspects that the used identity document has not been tampered with, but does not correspond to the person who is presenting it. The Debit Card use case is the withdrawal of money using a stolen debit card. In this case trace material is recorded by a small camera in the ATM. The Robbery use case is a robbery on a bank, shop or gas station. At those premises often CCTV surveillance cameras are mounted on a wall or ceiling.

The images and forensic use cases are shown for one subject in Figure 4. In particular, with the average interpupillary dis-



a) b) c) d) e)

Fig. 4. 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).



a) Upper facial parts b) Middle facial parts c) Lower facial parts

Fig. 3. Upper a), middle b), and lower c) parts of the face. Prefix U (resp. M and L) refers to the upper (resp. middle and lower) parts. (U1) Forehead hairline shape/symmetry/size, (U2) Hair/Forehead boundary shape, (U3) Cranial baldness shape/availability, (U4) Ridge structures shape/availability, (U5) Eyebrows shape/size/symmetry, (U6) Unibrow shape/availability, (U7-U11) Relative positions A-E. The relative positions are measured on both eyebrows. (M1) Fissure shape/size/symmetry, (M2) Upper Folds shape/availability/count, (M3) Superior Palpebral Furrow shape/availability, (M4) Lower Folds shape/availability/count, (M5) Inferior Palpebral Furrow shape/availability, (M6) Infraorbital Furrow shape/availability, (M7) Iris shape, (M8) Pupil shape, (M9) Caruncle shape, (M10) Cheekbone shape/availability, (M11) Dimple Cheek shape/availability, (M12) Nose shape/size/symmetry, (M13) Nasal Root shape/size, (M14) Nasal Body shape/size/symmetry, (M15) Nasal Tip shape/symmetry, (M16) Nasal Base size/deviation, (M17) Alae shape, (M18) Nostrils shape/size/symmetry, (M19) Outer Helix shape/symmetry/size, (M20) Inner Helix shape/size, (M21) Anti-Helix shape/size, (M22) Tragus shape/size, (M23) Anti-Tragus shape/size, (M24) Fissure angle, (M25) Nostril thickness, and (M26) Ear Protrusion. (L1) Philtrum Ridges width/symmetry, (L2) Philtrum Furrow width/symmetry, (L3) Upper Lip shape/symmetry, (L4) Upper Lip Tubercle shape, (L5) Upper Lip Creases shape, (L6) Lower Lip Outline shape/symmetry, (L7) Lower Lip Median Sulcus shape, (L8) Lower Lip Creases shape, (L9) Chin shape/size/symmetry, (L10) Chin Dimple shape/availability, (L11) Neck Boundaries size, (L12) Musculature shape/availability, (L13) Veins shape/availability, (L14) Double chin shape/availability, (L15) Laryngeal shape/size/availability, (L16) Jawline shape.

tance (IPD) in pixels, for each subject we have one annotated reference image (370px), and four annotated trace images: ID Card (35px), Debit Card (65px), and two for the Robbery use case: Robbery 1 (23px), and Robbery 2 (11px). The first two images are acquired by a photo camera, the latter three are extracted from CCTV footage. All images are color images, except the Debit Card.

V. Experimental setup

A. Annotation, Registration and Extraction of characteristic descriptors

The annotation in the ForenFace database contains landmarks and shapes. The latter are represented by Hermite splines. A Hermite spline is a piecewise third order polynomial parametric curve [6]. It is defined by the interpolation of the annotated points and, in the case of the ForenFace database, by assuming that the tangent at an annotated point is equal to the vector connecting the neighboring points.

Since the raw manual annotation data lacks a common coordinate system, we apply an affine transformation as a registration step. This registration maps pupil coordinates to fixed locations.

Characteristic descriptors are then extracted from the registered annotation. Shape descriptors are equidistantly subsampled from the corresponding Hermite spline. All other descriptors can be derived from Hermite splines and landmarks. Other descriptors like the B position of the eyebrow (Figure 3 U8) are derived from two types of annotation, in this case the eye fissure and eyebrow shapes.

B. Similarity score functions, score calibration, and score fusion

We use four different similarity score functions. Given the forensic context of our work, scores should be interpretable as evidential value. In some cases we can directly model the similarity score function as a log likelihood ratio:

$$s(x,y) = \log(L(x,y)) = \log\left(\frac{p\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_p\right)}{p\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_d\right)}\right). \quad (1)$$

Here, $p\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_p\right)$ (resp. $p\left(\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_d\right)$) models the joint probability distribution of trace x and reference y under the prosecutor (resp. defence) hypothesis. The prosecutor hypothesis states that the trace and the reference are from the same source, whereas the defence hypothesis states that the trace comes from a relevant population that does not include the suspect.

For low dimensional continuous descriptors like width, size, etc. we have $(x,y) \in \mathbb{R}^k \times \mathbb{R}^l$, $k,l \leq 2$. We assume a normal distribution (after substraction of the mean) with $\Sigma_p = \begin{pmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{xy}^T & \Sigma_{yy}\end{pmatrix}$ and $\Sigma_d = \begin{pmatrix}\Sigma_{xx} & 0\\ 0 & \Sigma_{yy}\end{pmatrix}$:

$$\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_p \sim \mathcal{N}(0,\Sigma_p) \text{ and } \begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_d \sim \mathcal{N}(0,\Sigma_d). \quad (2)$$

In this case (1) has a closed form, with $\Delta = \Sigma_d^{-1} - \Sigma_p^{-1}$:

$$l_N(x,y) = \frac{1}{2}\left(\log|\,\Sigma_d| - \log|\Sigma_p| + (x^T y^T)\Delta\begin{pmatrix}x\\y\end{pmatrix}\right). \quad (3)$$

For availability descriptors, that is, $(x,y) \in \{0,1\}\times\{0,1\}$, we assume a bivariate Bernoulli distribution:

$$\begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_p \sim \text{Bern}\begin{pmatrix}p_{00} & p_{10}\\ p_{01} & p_{11}\end{pmatrix} \text{ and } \begin{pmatrix}x\\y\end{pmatrix}|\mathcal{H}_d \sim \begin{pmatrix}\text{Bern}(q_x)\\ \text{Bern}(q_y)\end{pmatrix}. \quad (4)$$

Under this assumption, (1) reverts to

$$l_B(x,y) = \log\left(\frac{p_{xy}}{q_x^x(1-q_x)^{(1-x)}q_y^y(1-q_y)^{(1-y)}}\right). \quad (5)$$

We also define two similarity score functions that are not modelled as a log likelihood ratio. This is necessary when the probability distributions in (1) of a characteristic descriptor either cannot be easily modelled or its parameters cannot be estimated reliably.

The count similarity score function is applied on count descriptors and is given by

$$s_C(x, y) = -|x - y|. \tag{6}$$

We represent shapes in terms of pointclouds, 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 similarity score function is defined by

$$s_{Shape}(X, Y) = -\frac{1}{N_x} \sum_{i=1}^{N_x} d_{pc}^2(\mathbf{x}_i, Y) - \frac{1}{N_y} \sum_{i=1}^{N_y} d_{pc}^2(\mathbf{y}_i, X), \tag{7}$$

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

By assumption, the scores obtained from (3) and (5) are log likelihood ratios. The scores obtained by (6) and (7) are converted to log likelihood ratios by using the Pool of Adjacent Violators algorithm [9] on the set of scores. This algorithm constructs a monotonic transformation such that a similarity score s is mapped to an a posteriori probability $t = P(\mathcal{H}_p|s)$. From t the log likelihood ratio $l(s)$ can be derived:

$$l(s) = \log(L(s)) = \text{logit}(t) - \text{logit}(p(\mathcal{H}_p)). \tag{8}$$

If we assume independence of facial features, then score fusion by adding scores corresponds to the log likelihood ratio of the combined characteristic descriptors.

C. Experimental protocol

We use the train-test protocol specified by ForenFace. It specifies 50 randomly generated splits of 87 subjects into 67 train and 20 test subjects. Test scores for each round are aggregated. Parameters for (3) and (5) are estimated during the training phase. For the characteristic descriptors on which (6) or (7) are applied, during the training phase the transformation (8) is estimated, which is then applied to test scores for conversion into log likelihood values.

VI. Experimental Results and Discussion

Since the number of characteristic descriptors is large, we will restrict the presentation of results. In particular, we graphically present characteristic descriptors with the lowest EER within a facial category. This measure of discriminating power is chosen in accordance with [15]. Also, results of fusion based on the used similarity score function outside the facial categories are presented. Finally, results are compared to commercial face recognition systems.

A. Single and combined characteristic descriptors

Figure 5 a) shows the results in the ID Card use case. We find that for single characteristic descriptors two modalities have a similar lowest EER=0.24: the position of the lower ear landmark, and the shape of the jaw. Furthermore, the remainder of the characteristic descriptors have EER $\geq$ 0.28. The single best performing facial category is the composition of the face. This category encompasses all landmark positions and 4 distances, see Figure 2 a) H3-H23. When we combine all normal (l_N) scores within that category we obtain EER=0.16. In most other facial categories a combination based on normal scores also yields the lowest EER.

In the Debit Card use case (Figure 5 b)) we observe that some single characteristic descriptors have EER=0.29 to 0.31. They are in order of EER: the size of the nose, the hairline/forehead boundary, and the size of the ears. When considering the combination of characteristic descriptors then again the face composition normal combined yields the lowest EER (0.26).

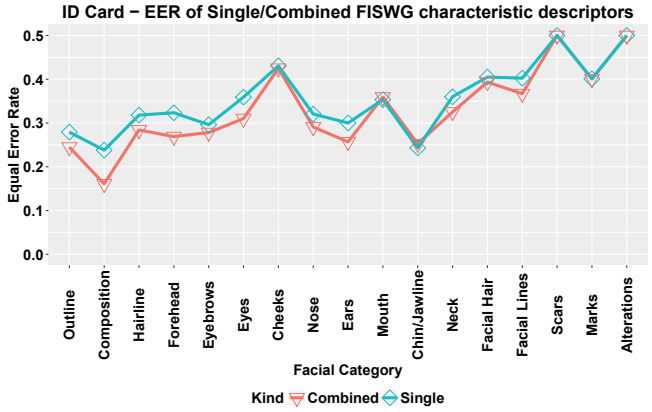
Apart from one specific landmark, in the ID Card use case mostly shape based features yield the best results. In the Debit Card use case simple measures like sizes have the highest discriminating power. This is probably caused by a low contrast, making a majority of shapes more difficult to discern. In both use cases the composition of the face yields the best combined results within a facial category. As indicated before, anthropometry in general has limited use in forensic evaluation. However, it seems that flexibility in the model underlying (3) helps to capture the relation between the trace and reference descriptors.

Figures 6 c) and d) indicate a significant reduction in discriminating power. We find for Robbery 1 and 2 one single characteristic descriptor performing relatively well: the hairline/forehead boundary (EER=0.31 resp. EER=0.38). The shape of the hair/forehead boundary seems actually somewhat resilient to harsh image conditions. This can be explained by the fact that even under challenging conditions this boundary is still visible due to its length and clear color or grayscale difference. In the work of Tome et al. [17] it is reported that in a similar situation the forehead area is the best performing facial area. We think that the discriminative nature of this area actually stems from the inclusion of this boundary.

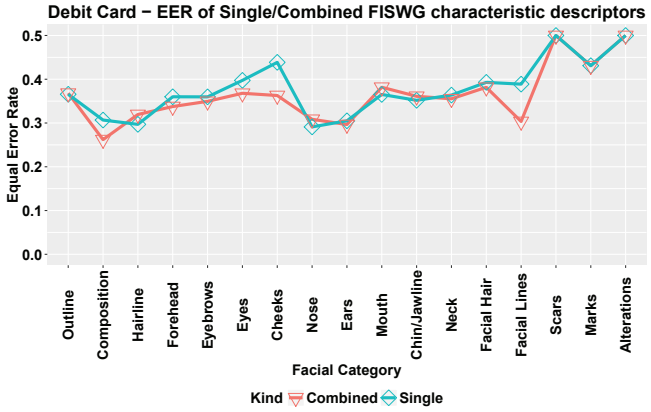
The EER of the characteristic descriptors rapidly pass the EER=0.40 mark. Other simple descriptors like the availability of facial hair, the size and availability of facial lines (especially the nasolabial lines, see Figure 2 D10), and size of the nose start to emerge as single and combined characteristic descriptors with the lowest EER.

B. Fusion outside a facial category

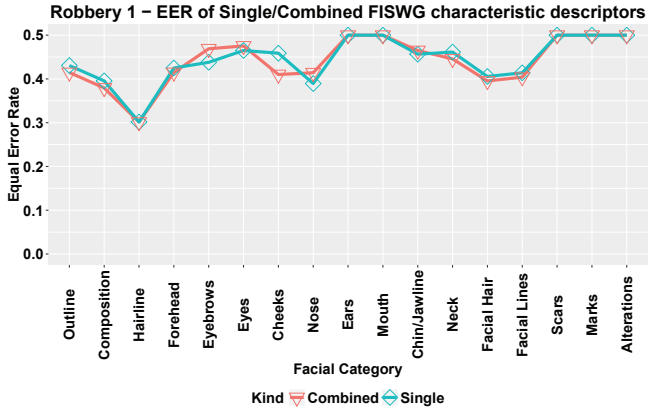
In Table I results of score fusion based on similarity score function are presented. This Table also includes the fusion of all scores. First of all, we observe that the fusion of count scores does not yield any satisfactory results. This has several causes. Upon inspection, we observed a mismatch between counts in trace and reference images, as counts seemingly are



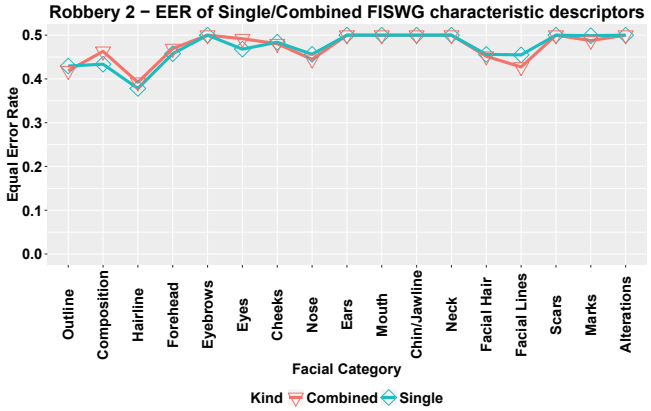
a) ID Card use case



b) Debit Card use case



c) Robbery 1 use case



d) Robbery 2 use case

Fig. 5. Lowest EER of single and combined characteristic descriptors within each facial category under different forensic use cases. Combined refers to score fusion within a facial category.

very sensitive to image size and conditions. Moreover, the score function s_C only measures count difference and does not take the count into account.

We observe that in the ID Card use case fusion of all normal scores is marginally better (EER=0.13) than that of the composition of the face (EER=0.16). This indicates that already a part of the discriminative power is contained in the composition of the face. A similar effect can be seen in the Debit Card use case.

Fusion of normal scores yields the highest discriminating power in the ID use case. This prevalence shifts towards fusion based on availability (l_B) features in the Debit Card and Robbery use case. This effect is to be expected for the latter use case, as only the availability descriptors is robust to severe image quality degradation.

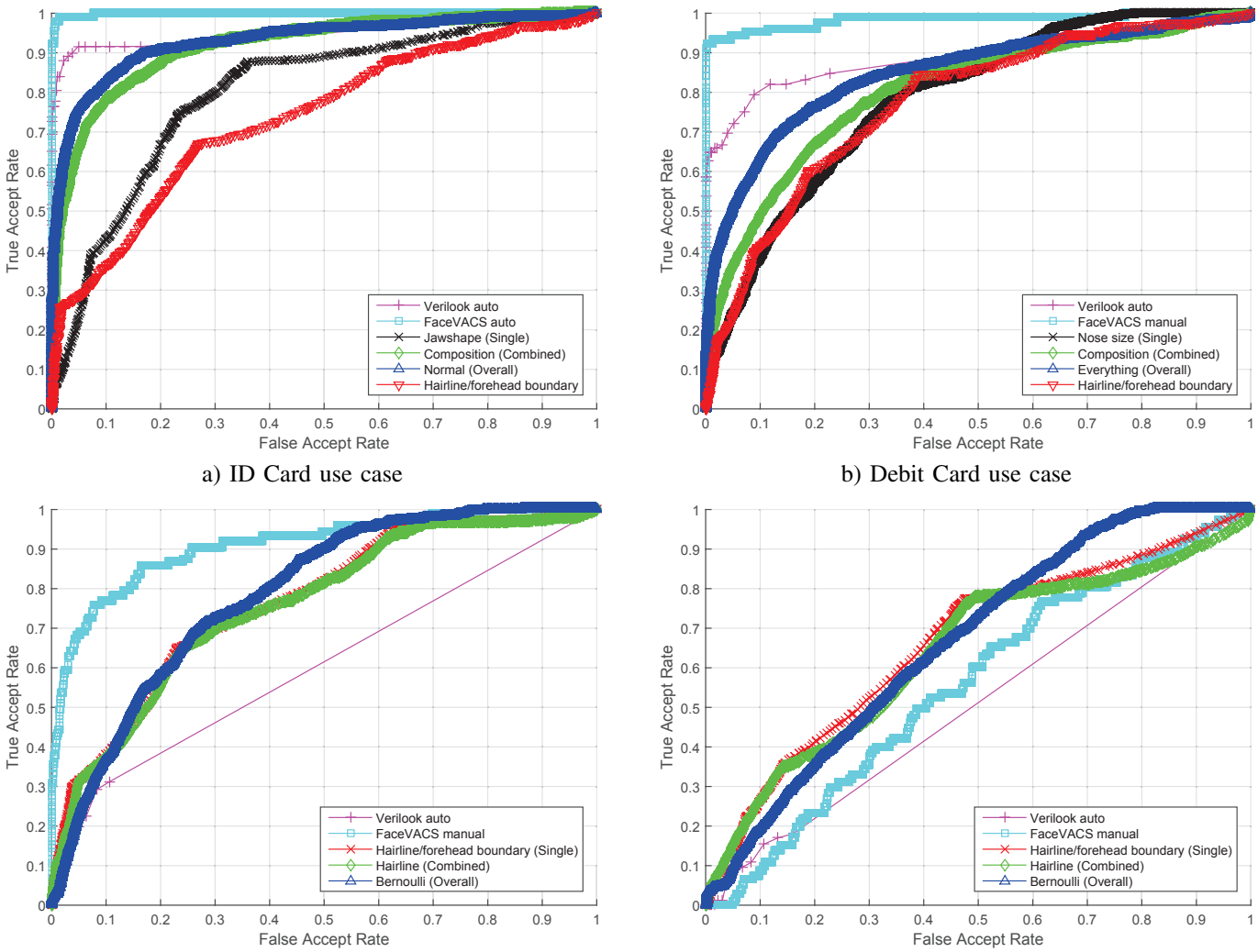
TABLE I
EER OF SCORE FUSION OUTSIDE A FACIAL CATEGORY AND UNDER DIFFERENT FORENSIC USE CASES.

	ID Card	Debit Card	Robbery 1	Robbery 2
Normal (l_N)	0.13	0.24	0.37	0.46
Bernoulli (l_B)	0.19	0.23	0.29	0.40
Count (s_C)	0.42	0.46	0.50	0.50
Shape (s_{Shape})	0.18	0.33	0.38	0.45
All	0.13	0.22	0.31	0.43

C. Commercial systems

We also test performance of two commercial systems using the same evaluation protocol. The Neurotec Verilook 6.0 [5] (with automatic eye coordinate detection) and Cognitec FaceVACS 9.1 [1] (with manually provided eye coordinates) systems are used for this purpose. Both systems use proprietary algorithms. The results are shown in Figure 6 for each forensic use case.

The characteristic descriptors are both in the ID Card and Debit Card use cases outperformed by commercial systems. The situation gets more interesting in the Robbery use cases. In both use cases, Verilook generates a large number of zero genuine and impostor scores, causing the large linear part in the ROC. In the Robbery 2 case, the discriminative power of both commercial systems are now essentially random and the shown characteristic descriptors have lower EER than the commercial systems. One could argue that the hairline/forehead boundary is not a very good biometric since its permanence



a) ID Card use case

b) Debit Card use case

c) Robbery 1 use case

d) Robbery 2 use case

Fig. 6. ROC's of commercial systems versus characteristic descriptors under different forensic use cases.

and collectability properties are challengeable. An alternative is the fusion of availability features. However, the reported EER's are not satisfactory from a biometric perspective.

VII. Conclusion

In this work we have investigated the discriminating power of a biometric system that uses FISWG characteristic descriptors as features under different forensic use cases. In every forensic use case we can find characteristic descriptors, either single or combined, that yield moderate to low discriminating power. However, in all but one forensic use case, the characteristic descriptors are outperformed by a commercial face recognition system. In the use case with low interpupillary distance (11px) we found that the hairline forehead boundary as a single characteristic descriptor and the combination of availability features perform somewhat better than both commercial systems. However, their discriminating power is low.

The presented results are averaged biometric results. In an ideal situation sufficient trace material of one subject is available, making a specific subject based comparison possible. Further research is needed to investigate the discriminitative power of specific FISWG characteristic descriptors in that situation.

Acknowledgment

We would like to thank Cognitec Systems GmbH. for supporting our research by providing the FaceVACS software. Results obtained for FaceVACS were produced in experiments conducted by University of Twente, and should therefore not be construed as a vendor's maximum effort full capability result.

REFERENCES

[1] "Cognitec FaceVACS website," http://www.cognitec.com/products.html, accessed: 2015-12-29.

[2] "FISWG Facial Image Comparison Feature List for Morphological Analysis," https://www.fiswg.org/document/viewDocument?id=50, accessed: 2014-10-24.

[3] "FISWG website," http://www.fiswg.org, accessed: 2014-04-22.

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

[5] "Neurotechnology Verilook website," http://www.neurotechnology.com/verilook.html, accessed: 2016-05-10.

[6] R. H. Bartels, J. C. Beatty, and B. A. Barsky, *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1987.

[7] J. P. Davis, T. Valentine, and R. E. Davis, "Computer assisted photo-anthropometric analyses of full-face and profile facial images," *Forensic Science International*, vol. 200, no. 13, pp. 165 – 176, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0379073810001842

[8] M. Evison and R. V. Bruegge, *Computer-aided forensic facial comparison*. Boca Raton, Florida, USA: Taylor and Francis Group, March 2010, edited book. Evison and Vorder Bruegge also author the introduction (pp. 1-9) and 'Problems and prospects' (pp.157-168). [Online]. Available: http://nrl.northumbria.ac.uk/5363/

[9] T. Fawcett and A. Niculescu-Mizil, "PAV and the ROC convex hull," *Machine Learning*, vol. 68, no. 1, pp. 97–106, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10994-007-5011-0

[10] M. Grgic, K. Delac, and S. Grgic, "Scface a surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11042-009-0417-2

[11] R. Haraksim, D. Ramos, D. Meuwly, and C. E. Berger, "Measuring coherence of computer-assisted likelihood ratio methods," *Forensic Science International*, vol. 249, pp. 123 – 132, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S037907381500047X

[12] A. K. Jain, B. Klare, and U. Park, "Face recognition: Some challenges in forensics," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, March 2011, pp. 726–733.

[13] ——, "Face Matching and Retrieval in forensics applications," *IEEE MultiMedia*, vol. 19, no. 1, pp. 20–20, Jan 2012.

[14] K. F. Kleinberg, "Facial anthropometry as an evidential tool in forensic image comparison." Ph.D. dissertation, University of Glasgow, 2008.

[15] D. Meuwly, D. Ramos, and R. Haraksim, "A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation," *Forensic Science International*, 2016/06/09. [Online]. Available: http://dx.doi.org/10.1016/j.forsciint.2016.03.048

[16] K. Ricanek Jr. and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, ser. FGR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 341–345. [Online]. Available: http://dx.doi.org/10.1109/FGR.2006.78

[17] P. Tome, J. Fierrez, R. Vera-Rodriguez, and D. Ramos, "Identification using face regions: Application and assessment in forensic scenarios," *Forensic Science International*, vol. 233, no. 13, pp. 75 – 83, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0379073813003976

[18] P. Tome, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, "Facial soft biometric features for forensic face recognition," *Forensic Science International*, vol. 257, pp. 271–284, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.forsciint.2015.09.002

[19] 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*, April 2013, pp. 1–4.

[20] 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 of the 35th WIC Symposium on Information Theory in the Benelux, Eindhoven, Netherlands*. Enschede: Centre for Telematics and Information Technology, University of Twente, May 2014, pp. 73–80.

[21] ——, "Beyond the eye of the beholder: on a forensic descriptor of the eye region," in *23rd European Signal Processing Conference, EUSIPCO 2015, Nice*. IEEE Signal Processing Society, September 2015, pp. 779–783.

[22] ——, "Examining the examiners: an online eyebrow verification experiment inspired by fiswg," in *International Workshop on Biometrics and Forensics, IWBF 2015 , Glovik, Norway*. USA: IEEE Computer Society, March 2015, pp. 1–6.