Plastic Surgery: An Obstacle for Deep Face Recognition?

Christian Rathgeb*, Didem Dogan*, Fabian Stockhardt*, Maria De Marsico†, and Christoph Busch*
*da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
†Dipartimento di Informatica, Sapienza Università di Roma, Italy
{christian.rathgeb,christoph.busch}@h-da.de
demarsico@di.uniroma1.it

Abstract

The impacts of plastic surgery on face recognition systems have been investigated in the past decade by many researchers. Diverse well-known face recognition approaches, e.g. based on PCA or LBP, have been benchmarked mostly on the web-collected IIITD plastic surgery face database. Generally, significant performance drops were reported when comparing facial images taken before and after plastic surgeries. On the one side, some researchers reported problems with said plastic surgery database, i.e. the presence of low quality images. On the other side, the applied methods no longer reflect the state-of-the-art in face recognition. This calls for evaluating the impact of plastic surgery on state-of-the-art deep face recognition systems anew considering high quality imagery of most relevant plastic surgeries.

This work introduces the new Hochschule Darmstadt (HDA) plastic surgery database of facial images taken before and after surgery. This database vastly complies with the quality requirements defined by the International Civil Aviation Organization (ICAO) for electronic travel documents and comprises face images of the five most frequently applied facial plastic surgeries. The HDA plastic surgery database, the IIITD plastic surgery database, and a non-surgery database, i.e. ICAO-compliant subsets of the FRGCv2 and FERET datasets, are used for comparative verification and identification evaluations which are conducted using the commercial Cognitec FaceVACS system and the open-source ArcFace system. The obtained results suggest that the impact of plastic surgery on deep face recognition systems is less significant than that observed for previously benchmarked methods.

1. Introduction

Facial recognition [32, 21, 17] has been a highly active research field in the past decades in which the use of deep learning has achieved a major breakthrough [27, 24]. The high generalization capabilities of deep neural networks paved the way for the use of facial recognition technologies in various application scenarios ranging from video-based surveillance to access control for mobile devices and automated border control (ABC). More specifically, significant performance enhancements have been reported on unconstrained face databases, *e.g.* the well-known Labeled Faces in the Wild (LFW) dataset [20]. In addition to the improved robustness reported for variations in pose, facial expression or image quality, the feasibility of deep face recognition has been demonstrated for other challenges, such as facial ageing [3] or recognition of children and newborns [8, 2]. More recently, the NIST reported auspicious performance rates of state-of-the-art face recognition systems for large-scale identification in the 2019 Face Recognition Vendor Test (FRVT) [12].

Figure 1. Examples of face images of Hollywood stars before (top row) and after (bottom row) the application of plastic surgery (images collected from YouTube).

The reverse of the medal in face recognition is also represented by the use of any kind of beautification, that can be either "volatile" or permanent [26]. In the first category, it is possible to mention makeup or retouching, the latter being nowadays a very common practice in gossip magazines or advertisement. As a consequence of digital beautification (*e.g*. by tools like Adobe Photoshop or GIMP) automatic face tagging might be hindered [4]. The second category encompasses all kinds of facial plastic surgery, see Fig. 1, which can concretely and significantly change the appearance of a person's face for either cosmetic purposes, to improve the facial appearance, or to correct and/or reconstruct the face from disfiguration due to illness or injury. While there are techniques that can support recognition in the presence of changes in pose, illumination and expression, or the capturing process can be even repeated in controlled and attended settings, facial plastic surgery makes the changes permanent and not reversible (unless already knowing the correct identity of the person). The decreased costs of advanced surgeries make related operations more affordable and thus widespread, therefore representing a significant problem for security controls.

In general, it is possible to distinguish local surgery, which aims at correcting single well localized defects and/or anomalies, from global surgery, which can affect the whole appearance of an individual face. In the former case, the possibility to recognize a person can depend on the possible mix of modifications and on their localization and extension. As a matter of fact, different face regions can affect recognition at a different extent, as the presented experiments will assess. In the case of global modifications, except for a surgery like skin peeling that only modifies the face texture, and especially if intentionally used to deceive controls, *e.g*. by criminals, it is extremely hard to devise possible countermeasures.

In summary, plastic surgery has been identified as another obstacle for robust and reliable face recognition since these types of operations might seriously change the appearance of individuals' faces [28, 26]. Their relevance is underlined by the huge amount of operations conducted worldwide, as reported by the International Society of Aesthetic Plastic Surgery (ISAPS) [14]. Over the past five years, out of more than ten million annually reported plastic surgeries approximately 40% have been performed on head and face [14]. The following five types of plastic surgery have been identified as the most popular ones:

1. *Eyebrow correction* is a surgical procedure to reposition the eyebrow, usually to create a more feminine or youthful appearance;
2. *Eyelid correction* is the plastic surgery operation for correcting defects, deformities, and disfigurations of the eyelids;
3. *Facelift* is a type of cosmetic surgery procedure used to give a more youthful facial appearance that usually involves the removal of excess facial skin, with or without the tightening of underlying tissues, and the redraping of the skin on the patient's face and neck;
4. *Nose correction* is a plastic surgery procedure for correcting and reconstructing the nose; it can be distinguished into reconstructive surgery which restores the form and functions of the nose and cosmetic surgery which alters/ improves the appearance of the nose;
5. *Facial bones correction* is a kind of surgery to aesthetically improve facial bones, *e.g*. jaw or cheek bones.

For more details on the mentioned types of plastic surgeries the reader is referred to [14]. Reportedly, eyelid corrections and facelifts together cover two thirds of all plastic surgeries done on faces followed by nose corrections which make up almost one quarter. Finally, eyebrow corrections and facial bones corrections both constitute only around five percent of all facial plastic surgeries [14]. It is important to note that these relative occurrences of types of facial plastic surgeries have to be taken into account when measuring their impact on face recognition systems. Further, it needs to be considered that the effects of the different surgeries can vary according to their extension. For instance, it is intuitive to expect that eyebrow and eyelids corrections can have a light or moderate impact on face geometry, while nose and facial bones corrections lead to more significant modifications. Regarding facelift, its impact can be expected to depend on the previous amount of excess skin. The experiments will assess the validity of such expectations.

Many researchers have investigated or tried to mitigate the influence of plastic surgery on different face recognition methods, as shortly summarized in Sect. 2. In order to complement and extend the conducted studies, this paper presents a new plastic surgery database collected from various web sources. This database consists of more than 600 vastly ICAO-compliant images which have been captured before and after the aforementioned most popular types of facial plastic surgeries. This means that both before and after images are good quality frontal face images with neutral expression. Similarly, ICAO-compliant subsets of the FRGCv2 and of the FERET database for a total of 900 subjects is used to benchmark face recognition performance when no plastic surgery has been applied. In contrast to the experimental setups of related studies, these datasets of high quality face images, which are referred to as *HDA plastic surgery database* and *non-surgery database*, respectively, allow for a clear isolation of the effects of plastic surgery on face recognition and facilitate comparability of results. Additionally, it is important to note that the availability of good quality facial data is realistic in present-day deployments of face recognition-based user authentication, *e.g*. ABC or smart-phone unlocking. For the sake of completeness and comparability, the widely used IIITD plastic surgery database is also used in experiments. State-of-the-art face recognition systems are employed for performance evaluations, *i.e*. the frequently deployed commercial Cognitec FaceVACS system [6] and the widely used open-source ArcFace algorithm [9]. They are evaluated on

Table 1. Most relevant works on the impact of plastic surgery on face recognition.

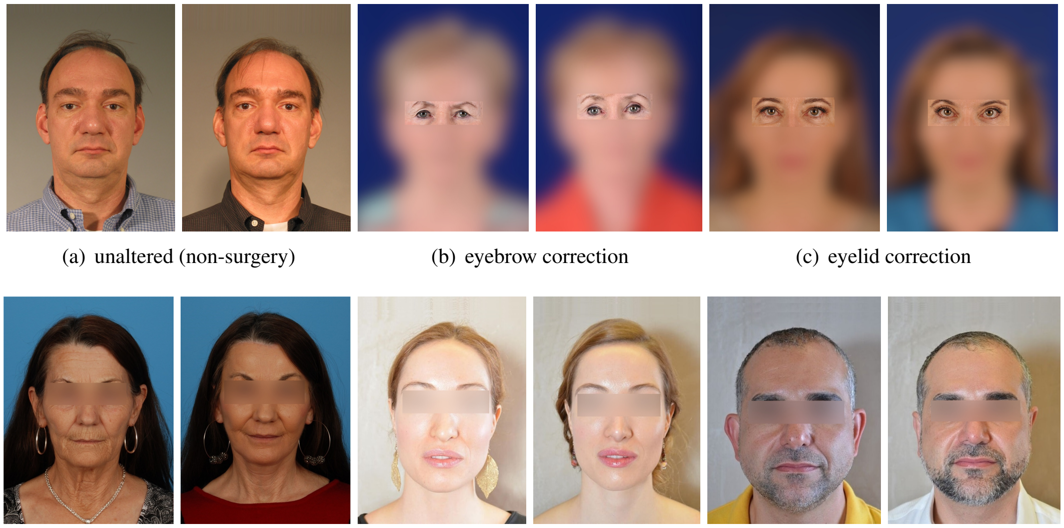
Year	Authors	Database	Method(s)	Performance rates		Remarks
				Unaltered	Surgically altered	
2009	Singh *et al.* [29]	Plastic surgery database (504 subjects)	PCA, FDA, GF, LFA, LBP and GNN	–	34.1% GMR at 0.1% FMR (GNN)	Improvements reported for algorithm fusion
2010	Singh *et al.* [28]	Plastic surgery database (900 subjects)	PCA, FDA, GF, LFA, LBP and GNN	84.1% R-1 (GNN)	54.2% R-1 (GNN)	–
2011	De Marsico *et al.* [7]	[28]	Image sub-region matching with localized correlation index	–	70% R-1, 20% EER	–
2012	Aggarwal *et al.* [1]	[28]	Part-wise fusion of PCA-features with sparse representation	–	77.9% R-1	Training on MBGC [25]
2012	Jillela and Ross [16]	[28]	Score-level fusion of COTS systems, LBP and SIFT on ocular region	–	87.4% R-1	Report of low-quality images in [28]
2012	Kose *et al.* [19]	Simulated nose alterations on FRGCv1 (275 subjects)	Image block-based PCA, LDA and CLBP	81.27% R-1, 72.39% GMR at 0.1% FMR for 2D data (CLBP), 83.32% R-1, 72.49% GMR at 0.1% FMR for 3D data (LDA)	76.33% R-1, 66.26% GMR at 0.1% FMR for 2D data (CLBP), 75.12% R-1, 60.35% GMR at 0.1% FMR for 3D data (LDA)	Evaluations on 2D and 3D face data
2013	Bhatt *et al.* [5]	[28]	Multiobjective evolutionary granular algorithm	89.87% R-1	87.32% R-1	Unaltered taken from combined heterogeneous database
2013	Sun *et al.* [30]	[28]	SSIM index weighted multi-patch LBP fusion scheme	–	77.55% R-1	–
2014	Feng and Prabhakaran [10]	[28]	Gabor and texture features for facial parts recognition	96.8% R-1	85.35% R-1	–
2015	Kohli *et al.* [18]	[28]	Recognition: region-based compact binary face descriptors, 2×COTS Detection: compact binary face descriptors and multiple projective dictionary learning	0.72% EER (COTS)	3.63% EER (detection + COTS)	Integration of detection scheme to verification system
2015	Moeini *et al.* [22]	[28]	Fusion of texture features and 3D face reconstruction methods	–	95.3% R-1, 10.8% EER	–
2018	Suri *et al.* [31]	[28]	DenseNet with color, shape and texture space classifier	–	91.75% R-1 ∼90% GMR at 0.1% FMR	Database division in training and test set

both plastic surgery databases as well as individually for each type of surgery. In summary, this work provides an up-to-date investigation on the impact of plastic surgery on deep face recognition in verification and identification mode including a constrained authentication scenario. Identification experiments reveal very high accuracy for the commercial as well as the open-source system achieving a Rank-1 Recognition Rate (R-1) of 99.16% and 99.51% on the IIITD plastic surgery database, respectively, to be compared with the results reported in Table 1. Significantly higher performance rates are obtained for the newly collected database comprising high quality imagery. Similar observations hold for the verification experiments considering False Non-Match Rates (FNMRs) at a relevant False Match Rate of 0.1%. Additionally, causes for decreases in recognition accuracy are discussed.

This article is organized as follows. Sect. 2 summarizes related work. Sect. 3 describes the used databases. Sect. 4 presents and discusses experimental results. Finally, Sect. 5 draws conclusions.

2. Related Works

Table 1 provides an overview of the most relevant works investigating the impact of plastic surgery on face recognition along with used databases, applied methods and obtained results. Pioneering work in the field was done by Singh *et al.* [29] who also provided the first plastic surgery database for face recognition research that was extended in [28]. This database, which was collected from the web, was used by different researchers who have since evaluated various face recognition methods on it. Usually, approaches were compared by reporting the obtained R-1 recognition rates in a closed-set scenario. It should be noted that in a real-world scenario fixed decision thresholds are more likely to be applied, and that even a first ranked similarity score can be below the system acceptance threshold. Nonetheless, over the past years a significant enhancement of face recognition performance on the mentioned database can be observed. While early works reported unpractical performance rates for real-world applications, *e.g.* neural network architecture based on 2-D Log Polar Gabor Transform (GNN) of [29], more recently proposed approaches achieved more exploitable results, *e.g.* deep-learning based method of [31]. Furthermore, it can be observed that many approaches, which were designed to be resilient to plastic surgery, process face images in a patch-wise manner, also referred to as "part-wise", "image block-wise" or "sub-region-wise", *e.g.* [7, 1, 19, 30]. The rationale of these schemes was to divide the extraction process into the different face patches, in order to better exploit the comparison results obtained over those ones not affected by plastic surgery. Additionally, multi-algorithm fusion strategies have been proposed to achieve a higher robustness to facial alterations caused by plastic surgery, *e.g.* [16, 22]. The advantage of a multi-algorithm fusion is that it increases the amount of extracted facial information, if fused feature extractors complement each other. Consequently, the resulting multi-biometric face recognition system is expected to achieve generally enhanced robustness. Other related works introduced plastic surgery detection schemes, *e.g.* [18]. For a comprehensive overview of published works in the field the interested reader is referred to [23, 26].



(a) unaltered (non-surgery) (b) eyebrow correction (c) eyelid correction

(d) facelift (e) nose correction (f) facial bones correction

Figure 2. Examples of facial portrait reference and probe images of (a) the non-surgery database and (b)-(f) the collected HDA plastic surgery database (images of the plastic surgery database have been anonymized to protect the individuals' privacy).

Table 2. Overview of the used databases.

Database	**Surgery**	**Image pairs**	
		Genuine	**Impostor**
IIITD plastic surgery	Dermabrasion	32	–
	Eyebrow correction	60	–
	Ear correction	74	–
	Eyelid correction	105	–
	Nose correction	192	–
	Skin peeling	73	–
	Facelift	308	–
	Others (mentoplasty, etc.)	56	–
HDA plastic surgery	Eyebrow correction	128	–
	Eyelid correction	131	–
	Facelift	98	–
	Nose correction	174	–
	Facial bones correction	107	–
Non-surgery	Non	900	404,550

3. Databases

The following subsections describe the databases used for the experimental evaluation, *i.e.* the IIITD plastic surgery face database (Sect. 3.1), the HDA plastic surgery database (Sect. 3.2), and the non-surgery database (Sect. 3.3). An overview of these databases is given in Table 2.

3.1. IIITD plastic surgery face database

The IIITD plastic surgery face database [28] is publicly available[1] through a collection of web links (due to copyright reasons). However, as this database was collected several years ago the majority of original images is not available anymore, possibly because they have been deleted or moved to different locations. Nevertheless, upon request face images cropped to 100×100 pixels have been kindly provided for this work by the Image Analysis and Biometrics Lab (IAB) of IIT Jodhpur[2]. For example images of the mentioned database the reader is referred to [7] and [16]. It is worth underlying that the methods presented in literature are often negatively affected by the presence of low quality images in this database. Jillela and Ross [16] reported duplicates and a varying image quality in this database, in particular with respect to inter-ocular distances and pose. While those variations certainly occur in real-world scenarios they hamper the isolation of the actual influence of plastic surgery. Eventually, it should be noted that the database of [28] is rather unbalanced with respect to the amount of images per type of plastic surgery (see Table 2) and contains images of surgeries which are less relevant for face recognition systems, *e.g.* otoplasty (ear surgery).

3.2. HDA plastic surgery face database

The above mentioned limitations motivated the collection of a new plastic surgery face database. Example images of both, the newly collected plastic surgery database and the additionally chosen non-surgery database, are shown in Fig. 2. The HDA plastic surgery database was collected from multiple web sources. Whenever possible, at least 100 image pairs were collected for each of the five most popular types of plastic surgeries that have been introduced in Sect. 1. Example images for the different plastic surgery types in this database are shown in Fig. 2(b)-Fig. 2(f). The HDA plastic surgery database consists of 540 female (85%) and 98 male (15%) subjects resulting in a total number of 638 subjects. Note that this gender imbalance is somewhat unavoidable as most plastic surgeries are applied to female

[1] http://www.iab-rubric.org/resources.html

[2] http://www.iab-rubric.org

patients [14]. It is also important to note that in many cases more than one type of surgery has been performed. Hence, it is difficult to find image pairs of subjects who underwent a single particular surgery.

Images were collected according to the prerequisites of the ICAO [13] for the production of passport portrait photos. The ICAO suggests face image data to be stored in accordance with the specifications established by the International Standard ISO/IEC 19794-5 [15]. These specifications ensure that faces are represented in sufficient quality. In particular, ICAO requires face images to be captured with frontal pose, neutral expression and sufficient inter-ocular distance.

3.3. Non-surgery face database

In order to compare the effects of plastic surgery on face recognition systems in the presence and absence of plastic surgeries additional (subsets of) publicly available face databases were used in the scientific literature, *e.g.* AR-Face in [28]. This modus operandi became firmly established because a direct comparison would require more than one image before the plastic surgeries, while only image pairs showing faces before and after plastic surgery are available on the web and, hence, also in the widely-used database of [28]. If an additional face image database is used it should exhibit properties similar to the used plastic surgery face image data. Otherwise comparisons might not be fair and, thus, obtained results might be misleading.

The non-surgery database used in this work contains a manually chosen subset of ICAO-compliant image pairs of the FRGCv2 and the FERET face database. A total number of 450 subjects were chosen from each of the two databases. An example image pair of the non-surgery database is shown in Fig. 2(a). The non-surgery database consists of 359 female (39.9%) and 541 male (60.1%) subjects. This database was exclusively used to obtain impostor comparisons which were then combined with genuine comparison of all used databases. Thereby, a fixed decision threshold which was estimated from this single impostor score distribution can be employed in verification experiments.

4. Evaluation

The next subsection describes the experimental setup (Sect. 4.1). Subsequently, the obtained results are reported and discussed in detail (Sect. 4.2).

4.1. Experimental setup

Two different face recognition systems were evaluated, *i.e.* the Cognitec FaceVACS [6] and the ArcFace algorithm [9]. While the first system is a frequently deployed commercial product, the latter represents an open-source algorithm which is widely used in the research community. Both algorithms rely on the use of deep learning[3] [9]. Given a pair of face images, the Cognitec FaceVACS returns a similarity score in the range $[0, 1]$ (*i.e.* high values indicate high similarity). Distance scores produced by the ArcFace algorithm were normalized to the same range.

Table 3. Identification rates for both face recognition systems on the used databases considering all types of plastic surgeries.

System	Database	R-1 (%)
Cognitec FaceVACS	Non-surgery	99.89
	HDA plastic surgery	100.0
	IIITD plastic surgery	99.16
ArcFace	Non-surgery	99.0
	HDA plastic surgery	99.84
	IIITD plastic surgery	99.51

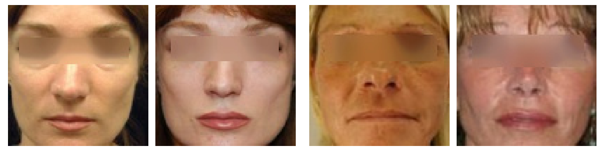


Figure 3. Example image pairs before and after surgery which failed to be identified on the IIITD plastic surgery face database (images have been anonymized to protect the individuals' privacy).

Biometric (closed-set) identification performance was estimated in terms of R-1 in order to achieve comparability with published works. Subsequently, biometric verification performance was evaluated in terms of False Non-Match Rate (FNMR) and False Match Rate (FMR). More precisely, the FNMR at a FMR of 0.1%, referred to as $\text{FNMR}_{0.1}$, is reported which represents the operation point recommended in the guidelines of European Agency for the Management of Operational Cooperation at the External Borders (FRONTEX) [11]. As the overlaps of genuine and impostor score distributions have been found to be minimal (see Sect. 4.2) Detection Error Trade-Off (DET) curves are not presented. Rather than reporting Equal Error Rates (EERs) probability density distributions are shown and as a measure of decidability $d' = |\mu_g - \mu_i| / \sqrt{\frac{1}{2}(\sigma_g^2 + \sigma_i^2)}$ is reported, where μ_g and μ_i represent the means of the genuine and the impostor score distributions and σ_g and σ_i their standard deviations, respectively.

4.2. Performance evaluation

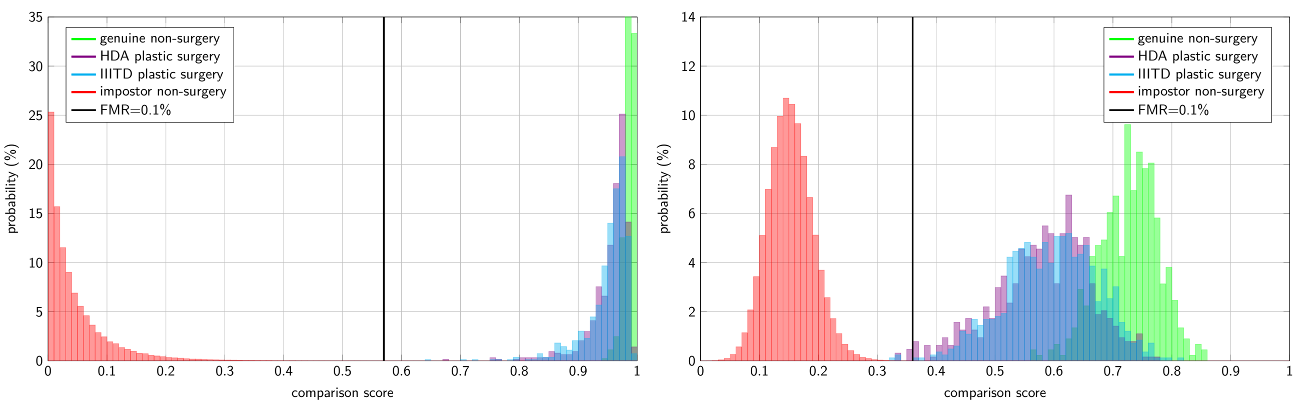
Table 3 lists the R-1 identification performance of both face recognition systems obtained on the used databases. It can be observed that the used state-of-the-art face recognition systems significantly outperform published approaches, *c.f.* Table 1. Fig. 3 shows examples of rare cases of the IIITD plastic surgery face database which did not result in correct identification.

For verification performance evaluations all genuine comparisons were calculated on each database for both face recognition systems. To obtain the fixed thresholds for the

[3] Information provided by Cognitec Systems GmbH.

Table 4. Verification performance results for both face recognition systems on the considered databases.

System	Database	Distribution	Surgery	Mean (μ)	Std. Dev. (σ)	d'	FNMR$_{0.1}$ (%)
Cognitec FaceVACS	Non-surgery face database	Impostor	Non	0.046	0.054	–	–
		Genuine	Non	0.985	0.007	24.17	0.0
	HDA plastic surgery face database	Genuine	Eyebrow correction	0.946	0.044	18.14	0.0
			Eyelid correction	0.962	0.029	20.96	0.0
			Facelift	0.962	0.025	21.46	0.0
			Nose correction	0.951	0.077	13.53	0.0
			Facial bones correction	0.933	0.095	11.39	0.0
			All	0.951	0.062	15.39	0.0
	IIITD plastic surgery face database	Genuine	Dermabrasion	0.961	0.026	21.34	0.0
			Eyebrow correction	0.945	0.038	18.97	0.0
			Eyelid correction	0.962	0.021	22.12	0.0
			Nose correction	0.940	0.048	17.34	0.0
			Skin Peeling	0.964	0.024	21.78	0.0
			Facelift	0.947	0.036	19.42	0.0
			All	0.950	0.037	19.25	0.0
ArcFace	Non-surgery face database	Impostor	Non	0.195	0.036	–	–
		Genuine	Non	0.726	0.050	12.16	0.0
	HDA plastic surgery face database	Genuine	Eyebrow correction	0.574	0.075	6.43	1.56
			Eyelid correction	0.604	0.071	7.22	0.0
			Facelift	0.606	0.077	6.79	0.0
			Nose correction	0.569	0.074	6.36	0.0
			Facial bones correction	0.933	0.095	4.36	0.0
			All	0.580	0.084	5.95	0.31
	IIITD plastic surgery face database	Genuine	Dermabrasion	0.621	0.048	9.99	0.0
			Eyebrow correction	0.573	0.072	6.61	0.0
			Eyelid correction	0.614	0.064	7.97	0.0
			Nose correction	0.558	0.077	6.03	0.0
			Skin Peeling	0.631	0.073	7.49	0.0
			Facelift	0.599	0.071	7.16	0.0
			All	0.595	0.078	6.57	0.36



(a) Cognitec FaceVACS (b) ArcFace

Figure 4. Score distributions for the used databases for all types of plastic surgeries.

FNMR$_{0.1}$ values for each face recognition system, impostor comparisons were obtained from the non-surgery database. An overview of the number of performed genuine and impostor comparisons is given in Table 2. Obtained verification performance rates are summarized in Table 4. For the IIITD plastic surgery face database only surgeries for which at least ten comparisons were available have been included. In order to estimate the FNMR$_{0.1}$, fixed decision thresholds of 0.57 and 0.36 were estimated for the commercial system and the open-source system, respectively. Comparison score distributions for both face recognition systems considering all types of plastic surgeries are shown in Fig. 4.

It can be observed that the commercial face recognition system outperforms the open source one, *i.e.* significantly higher d' values are obtained while a zero FNMR$_{0.1}$ is achieved across most types of surgeries. Furthermore, considering the lowest d' values of both systems, it can be concluded that facial bones corrections have the most severe impact, followed by nose corrections and eyebrow corrections. The least impact is observed for facelifts and eyelid corrections.

With respect to facial bones corrections it was found that these surgeries are frequently performed as part of feminisation operations. The aim of these types of operations is to create a more feminine facial appearance. Fig. 5 illustrates two examples of facial bones corrections as part of feminisation operations. While the left image pair of Fig. 5 yields a good comparison score, the right one reveals a score

Figure 5. Examples of facial bones corrections as part of feminisation operations before and after surgery (images have been anonymized to protect the individuals' privacy).

(a) eyelid (b) eyebrow

(c) nose correction (d) facelift

Figure 6. Examples of problematic cases for different types of plastic surgeries (images have been anonymized to protect the individuals' privacy).

closer to the impostor distribution for both face recognition systems. Additionally, the sex prediction algorithm of the commercial system changes its estimation from male to female in the latter image.

Focusing on the other types of surgeries, their effect on face recognition systems obviously depends on how severely they change a subject's facial appearance. Fig. 6 schows examples for different types of plastic surgeries which were found to reveal inferior comparison scores for both face recognition systems. Besides facial bone corrections, eyebrow corrections may significantly alter the periocular region. Hence, they can have a great impact on face recognition, which also holds for nose corrections. In contrast, facelifts result in a facial smoothing across the entire face. That is, of all considered types of plastic surgeries facelifts have shown the least impact on both face recognition systems.

At this point, it is important to recall that the two types of plastic surgeries which tend to have the greatest impact on recognition accuracy, *i.e.* facial bones correction and eyebrow correction, are the ones which are rarely performed. More precisely, together these types of plastic surgeries make up less than ten percent of all surgeries performed on the face.

5. Conclusions

The influence of plastic surgery on face recognition has been investigated by many research groups in the past decade. Based on results obtained from a single database [28] it has been concluded that plastic surgeries significantly decrease the recognition performance of face recognition systems. Since researchers have pointed out different problems with the mentioned database [16] a new plastic surgery database has been collected in this work. This database contains images captured before and after the five most popular plastic surgeries performed on the face. Images of this database are vastly ICAO-compliant with respect to face image quality and were used together with a public non-surgery database of equal quality.

In the presented experiments, two types of face recognition systems have been evaluated, the commercial Cognitec FaceVACS system [6] and the widely used open-source ArcFace algorithm [9]. Both state-of-the-art face recognition systems which rely on deep learning have been found to be extremely robust to facial alterations induced by plastic surgery. Compared to previously published approaches significant performance gains have been achieved, *i.e.* R-1s above 99% on the IIITD plastic surgery face database. Experimental results on the newly created high-quality plastic surgery database also reveal that once distortions from low-quality face images are ruled out, as it happens in ICAO-compliant images, most plastic surgeries do not cause significant errors by deep face recognition systems. For both types of systems facial bones corrections have the most severe impact on face recognition accuracy, followed by eyebrow corrections. Further, the results suggest that there is some performance gap between the commercial system and the open-source system. However, at a practically relevant threshold which yields a FMR of 0.1% both systems almost maintain a perfect verification performance across all plastic surgeries. On the contrary, further investigations and data collections are required to analyse whether face alterations entailed by plastic surgery can further degrade a face recognition system's performance when they co-occur with other factors, *e.g.* variations in pose or illumination.

Acknowledgments

This research work has been partly funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] G. Aggarwal, S. Biswas, P. J. Flynn, and K. W. Bowyer. A sparse representation approach to face matching across plastic surgery. In *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, pages 113–119, 2012.

[2] L. Best-Rowden, Y. Hoole, and A. Jain. Automatic face recognition of newborns, infants, and toddlers: A longitudinal evaluation. In *2016 Int'l Conf. of the Biometrics Special Interest Group (BIOSIG)*, pages 1–8, 2016.

[3] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(1):148–162, 2018.

[4] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Trans. on Information Forensics and Security*, 11(9):1903–1913, 2016.

[5] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa. Recognizing surgically altered face images using multiobjective evolutionary algorithm. *IEEE Trans. on Information Forensics and Security*, 8(1):89–100, 2013.

[6] Cognitec: the face recognition company. FaceVACS Engine 9.3. http://www.cognitec.com/technology.html, 2019.

[7] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler. Robust face recognition after plastic surgery using local region analysis. In *Proc. of the 8th Int'l Conf. on Image Analysis and Recognition - Volume Part II*, ICIAR'11, pages 191–200, 2011.

[8] D. Deb, N. Nain, and A. K. Jain. Longitudinal study of child face recognition. In *2018 Int'l Conf. on Biometrics (ICB)*, pages 225–232, 2018.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699. IEEE, June 2019.

[10] R. Feng and B. Prabhakaran. A novel method for post-surgery face recognition using sum of facial parts recognition. In *IEEE Winter Conf. on Applications of Computer Vision*, pages 1082–1089, 2014.

[11] Frontex. Best Practice Technical Guidelines for Automated Border Control (ABC) Systems, 2016.

[12] P. Grother, M. Ngan, and K. Hanaoka. Face Recognition Vendor Test (FRVT) Part 2: Identification (NISTIR 8271). Technical report, National Institue of Standards and Technology (NIST), 2019.

[13] International Civil Aviation Organization. ICAO doc 9303, machine readable travel documents – part 9: Deployment of biometric identification and electronic storage of data in MRTDs (7th edition). Technical report, ICAO, 2015.

[14] International Society of Aesthetic Plastic Surgery. ISAPS Global Statistics. https://www.isaps.org/medical-professionals/isaps-global-statistics/, December 2019.

[15] ISO/IEC JTC1 SC37 Biometrics. *Information technology – Biometric data interchange formats – Part 5: Face image data*, 2005.

[16] R. Jillela and A. Ross. Mitigating effects of plastic surgery: Fusing face and ocular biometrics. In *2012 IEEE Fifth Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 402–411, 2012.

[17] M. Kawulok, M. E. Celebi, and B. Smolka. *Advances in Face Detection and Facial Image Analysis*. Springer, 1st edition, 2016.

[18] N. Kohli, D. Yadav, and A. Noore. Multiple projective dictionary learning to detect plastic surgery for face verification. *IEEE Access*, 3:2572–2580, 2015.

[19] N. Kose, N. Erdogmus, and J. Dugelay. Block based face recognition approach robust to nose alterations. In *2012 IEEE Fifth Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 121–126, 2012.

[20] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, , and G. Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.

[21] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition*. Springer London, 2011.

[22] A. Moeini, K. Faez, and H. Moeini. Face recognition across makeup and plastic surgery from real-world images. *Journal of Electronic Imaging*, 24(5):1 – 12 – 12, 2015.

[23] M. Nappi, S. Ricciardi, and M. Tistarelli. Deceiving faces: When plastic surgery challenges face recognition. *Image and Vision Computing*, 54:71 – 82, 2016.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition, 2015.

[25] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, D. Bolme, K. W. Bowyer, B. A. Draper, G. H. Givens, Y. M. Lui, H. Sahibzada, J. A. Scallan, and S. Weimer. Overview of the multiple biometrics grand challenge, 2009.

[26] C. Rathgeb, A. Dantcheva, and C. Busch. Impact and detection of facial beautification in face recognition: An overview. *IEEE Access*, 7:152667–152678, October 2019.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering, jun 2015.

[28] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyezdan. Plastic surgery: A new dimension to face recognition. *IEEE Trans. on Information Forensics and Security*, 5(3):441–448, 2010.

[29] R. Singh, M. Vatsa, and A. Noore. Effect of plastic surgery on face recognition: A preliminary study. In *2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pages 72–77, 2009.

[30] Y. Sun, M. Tistarelli, and D. Maltoni. Structural similarity based image quality map for face recognition across plastic surgery. In *2013 IEEE Sixth Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2013.

[31] S. Suri, A. Sankaran, M. Vatsa, and R. Singh. On matching faces with alterations due to plastic surgery and disguise. In *2018 IEEE Ninth Int'l Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8, 2018.

[32] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition. *ACM Computing Surveys*, 35(4):399–458, dec 2003.