Face Recognition Systems under Morphing Attacks: A Survey

ULRICH SCHERHAG¹, CHRISTIAN RATHGEB¹², JOHANNES MERKLE², RALPH BREITHAUPT³, CHRISTOPH BUSCH¹

¹da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {ulrich.scherhag, christian.rathgeb, christoph.busch}@h-da.de
²secunet Security Networks AG, Essen, Germany, johannes.merkle@secunet.com
³Federal Office of Information Security (BSI), Bonn, Germany, ralph.breithaupt@bsi.bund.de

Corresponding author: Ulrich Scherhag (e-mail: ulrich.scherhag@h-da.de).

ABSTRACT Recently, researchers found that the intended generalisability of (deep) face recognition systems increases their vulnerability against attacks. In particular, attacks based on morphed face images pose a severe security risk to face recognition systems. In the last few years, the topic of (face) image morphing and automated morphing attack detection has sparked the interest of several research laboratories working in the field of biometrics and many different approaches have been published. In this work, a conceptual categorisation and metrics for an evaluation of such methods is presented, followed by a comprehensive survey of relevant publications. Additionally, technical considerations and trade-offs of the surveyed methods are discussed along with open issues and challenges in the field.

INDEX TERMS Biometrics, face morphing attack, face recognition, image morphing, morphing attack detection.

I. INTRODUCTION
Automated face recognition [1], [2] represents a longstanding field of research in which a major break-through has been achieved by the introduction of deep neural networks [3], [4]. Due to the high generalization capabilities of deep neural networks specifically and recognition systems in general, the performance of operational face recognition systems in unconstrained environments, e.g., regarding illumination, poses, image quality or cameras, improved significantly. Resulting performance improvements paved the way for deployments of face recognition technologies in diverse application scenarios, ranging from video-based surveillance and mobile device access control to Automated Border Control (ABC). However, recently researchers found that the generalizability of (deep) face recognition systems increases their vulnerability against attacks, e.g., spoofing attacks (also referred to as presentation attacks) [5]. An additional attack vector enabled by the high generalization capabilities is a specific attack against face recognition systems based on morphed face images, as introduced by Ferrara et al. [6].

A. FACE MORPHING ATTACK
Image morphing has been an active area of image processing research since the 80s [7], [8] with a wide variety of application scenarios, most notably in the film industry. Morphing techniques can be used to create artificial biometric samples, which resemble the biometric information of two (or more) individuals in image and feature domain. An example of a morphed face image as the result of two non-morphed, i.e., bona fide [9], face images, is depicted in Fig. 1. The created morphed face image will be successfully verified against probe samples of both contributing subjects by state-of-the-art face recognition systems. This means, if a morphed face image is stored as reference in the database of a face recognition system, both contributing subjects can be successfully verified against this manipulated reference. Thus, morphed
In many countries, the face image used for the ePassport issuance process is provided by the applicant in either analog or digital form. In a face morphing attack scenario, a wanted criminal could morph his face image with one of a lookalike accomplice. If the accomplice applies for an ePassport with the morphed face image, he will receive a valid ePassport equipped with the morphed face image. It is important to note, that morphed face images can be realistic enough to fool human examiners [10], [11]. Both, the criminal and the accomplice could then be successfully verified against the morphed image stored on the ePassport, as visualized in Fig. 2. This means, the criminal can use the ePassport issued to the accomplice to pass ABC gates (or even human inspections at border crossings). The risk posed by this attack, to the accomplice to pass ABC gates, or even human inspections at border crossings), is amplified by the fact that realistic morphed face images can be generated by non-experts employing easy-to-use face morphing software which makes the prerequisites of the International Civil Aviation Organization (ICAO) [13] for the issuance of passport portrait photos have to be met. These specifications ensure that all faces are represented equally with respect to resolution, exposure, etc. Semi-profile recordings can indeed be partially corrected, but then there is usually information missing of the far side of the face. Furthermore, the quality of the source images has a direct influence on the result. The quality of the morph cannot be expected to be higher than that of the source images. Distortions and scaling usually negatively affect quality during the process chain. The quality of morphed face images is further discussed in Sect. III.

In general, the morphing process of face images can be divided into three steps. First, a correspondence between the contributing samples is determined. In a second step, called warping, both images are distorted, such that the corresponding elements of both samples are geometrically aligned. Finally, the colour values of the warped images are merged, referred to as blending, in order to create the morphed face image. Said processing steps are described in detail in the following subsections, along with post-processing, studies on human perception of morphed face images and a summary of available research resources.

1FaceMorpher, Luxand: http://www.facemorpher.com/
2WinMorph, DebugMode: http://www.debugmode.com/winmorph/
3FantaMorph, Abrasoft: http://www.fantamorph.com/
The most common way of determining correspondences is to manually define the coordinates of prominent characteristics, e.g., eyes, eyebrows, tip of the nose, etc., as for instance done in the morphing process of [6] and [14]. The manual annotation of images is very accurate (if done properly), but time consuming. More convenient is the automated detection of landmarks. The established approach for landmark detection is to detect each point separately, e.g., utilizing geometric features [15]. A more sophisticated solution is to fit a predefined model, e.g., active shape models [16] or elastic bunch graph models [17], [18] to the face image, whereas the fitting of the model is the key issue. Zanella and Fuentes propose an untrained generic model, which is fit to the contours of a binary image using evolutionary strategies [19]. Saragih et al. [20] propose a principled optimization strategy where a non-parametric representation of the landmark distributions is maximized within a hierarchy of smoothed estimates. Further algorithms train multiple regression trees for landmark detection [21], [22], of which the method of Kazemi and Sullivan [22] was further implemented in the widely used dlib landmark detector [23]. For detailed information and benchmarks of different automated landmark detection approaches the reader is referred to [24].

B. WARPING

If the landmarks are determined, the image should be distorted in a manner, that corresponding landmarks are aligned. A straightforward method for morphing is scattered data interpolation [25]. The landmarks, also called control points, are moved to a new position, the new position of all intervening pixels is interpolated based on the nearby control points. More advanced morphing techniques take the correlation between the landmarks into account. For example, Sederberg et al. [26] propose a grid or mesh-based warping technique called Free-Form Deformation (FFD), which was extended by Lee et al. to multi-level FFD [27]. The whole image is considered as a grid, which is deformed by the flow of the landmarks. Another approach is field morphing introduced by Beie and Neely [28], where grid lines are controlling the metamorphosis of the image in the transformation. In particular, for manual morphing this approach has advantages, as the user can position lines instead of points. For automatic morphing the lines can be derived from detected landmarks. In the work of Schäfer et al. [29] the moving least squares are minimized in order to estimate the optimal affine transformation. This approach can be employed to optimize different warping methods based on landmarks or lines. Choi et al. proposes a morphing process by simulating the image as a mass spring system [30]. Thus, each translated landmark influences nearby pixels and landmarks.

Most state-of-the-art morphing algorithms, e.g., as used for the morph-creation in [31]–[38], do not consider the image as a grid, but apply a Delaunay triangulation on the landmarks in order to determine non overlapping triangles, as depicted in Fig. 3. Delaunay triangulations maximize the minimum angle of each triangle in the triangulation and can be calculated efficiently. Subsequently, the triangles of both contributing images are distorted, rotated and shifted until an alignment is achieved.

The first step in traditional approaches for creating a morph between a pair of face images $I_0$ and $I_1$ is to define a map $\phi$ from $I_0$ to $I_1$. The contribution of each subject to the warping process is defined by an $\alpha_w$-value, whereas an $\alpha_w = 0$ would be the landmark-position of the first subject, $\alpha_w = 1$ the landmark-position of the second subject and an $\alpha_w$ between 0 and 1 any combination of both. The impact of different $\alpha_w$-values on the resulting face morph can be seen by analysing the first versus the last row of Fig. 4. One issue that might occur are disocclusions which refers to regions in the object space that are visible in $I_0$, but disappear in $I_1$ as described by Liao in [39]. For disocclusions in $I_0$, the map $\phi$ is typically undefined, for disocclusions in $I_1$ it is discontinuous. To obtain a more complete representation, one can introduce a second map from $I_1$ back to $I_0$. Maintaining consistency between the two maps during an optimization process becomes quite expensive [39]. One approach solving this issue is proposed by Wu et al. [40]. The images are warped forward and backward in order to obtain a complete mapping $\phi$. In addition, to obtain a more natural warping, the face images are projected into a 3D space and an energy
function is minimized to avoid ghost and blur artefacts. Seitz et al. [41] also proposes a projection into 3D-space, in order to consider perspective effects during the morphing process. Another technique for morphing in 3D-space is given by Yang et al. in [42]. In order to recover the face geometry, the 2D face image is projected on a pre-learned 3D face mask. In particular, for variances in pose and expression this approach promises a higher quality.

Further, some warping algorithms do not need previously detected landmarks. Bichsel et al. [43] propose to employ the Bayesian framework in order to determine the optimal mapping function.

C. BLENDING

After the alignment of the two contributing images, the two arranged textures are combined using blending, usually over the entire image region. The most frequent way of blending for face morph creation is linear blending, i.e. all colour values at same pixel positions are combined in the same manner. Similar to the warping process the contribution to the blending of each image can be weighted by an $\alpha_b$-value, e.g. $\alpha_b = 0.5$ for averaging. The impact of a changing $\alpha_b$-value to the morphed image can be seen in Fig. 4 on the vertical axis.

D. FURTHER APPROACHES

There are, however, some morphing algorithms, where a subdivision into the steps described above is not feasible. In [44], a morphing approach is proposed using generative morphing to combine warping and blending. The resulting morphed image is regenerated from small pieces of the source images. Korshunova et al. [45] propose to train a Convolutional Neural Network (CNN) to swap the face image of one subject with the face of a second one. A huge disadvantage of this method is, that a new network has to be trained for each

FIGURE 4: Matrix of the two variables in a morphing process (blending and warping). This morph sequence was created using dlib for landmark detection, Delaunay triangulation and linear affine transformation for warping and linear blending.
subject.

Beside the morphing of samples in image domain, it is possible to morph in feature domain, as e.g., shown in [46] for minutiae sets and in [47] for iris-codes. It would be feasible to also morph face representations in feature domain, e.g., by averaging the feature vector of a CNN [49]. In order to use the morphed feature vector in a face recognition system, a face image can be reconstructed from the feature domain, as shown in [49]. However, it is most likely, that the reconstructed morphed face image only works for the same feature space, meaning an attack against the same face recognition system, as used for creation of the morphed feature vector.

E. POST-PROCESSING

After the creation of the morphed face image, the image might be further processed and altered. In order to obscure the image manipulation, the image quality might be enhanced or reduced on purpose.

In particular, the automated creation of morphed face images can lead to morphing artefacts. Missing or misplaced landmarks might cause shadow or ghost artefacts, as they can be seen in Fig. 5 (a). This issue can be tackled by swapping the facial area of the morphed face image with an adapted outer area of one of the subjects [35], [50]. Artefacts in the hair region can be concealed by an interpolation of the hair region as proposed by Weng et al. [51]. Further, unnatural colour gradients and edges might occur, due to inappropriate interpolation methods, which can be removed by blurring or sharpening. Due to the averaging during the blending process, the histograms of the colour values might get narrow. This artefact can be avoided by an adaptation of the colour histogram, e.g. by using histogram equalization or an adaption of lumination, in order to achieve realistic histogram shapes. Examples for sharpening and histogram equalization are depicted in Fig. 5 (b) and (c).

In addition to the removal or reduction of morphing artefacts, further post-processing steps might be carried out, which can sometimes be unavoidable, i.e., printing and scanning of the image, in order to use it as a passport photo. Even with high-end photo printer in the processing pipeline, some information contained in the face image signal will always be lost in the process, masking or reducing morphing artefacts, as described in [35]. Once the image has been submitted to a passport application office, it has to be scanned again. Again, information can be lost, helping to hide or reduce erroneous artefacts.

Further, information from or trace of the morphing process can be lost when the image format is changed. By storing the image in a lossy format, high-frequency information is eliminated from the signal permanently. If the image is loaded and stored multiple times as part of the process chain, the accumulated compression error can significantly degrade the image quality.

FIGURE 5: Examples of different post-processing methods likely to be applied by an attacker to conceal the morphing process.

III. QUALITY ASSESSMENT OF FACE MORPHS

Generally speaking, automatically generated databases of morphed face images are expected to differ in quality from real world attack scenarios. Automatically generated morphs might reveal artefacts, which can be avoided when the attacker is producing only one single high quality morph between himself and his accomplice and manually optimising the resulting image. When aiming to develop a robust detection algorithm on such an automatically generated database, it is crucial to assure high quality of morphed face images. Otherwise, it is likely that a trained classifier might strongly rely on these specific artefacts.

As described by Scherhag et al. [52], it is difficult to define objective metrics for quality assessment of face morphs due to the large number of contributing factors. Basically, the output image of the algorithms can be evaluated according to the criteria summarized in the following subsections.

A. IMAGE QUALITY

Each processing step affects the quality of an image. In particular, factors such as image size, sharpness, colour saturation, aspect ratio and the overall natural appearance of the face image should be influenced as little as possible by the morphing algorithm. The minimum requirements for these factors can be found in the specifications for passport images of the ICAO [13]. Thus, for example, the minimum resolution of the facial image is set to an inter eye distance of 90 pixels. If a picture deviates from these minimum requirements, it is no longer accepted in countries that comply with ICAO recommendations to produce a passport or other machine readable travel documents (e.g., citizen cards). Furthermore, the image quality may be affected by compression of the image. In the case of lossy compression, the storage of high-frequency information is deliberately omitted in order to increase the compression rate. At high compression rates, however, this can lead to elimination of details and compression artefacts in the image. Since poor image quality usually results from lack of information, for example, too few pixels or too little high-frequency information, it is often difficult to improve the quality later.
Quality metrics for images can be used to objectively evaluate the output images based on quality measures derived from the signal. Since no reference image is available in the evaluation of the output image, the classical image quality determination methods, such as signal-to-noise ratio or mean square deviation, are not feasible. For the selection of the quality metric, the quality properties to be considered have to be determined. The metric proposed by Farias and Mitra [53] evaluates the occurrence of image artefacts, such as block artefacts, blur or noise. If the authentic appearance of a submitted passport image is to be evaluated for the human observer, then metrics are recommended that take into account the human perception, i.e., factors like sharpness [54] or perceptual quality [55] of the image. Another option is the automated assessment of the naturalness of the image using some no-reference image quality metrics, e.g., Blind / Referenceless Image Spatial Quality Evaluator (BRISQUE) [56]. Fig. 6 shows examples of BRISQUE values where low values indicated high quality and vice versa. On the left a non-morphed face image is shown, the associated BRISQUE value of 21 corresponds to a high quality. The middle image is a high quality morph without compression, the BRISQUE value is slightly worse. The image on the right shows the same morph with JPEG compression. Even if no artefacts are visible, the BRISQUE value is strongly influenced by the compression.

B. MORPHING ARTEFACTS

Morphing artefacts as illustrated in Fig. 7 (right) can appear in the image during the multi-step morph process. Within landmark-based methods artefacts are usually caused by the absence or misplacement of landmarks. As a result, the corresponding image areas are not transformed correctly so that they do not completely overlap. This creates shadow-like, semi-transparent areas, so-called ghost artefacts. Fig. 7 depicts a manual morphed face image and an automatically generated morph comprising said artefacts. On the right, one can see a morphed facial image with poorly placed landmarks. Especially, in the region of the neck, but also on the hair and ears, strong ghost artefacts can be observed. The iris proved to be particularly susceptible to artefacts because algorithms for automatic landmark determination are usually not able to provide the iris with correct landmarks. As a workaround, the located left and right eye corner could approximate the iris center half way between the two corners. Furthermore, shadow effects may occur in facial hair (e.g., beards and eyelashes), in differently pigmented areas (e.g., liver spots, tattoos), or by glasses and jewellery. Morph artefacts, which are caused by landmark-based morphing, can usually be remedied by manual post-processing in image processing programs as shown by Ferrara et al. [6]. An additional cause of artefacts may be the differences in the source images or inappropriate interpolation methods, which can lead to unnatural colour gradients and overly hard edges in the target images. Further artefacts induced by morphing may be low contrast and blur of the images, which may result from the averaging and interpolation of pixel positions and colour values. Another type of morph artefact may be generated using machine learning to create the morphed facial images. Due to the opacity of the process of the training algorithms, the errors might be difficult to narrow down or classify. Some of the potential mistakes are missing or deformed facial features, blurred areas and ghost artefacts. The emergence of such artefacts can be reduced by appropriate learning methods and a large number of training data. Due to the high agility of the relevant research area, a rapid improvement in the quality of morph images that can be achieved by the application of machine learning can also be expected.

C. PLAUSIBILITY OF FACE MORPHS

The quality of a morph can also be assessed by how plausible the image appears as a facial image. Here, on the one hand, the natural appearance of the produced image plays a role, and on the other hand, the similarity of the morph with the contributing data subject. The natural appearance can be adversely affected by strong artefacts. In addition, the

FIGURE 6: Examples of BRISQUE scores for quality estimation (low values indicate high image quality). The BRISQUE score of bona fide (a) and uncompressed morphed images (b) are close to each other, the score of a JPEG compressed morphed image (c) is noticeably higher.

FIGURE 7: Comparison between a manually created high quality (left) and an automatically created low quality face morph (right).
TABLE 1: Overview of publicly available morphing tools.

<table>
<thead>
<tr>
<th>Developer</th>
<th>Software</th>
<th>Platform</th>
<th>Method</th>
<th>Automatic</th>
<th>Manual Effort</th>
<th>Required Skills</th>
<th>Parameters</th>
<th>Expected Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morpheus</td>
<td>Photo Morpher</td>
<td>Win 7 / MacOS</td>
<td>landmarks</td>
<td>no</td>
<td>medium</td>
<td>positioning of landmarks</td>
<td>α</td>
<td>no outer region</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>minor shadow artefacts</td>
</tr>
<tr>
<td>DebugMode</td>
<td>WinMorph</td>
<td>Win 7</td>
<td>probably landmarks</td>
<td>no</td>
<td>medium</td>
<td>positioning of landmarks</td>
<td>α</td>
<td>minor shadow artefacts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>issues in hair regions</td>
</tr>
<tr>
<td>Abrasoft</td>
<td>Fantamorph</td>
<td>Win 7-10 MacOS</td>
<td>landmarks</td>
<td>landmark detection</td>
<td>low</td>
<td>no</td>
<td>α</td>
<td>high quality for manual morphs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>minor shadow artefacts</td>
</tr>
<tr>
<td>Luxand Inc.</td>
<td>FaceMorpher</td>
<td>Win 7-10</td>
<td>landmarks</td>
<td>landmark detection</td>
<td>low</td>
<td>no</td>
<td>α</td>
<td>shadow artefacts</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>blurry</td>
</tr>
<tr>
<td>Adobe</td>
<td>After Effects</td>
<td>Win 7-10 MacOS</td>
<td>lines</td>
<td>no</td>
<td>very high</td>
<td>operate Adobe After Effects</td>
<td>α₃, α₄, α₅</td>
<td>very high quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>minor shadow artefacts</td>
</tr>
<tr>
<td>Adobe</td>
<td>Photoshop + Morph Animation</td>
<td>Win 7-10 MacOS</td>
<td>landmarks</td>
<td>rough shape</td>
<td>high</td>
<td>operate Adobe Photoshop</td>
<td>α₃, α₄, α₅</td>
<td>high quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>minor shadow artefacts</td>
</tr>
<tr>
<td>PiVi &amp; Co.</td>
<td>MixBooth</td>
<td>Android / iOS</td>
<td>swapping</td>
<td>no</td>
<td>low</td>
<td>no</td>
<td>no</td>
<td>low resolution</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>unrealistic morphs</td>
</tr>
<tr>
<td>Moment Media</td>
<td>FaceFusion</td>
<td>iOS</td>
<td>landmarks</td>
<td>morph process</td>
<td>very low</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
</tbody>
</table>

Open Source Software:

<table>
<thead>
<tr>
<th>The blender project</th>
<th>blender</th>
<th>Win 7-10 MacOS/Linux</th>
<th>manual mesh warping</th>
<th>via plugins</th>
<th>high</th>
<th>operate blender</th>
<th>inf.</th>
<th>nearly faultless (manual morphing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenCV team</td>
<td>OpenCV</td>
<td>Win 7-10 MacOS/Linux</td>
<td>landmarks + triangulation + warping</td>
<td>full automatic</td>
<td>implementation</td>
<td>Python commandline</td>
<td>inf.</td>
<td>limited by landmark detection issues e.g., for pupils</td>
</tr>
<tr>
<td>Alyssa Quek</td>
<td>FaceMorpher</td>
<td>MacOS / Linux</td>
<td>landmarks + triangulation + warping</td>
<td>full automatic</td>
<td>low</td>
<td>Python commandline</td>
<td>α, blur</td>
<td>good quality</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>no outer region</td>
</tr>
<tr>
<td>The GIMP-Team</td>
<td>GIMP + GAP</td>
<td>Win 7-10 MacOS/Linux</td>
<td>landmarks + triangulation + warping</td>
<td>automation via API</td>
<td>medium (if manual)</td>
<td>positioning of landmarks (if manual)</td>
<td>α</td>
<td>good quality easy to postprocess</td>
</tr>
<tr>
<td>Atsushi Nitanda</td>
<td>VAEGAN</td>
<td>Theano + Python</td>
<td>DNN</td>
<td>full automatic</td>
<td>very low</td>
<td>Theano + Python</td>
<td>no</td>
<td>deep learning artefacts</td>
</tr>
<tr>
<td>Michael Gourlay</td>
<td>gkmorph</td>
<td>Win 7-10 MacOS/Linux</td>
<td>landmarks + mesh warping</td>
<td>no</td>
<td>medium</td>
<td>positioning of landmarks</td>
<td>α</td>
<td>very detailed partly too sharp</td>
</tr>
</tbody>
</table>

similarity of the contributing subjects, e.g., with respect to gender, ethnicity or age group, influences the plausibility of the resulting morph. e.g., the morph depicted in Fig.1 appears less plausible since the age gap between the two contributing subjects is more than 20 years. Thus, it is recommended to select similar subjects as a basis. An approach for an automatic selection of suitable subjects is given in [57].

**D. HUMAN PERCEPTION OF MORPHED FACE IMAGES**

The issue of morphed face images in face comparison scenarios (e.g., border control) does not only affect automated face recognition systems, but also human observers. In general, humans are rather weak in recognizing unfamiliar faces as reported by Megreya and Burton [58] and Bruce et al. [59], independent of comparing two face images or a face image to a live data subject [60]–[62]. In particular, for border control scenarios, it is of relevance, that the difficulty to successfully verify a subject against its reference face image increases with the age of the taken image [63]. Depending on the individual, the face comparison capabilities vary. Hereby it is not relevant, if the human examiner is a border guard or an untrained student, the ratio of false negative to false positive remains the same [64], thus, it is uncertain whether a human expert can effectively detect morphed face images unless he is explicitly trained on morphing attacks. Recently, it has been shown that training makes a huge difference for a human observer. In [65], Robertson et al. showed, that without the knowledge of the morphing issues, a human observer would accept 68% of morphed images created with an α factor of 0.5. After a briefing, the false acceptance rate of morphed images dropped as low as 21%. Further, examiners that are better in distinguishing faces have a higher success chance to detected morphed face image [11]. Another parameter to consider is the weight (α) of the two subjects contributing to the morphed face image which represents a key factor in morphing attack scenario [66]. The role of the two subjects could be asymmetric, since the accomplice has to fool a human examiner, e.g., at the passport application office, and the criminal must fool the face verification algorithm, e.g., at an ABC gate. A higher weight of the accomplice is expected to hamper a successful detection of the morphed face image by a human examiner during presentation at enrolment, e.g., at the time of the passport issuance.

**IV. MORPHING SOFTWARE**

Table I lists available proprietary/open source morphing software and their properties. Applications were considered for the common desktop operating systems (Windows, Linux, Mac) and mobile operating systems (Android, iOS). Ex-
cluded from the list are web services available on the internet. These web services provide an easy way to manually create morphed images. However, firstly, an automated generation of face morphs is difficult and secondly, it is unclear how the uploaded images are processed and stored, which would make it impossible for researchers to upload face images of their models/volunteers and to comply with privacy regulations at the same point in time.

In order to enable well-founded and efficient experiments, it is generally advisable to use applications that can produce morphs in an automated manner in good quality without manual post-processing. Open source algorithms have the advantage that they can be much better automated and adapted to the needs than commercial applications. For commercial programs, automation is generally more difficult to achieve.

A. MORPHING MORE THAN TWO FACE IMAGES

The procedures described above for morphing two face images are easily extended to any number of source images. The contributing images may be weighted similarly to the $\alpha$-factor, each image having its own factor such that the sum of the factors is 1. The more images included in an equally weighted morph, the smaller the weights will be. The more subjects are contributing to the image, the higher is the risk of quality issues described in Sect. III. Furthermore, the morphing of more than two images can also be done iteratively in pairs, i.e. the morphs are used as source images for the next morph process. Generally, no difference is visually discernible between the morphs created by both methods, i.e., the difference between artefacts resulting from a direct or iterative morphing process is below the perception threshold of a human observer. For this reason, the representation of sample images is omitted.

V. METRICS FOR MORPHING ATTACK EVALUATIONS

Standardized metrics are vital to enable direct benchmarks and comparative assessments of proposed methods. Regarding the topic of face morphing attacks efforts to define evaluation metrics for morphing attack detection and vulnerability analysis have already been made, e.g., in [33], [52]. Metrics suggested by Scherhag et al. [52] are briefly summarized in the following subsections.

A. VULNERABILITY ASSESSMENT

In their well-established guidelines Mansfield and Wayman [67] recommended that all comparisons in a biometric system’s evaluation should be uncorrelated. That is, the samples compared to the morphed face images should not be the same as the ones used for the morphing process since such a comparison would ignore the natural biometric variance.

Regarding evaluation metrics the Impostor Attack Presentation Match Rate (IAPMR) introduced in ISO/IEC 30107-3 on Presentation Attack Detection evaluation [9] represents a standardized metric for attack success evaluation:

$$IAPMR = \frac{1}{M} \sum_{n=1}^{M} \left\{ \left[ \min_{m=1,\ldots,N_m} S_m^n > \tau \right] \right\},$$

where $\tau$ is the decision threshold, $S_m^n$ is the morph comparison score of the $n$-th subject of morph $m$, $M$ is the total number of morphed images and $N_m$ the total number of subjects constituting to morph $m$. Decisions of human examiners could be integrated to the above equation to evaluate a scenario with human inspection in the loop. Further, Scherhag et al. [52] proposed adaptations of the metric for evaluations where multiple samples of one subject are compared to one morphed face image.

$MMPMR$, as well as $IAPMR$, are directly dependent on the threshold $\tau$ of the biometric system. In order to achieve a more generalized metric in relation to the False Non-Match Rate ($FNMR$) of the system, Scherhag et al. propose to compute the difference between $1 – FNMR$ and $MMPMR$ or $IAPMR$, respectively. The Relative Morph Match Rate ($RMMR$) is defined as follows:

$$RMMR(\tau) = 1 + \left( MMPMR(\tau) - (1 - FNMR(\tau)) \right) = 1 + \left( MMPMR(\tau) - TMR(\tau) \right).$$

Different relevant examples for combinations of score distributions, thresholds and resulting $RMMR$ values are depicted in Fig. 8.

Gomez-Barrero et al. [68], [69] proposed a theoretical framework to predict the vulnerability of biometric systems to attacks based on morphed biometric samples. Further, key factors which take a major influence on a system’s vulnerability to such attacks have been identified, e.g., the shape of mated (genuine) and non-mated (impostor) score distributions or the False Match Rate ($FMR$) the system is operated at.

B. DETECTION PERFORMANCE REPORTING

Given multiple procedures for preparing morphed images and/or multiple morph detectors these can be benchmarked...
employing metrics defined in [9], in particular, Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). The APCER is defined as the proportion of attack presentations using the same presentation attack instrument species incorrectly classified as bona fide presentations in a specific scenario. The BPCER is defined as the proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario. Further, the BPCER-10 and BPCER-20 representing the operation points related to an APCER of 10% and 5%, respectively, can be used to rank the tested morphing attack detection mechanisms. Additionally, it is recommend to plot the BPCER over the APCER in a Detection Error Tradeoff (DET) curve. In order to achieve reproducible and comparable performance evaluations of morphing attack detection systems, a common comprehension of the training and testing methodology is needed. In general, the standards defined in ISO/IEC 19795-1 on biometric performance testing and reporting [70] should be followed, e.g., a disjoint subdivision of the data into training and testing set. In particular a strict separation of the morphed samples with respect to the originating subjects is important, in order to avoid an unrealistic high detection performance. It should be noted, that one morphed sample is related to at least two subjects and each subject might contribute to several morphing samples.

VI. FACE MORPHING ATTACK DETECTION
Proposed approaches can be coarsely categorized with respect to the considered morphing attack detection scenario. The two classes of detection methods, i.e., no-reference and differential, are described in the following subsection. Subsequently, the state-of-the-art with respect to morph detection algorithms is surveyed.

A. DETECTION SCENARIOS
Two automated morph detection scenarios depicted in Fig. 9 can be distinguished:

- No-reference morphing attack detection: the detector processes a single image, e.g., an off-line authenticity check of an electronic travel document (this scenario is also referred to as single image morphing attack detection or forensic morphing attack detection);

- Differential morphing attack detection: a trusted live capture from an authentication attempt serves as additional source of information for the morph detector, e.g., during authentication at an ABC gate (this scenario is also referred to as a image pair-based morphing attack detection). Note that all information extracted by no-reference morph detectors might as well be leveraged within this scenario [38].

B. STATE-OF-THE-ART
In the past years, numerous approaches to automated face morphing attack detection have been proposed. Published methods and their properties are summarized in Table 2. In some works, more than one system was presented, in such cases only those approaches, which were reported to reveal best morphing attack detection performance are listed. The majority of works assume the challenging no-reference scenario while some implement a differential morphing attack detection. Despite promising results reported in many works, a reliable detection of morphed face images still represents an open research challenge. It is important to note that the generalizability/robustness of published approaches has not been shown. So far, there are no publicly available large-scale databases of bona fide and morphed face images and no publicly available morph detection algorithms, which allow...
for a comprehensive experimental evaluation. Hence, the vast majority of methods has been mostly trained and tested on different in-house databases. In addition, face morph detection methods are mostly trained and tested on a single database using a single morph generation algorithm. Further, the likely appliance of image post-processing techniques by an attacker, e.g., image sharpening, is neglected in most works. Due to these facts, a comparison of published approaches in terms of reported detection performance would potentially be misleading and is purposely avoided in this survey. However, planned benchmark tests, e.g., by the National Institute of Standards and Technology (NIST) [71], are expected to facilitate a meaningful quantitative comparison of published approaches in the near future.

1) No-reference morphing attack detection

Several researchers have suggested the use of general purpose image descriptors, e.g., Local Binary Patterns (LBP) [102] or Binarized Statistical Image Features (BSIF) [103], which have been employed widely for biometric recognition. Ramachandra et al. [14] proposed a no-reference detection system based on a Support Vector Machine (SVM) trained on extracted BSIF-features of grayscale images. For training and evaluation of the SVMs an in-house database of morphed face images was created. On a derive version of the same database, Scherhag et al. [36] investigated the accuracy of morphing detection on printed and scanned images employing the proposed algorithm. Furthermore, a Probabilistic Collaborative Representation Classifier (Pro-CRC) [104] trained on LBP-feature extracted from the colour channels was proposed in [72]. As database an in-house database based on FRGCv2 [73] was used. The authors focus on the differences between morphed and averaged images in the evaluation. In [48] the suitability of LBP features for the detection of morphs generated by Generative Adversarial Networks (GANs) was tested.

The features extracted by texture descriptors can be further processed. A more complex method for morphing detection is proposed in [75], [76], where a Vietoris–Rips complex is built of the responses of uniform LBP extractors on the image. In [100], a high detection performance was shown for a linear SVM trained on high-dimensional LBP features [105] extracted from the FEI database [7]. Agarwal et al. [74] propose to train an SVM with Weighted Local Magnitude Pattern. Similar to LBP, the proposed descriptor encodes the differences between a center pixel and its neighbors. However, instead of binarizing them, it assigns the weights inversely in proportion to the difference from the center pixel. Depending on the feature representation of texture descriptors the inputs of classifiers have to be adapted. E.g., for Scale-Invariant Feature Transform (SIFT) [106] the number of extracted keypoints has been shown to be suitable for the task of morph detection [38], [78]. A score-level fusion of multiple image descriptors might even improve the detection rate [79]. Therefore, LBP, BSIF, SIFT, Speeded Up Robust Features (SURF) [107], Histogram of Oriented Gradients (HOG) [108] and the deep features of Openface [109] were fused and evaluated in [79].

In particular, in the no-reference scenario, classifiers may overfit to distinct micro texture features. These can be dataset-specific features, which are altered or introduced by the applied morphing process. In particular the combination of features reflecting different information, e.g., LBP and SIFT, leads to improvements. It has been shown that the performance of morph detectors based on general purpose image descriptors might significantly decrease if training and test images stem from a different source, i.e., face database [37], [82]. In order to adapt the no-reference general purpose image descriptors a differential scenario, differences between feature vectors can (additionally) be employed [38].

During the morphing process, not only the texture, but the whole signal of the image is manipulated. Thus, a further detection approach is to analyze the changes in noise patterns, e.g., Photo Response Non-Uniformity (PRNU) [84]. Therefore, the PRNU-patterns, that are originating from the camera and which are distinct not only for each model but also for each single camera, are extracted from a face image, the discrete Fourier magnitudes are computed. Subsequently, the mean and variance are derived from the resulting histogram. A very similar approach was presented in [86]. Recently, an improved version of this scheme based on PRNU variance analysis across image blocks was proposed in [85].

Morphing attack detection methods based on continuous image degradation have been proposed in [78], [110], [111]. The basic idea behind these methods is to continuously degrade the image quality, e.g., by using JPEG compression, to create multiple artificial self-references of a face image. The distances from these references to the original image are then analysed for morph detection. Ramachandra et al. [89] proposes the analysis of high frequencies in grayscale images. Therefore, the images are converted to grayscale according their luminance, a steerable pyramid is build and a Collaborative Representation Classifier (CRC) is trained on the high frequencies. The employed database was printed and scanned, but no further post-processing was tested. An alternative to handcrafted feature extractors is to employ statistical machine learning on the unprocessed image in order to distinguish between morphed and bona fide images. Ramachandra et al. [94] proposed to adapt two CNNs (VGG19 [112] and AlexNet [113]) by transfer-learning and combine the intermediate features to train a CRC. In [35], three CNNs, namely VGG19, AlexNet and GoogLeNet [114], are benchmarked as pre-trained and non-pre-trained models regarding their morph detection capabilities. Again, with these methods there is a potential problem of overfitting. In particular, resulting deep classifiers may favour image locations where artefacts, e.g., shadows around the iris region, are likely to appear due to an imperfect automated morph creation process, as described in Sect. III-B. As an attempt to avoid overfitting, Seibold et al. [99] trained a VGG19-net on a set of diverse images with two different databases, morphing algorithms and post-processings (mo-
<table>
<thead>
<tr>
<th>Publication</th>
<th>Approach</th>
<th>Scenario</th>
<th>Morph Algorithms</th>
<th>Source Face Database</th>
<th>Post-Processing</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>[74]</td>
<td>BSIF + SVM</td>
<td>no-reference</td>
<td>GIMP/GAP + triangulation + blending (+ swapping)</td>
<td>Utrecht, ARface [81]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[96]</td>
<td>BSIF + SVM</td>
<td>no-reference</td>
<td>GIMP/GAP + triangulation + blending (+ swapping)</td>
<td>FRGCv2 [73]</td>
<td>print and scan</td>
<td>fixed database of [14]</td>
</tr>
<tr>
<td>[73]</td>
<td>multi-channel-LBP + ProCRC</td>
<td>no-reference</td>
<td>OpenCV + triangulation + blending (+ swapping)</td>
<td>PUT [98], scFace [99]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[74]</td>
<td>WLM + SVM</td>
<td>no-reference</td>
<td>Snapchats + triangulation + blending (+ swapping)</td>
<td>in-house</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[72]</td>
<td>ULP + RIPS + KNN</td>
<td>no-reference</td>
<td>FRGCv2 [73]</td>
<td>Utrecht [77]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[78]</td>
<td>image degradation</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>in-house</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[79]</td>
<td>general purpose image descriptors + score-level fusion</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>FRGCv2 [73]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[80]</td>
<td>HOG + SVM</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>FERET [80], ARface [81]</td>
<td>-</td>
<td>cross database performance evaluation</td>
</tr>
<tr>
<td>[82]</td>
<td>LBP + SVM</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>FRGCv2 [73], FERET [80]</td>
<td>-</td>
<td>cross database performance evaluation</td>
</tr>
<tr>
<td>[83]</td>
<td>LBP + SVM</td>
<td>no-reference</td>
<td>MorGAN [40]</td>
<td>CelebA [84]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[85]</td>
<td>PRNU analysis</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>FRGCv2 [73]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[86]</td>
<td>SPN analysis</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>Utrecht, FEI [87]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[87]</td>
<td>double-compression artefacts</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>Utrecht, FEI [87]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[88]</td>
<td>double-compression artefacts</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>Utrecht, FEI [87]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[89]</td>
<td>reflection analysis</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>in-house</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[90]</td>
<td>luminance component + steerable pyramid + ProCRC</td>
<td>no-reference</td>
<td>unclear</td>
<td>extended [72]</td>
<td>print and scan</td>
<td>-</td>
</tr>
<tr>
<td>[91]</td>
<td>landmark angles</td>
<td>differential</td>
<td>OpenCV</td>
<td>ARface [81]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[92]</td>
<td>Demorphing</td>
<td>differential</td>
<td>GIMP/GAP</td>
<td>ARface [81]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[93]</td>
<td>VGG19 + AlexNet + ProCRC</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>BU-4DFE [90], CFD [97], FEI [87], FERET [80], PUT [98], scFace [99]</td>
<td>print and scan</td>
<td>-</td>
</tr>
<tr>
<td>[94]</td>
<td>VGG19</td>
<td>no-reference</td>
<td>triangulation + blending (+ swapping)</td>
<td>Utrecht [77], in-house</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[95]</td>
<td>high-dim. LBP + SVM</td>
<td>no-reference</td>
<td>ULBP + RIPS + KNN</td>
<td>Multi-PIE [101]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Focusing on the no-reference scenario diverse approaches related to media forensics have been presented. In different works, the detection of JPEG double-compression artefacts has been suggested for the purpose of morph detection [32], [33]. However, the presence of such artefacts implies a strong assumption on the image format of face images used for morph generation as well as the resulting morphed face image. The ICAO suggests face image data to be stored in accordance with the specifications established by the International Standard ISO/IEC 19794-5 [115]. More specifically, the ICAO requires face images to be stored in electronic travel documents at an average compressed sizes of 15kB to 20kB in JPEG or JPEG 2000 format [13]. However, JPEG 2000 is the de-facto-standard for electronic travel documents, as it maintains a higher quality when compressing face images to 15 kB. Hence, depending on the image size and the employed compression algorithm the detection of JPEG double-compression artefacts might not be feasible. In [88], a morph detection method based on reflection analysis in face images is presented. The lightning direction is estimated based on reflections detected in the eyes of a potentially morphed image. Subsequently, reflections on the nose of the face are analysed. However, ISO/IEC standard requires hot spots and specular reflections to be absent in face images used in electronic travel documents. In particular, diffused lighting, multiple balanced sources or other lighting methods shall be used, i.e., a single bare “point” light source like a camera mounted flash is not acceptable for imaging [115].
2) Differential morphing attack detection

Morphing detection algorithms based on general purpose image descriptors, signal or quality analysis are mostly non-reference algorithms, but can be adapted to differential morphing attack detection scenarios. However, there are some algorithms, that can solely be used in differential scenarios, as they require a trusted live capture. In [90], a morph detection algorithm based on landmark positions and angles is introduced. Therefore, the landmarks between both, the passport image and the trusted live capture are determined, the angle between all combinations of landmarks per image are computed and compared over both images. Due to the high intra-class variance of landmarks, the detection performance of this algorithm is rather moderate.

Another differential morph detection method referred to as de-morphing was proposed by Ferrara et al. [91]. In this approach a trusted live capture is aligned to a potential morph and “subtracted” from it in the image domain by applying a reverse morphing operation. The resulting image is then compared against the trusted live capture. The assumption is, that, if two subjects are morphed into one image, and one of the subjects is subtracted, the second subject remains. If there is only one subject in the image, this subject will remain after the subtraction. Thus, a morph is detected if the biometric decision changes from “accept” to “reject” when using the de-morphed image as reference. Robustness of de-morphing against slight face pose variations has been confirmed in [92]. Nevertheless, the authors indicate that in an ABC scenario the performance of de-morphing might degrade due to potential variations of quality and environmental conditions.

VII. ISSUES AND CHALLENGES

Several open issues and challenges exist in research related to face morphing and face morphing attack detection. The most relevant issues and challenges, which have already been pointed out throughout this survey, can be briefly summarized as follows:

- **Quality**: the automated generation of high-quality face morphs remains a challenging issue and of utmost importance in order to enable statistically significant testing of developed morphing attack detection methods under realistic conditions, see Sect. [III]
- **Comparability/benchmarks**: the lack of publicly available large-scale databases comprising bona fide as well as morphed face images and open-source face morphing attack detection software prevents from a meaningful comparative benchmark of the current state-of-the-art in this field, see Sect. [V]
- **Result reporting**: while first efforts have been made to apply standardized metrics for reporting the performance of morphing attack detection mechanisms equivalent measures for the vulnerability of face recognition systems w.r.t morphing attacks are non-existent; however, these would be vital in order to enable an unambiguous comparisons of proposed approaches, see Sect. [V]
- **Over-fitting/robustness analysis**: like any other image-based classification task, approaches to morphing attack detection are prone to overfitting, i.e., rigorous evaluations including face morphs from unseen databases created by unseen morphing techniques are necessary, see Sect. [VI]
- **Print-scan databases**: to simulate real-world scenarios where potentially morphed portrait images are printed and scanned, publicly available large-scale databases of printed and scanned bona fide and morphed face images are required, see Sect. [VI]

VIII. CONCLUSION

This survey provides a comprehensive overview of published literature in the field of (face) image morphing and face morphing attack detection as well as a detailed discussion of open issues and challenges. The research in this important field is only in its infancy while not being limited to face recognition systems. The feasibility of morphing biometric samples has also been shown for other biometric characteristics, e.g. fingerprint [46], [116] or iris [47], which might as well be morphed in feature domain. The possibility of morphing biometric features and subsequently reconstructing a biometric sample from morphed feature vectors underlines the importance of data protection mechanisms, i.e. biometric template protection [117], [118] or conventional cryptographic techniques [119], [120]. Similar to face, for other characteristics certain aspects require more in-depth analysis, e.g., biometric quality estimation of (morphed) fingerprint [121], [122] or iris samples [123], [124], respectively. The reported face image morphing attack detection accuracy is yet not reflecting generalization to datasets incorporating the real world variety of capture conditions. This will change, once benchmark portals such as the NIST Face Recognition Vendor Test (FRVT) MORPH competition [71] are established. Nevertheless, robust algorithms must also anticipate the large variety of image post-processing as well as printing and scanning technology that could be used in the governmental procedures for the application of electronic travel documents. Morphing attack detection mechanisms that are robust against all those factors, will require a significant amount of future research.

ACKNOWLEDGMENT

This work was partially supported by the German Federal Ministry of Education and Research (BMBF), by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the Center for Research in Security and Privacy (CRISP) as well as by the Federal Office of Information Security (BSI) within the FACETRUST project.

REFERENCES


—“Predicting the vulnerability of biometric systems to attacks based on morphed biometric information,” IET Biometrics, vol. 7, no. 4, pp. 208–223, apr 2018.


A. M. Megreya, A. Sandford, and A. M. Burton, “Matching face images taken on the same day or months apart: the limitations of photo ID,” Applied Cognitive Psychology, vol. 27, no. 6, pp. 700–706, oct 2013.


