Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting

U. Scherhag*, A. Nautsch*, C. Rathgeb*, M. Gomez-Barrero*, R.N.J. Veldhuis†, L. Spreeuwers†, M. Schils†, D. Maltoni‡, P. Grother§, S.Marcel¶, R. Breithaupt‖ , R. Raghavendra**, C. Busch*
* da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
Email: {ulrich.scherhag, andreas.nautsch, christian.rathgeb, marta.gomez-barrero, christoph.busch}@h-da.de
† Services, Cybersecurity and Safety Research Group, University of Twente, The Netherlands
Email: {r.n.j.veldhuis, l.j.spreeuwers}@utwente.nl, m.schils@student.utwente.nl
‡ DISI, University of Bologna, Italy
Email: davide.maltoni@unibo.it
§ National Institute of Standards and Technology
Email: patrick.grother@nist.gov
¶ Idiap Research Institute, Martigny, Switzerland
Email: marcel@idiap.ch
‖ Bundesamt für Sicherheit in der Informationstechnik (BSI), Bonn, Germany
Email: ralph.breithaupt@bsi.bund.de
** Norwegian Biometrics Laboratory, NTNU, Gjøvik, Norway
Email: raghavendra.ramachandra@ntnu.no

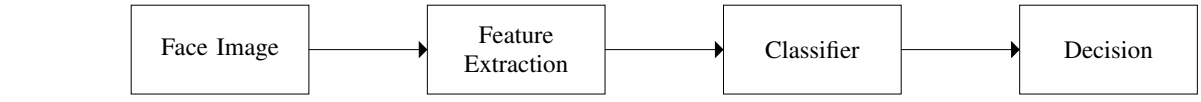
***Abstract*—With the widespread deployment of biometric recognition systems, the interest in attacking these systems is increasing. One of the easiest ways to circumvent a biometric recognition system are so-called presentation attacks, in which artefacts are presented to the sensor to either impersonate another subject or avoid being recognised. In the recent past, the vulnerabilities of biometric systems to so-called morphing attacks have been unveiled. In such attacks, biometric samples of multiple subjects are merged in the signal or feature domain, in order to allow a successful verification of all contributing subjects against the morphed identity. Being a recent area of research, there is to date no standardised manner to evaluate the vulnerability of biometric systems to these attacks. Hence, it is not yet possible to establish a common benchmark between different morph detection algorithms. In this paper, we tackle this issue proposing new metrics for vulnerability reporting, which build upon our joint experience in researching this challenging attack scenario. In addition, recommendations on the assessment of morphing techniques and morphing detection metrics are given.**

***Index Terms*—Biometrics, Morphing, Performance Reporting, Attack Detection**

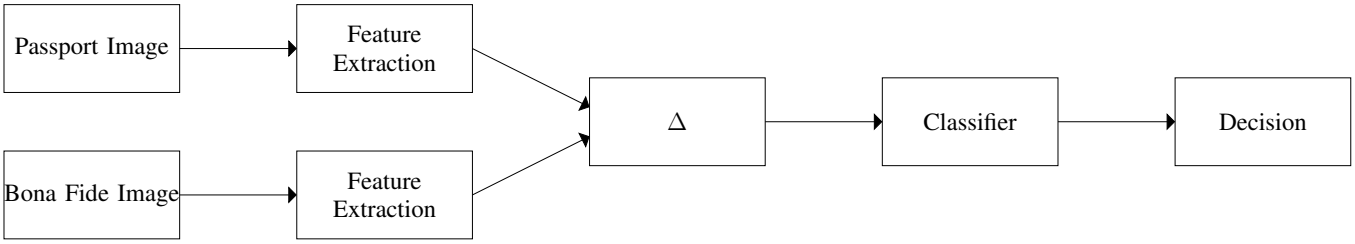
I. Introduction

Biometrics refers to the automated recognition of individuals based on their biological and behavioural characteristics [1]. Due to the strong link between subjects and their biometric samples, the wide acceptance, and their user convenience, biometric systems become increasingly popular. Even though the security of biometric systems is increasing, recent research revealed a security gap to subvert the unique link between a biometric sample and its subject. By enrolling an artificial sample, generated by merging samples of two or multiple subjects in image or feature domain, the contributing subjects might be verified successfully against the manipulated reference. This can be done, for instance, in the passport application process, where in most countries the applicant brings his own printed photograph. This way, the unique link between individuals and their biometric reference data is annulled. The feasibility of such *morphing attacks* was first shown for face recognition systems [2], [3] and most recently for fingerprint [4] and iris [5] recognition systems. The remainder of the paper will focus on the face case study, as it most widely studied and allows for a comprehensible visual explanation of the morphing process and the occurring difficulties.

To prevent the aforementioned morphing attacks, an automatic detection of morphs is required. Focusing on the workflow of a generic biometric system, two morph detection tasks can be distinguished: (1) detection during enrolment, e.g. the passport application process, where the detector processes a single image, referred to as no-reference morphing detection and depicted in Figure 1(a); and (2) detection at the time of authentication, e.g. the usage of Automated Border Control (ABC) gates at borders, where a live capture from an authentication attempt serves as additional source of information for the morph detector, referred to as differential morphing detection and depicted in Figure 1(b). Moreover, two attack scenarios can be distinguished: (i) an attacker could try to attack a fully-automated biometric system or (ii) a semi-automated system with human examiners in the loop. In the latter case, the role of subjects contributing to a morphed image might be asymmetric, i.e., some subjects might have to



(a) no-reference morphing detection scheme



(b) differential morphing detection scheme

Fig. 1. Difference between no-reference and differential morphing detection schemes

pass the human inspection while others have to pass biometric recognition systems.

The metrics and terminology defined in ISO/IEC 30107-3 on Presentation Attack Detection evaluation [6] strongly relate to morphing attacks. However, those metrics only apply to one subject per attack. On the contrary, for morphing attacks success is achieved if multiple subjects bypass the system for a single sample, i.e., more than one biometric decisions have to be considered. Thus, only parts of this standard can be employed for evaluating morphing detection, while other metrics, e.g. the Impostor Attack Presentation Match Rate (IAPMR), need to be adapted.

Even if there exist some works dealing with morphing [2]–[5], [7] or morphing detection [8], [9], no common understanding for morphing attacks and morphing detection has been developed yet. During the creation of morphing databases and the design of morphing detection algorithms we observed multiple pitfalls, which are summarized in this paper. In order to allow a common evaluation of the attack success rate, we propose new metrics, in particular the Mated Morph Presentation Match Rate (MMPMR) and Relative Morph Match Rate (RMMR). With this proposal, we intend to spark a discussion within the research community and awaken the interest of the ISO/IEC biometrics standardisation committee to compose a comprehensive list of requirements that need to be taken into account when evaluating morphing attacks.

The remainder of this paper is organized as follows: In Section II we present observations from our work on creating morphed images, based on which diverse recommendations are given. Section III proposes metrics to evaluate the vulnerability of biometric systems to morphing attacks. In Section IV recommendations for morphing detection and morphing detection evaluation are given.

II. Recommendations on Morph Generation

Morphing attack samples generated for research databases may differ from real world attack samples. In order to achieve significant evaluation results, a large number of attack samples has to be created, which can be achieved by automated methods. For the sake of realistic attack scenarios, four major factors have to be considered: (1) the morph quality, (2) the similarity of the morphed subjects, (3) the consistent quality of the database and (4) weights used to generate the morph. All of these factors will be discussed in detail in the subsequent paragraphs.

(1) The real world attacker has the option to spend much time on one single morph, which might reveal a higher quality compared to automatically generated images, depending on the goodness of the automatic morphing process and the skills of the attacker. Figure 2(c) shows a high quality morphed face image attack sample generated using FantaMorph [10], whereas an example for a low quality morph is depicted in Figure 2(e). Both images are successfully verified against the contributing subjects, but Figure 2(e) contains a huge amount of obvious morphing artefacts which can be easily detected by human observers or common pattern recognition algorithms. As the attacker is willing to do his best to circumvent the system, the best conditions should be expected for the attacker. In particular, for face image morphing attacks in border control scenarios, the image needs to fulfill specific quality requirements, defined in [11], in order to be accepted as a biometric passport sample. Thus, assuming the preserved chain-of-trust of the passport creation process, the appearance of obvious artefacts should be minimized during the morphing process. However, for evaluation purposes morphs with lower quality might be of interest as well. In order to achieve a significant evaluation on a database containing multiple quality levels of morphing attacks, a clear labeling of the data is mandatory. For automatically generated morphs, a consistent quality per algorithm is assumed. Manually generated morphs, however, might vary in quality, requiring a quality metric to ensure a proper labeling. The definition of the quality metric depends on the specific scenario and is not in the scope of this paper.

(2) As motivated in [7], morphs of two subjects yielding a high chance of both being positively matched, referred to as *lookalike morphs*, are more relevant than morphs of two subjects highly differing in appearance, referred to as *non-lookalike morphs*. One option is selecting subjects to

(a) Subject 1 (b) Subject 2

(c) manual high quality morph (d) automated morph (e) low quality morph

Fig. 2. Differences between morphing qualities

(a) uncompressed face image of original subject BRISQUE = 21.0 (b) uncompressed morphed face image BRISQUE = 29.1 (c) compressed morphed face image BRISQUE = 50.0

Fig. 3. Impact of JPEG-compression

(a) eye of manual morph (b) eye of autom. morph

Fig. 4. Differences between different morphing techniques

be morphed according to the similarity score returned by a face recognition algorithm. However, in a real world scenario, realistic lookalike morphs are necessary to fool human experts [3], e.g. when applying for an ID document. A high number of non-lookalike or bad quality morphs might reduce the impact of the attack on the recognition system and at the same time artificially increase the detection performance of the morph detector. In order to achieve realistic combinations of subjects, a clear pre-categorization of the subjects according to soft-biometric attributes is recommended, e.g skin, hair, gender, or age in case of face images. The definition of the different categories falls out of scope of this work and needs to be addressed in further research. Employing such a classification scheme, the total number of morphs can be divided into subsets representing different relevance-classes.

(3) For verification purposes, training on images with different quality and resolution leads to a higher recognition accuracy and robustness to different scenarios. However, for morphing detection, it is important to obtain an equal quality for bona fide and morphed samples, in order to avoid bias towards different quality levels on the morphing detection algorithm. To illustrate this fact, Figure 3 depicts the impact of JPEG-compression on morphed face images. For quality estimation, the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [12] was employed. While the image of the originating subject, as well as the morphed face image depicted in Figure 3(b) are uncompressed, Figure 3(c) is JPEG-compressed. Even if no visual difference can be observed between these images, the BRISQUE-score indicates a severe quality loss of the image due to the compression.

So far, we have observed that common machine learning algorithms, e.g. Support Vector Machines (SVM), trained on local image descriptors, e.g. Binarized Statistical Image Features (BSIF) [13] or Local Binary Patterns (LBP) [14], might classify the images according to such compression artefacts or image quality differences, if present, instead of attributes related to the morphing process. The same applies for other artefacts introduced by some post-processing, e.g. rescaling, image-optimization or rotation. However, distinct artefacts might disappear if an image is printed and scanned.

The aforementioned quality issues are more likely to appear in automatically generated morphs, which are thus expected to differ in quality from manual morphs. Figure 2(c) depicts a manual morphed face image and Figure 2(d) an automatically generated morph. As emphasized in Figure 4(a) and 4(b), the automatically generated morph reveals small shadow artefacts which can be avoided in the manual morph.

(4) Another factor to consider is the weights of multiple subjects in a morph. This is a key factor in scenarios where humans are required to check the morph against a live subject. For border control scenarios, it could be feasible to generate a morphed image with a high weight for the accomplice applying for the passport. This way, the accomplice has a high chance of deceiving the officer at enrolment time and the criminal will still be able to be successfully verified by the ABC gate.

III. Assessment of Vulnerability to Morphing Attacks

To assess the vulnerability of a specific biometric system to morphing attacks, a standardized methodology is needed. In general, it is crucial to follow the guidelines proposed in [15], where it is recommended that all comparisons in one evaluation should be uncorrelated. In particular, 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 [6] represents a standardized metric for attack success evaluation:

> **IAPMR:** in a full-system evaluation of a verification system, the proportion of impostor attack presentations using the same Presentation Attack Instrument (PAI) species in which the target reference is matched.

However, for the evaluation of morphing attacks, the aforementioned IAPMR metric presents some drawbacks, as a morphing attack can only be considered successful if all contributing subjects are successfully matched against the morphed sample. To avoid a confusion of the wording impostor, which is used for zero effort impostors, the comparison of a morphed sample to another independent sample of one contributing subject will be referred to as *mated morph comparison*. Motivated by the international standard ISO/IEC 30107-3 [6], we propose a new metric for the evaluation of the impact of a morphing attack in a full-system evaluation, referred to as Mated Morph Presentation Match Rate (*MMPMR*).

If the recommendations of [15] are considered, only one mated morph comparison per subject is possible. As the morphing attack succeeds if all contributing subjects are verified successfully, only the minimum (for similarity scores) or maximum (for dissimilarity scores) of all mated morph comparisons of one morphed sample are of interest. The *MMPMR* for similarity scores is accordingly defined as:

$$\textit{MMPMR}(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left\{ \left[\min_{n=1,\ldots,N_m} S_m^n \right] > \tau \right\}, \quad (1)$$

where τ is the verification threshold, S_m^n is the mated 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 contributing to morph m. The following examples are for similarity scores.

If, due to a lack of data, the recommendation in [15] is not met, and multiple samples of one subject are compared to one morphed image, there are two possibilities to adapt the metric. For smaller number of samples, multiple comparisons can be understood as multiple authentication attempts per subject. For instance, for face image morphing attacks in a border control scenario, the attacker is able to conduct several authentication attempts and will be successfully verified, as long as one attempt is above the threshold of the biometric system. Thus, the metric can be extended by only considering the maximum (for similarity scores) or minimum (for dissimilarity scores) over all mated morph comparisons of one subject, referred to as *MinMax-MMPMR* and depicted in Figure 5(a).

$$\textit{MinMax-MMPMR}(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left\{ \left(\min_{n=1,\ldots,N_m} \left[\max_{i=1,\ldots,I_m^n} S_m^{n,i} \right] \right) > \tau \right\}, \quad (2)$$

where I_m^n is the number of samples of subject n within morph m. *MinMax-MMPMR* also models the case where N_m subjects launch single attacks to several biometric authentication systems ($I_m^n = 1$).

However, for larger number of probe sample per subject, the MinMax approach is prone to falsely increase the number of accepted subjects. Thus, we propose a probabilistic interpretation, by calculating the proportion of accepted attempts per subject and multiply the probabilities of all contributing subjects (i.e., joint probability). The mated morph acceptance rate is calculated over all contributing subjects, referred to as *ProdAvg-MMPMR*, in analogy to the *MinMax-MMPMR* and depicted in Figure 5(b):

$$\textit{ProdAvg-MMPMR}(\tau) = \frac{1}{M} \cdot \sum_{m=1}^{M} \left[\prod_{n=1}^{N_m} \left(\frac{1}{I_m^n} \cdot \sum_{i=1}^{I_m^n} \left\{ S_m^{n,i} > \tau \right\} \right) \right]. \quad (3)$$

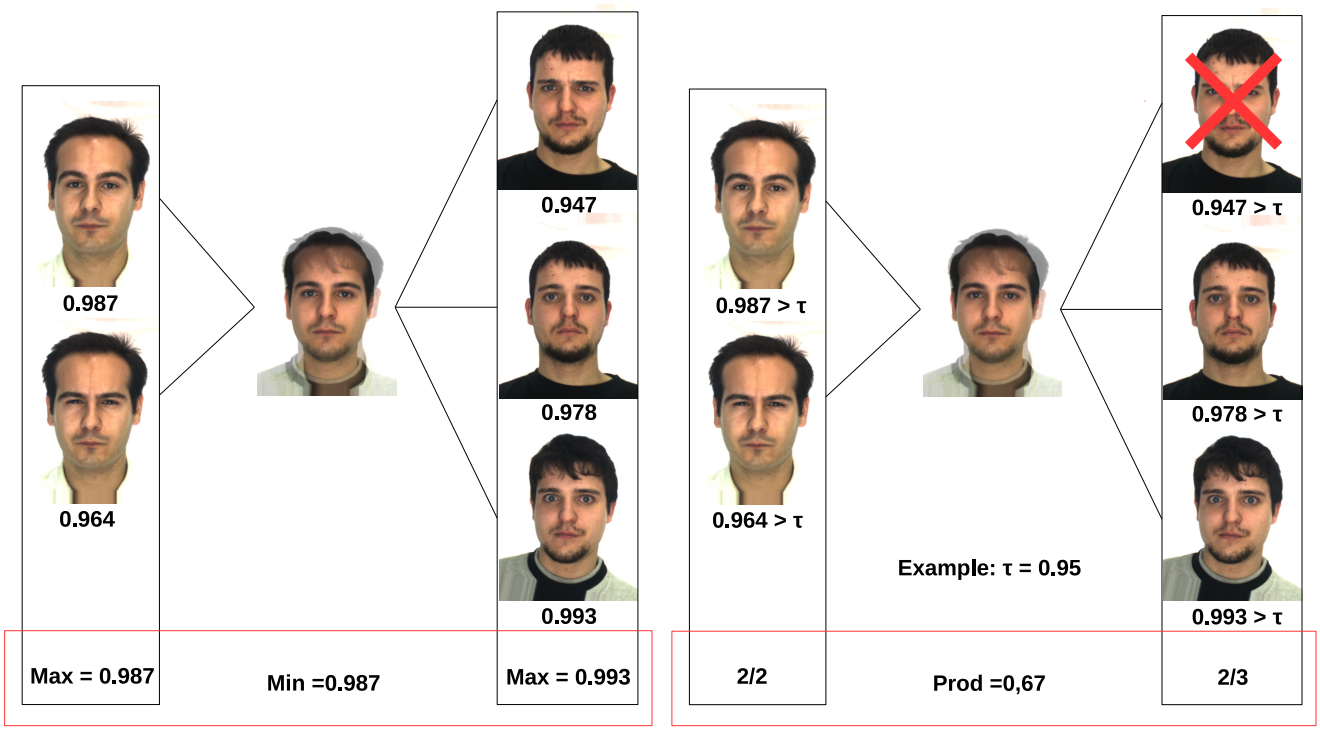
MMPMR, as well as IAPMR, are directly dependent on the threshold τ of the biometric system. In order to achieve a more generalized metric, we propose to compute the difference between *MMPMR* or IAPMR and $1-FNMR$, respectively. The Relative Morph Match Rate (*RMMR*) is defined as follows:

$$\begin{aligned} \textit{RMMR}(\tau) &= 1 + (\textit{MMPMR}(\tau) - (1 - \textit{FNMR}(\tau))) \\ &= 1 + (\textit{MMPMR}(\tau) - \textit{TMR}(\tau)). \end{aligned} \quad (4)$$

For presentation attacks as described in [6], $MMPMR$ can be replaced by $IAPMR$.

Figure 6 depicts different *RMMR* examples for combinations of distributions and thresholds. If $MMPMR$ and $1-FNMR$ are equal sized, the $RMMR$ will be 1 (Figure 6(a) and 6(e)). For a more restrictive threshold, the $RMMR$ will decrease (Figure 6(b) and 6(c)), until the threshold reaches a point where the $FNMR$ increases (see Figure 6(d)). For a scenario in which all mated morph comparisons are rejected (as depicted in Figure 6(d)), the distribution of the comparison scores is not of interest. Even for a mated morph distribution far below the impostor comparisons, the *RMMR* would remain the same. If the score distribution of mated morphs is bigger than $1 - FNMR$, the *RMMR* will be above 1 (see Figure 6(f)). Note that the latter "extreme" case is considered as unrealistic, since the $RMMR$ is assumed to be upper-bounded by $1 - FNMR$.

For security assessment scenarios, a morphed sample is a threat as soon as more than one subject is successfully verified against it. For these assessments only the two most



(a) MinMax-MMPMR (Eq. (2))

(b) ProdAvg-MMPMR (Eq. (3))

Fig. 5. Examples for the computation of *MMPMR*

successful subjects are considered for the *MMPMR* estimation. If more subjects are successfully verified against the morphed reference, the attack can be considered as more severe, thus the amount of successful mated morph comparisons should be reported as the weight of the attack.

IV. Morphing Detection Performance Reporting

Multiple procedures for creating morphed images and/ or multiple morph detectors can be independently benchmarked employing the metrics defined in [6]. In particular, the Attack Presentation Classification Error Rate (APCER) and the Bona Fide Presentation Classification Error Rate (BPCER) should be computed, and visualised in a Detection Error Trade-Off (DET) curve.

In addition, in order to achieve reproducible and comparable performance evaluations of morphing detection systems, for each procedure or detector 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 [16] should be followed, e.g. a disjoint subdivision of the data into training and testing set. More specifically, 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.

As described in Section II, when aiming to develop and test a robust detection algorithm, it is crucial to ensure, that the feature extractor is not based on artefacts present on low quality morphs. Otherwise, it is likely that a trained classifier might strongly rely on these specific artefacts. As a consequence, if different quality levels of morphed samples should be examined, these should be evaluated separately according the quality labels defined during the database creation process. For Example, if a morphing detection system is trained on a mixture of low and high quality morphs, the evaluation should be conducted separately on low and high quality morphs to ensure reliable performance measures for the different attack classes.

Finally, for morphed face image attacks in border control scenarios, the employment of comprehensible detection algorithms is strongly recommended. In order to achive justifiable and reliable results of the detection system, the system should reveal morphing-specific information, thus a clear separation of frontend (feature extractor) and backend (classifier) is needed: back-end classifiers shall be based on discriminative features subjective to morphing detection (not necessarily biometric recognition), thus a depending front-end must be employed, extracting features in a morphing-discriminant space. On assessing non morphing specific information, classifier training may be mislead regarding nuisance attributes, e.g. processing artefacts introduced by compression or scaling. In order to avoid opaque results in algorithm benchmarking, we strongly advise against algorithms, not encapsulating a clear distinction between front- and back-end, when aiming at sensitive operational real world scenarios, such as border control.

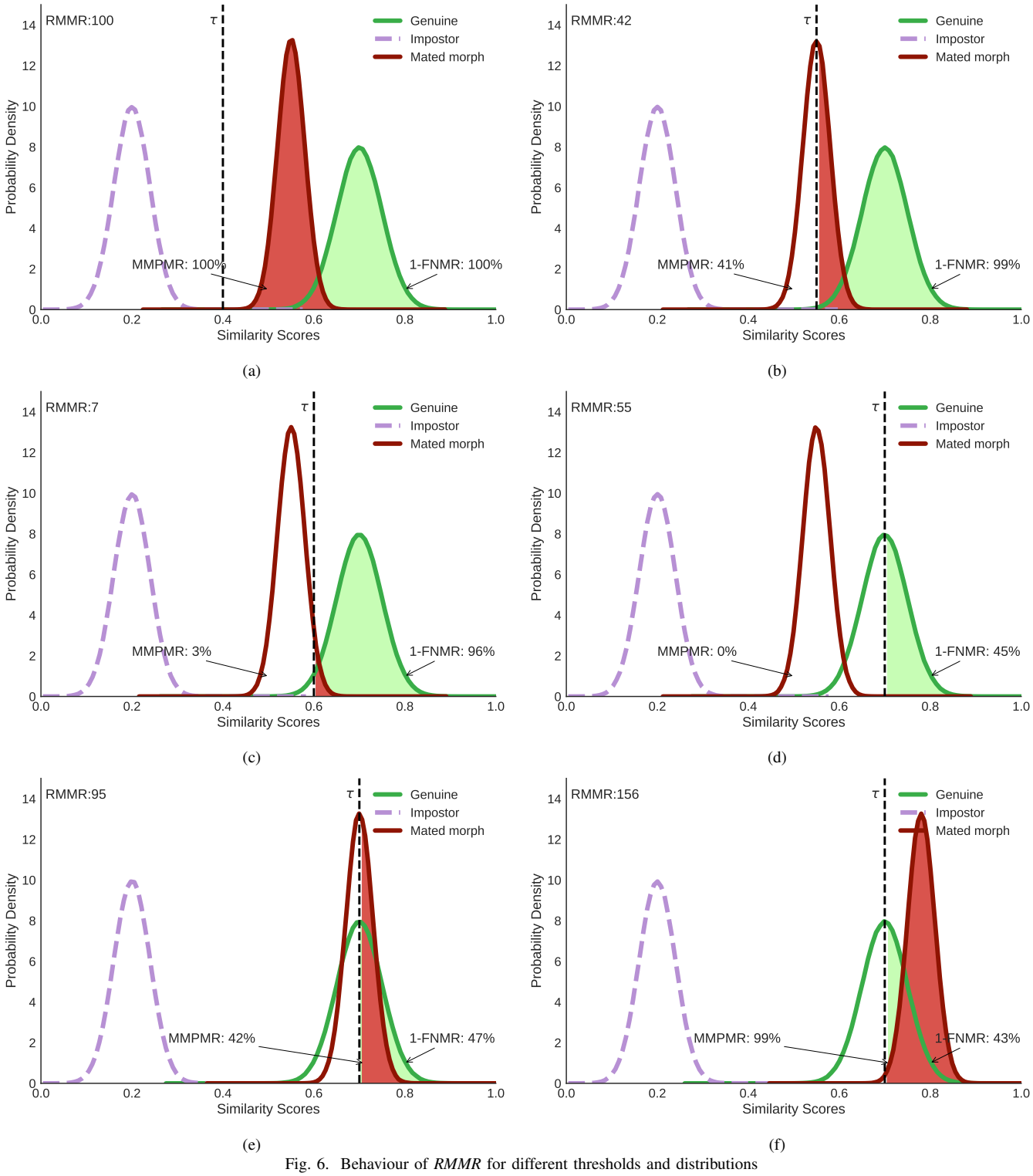


Fig. 6. Behaviour of *RMMR* for different thresholds and distributions

V. Conclusion

During the creation of morphing samples, multiple pitfalls have to be avoided. To that end, we have presented a summary of the observations we made so far. The key issues are the morph quality, the similarity of subjects and the consistent quality of the database. In order to allow a fair evaluation of biometric systems' vulnerability to morphing attacks, we propose new metrics, i.e. *MMPMR* and *RMMR*. Further, our experiences and considerations regarding morph detection and morph detection evaluation are summarized. The paper focuses on morphing attacks on face recognition systems, but the considerations and metrics are applicable for other modalities as well. To facilitate the use of the proposed metrics, an implementation of the evaluation metrics is provided in [17].

Acknowledgment

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) as well as by the Hessen State Ministry for Higher Education, Research, the Arts (HMWK) within CRISP (www.crisp-da.de) and the BioMobile II project (no. 518/16-30).

References

[1] International Organization for Standardization, "Information technology – Vocabulary – Part 37: Biometrics," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 2382-37:2012, 2012.

[2] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014.

[3] ——, *Face Recognition Across the Imaging Spectrum*. Springer International Publishing, 2016, ch. On the Effects of Image Alterations on Face Recognition Accuracy.

[4] M. Ferrara, R. Cappelli, and D. Maltoni, "On the feasibility of creating double-identity fingerprints," *IEEE Trans. on Information Forensics and Security*, vol. 12, no. 4, 2017.

[5] C. Rathgeb and C. Busch, "On the feasibility of creating morphed iris-codes," in *Proc. Int. Joint Conf. on Biometrics (IJCB)*, 2017, pp. 1–6.

[6] International Organization for Standardization, "Information Technology – Biometric presentation attack detection – Part 3: Testing and reporting," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC FDIS 30107-3:2017, 2017.

[7] M. Gomez-Barrero, C. Rathgeb, U. Scherhag, and C. Busch, "Is Your Biometric System Robust to Morphing Attacks?" in *Int. Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2017.

[8] R. Raghavendra, K. B. Raja, and C. Busch, "Detecting Morphed Face Images," in *8th IEEE Int'l Conf. on Biometrics: Theory, Applications, and Systems*, 2016.

[9] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the Vulnerability of Face Recognition Systems Towards Morphed Face Attacks," in *Proc. International Workshop on Biometrics and Forensics (IWBF)*, 2017.

[10] Abrasoft, "FantaMorph," http://www.fantamorph.com, 2017, accessed: 2017-04-18. [Online]. Available: http://www.fantamorph.com

[11] International Organization for Standardization, "Information Technology – Biometrics – Biometric Data Interchange Formats – Face Image Data," JTC 1 /SC 37, Geneva, Switzerland, Tech. Rep. 19794-5, 2005.

[12] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Blind/referenceless image spatial quality evaluator," in *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2011.

[13] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," *21st Int'l Conf. on Pattern Recognition (ICPR)*, 2012.

[14] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, 2002.

[15] A. J. Mansfield and J. L. Wayman, "Best practices in testing and reporting performance of biometric devices," 2002.

[16] International Organization for Standardization, "Information technology – Biometric performance testing and reporting – Part 1: Principles and framework," JTC 1/SC 37, Geneva, Switzerland, ISO/IEC 19795-1:2006, 2006.

[17] "Morphing vulnerability reporting," 2017, online available: https://github.com/dasec/mvr.