Towards Robust Evaluation of Face Morphing Detection

Luuk Spreeuwers
Data Management and Biometrics Group
Faculty of EEMCS
University of Twente
7500 AE Enschede, The Netherlands
Email: l.j.spreeuwers@utwente.nl

Maikel Schils
Email: maikel.schils@gmail.com

Raymond Veldhuis
Data Management and Biometrics Group
Faculty of EEMCS
University of Twente
7500 AE Enschede, The Netherlands
Email: r.n.j.veldhuis@utwente.nl

***Abstract*—Automated face recognition is increasingly used as a reliable means to establish the identity of persons for various purposes, ranging from automated passport checks at the border to transferring money and unlocking mobile phones. Face morphing is a technique to blend facial images of two or more subjects such that the result resembles both subjects. Face morphing attacks pose a serious risk for any face recognition system. Without automated morphing detection, state of the art face recognition systems are extremely vulnerable to morphing attacks. Morphing detection methods published in literature often only work for a few types of morphs or on a single dataset with morphed photographs. We create face morphing databases with varying characteristics and how for a LBP/SVM based morphing detection method that performs on par with the state of the art (around 2% EER), the performance collapses with an EER as high as if it is tested across databases with different characteristics. In addition we show that simple image manipulations like adding noise or rescaling can be used to obscure morphing artifacts and deteriorate the morphing detection performance.**

I. Introduction

A morphed face image is a combination of two or more face images, created in a way that all contributing subjects are verified successfully against the morphed image. Suppose A' and B' are images of two distinct subjects A and B, shown in Figure 1a and 1b. With face morphing, the two images are combined to create attack sample M, see Figure 1c. If we perform identification tasks with state of the art facial recognition software, a good morph will generate high scores for comparisons between morph M and templates of subjects A and B. It is obvious that face morphing poses a severe threat to all processes where face recognition is used to establish the identity of subjects, as first reported in [1]. Also human face recognition is vulnerable, as reported by Robertson et al. [2].

Automated morphing attack detection can be the solution to this problem. The morphing process leaves certain traces in the morphed image because the image is locally stretched or compressed and the images are combined. In high quality morphs, these texture differences are not visible to humans. Automated morphing attack detection scenarios can be subdivided into two types; morphing attack detection with or without a reference sample. The scenario with reference sample means that apart from the morphed image, also an image of one of the original contributing subjects is available, which in principle makes morphing detection simpler. In this research we address automated morphing attack detection without reference sample.

Most of the published methods for face morphing detection are developed and tested using a single database with morphed and bona fide samples and often good detection results are reported. However, the use of a single database and therefore a single, specific way to generate morphed images, may result in a morphing detection method that works well only for this specific type of face morphing. An example is morphing detection based on so-called double JPEG compression detection - detection of artifacts that occur because the morphed images are created from JPEG compressed images and compressed again when they are stored. Such a method will fail to detect morphed images if they are stored uncompressed.

The aim of this paper is to investigate robust evaluation of morphing detection methods using cross database testing and sensitivity to simple morphing disguise techniques. The paper is based on the Master's thesis of Maikel Schils [3].

(a) Face image A' (b) Face image B' (c) Morph M_m

Fig. 1: Bona fide face samples and manual face morph

In the remainder of this paper, first a brief overview of related work on face morphing detection is presented. Next, the creation of 4 datasets with morphed face images is described that are used in the development and testing of morphing detection methods. Multiple datasets are required to investigate cross dataset performance of morphing detection.

ISBN 978-90-827970-1-5

Subsequently, a morphing detection method based on Local Binary Patterns (LBP) and a Support Vector Machine (SVM) is presented of which we will show that it performs close to the state of the art as reported in literature. Next, two approaches to disguise morphing: adding nose and scaling images, are presented for which we will investigate morphing detection robustness. Then, experiments and results are presented concerning within and cross database performance of morphing detection and robustness against morphing disguise and finally, conclusions are presented.

II. Related Work

In order to evaluate the performance of morphing detection methods, the following metrics were introduced in ISO/IEC FDIS 301073 [4]:

Attack Presentation Classification Error Rate (APCER) Proportion of attack presentations incorrectly classified as bona fide presentations.

Bona Fide Presentation Classification Error Rate (BPCER) Proportion of bona fide presentations incorrectly classified as presentation attacks.

A bona fide sample refers to a non-morph and an attack sample refers to a morph.The trade-off between APCER and BPCER can be represented in a Detection Error Trade-Off (DET)-curve and also Equal Error Rates (EER) can be reported.

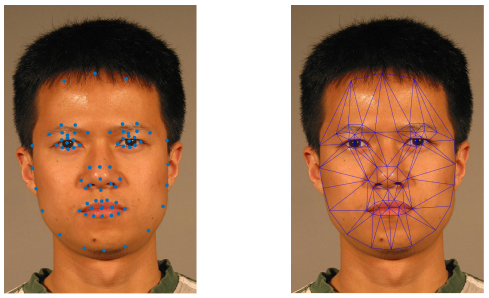
Currently, most published work on face morphing detection is based on textural feature classifiers, e.g. LBP features followed by an SVM classifier. Tested on single datasets of morphed face images good results are reported in literature. Creation of good datasets with morphed face images is one of the most important steps in the development of reliable face morphing detection methods. In [5] 450 morphed faces are created manually from a database comprised of 110 subjects. The face region is detected with Viola Jones detection. Various features like LBP, LBQ, 2DFFT (Fourier Transform) and BSIF filters are extracted. The combination of BSIF with 7x7 and 12bit and SVM yields an Attack Presentation Classification Error Rate (APCER) of 1.73%. BSIF filters [6] are trained by utilising statistics of natural images. The BSIF performance exceeds that of all other features, the next best feature is LBQ with an average classification error of 20.23% APCER. The dataset of 450 morphs was split into three subsets; training, testing and validation. A problem with the database however is that these sets are not split according to the original 115 subjects. This means a morph in the training set may share a contributing subject with a morph in the test or validation set. In [7] the experiments from [5] are repeated, but instead the morphing detection process at a passport control is simulated by printing and scanning the face images. Morphing attack detection performance was analysed before and after printing and scanning. It is found that printing and scanning images adds noise and granularity, causing a loss in morphing attack detection performance. The dataset was properly split into training and testing sets without overlapping subjects. The reported performances are in the order of 40% BPCER at 10% APCER.

III. Creation of morphing datasets

For experiments with morphing attack detection a large number of face morph images is required. We use automated morphing algorithms to quickly generate morphs. The dataset is split in a part for training and a part for testing with no overlap in subjects.

A. Creating Morphs

To create a face morph, the first step is to extract landmarks from both faces. For manual morphing the landmarks can be selected by hand, for automated morphing we use an existing landmark localisation algorithm. For morphing it is critical to know which parts in the image of one contributing subject correspond to the parts of another. Therefore it is vital that landmarks are accurately extracted, if they are placed incorrectly, it can lead to extremely poor morphs. There are several landmark localisation algorithms available. We found that Stasm [8] and dlib [9] result in high quality morphs. Figure 2a shows Stasm landmarks on a face sample A'. A triangular mesh is defined over the landmarks using Delaunay Triangulation [10] (Figure 2b). Now each triangle can be related to its corresponding triangle from the other contributing image. The triangles are morphed toward average triangles located in the final morph M_a using an affine transformation.



(a) Stasm Landmarks (b) Delauney Triangulation

Fig. 2: Initial steps of the morphing process

A blending value α defines the weight of contribution of the involved subjects. There are two common ways of selecting α: either we set $\alpha = 0.5$ so that both subjects contribute equally to the morph or face recognition software is used to set α so that the morph generates approximately the same comparison score for both contributing subjects.

The automatically generated morphs normally suffer from artifacts near the boundaries of the face and around the eyes, nose and mouth, because of the limited number of landmarks. In our research on morphing detection, we only used the inner part of the face.

ISBN 978-90-827970-1-5 © EURASIP 2018

When creating morphed face images, it is vital to save them in a lossless format like '.png' to ensure the morphing detection methods do not detect compression artifacts.

B. Datasets

We created four datasets with images of different quality and properties, originating from different facial datasets: FRGC [11], ARF [12], Feret colour and Feret gray [13].

An overview of the created datasets with information on resolution (Inter Occular Distance, IOD), number of training and testing images is given in Table I.

Datasets				
Dataset:	FRGC	ARF	Feret Colour	Feret Gray
Resolution (IOD, pix)	129	177	177	60
Morph Train	500	500	750	500
Non-Morph Train	150	150	250	200
Morph Test	500	500	750	500
Non-Morph Test	100	100	100	200

TABLE I: Characteristics of the Datasets, resolution is given in pixels Inter Ocular Distance (IOD)

Note that the resolution of the Feret Gray dataset is much lower than the resolution of the other datasets. This may impact morphing detection performance. Care was taken to use different subjects for each of the subsets: Morph Train, Non-Morph Train, Morph Test and Non-Morph Test.

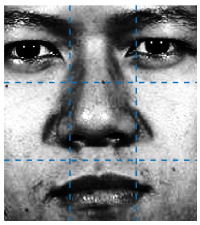
IV. Texture based face morphing detection

Even though BSIF filters perform better in literature, we chose to use LBP as it is not trained and shows results close to that of BSIF. With the use of landmarks the face region as shown in Figure 3a is extracted and resized to a fixed size. The face region is cut off at the top of the eyebrows and somewhat below the mouth. With this region we ensure that the sides of the face which often contain obvious morphing artifacts are not present in the face image. We convert the image to gray scale and apply histogram equalisation, enhancing image contrast (Figure 3b). Using the FRGC database we performed a parameter sweep for LBP parameters: uniform/non-uniform LBP, number of neighbours n and radius r. We find that uniform LBP features with "standard" parameters, ($n = 8$, $r = 1$) and a 3x3 histogram result in a good performance. Increasing the number of histograms; e.g. 4x4 or 5x5 layout, only slightly increases the performance but also the dimensionality of the feature space increases. We therefore decided to use the "standard" parameters. For uniform LBP, a single histogram contains 59 feature values, which means for a 3x3 layout the feature space has 531 dimensions. The SVM classifiers are trained on between 650 and 1,000 samples.

V. Morphing disguising

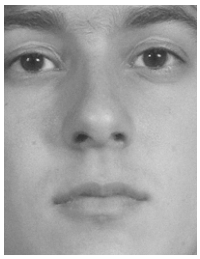
As pointed out earlier, often morphing detection methods are trained on a single database with morphed images. This may result in a morphing detection method that only detects a certain property of the morphing creation process. If the morphing creation process is slightly disturbed, these methods will fail. Here, we investigate two simple ways do disguise the morphing process: adding Gaussian noise to the image and rescaling. In the first approach, a small amount of Gaussian noise is added to the image, masking certain noise characteristics of the morphing process that a morphing detection method may have learnt. The noise is kept small, such that to the human eye it is barely noticeable, see Figure 4.

(a) Original cropped face

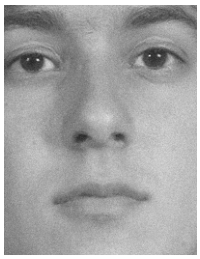


(b) Gray, histogram equalised cropped face

Fig. 3: Region Of Interest for LBP operator, the dashed lines show the areas for which local LBP histograms are obtained



(a) Example of morph with $\sigma = 0.01$



(b) Example of morph with $\sigma = 0.025$

Fig. 4: Morphs with added Gaussian noise. The gray level range of the image is 0..1

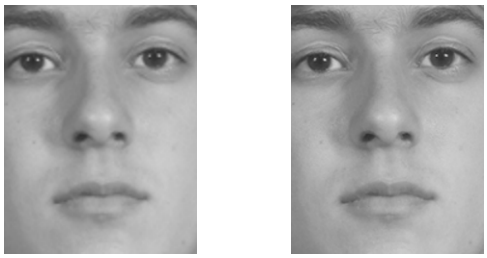
In the second approach, the image is down-sampled using a scaling factor s and then up-sampled again to its original resolution. In this way, some of the higher spatial frequencies are lost also masking the typical noise characteristics of morphed images. Examples of down-up scaled images are shown in Figure 5. Again the manipulation is barely noticeable to the human eye.

VI. Experiments and Results

In order to investigate the robustness of the LBP/SVM morphing detector, we present the following experiments:

1. Within dataset performance
2. Cross dataset performance
3. Mixed dataset performance
4. Robustness against additive Gaussian noise
5. Robustness against down-up scaling

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) Example of a morph with $s = 0.8$

(b) Example of a morph with $s = 0.5$

Fig. 5: Down-up scaled morphs to disguise morphing

A. Within dataset performance

With this experiments we investigate if the morphing detection method we used performs in line with the state of the art. Furthermore, we use the performance as a baseline to compare the results of the other experiments with.

For each of the databases listed in Table I the SVM of the morphing detector was trained on features extracted from the training set and the morphing detection was determined using the test set.

The results are shown in the form of a DET-curve in Figure 6. We can observe that the performance for 3 of the 4 datasets is similar (EER 2.5%-5%), while for the low resolution Feret Gray set the results are poorer (EER=17%). The reason for the poorer results are likely that the image quality (resolution) of the Feret Gray dataset is significantly lower.

The performance on the other datasets is in line with results reported in literature (EER=1.7% in [5]).

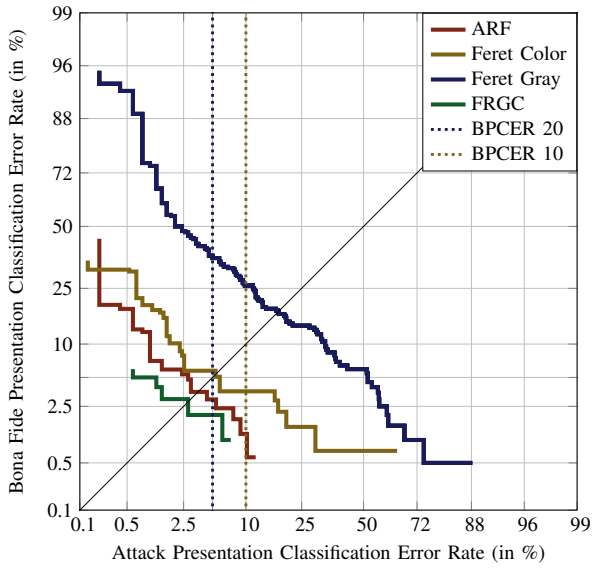


Fig. 6: DET-Curve of LBP experiments on all Databases.

B. Cross dataset performance

Next the cross dataset morphing detection performance is determined. In this experiment the SVMs are trained using the binary pattern features of the one dataset and tested using the test set of another dataset. The experiments were only conducted for the FRGC and ARF datasets and are shown in Table II.

Training set	FRGC	ARF
Test set	ARF	FRGC
EER	80%	79%

TABLE II: Cross dataset morphing detection performance

The cross dataset performances were much worse than the within dataset performances, suggesting that indeed the morphing detector learnt features very specific for the dataset it was trained on.

C. Mixed dataset performance

In this experiment the SVMs are trained using 50% of both of the datasets FRGC and ARF and tested using the test set of both datasets. The results are given in Table III.

Training set	FRGC+ARF
Test set	FRGC+ARF
EER	35%

TABLE III: Mixed dataset morphing detection performance

The mixed dataset performance is better than the cross dataset performances, suggesting that if multiple datasets are used for training, the morphing detector becomes more robust. The performance is still much worse than the within dataset performance, though.

D. Robustness against additive Gaussian noise

In this experiment, we add Gaussian noise to the morphed images in order to disguise artifacts generated by the morphing process. The standard deviation of the noise was varied from 0.004 to 0.027, where the gray level range was normalised to 0..1. Only within database performance is reported.

The results are depicted in Figure 7. We can observe that for small σ of the noise, the EER of the morphing detection is still around 5%, close to the baseline experiment. When the noise increases, the EER increases to above 20% for $\sigma = 0.027$. Note that even this noise will not be observed by human inspection, so it seems morphing artifacts can quite successfully be disguised by adding a bit of noise to the morphed images.

The experiments were done several times for different divisions of the data in training and test sets. The error bars show the minimum and maximum EER values obtained.

E. Robustness against scaling

In this experiment, the original face images are first down-scaled with a factor s and then up-scaled again to their original resolution. In this way, some fine detail, i.e. high spatial frequency information is lost. Since morphing also influences (high) frequency contents of the face images, it is likely that traces caused by morphing can be obscured by this

ISBN 978-90-827970-1-5 © EURASIP 2018

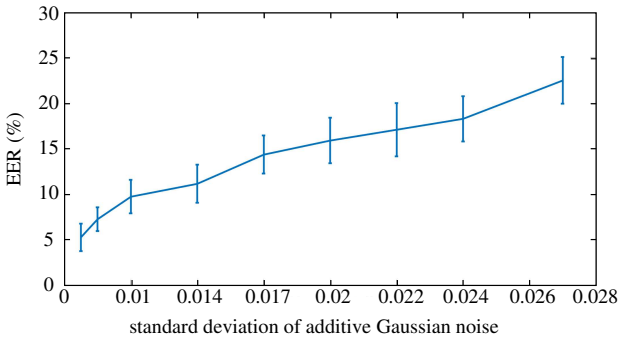


Fig. 7: Morphing detection performance for added Gaussian noise

down-up scaling of the image. We investigated the impact on the morphing detection performance for a scaling range of $s = 0.5..0.95$. Only within database performance is reported.

The results are depicted in Figure 8. We can observe that for $s = 0.95$, i.e. hardly any high frequency information is lost, the EER of the morphing detection is still around 5%, close to the baseline experiment. When the down scaling factor is lower, the EER increases to above 12% for $s = 0.5$. Note that even for this scaling factor, the difference to the original image will not be observed by human inspection, so it seems morphing artifacts can successfully be disguised by down-up scaling as well.

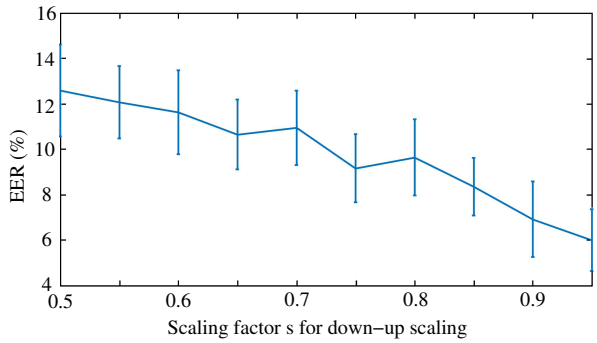


Fig. 8: Morphing detection performance for down and up scaling with scaling factor s.

The experiments were done several times for different divisions of the data in training and test sets. The error bars show the minimum and maximum EER values obtained.

VII. Conclusion

Face morphing, the blending of two face images of distinct subjects into one that resembles both subjects, poses a serious threat to face recognition. In several publications it is claimed that reliable morphing detection is possible. We noticed that often morphing detection methods are developed using a single database with morphed face images. In this paper we show that this results in morphing detection that only works well for a single type of morph or database. Using a LBP/SVM based morphing detection method that performs on par with the state of the art (around 2% EER) within a single dataset, we show that for cross database testing, the performance collapses resulting in an EER as high as 80%. Experiments with mixed datasets suggest that morphing detection can be made more robust if trained on multiple datasets. In addition, we show that the morphing artifacts that are used as features for detection, can be obscured by simple image manipulations like adding Gaussian noise or down-up scaling the morphed images. The EER for within database detection increased from below 5% to above 20% for adding noise and above 12% for down-up scaling. In both cases the manipulation were almost invisible to the human observer.

We therefore argue that morphing detection methods should be tested extensively on multiple datasets obtained from different sources and morphing methods and a range of image manipulations.

References

[1] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, Sept 2014, pp. 1–7.

[2] D. Robertson, R. Kramer, and A. Burton, "Fraudulent id using face morphs: Experiments on human and automatic recognition." *PLoS ONE*, vol. 12, no. 3, 2017.

[3] M. Schils, "Towards a structured approach for face morphing detection." University of Twente, Master EE Biometrics and Computer Vision, July 2017.

[4] "Information technology-biometric presentation attack detection-part 3: Testing and reporting, jtc 1/sc 37," ISO/IEC FDIS 30107-3:2017, Geneva, Switzerland, 2017.

[5] R. Raghavendra, K. B. Raja, and C. Busch, "Detecting morphed face images," in *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Sept 2016, pp. 1–7.

[6] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 1363–1366.

[7] 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 *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, April 2017, pp. 1–6.

[8] S. Milborrow and F. Nicolls, "Active Shape Models with SIFT Descriptors and MARS," *VISAPP*, 2014.

[9] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[10] D. T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, Jun 1980. [Online]. Available: https://doi.org/10.1007/BF00977785

[11] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 947–954. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.268

[12] A. M. Martinez and R. Benavente, "The AR Face Database," CVC, Tech. Rep., Jun. 1998.

[13] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms." *Image Vision Comput.*, vol. 16, no. 5, pp. 295–306, 1998.

ISBN 978-90-827970-1-5 © EURASIP 2018