IET Biometrics

Special Issue: Selected Papers from the International Workshop on Biometrics and Forensics (IWBF2017)

IET Journals
The Institution of Engineering and Technology

Extended *StirTrace* benchmarking of biometric and forensic qualities of morphed face images

ISSN 2047-4938
Received on 29th July 2017
Revised 18th January 2018
Accepted on 20th January 2018
E-First on 15th March 2018
doi: 10.1049/iet-bmt.2017.0147
www.ietdl.org

Tom Neubert[1] ✉, Andrey Makrushin[1], Mario Hildebrandt[1], Christian Kraetzer[1], Jana Dittmann[1,2]

[1]*Department of Computer Science, Otto-von-Guericke-University Magdeburg, Research Group Multimedia and Security, P.O. Box 4120, 39016 Magdeburg, Germany*

[2]*Department of Applied Computing, University of Buckingham, School of Science, Buckingham MK18 1EG, UK*

✉ *E-mail: tom.neubert@iti.cs.uni-magdeburg.de*

Abstract: Since its introduction in 2014, the face morphing forgery (FMF) attack has received significant attention from the biometric and media forensic research communities. The attack aims at creating artificially weakened templates which can be successfully matched against multiple persons. If successful, the attack has an immense impact on many biometric authentication scenarios including the application of electronic machine-readable travel document (eMRTD) at automated border control gates. We extend the *StirTrace* framework for benchmarking FMF attacks by adding five issues: a novel three-fold definition for the quality of morphed images, a novel FMF realisation (combined morphing), a post-processing operation to simulate the digital image format used in eMRTD (passport scaling 15 kB), an automated face recognition system (VGG face descriptor) as additional means for biometric quality assessment and two feature spaces for FMF detection (keypoint features and fusion of keypoint and Benford features) as additional means for forensic quality assessment. We show that the impact of *StirTrace* post-processing operations on the biometric quality of morphed face images is negligible except for two noise operators and passport scaling 15 kB, the impact on the forensic quality depends on the type of post-processing, and the new FMF realisation outperforms the previously considered ones.

1 Introduction

This paper is based on the workshop publication [1] and significantly extends the state-of-the-art of benchmarking face morphing forgery (FMF) attacks.

The *StirTrace* framework is originally designed in [2] as a set of image processing operations to simulate artefacts on images presenting artificial sweat printed fingerprints. The idea was to benchmark both biometric matching performance as well as forensic detection performance with distorted fingerprint images. In [1], the *StirTrace* framework is adopted for detection of FMF by including new face-specific operations and excluding unnecessary fingerprint-specific operations.

The FMF attack has been known to a broader audience since the publication of Ferrara *et al.* [3]. This attack addresses biometric authentication based on face images. The target is the enrolment phase by creating artificially weakened templates which can be successfully matched against multiple persons. The attack consists of manipulating a face image followed the submission of the manipulated image into enrolment. It was originally designed to have a biometric passport (also referred to as an electronic machine-readable travel document eMRTD) in mind. Ferrara *et al.* [3] demonstrate how wanted persons could easily bypass border security controls without the risk of using stolen or forged passports. In fact, the attack poses a severe threat to many biometric authentication scenarios. Such attacks highlight the fact that image tampering detection is of great importance not only for a media forensic community but also for a biometric community and, therefore, demanding interdisciplinary research for detecting and countering FMF.

In order to detect FMF, the established approaches from digital image forensics can be utilised [4]. However, the forensic traces in images are known to be fragile against many image post-processing operations. The *StirTrace* framework simulates post-processing operations (anti-forensics) which are likely to be applied to morphed face images and thus, supports the robustness evaluation of FMF detection approaches.

We benchmark the impact of *StirTrace* post-processing operations on biometric and forensic qualities of different FMF realisations. The number of morphed face images under consideration has been increased from 86,614 in [1] to 125,248.

The contribution of this paper is the extension of the benchmarking presented in [1] and can be summarised as follows:

- C1: We propose a novel three-fold quality definition for benchmarking using visual, biometric and forensic qualities of morphed face images generated by different FMF realisations (Section 3.1). In addition to this theoretical contribution, we propose metrics for two of these qualities (biometric and forensic). The metrics are the recognition performances of an automated face recognition (AFR) system and FMF detector correspondingly. These metrics are used in practical benchmarking (for all benchmarking goals G1–G4, see Section 4.1) of three FMF realisations (see below).
- C2: We extend the benchmarking by introducing a third novel FMF realisation (combined morphing (FMF_{Bi}); Section 3.2). It is designed to improve the visual quality of complete morphing (FMF_C) [4] and biometric quality of splicing morphing (FMF_S) [4]. The results (addressed in G1) show that FMF_{Bi} and FMF_C clearly outperform FMF_S in terms of biometric quality, while the forensic quality varies depending on the FMF detector.
- C3: We introduce the concept of passport scaling 15 kB as a novel *StirTrace* post-processing operation for benchmarking in order to simulate the relevant image editing operation of cropping and scaling to the format used in eMRTD, especially International Civil Aviation Organization (ICAO) compliant travelling passports (Section 3.3). This operation significantly deteriorates the image quality. It is identified in this paper for the first time as a necessary requirement for benchmarking FMF realisations. The benchmarking results (addressed in G2) show that, if grouping the *StirTrace* post-processing operations into classes with negligible impact or significant impact, the passport scaling 15 kB (together with noise addition operations) has to be

considered as belonging to the class with significant impact for all three benchmarked FMF realisations.

- C4: We introduce a second AFR system as a metric for the biometric quality to extend *StirTrace* benchmarking (Section 3.4). This new AFR system makes use of the VGG face descriptor based on deep convolutional neural networks BM_{VGG} (DCNNs), representing the current trend in face recognition research and aiming at differently measuring the biometric quality of FMF realisations and thereby validating the benchmarking results obtained with a commercial off-the-shelf system BM_{Lux} in [1]. The results (addressed in G3) show similar trends for both ways of assessing the biometric quality for all three FMF realisations, establishing that, in terms of biometric quality, FMF_C outperform FMF_{Bi} which is better than FMF_S. Furthermore, most of the *StirTrace* post-processing operations have a negligible influence on the biometric quality.
- C5: We introduce the concept of fusion of FMF detectors by transferring from other disciplines like multi-biometrics (Section 3.5). The concept is included into benchmarking by using the established idea of feature-level fusion FS_{Combo} to create a new FMF detector combining existing feature spaces: keypoint features FS_{KP} from [5] and the Benford features FS_{BF} from [4], aiming at differently measuring the forensic quality of FMF realisations. While the concept is promising, the benchmarking results (addressed in G4) show that the existing individual feature spaces outperform FS_{Combo} for all three FMF realisations. This indicates that for the complex image forensic task of detecting face morphs a simplistic fusion approach tested here is not sufficient indicating the need for investigation of feature selection strategies.

2 State-of-the-art

In [3, 6], a vulnerability of AFR systems to the FMF attack is demonstrated without proposing security mechanisms to prevent or detect this attack. The authors manually generate morphed face images and demonstrate their visual quality in an experiment with human observers. The manual generation of morphed faces images with GIMP/GAP (www.gimp.org) can be seen as a shortcoming in research. This exemplary FMF realisation lacks reproducibility and can be hardly applied for the generation of a large number of morphed face images while abundant training data is a prerequisite for designing FMF detection mechanisms based on pattern recognition.

A significant effort to structure the research on FMF generation and detection is made by Kraetzer *et al.* in [5]. The authors propose a new modelling approach for FMF attacks. In particular, they propose a life-cycle model for photo-ID documents and extend it by an image editing history model allowing for a precise description and comparison of FMF realisations as a foundation for performing forensics FMF detection. Based on recent publications addressing FMF, we assume very high visual and biometric qualities of morphed face images, but the forensic quality (i.e. the amount of tell-tale traces and used anti-forensics) is still in question.

2.1 Automated face morphing

Automated face morphing is required to address the issue of missing training data for designing FMF detectors based on pattern recognition. Makrushin *et al.* in [4] introduce two FMF realisations for automated generation of morphed face images (or morphs). The strategies are based on warping and subsequent alpha-blending as proposed in [7]. The common steps in both strategies are the localisation of facial landmarks (keypoints) describing parts of the face such as eyes, nose, mouth and face contour, triangulation based on these landmarks, warping of triangles to some average position so that for all faces the parts of the face are located at the same position, and alpha-blending of the resulting images. The alpha is set to 0.5 making alpha-blending equal to average. The morphs generated by the first FMF realisation are referred to as complete morphs. The only addition to the common steps is that the set of facial landmarks is extended by additional 20 landmarks on the image borders, as proposed in [8] to cover the complete image. A complete morph has an average geometry and an average texture of the original faces resulting in its high biometric quality. However, a complete morph often has ghosting artefacts as a result of the blending of not perfectly aligned parts of the face resulting in its low visual quality. In order to improve the visual quality of complete morphs, the manual retouching is required as in [3, 9].

The objective of the second FMF realisation is the generation of visually faultless facial morphs, called splicing morphs. The name is motivated by the common challenge of splicing detection in image forensics (see, e.g. [10]). In the splicing process, a fragment of an image is inserted into another image so that the result of the manipulation is invisible to the naked eye. Splicing morphs are designed to avoid ghosting artefacts usually presented in the hair region, which is done by warping and blending of only face regions and inserting the blended face into one of the original face images. The background, hair and torso regions remain untouched. Comparing splicing with complete morphs, there are three additional steps required: cutting of the face region, inverse warping of the face region, and concealing the transition between the inserted face and the rest of the image. The disadvantage is that a splicing morph inherits the face shape of only one individual and thus is expected to match only one person well reducing its biometric quality. Moreover, if skin colour is significantly different for both persons, a morph may look unrealistic. Both complete and splicing morphs are a part of the benchmark in [1].

The novel FMF realisation proposed in this paper and called combined morph is designed to avoid the shortcomings of both aforementioned FMF realisations. The realisation details are presented in Section 3.2.

2.2 Benchmarking using StirTrace

The *StirTrace* framework [2] (https://sourceforge.net/projects/stirtrace) is designed for benchmarking the robustness of detection approaches based on pattern recognition against various artefacts in digitised forensics. The first versions are tailored for benchmarking fingerprint forgery detection. With respect to the challenge of FMF as described in [3], a robust detection mechanism is crucial. Anticipating the arms race between FMF realisations and FMF detectors, we benchmark in [1] the FMF detector from [4] using *StirTrace* to simulate different post-processing operations. This can be seen as image filtering applied to morphed face image to hide morphing artefacts. In [1], 3940 morphed face images are generated with two FMF realisations from [4]. Based on this dataset, 86,614 *StirTrace* samples are generated to evaluate the influence of post-processing operations on the FMF detector and on the face matcher performance, determining the impact of the post-processing on the biometric and forensic qualities of the forgeries. In particular rotation, additive Gaussian, uniform and salt & pepper noises, median filtering, removal of lines/columns, horizontal stretching, rescaling, cropping, double scaling and passport scaling are considered as post-processing operations. These post-processing operations can be divided into legitimate (e.g. scaling, cropping) and illegitimate (e.g. stretching, additive noises, median filtering) regarding ICAO requirements for photo-ID [11]. The passport scaling is used to simulate a common image size of 413×531 pixels, aspect ratio and contents for photo-IDs. For the biometric matcher, the influence of the additional post-processing is negligible. Considering the fact that the FMF realisations will at some point include anti-forensics to conceal traces, such benchmarking is strictly necessary to establish the limits of any proposed FMF detector.

3 Benchmarking concept

The main contribution of this paper is the extension of the benchmarking presented in [1] with a novel definition of the quality of FMF (C1) and new components (C2–C5). The overall design goal of the benchmarking is the creation of an automated processing pipeline to create large quantities of high-quality FMF samples and performance evaluation of biometric face matchers as well as dedicated FMF detectors with several post-processing operations (i.e. potential anti-forensics).

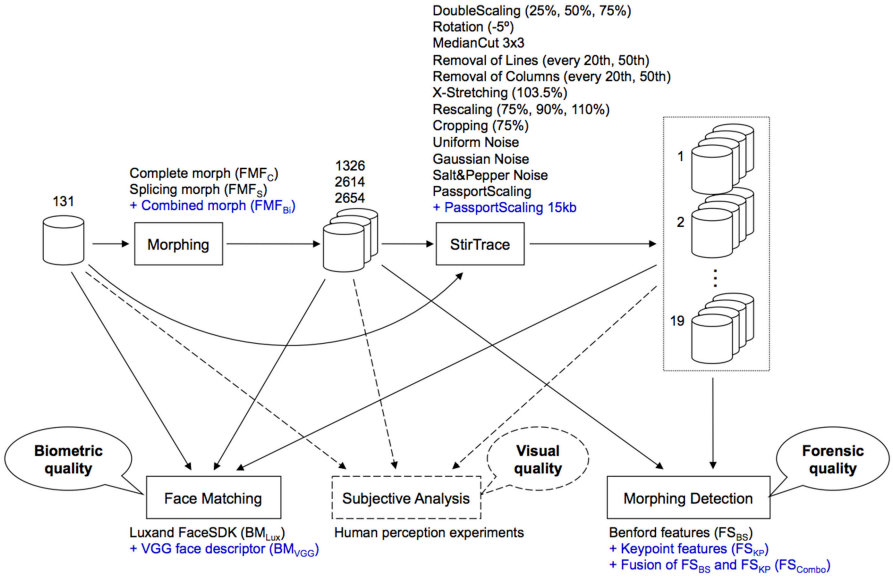

Fig. 1 *Benchmarking concept: morphed face images are created from the original set of 131 Utrecht ECVP face images (1326 complete morphs, 2614 splicing morphs and 2652 combined morphs); 19 StirTrace post-processing operations are applied to original and morphed images; the biometric quality of morphed images is measured using two AFR systems by matching morphed images against original images; the forensic quality of morphed images is measured using FMF detectors by their TPR values; the visual quality of morphed images can be measured only subjectively in an experiment with human observers (blue: new components in comparison to [1], dashed lines: disregarded parts)*

All components of the concept are illustrated in Fig. 1. The new components are marked in the figure in blue and can be summarised as follows:

- one morphing approach (C2; combined morph FMF_{Bi}),
- one *StirTrace* post-processing operation (C3, passport scaling 15 kB),
- one AFR system (C4; VGG face descriptor BM_{VGG}) and
- two feature spaces for morphing detection (keypoint features FS_{KP}, fusion FS_{Combo} of Benford and keypoint features (C5)).

By benchmarking different FMF realisations with various post-processing operations, we estimate how dangerous a specific FMF realisation is, or more precisely, how probable it is that the morphed image will be successfully used for a user authentication.

3.1 Definition of the quality of morphed face images

In our considerations, the quality of morphed face images is three-fold:

i. *visual quality*: morphs should be visually faultless – meaning that there are no visible artefacts and the morphed face is visually similar to faces of both persons;
ii. *biometric quality*: morphs should be successfully verified by any AFR-system against any photographs of both persons;
iii. *forensic quality*: morphs do not include forensic traces of illegitimate image editing operations.

We consider only the first and the second quality issues in our automated morph generation pipelines. The third quality issue is addressed by applying *StirTrace* post-processing operations to morphed images. In contrast, we evaluate only the second and the third quality issues in this paper, because the first one can be only subjectively evaluated in an experiment with human observers. Such experiments have been already conducted in [4, 6, 12] showing what not only manually but also automatically generated face morphs cannot be recognised as such by human observers with a sufficient degree of certainty. Nonetheless, the evaluation of the visual quality of combined morphs proposed in this paper is a subject of future work.

3.1.1 Metric for the biometric quality: The link between biometric quality and a particular AFR system allows for interpreting the biometric quality in terms of error rates of such a system. The disadvantage is that the biometric quality depends on a selected AFR system. Since AFR systems are evaluated by means of standard biometric error rates: FAR (false acceptance rate) and FRR (false rejection rate) the same metrics define the biometric quality. The higher the error rates are, the higher the biometric quality of morphed images is. Taken a reference to an automated border control (ABC) scenario, an AFR system is calibrated for achieving the FAR of less or equal to 0.1%. The performance is evaluated by means of FRR for genuine images as well as MAR (morph acceptance rate) for morphed face images, as proposed in [6]. The same metrics are used for the 'FVC ongoing Face Morphing Benchmarking Challenge' [13]. Motivated by Makrushin *et al.* [4] we replace MAR by rMAR (realistic morph acceptance rate) for more realistic performance evaluation. All in all, the biometric quality is described by the tuple (FRR, rMAR), where FRR is a characteristic of an AFR system and is constant for all FMF realisations. The error rates at 0.1% FAR are referred to as FRR_{1000}, MAR_{1000} and $rMAR_{1000}$.

3.1.2 Metric for the forensic quality: Similar to the biometric quality, the forensic quality is defined in relationship to a particular FMF detector. Therefore, we link the forensic quality to the error rates of a morph detection system. We consider morphed images as positive samples and original images as negative samples. The common error rates are FPR (false positive rate) as a relative number of original images falsely detected as morphs and FNR (false negative rate) as a relative number of falsely missed morphed face images. The higher the error rates are, the higher the forensic quality of morphed images is. For better interpretability, we do not use FPR and FNR but the opposites, namely TPR (true positive rate) for morphed images and TNR (true negative rate) for original images. The forensic quality is described by the tuple (TNR, TPR).

3.2 FMF realisations

We benchmark three FMF realisations: complete morphs (FMF_C), splicing morphs (FMF_S) and the novel combined morphs (FMF_{Bi}). The exemplary morphed face images for all three strategies are illustrated in Fig. 2. The generation pipeline of complete and splicing morphs is originally presented in [4] and used as a part of the benchmark in [1]. A brief description of the strategies can be found in Section 2.2. The drawback of a complete morph is its low visual quality due to ghosting artefacts. The drawback of a splicing

morph is its low biometric quality because the face shape is inherited from one face image and, therefore, the morphed image is expected to match well with only one person. Moreover, the visual quality drastically declines if skin colour is significantly different in the original images.

A combined morph is designed to avoid the drawbacks of both aforementioned FMF realisations. In the generation pipeline, we first geometrically align face images by means of in-plane rotation, scaling and cropping. Afterwards, we localise 68 facial landmarks with the dlib programming library (http://dlib.net) and apply the Delaunay triangulation. Next, we warp complete images to an average position and blend the warped images. Then, we cut a face region from the blended image and splice it into one of the warped images. In order to have a seamless transition between the facial region and the rest of the image, we apply the fast implementation of Poisson image editing [15] from [16]. An important difference between complete and splicing morphing strategies is the alignment of original images prior to warping. Having done this, we assure that warping does not lead to drastic distortion of face geometry. As a result, the warped face image has a realistic appearance and can be used as a target for face splicing. As before, the blending parameter alpha is set to 0.5. The pipeline for generation of combined morphs is presented in Fig. 3.

A morphed face generated by this FMF realisation has an average geometry and an average texture, has no major ghosting artefacts, and the skin colour has no influence on the visual quality of morphs. As a result, the combined morphs are of high visual as well as of high biometric quality.

3.3 Post-processing with StirTrace

In this paper, we use the same selection of post-processing operations as used in [1] (see Section 2.2). In addition, we introduce *passport scaling 15 kB*, in order to follow the ICAO standard limiting the maximum facial image size within eMRTD to 15 kB. For this post-processing operation, we apply a JPEG compression with a variable quality level to the passport scaled images resulting in an additional set of images. With the given constraints regarding the number of pixels in each image, the resulting quality level varies between 50 and 55. This additional lossy JPEG compression can be considered as an additional filter which is utilised in combination with the down scaling and cropping of the passport scaling.

If the combination of compressing and scaling of the image hides particular artefacts caused by the FMF, then the image might be mistakenly considered as a genuine image by FMF detectors, which we consider as an increased forensic quality. The *StirTrace* simulation supports the systematic benchmarking of the limitations of forensic FMF detectors.

3.4 AFR systems for evaluation of biometric quality

The biometric quality of our morphed face images is evaluated with two types of AFR systems: commercial off-the-shelf (COTS) software and academic open-source software based on a DCNN which is a recent trend in face recognition.

As a proponent of COTS systems, the Luxand FaceSDK 6.1 (BM_{Lux}) [17] is selected as in [1, 3, 4, 6]. The selection of the Luxand FaceSDK is motivated by the fact that the SDK has been available in the market for a long time, has a solid recognition performance and does not rely on DCNN for feature learning.

The academic open-source software is represented by the VGG face descriptor (BM_{VGG}) based on the VGG-Very-Deep-16 CNN architecture. We make use of the VGG-face classification model trained using the Caffe framework [18] with 2.6 million face images [19]. The VGG has shown a solid verification performance in the test with the Labeled Faces in the Wild dataset [18] and currently belongs to the leading AFR systems. The input of the neural network is a colour face image of 224 × 224 pixels. The last but one layer of the network returns a 4096-dimensional feature vector describing a face. Prior to feature extraction with the VGG-face model, a face in the image is localised and the face patch is cut and scaled to 224 × 224 pixels. We use the Stasm software library [20] to localise faces. While cutting a face from the image, the size of a face patch is extended by 20%. All VGG feature vectors are normalised in L2. The similarity between face images is computed as one minus the normalised Euclidean distance between the normalised VGG feature vectors.

Both AFR systems operate in verification mode providing us with a similarity score for each pair of face images. The matching scores range from 0 to 1 with higher values for more similar faces. We calculate error rates at the decision threshold with which the FAR does not exceed 0.1% as prescribed in best practice technical guidelines for ABC systems [21]. The corresponding decision threshold suggested in the Luxand documentation is 0.999. The decision thresholds for VGG has been empirically estimated based on the Utrecht ECVP face database and yields 0.49.

3.5 Forensic feature spaces for FMF detection

Three feature spaces are addressed in our benchmark: Benford features (FS_{BF}), keypoint features (FS_{KP}), and the combination of Benford and keypoint features (FS_{Combo}).

3.5.1 Benford features FS_{BF}*:* Application of Benford features to FMF detection is demonstrated in [4] and is already a part of the benchmark in [1]. The motivation for using Benford features stems from [22] with an assumption that natural and artificially generated images have different distributions of DCT coefficients. Thus, the first digits of DCT coefficients (Benford features) should be differently distributed in original and morphed images. All nine Benford features are utilised.

3.5.2 Keypoint features FS_{KP}*:* FMF detector based on keypoint features is introduced in [5] and based on the idea that the blending operation in the morphing process causes a reduction of face details. In morphed face images, the number of corners (keypoints) and edges should become significantly lower. This reduction of details is quantified by the number of detected Scale-invariant feature transform (SIFT) [23], Speeded up robust features (SURF) [24], Oriented FAST and rotated BRIEF (ORB) [25], Features from accelerated segment test (FAST) [26] and Adaptive and generic corner detection based on the accelerated segment test (AGAST) [27] keypoints and CannyEdge [28], SobelX [29] and SobelY [29] edge-pixels [5] in the face region, eight features in total. The number of detected keypoints and edge pixels is normalised for every feature with the natural logarithm of the face pixels because the number of detected keypoints and edges increases non-linearly with the face size. Furthermore, we assume that the JPEG compression applied to original images leads to a higher reduction of details than that applied to morphed images. Hence, we build a lossy self-reference of each sample by applying JPEG compression with the quality set to 75. Then, the same features are extracted from the face region of the compressed image. In order to quantify the quality loss, we introduce next eight features representing the difference between features extracted from original and lossy images. The final feature space has 16 dimensions.

3.5.3 Combined features FS_{Combo}*:* The third feature space FS_{Combo} is a result of feature-level fusion [30] of the feature spaces FS_{BF} and FS_{KP}. The fusion is performed by a simple concatenation of the vectors, i.e. the dimensionality of the new vector FS_{Combo} is equal to the sum of the dimensionalities of FS_{BF} and FS_{KP}. Following the argumentation in [30] on information fusion, we assume that the larger feature space allows for a more accurate classification. This assumption is contradicted in the experiments performed in Section 5.

3.6 Evaluation strategies

One minor change regarding the choice of the classifier is made. While in [1] the support vector machine (SVM) classifier is used [31], here we switch to the Naive Bayes classifier because it achieves comparable classification accuracy while allowing for forensically more mature analyses, because it also reports the

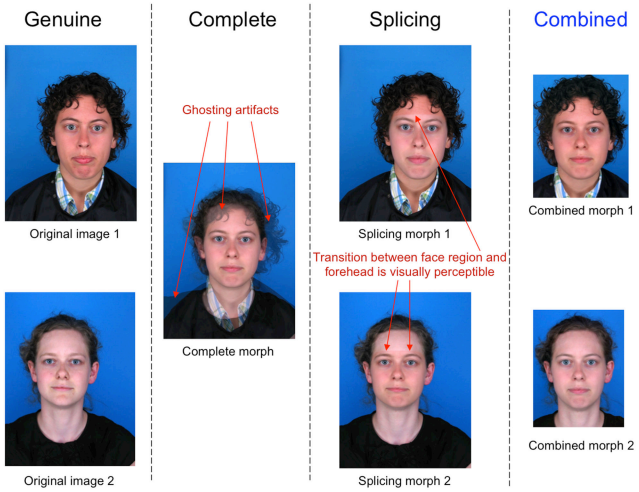

Fig. 2 *Example of morphed face images: complete, splicing and combined approaches; original images are from Utrecht ECVP face database [14] (blue: new in comparison to [1])*

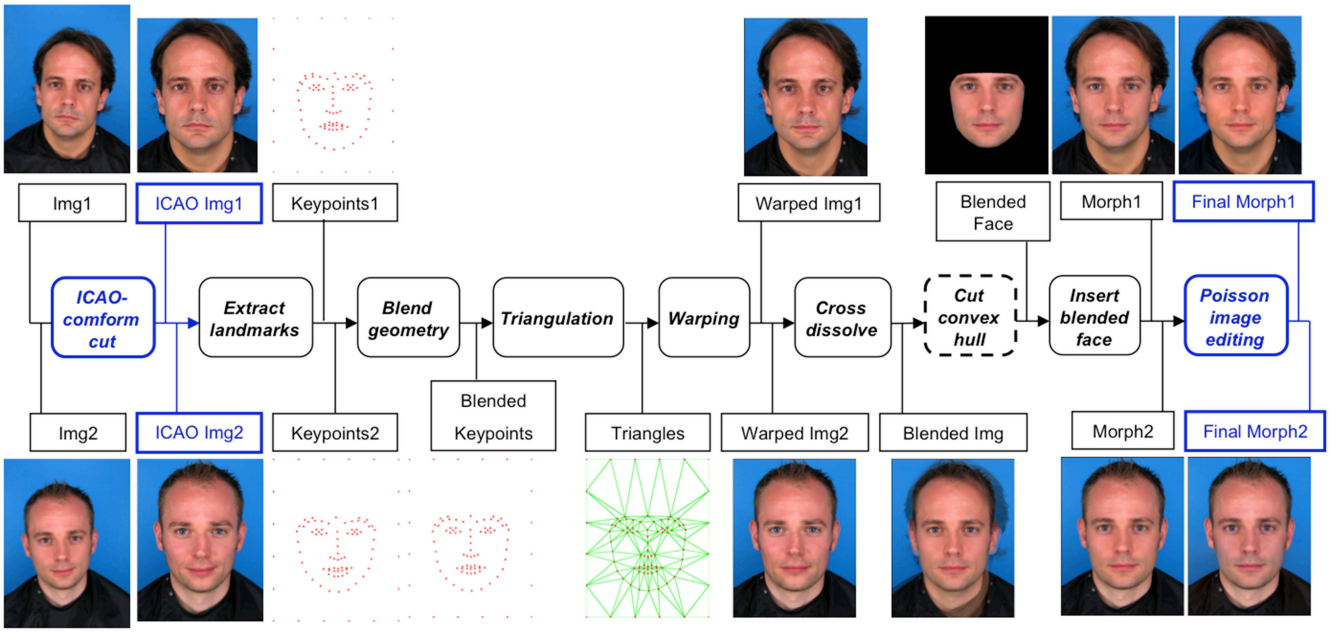

Fig. 3 *Pipeline for generation of combined morphs (FMF_{Bi}); original images are from Utrecht ECVP face database [14] (blue: new operations in comparison to complete and splicing morphs, dashed lines: the operation was moved)*

likelihood ratios (see e.g. [32]). In forensics, likelihood ratios are important to express uncertainties and give a judge in court a sense about the reliability of a decision made by a classifier. However, evaluating likelihood ratios to investigate the impact of *StirTrace* post-processing on both FRRs and morph acceptance rates is a subject of our future work.

4 Benchmarking experiments

4.1 Benchmarking goals

The benchmarking in this paper has four main goals:

- G1: Comparison of the novel FMF realisation (FMF_{Bi}) with the existing FMF realisations (FMF_C, FMF_S) regarding the biometric quality (benchmarking results in Section 5.1).
- G2: Comparison of the novel *StirTrace* post-processing operation (passport scaling 15 kB) with the existing ones to establish the impact on the biometric and forensic qualities of the morphed face images generated by all three FMF realisations (benchmarking results in Section 5.2). G2 is divided into two sub-goals: in G2.1 we focus on the biometric quality and in G2.2 on forensic quality.
- G3: Check whether the expected biometric quality of the morphed face images generated by three FMF realisations is consistent (i.e. provide similar tendencies for the ranking) with the AFR systems (the existing one BM_{Lux} and the one based on the VGG face descriptor BM_{VGG}) (benchmarking results in Section 5.3).
- G4: Check whether the expected forensic quality of the morphed face images generated by three FMF realisations is consistent (i.e. provide similar tendencies for the ranking) with the FMF detectors (the two existing ones M_{BF} and M_{KP} and the one based on the feature-level fusion of these M_{Combo}) (benchmarking results in Section 5.4).

4.2 Benchmarking setup

4.2.1 Training data and setup for FMF detectors: We design three classification models (FMF detectors) M_{BF}, M_{KP} and M_{Combo}, one for each forensic feature space. The models have two classes: 'Genuine' and 'Morphing'. The class 'Genuine' consists of a set of original images, as well as post-processed original images. Four post-processing operations allowed by the ICAO standard [11] with randomly chosen parameters are utilised:

- Cropping: from 60 to 80 pixels at every side (left, right, top, bottom);
- Rotation: angle from − 3° to 3°;
- Scaling: factor from 0.8 to 1.2;
- PassportScaling: fixed resolution of photoID documents 413 × 531.

The other class 'Morphing' is comprised of splicing morphs in native and passport scaled resolutions. We use splicing morphs to make our results comparable to that in [1]. The original images are from the Utrecht ECVP face database [14]. The database consists of 131 facial images of 67 persons. Both classes 'Genuine' and 'Morphing' consist of 655 images ('Genuine': 131 samples + 131 samples × 4 post-processing operations; 'Morphing': 524 randomly selected splicing morphs in native resolution + 131 randomly selected passport scaled splicing morphs). The balanced training set with the equal number of samples in classes should lead to unbiased classification. The training of the models and the classification of the test samples are carried out with WEKA data mining software 3.8.0 [31] using a Naive Bayes classifier with default parameterisation.

4.2.2 Test data for benchmarking: The images of 52 persons with a neutral facial expression from the Utrecht database are the basis for the creation of our FMF test sets. There are 1326 complete morphs, 2614 splicing morphs and 2652 combined morphs generated. The algorithm failed to generate the remaining 38 splicing morphs due to errors in the warping procedure. The datasets are summarised in Table 1.

Since 524 splicing morphs are used for training, the dataset FMF_S contains only 2102 test samples without StirTrace. 19 *StirTrace* post-processing operations are applied to each test sample.

5 Benchmarking results and discussion

As described in Section 3.1.1, the recognition performance of an AFR system with regard to morphed face images is defined by the tuple (FRR, MAR) or (FRR, rMAR). Since the FRR describes the ability of the system to correctly verify genuine images, the FRR value is the same for all FMF realisations. For the sake of completeness, we report the FRR values for the Utrecht database images with both BM_{Lux} and BM_{VGG}. For BM_{VGG}, the FRR_{1000} is always 0% with and without *StirTrace* post-processing. For BM_{Lux}, the FRR_{1000} is 0% except for the following *StirTrace* post-processing operations: scaling to 75% (FRR_{1000} is 0.37453%), rotation by minus five degrees, additive Gaussian noise, cropping to 75%, double scaling and the removal of every 20th column (for all these cases the FRR_{1000} is 0.74906%). The biometric quality of FMF realisations should be compared by comparing the MAR or rMAR values (see Table 2).

Table 1 Summary of test datasets for all benchmarking goals

Dataset	Abbreviation	Samples	*StirTrace* samples
original dataset	DSO	131	2.489
FMF complete morphs	FMF_C	1.326	25.194
FMF splicing morphs	FMF_S	2.614	49.666
FMF combined morphs	FMF_{Bi}	2.652	50.388
total	—	*6.723*	*127.737*

Table 2 Comparison of biometric qualities of the FMF realisations FMF_C, FMF_S and FMF_{Bi} described in terms of morph acceptance rates MAR_{1000} and $rMAR_{1000}$ of biometric matchers BM_{Lux} and BM_{VGG} before and after applying *StriTrace* post-processing operations. Higher MAR_{1000}/$rMAR_{1000}$ values imply higher biometric quality

Filter	Parameter	Benchmarking goals G1, G2.1 and G3											
Metric		BM_{Lux}			BM_{VGG}			BM_{Lux}			BM_{VGG}		
		MAR_{1000}						$rMAR_{1000}$					
Dataset		FMF_C, %	FMF_S, %	FMF_{Bi}, %	FMF_C, %	FMF_S, %	FMF_{Bi}, %	FMF_C, %	FMF_S, %	FMF_{Bi}, %	FMF_C, %	FMF_S, %	FMF_{Bi}, %
baseline (without *StirTrace* post-processing)		**83.26**	**53.77**	**72.83**	**51.88**	**47.22**	**46.91**	**55.20**	**3.21**	**28.17**	**20.21**	**3.18**	**9.58**
PassportScaling	413 × 531	81.95	53.38	70.25	51.80	47.19	46.76	52.26	2.68	24.17	20.29	3.18	9.43
PassportScaling15 kB (G2.1)	15 kb	***79.63***	***52.64***	***66.67***	***51.72***	***47.03***	***46.41***	***47.74***	***2.07***	***19.72***	***20.21***	***3.03***	***9.65***
DoubleScaling, %	25	82.58	53.89	72.10	51.97	47.15	46.68	53.62	3.40	27.68	20.66	3.10	9.35
	50	83.31	53.88	72.69	51.99	47.24	46.72	55.35	3.33	27.90	20.44	3.25	9.43
	75	83.72	53.88	72.47	51.65	47.13	46.92	56.34	3.25	27.87	20.51	3.18	9.47
Rotation, deg.	−5	83.51	53.83	72.88	51.49	47.27	46.40	54.98	3.33	28.39	20.14	3.18	9.20
MedianCut	3	82.70	53.71	72.21	51.84	47.21	46.78	53.47	3.18	26.81	20.51	3.21	9.62
removal of lines	20	76.20	53.01	66.91	51.44	47.12	46.18	41.86	2.56	21.46	20.21	2.98	8.94
	50	82.39	53.60	71.57	51.65	47.17	46.54	53.39	3.18	26.76	20.66	3.02	9.31
removal of columns	20	75.29	53.47	67.91	51.78	46.88	46.04	37.18	2.68	20.85	19.91	3.18	9.16
	50	82.35	53.68	71.50	52.12	47.16	46.67	52.04	2.87	25.68	20.66	3.18	9.62
X-stretching	1.035	79.76	53.32	70.17	51.70	47.07	46.60	47.21	3.10	25.30	20.06	2.91	9.24
rescaling, %	75	82.56	53.59	71.86	51.91	47.13	46.83	52.72	2.87	25.98	20.44	3.25	9.62
	90	83.11	53.88	72.52	51.88	47.05	46.75	54.30	3.37	27.98	20.74	3.21	9.58
	110	83.59	53.87	72.81	51.89	47.15	46.77	55.43	3.21	28.51	20.81	3.18	9.58
cropping, %	75	83.76	53.79	72.09	51.55	47.05	46.47	56.18	3.18	26.70	20.36	3.10	9.24
Gaussian noise	3	***74.28***	***52.21***	***65.31***	***48.28***	***46.38***	***44.78***	***39.14***	***1.87***	***17.76***	***17.42***	***2.87***	***8.30***
uniform noise	3	83.95	54.04	72.62	51.67	47.06	46.78	56.94	3.48	27.79	20.06	3.18	9.65
salt & pepper noise	3	***58.39***	***50.94***	***60.56***	***45.50***	***45.68***	***43.99***	***21.27***	***0.69***	***13.39***	***16.67***	***2.68***	***7.47***

Table 3 Comparison of forensic qualities of the FMF realisations FMF_C, FMF_S and FMF_{Bi} described in terms of TPR and TNR of FMF detectors M_{BF}, M_{KP} and M_{Combo} before and after applying *StriTrace* post-processing operations. Morped face images are positive samples and genuine face images are negative samples. Lower TPR values imply higher forensic quality

Filter	Parameter	Benchmarking goals G2.2 and G4											
Metric		M_{BF}				M_{KP}				M_{Combo}			
Dataset		TNR DSO, %	TPR FMF_C, %	TPR FMF_S, %	TPR FMF_{Bi}, %	TNR DSO, %	TPR FMF_C, %	TPR FMF_S, %	TPR FMF_{Bi}, %	TNR DSO, %	TPR FMF_C, %	TPR FMF_S, %	TPR FMF_{Bi}, %
baseline (without *StirTrace* post-processing)		—	99.92	97.55	89.89	—	80.92	99.52	87.14	—	86.04	99.76	87.59
PassportScaling	413 × 531	—	31.82	24.12	62.89	—	82.65	98.12	98.52	—	59.95	72.19	90.72
PassportScaling15 kB (G2.2)	15 kb	100.00	***0.00***	***0.00***	***0.00***	58.77	93.30	88.89	98.34	100.00	***0.00***	***0.00***	***0.03***
DoubleScaling, %	25	0.00	99.77	100.00	0.00	54.43	99.32	100.00	100.00	45.80	100.00	100.00	99.84
	50	2.29	100.00	100.00	0.00	78.62	88.31	99.84	90.91	76.33	91.55	99.92	88.65
	75	100.00	100.00	100.00	0.00	85.49	84.16	99.65	88.08	84.73	88.89	99.84	85.89
rotation, deg.	−5	49.61	99.92	96.09	100.00	74.80	89.96	99.54	95.81	68.70	92.68	99.69	95.69
MedianCut	3	0.00	100.00	100.00	0.00	54.19	99.62	100.00	99.96	46.56	84.69	100.00	99.39
removal of lines	20	61.06	99.92	98.05	100.00	83.96	77.14	91.81	86.31	82.44	86.34	94.56	83.29
	50	59.54	99.84	98.05	100.00	83.20	81.90	96.05	88.57	86.25	87.02	97.62	86.72
removal of columns	20	60.30	99.92	98.05	100.00	83.96	80.61	88.78	86.65	87.02	84.01	93.72	84.31
	50	64.12	99.92	98.05	100.00	84.73	77.07	98.01	85.33	84.73	91.32	99.27	82.73
X-stretching	1.035	22.90	100.00	100.00	0.00	77.86	87.40	99.84	91.44	75.57	84.23	99.92	89.12
rescaling, %	75	87.02	96.07	88.98	1.47	88.54	80.54	98.60	86.04	90.83	86.80	99.00	85.10
	90	51.90	99.92	98.00	0.00	83.49	81.90	99.38	86.23	83.20	86.80	99.69	84.76
	110	22.13	100.00	100.00	0.00	85.49	81.59	99.57	86.68	81.67	92.15	99.88	81.03
cropping, %	75	76.33	98.86	97.85	0.00	74.80	89.74	99.96	94.75	74.04	92.15	99.96	93.51
Gaussian noise, uniform noise, salt & pepper noise	3	100.00	***0.00***	***0.00***	***0.00***	100.00	***0.00***	***0.00***	***0.00***	100.00	***0.00***	***0.00***	***0.00***

The classification performance of an FMF detector with regard to morphed face images is defined by the tuple (TPR, TNR) (see Section 3.1.2). Since the TNR describes the ability of the system to correctly miss genuine images, the TNR value is the same for all FMF realisations. Hence, the forensic qualities of FMF realisations should be compared by comparing the TPR values (see Table 3). The TNR values of FMF detectors M_{BF}, M_{KP} and M_{Combo} are shown in columns 3, 7 and 11 of Table 3 correspondingly.

5.1 Results for benchmarking goal G1

If the results on the biometric quality are compared in Table 2 using the 'higher-the-more-dangerous' metrics MAR_{1000} and $rMAR_{1000}$, in three of the four comparisons on material without *StirTrace* influences (first row in the table) FMF_{Bi} is ranked second, always being outperformed by the complete morphs FMF_C. Only for the MAR_{1000} measured using the BM_{VGG}, FMF_{Bi} shows the lowest biometric quality but in this test case, the results (51.88, 47.22 and 46.91% resp.) are close together.

5.2 Results for benchmarking goal G2

The main conclusion is on the one hand, that *StirTrace* post-processing operations have no significant influence on the biometric quality of the morphed face images which is reflected in the relatively low variations of MAR_{1000} and $rMAR_{1000}$ values (see Table 2). Only passport scaling 15 kB and noise operators have a not negligible influence to MAR_{1000} and $rMAR_{1000}$.

On the other hand, *StirTrace* post-processing (especially passport scaling 15 kB and noise) has a significant impact on forensic quality of the morphed face images which is reflected in changing TPR values (see Table 3).

5.2.1 Results for benchmarking goal G2.1: All *StirTrace* post-processing operations have a negligible influence on the biometric quality of FMF realisations, except for passport scaling 15 kB and two noise operators (Gaussian and salt & pepper). These post-processing operations decrease the MAR_{1000} and $rMAR_{1000}$ on BM_{Lux} and BM_{VGG} (only the two noise operators) for all three FMF realisations significantly (see values highlighted in bold italic type in Table 3). For all other operations, both AFR systems are robust.

5.2.2 Results for benchmarking goal G2.2: StirTrace post-processing operations have a significant impact on the TPR of all FMF detectors. The passport scaling 15 kB clearly leads to a decrease of TPR of the FMF detectors M_{BF} and M_{Combo} (see values highlighted in bold italic type in Table 3). The addition of any kind of noise leads to critical results because all samples (original and morphed) are classified as 'genuine' by all FMF detectors (see values highlighted in bold italic type in Table 3). Thus, the forensic quality of our morphs increases if post-processing operations like passport scaling or noise are applied to the data because TPR of the FMF detectors drops down. All in all, all FMF detectors react sensibly to post-processing operations.

5.3 Results for benchmarking goal G3

The biometric quality is consistent for both AFR systems. Face morphs with the highest biometric quality are generated by FMF_C, followed by FMF_{Bi} and then by FMF_S. For the BM_{Lux}, the baseline $rMAR_{1000}$ values are 55.20, 28.17 and 3.21% correspondingly. For the BM_{VGG}, the baseline $rMAR_{1000}$ values are 20.21, 9.58 and 3.18%. Based on these values we can say that BM_{VGG} is generally more robust to morphed face images.

5.4 Results for benchmarking goal G4

The expected forensic quality of the morphed face images generated by three FMF realisations differs between the three FMF detectors. M_{BF} delivers the highest TPR (92.72%), followed by M_{Combo} (89.51%) and then by M_{KP} (88.14%) on morphed face images. Especially, the morphed face images created with the

FMF_C realisation, are detected more accurately by M_{BF} (99.92%) than by M_{KP} (80.92%) and M_{Combo} (86.04%). Thus, we can assert that a feature level fusion, does not lead to a higher TPR, but at least a feature selection should be investigated in future work. FMF_{Bi} has the lowest average TPR over all FMF detectors (88.20%) without *StirTrace*, so we could argue, that FMF_{Bi} has the highest forensic quality, followed by FMF_C (88.96%) and then by FMF_S (98.94%). Notice that all FMF detectors are trained with samples generated by FMF_S. For FMF_{Bi} and FMF_S, the classification results of the FMF detectors are nearly similar and the TPR varies not significantly (>3%) without *StirTrace*.

6 Conclusions and future work

Our benchmarking concept addresses four issues: automated generation of face morphs, simulation of anti-forensics by means of applying *StirTrace* post-processing, biometric matching of morphed and original faces to estimate the biometric quality of morphs, and forensic detection of morphs to estimate their forensic quality.

We discuss in this paper the following components:

- a three-fold definition of the benchmarking quality of morphed images (C1),
- an approach to automatically generate combined morphings (C2),
- a *StirTrace* operation (passport scaling 15 kB) to make a digital photograph fit into eMRTD (C3),
- a widely established open-source AFR system based on DCNN (C4) and
- a concept of fusion systems to FMF detection (C5).

Based on the results from Section 5, the FMF_C generates the morphed images with the highest biometric quality. However, considering the strong amount of visual artefacts imposed by the FMF_C, the resulting low visual quality negate the high biometric quality of this FMF realisation. Hence, regarding G1 we conclude that the FMF_{Bi}, newly introduced in this paper, present the most plausible choice of the three FMF realisation in our benchmarking. For G2, the main conclusion is that *StirTrace* post-processing (especially passport scaling 15 kB and noise) has more impact on forensic quality of the morphed face images than on biometric quality. Only the passport scaling 15 kB and noise operators have a significant influence on the biometric quality, while nearly all post-processing operation influence the forensic quality. Furthermore, for benchmarking goal G3 we can assert that the biometric quality is consistent for both considered AFR systems. Face morphs with the highest biometric quality are generated by FMF_C, followed by FMF_{Bi} and then by FMF_S. The last benchmarking goal G4 shows that the expected forensic quality differs between the FMF detectors. FMF_{Bi} delivers the best forensic quality, because it has the lowest average TPR over all FMF detectors.

In our future work, we will concentrate on designing more robust features for FMF detectors as well as on extending our benchmarking framework by new FMF realisations, AFR systems and anti-forensic operations. Based on the results on M_{Combo}, we see the need for a feature selection in feature work. Furthermore, likelihood ratios should be used to better express uncertainties of a decision made by a classifier.

7 Acknowledgments

The work in this paper was funded in part by the German Federal Ministry of Education and Research (BMBF) through the research program ANANAS under the contract no. FKZ: 16KIS0509K.

8 References

[1] Hildebrandt, M., Neubert, T., Makrushin, A., *et al.*: 'Benchmarking face morphing forgery detection: application of stirtrace for impact simulation of different processing steps'. 5th Int. Workshop on Biometrics and Forensics, IWBF 2017, Coventry, UK, 4–5 April 2017, pp. 1–6

[2] Hildebrandt, M., Dittmann, J.: 'Stirtracev2.0: enhanced benchmarking and tuning of printed fingerprint detection', *IEEE Trans. Inf. Forensics Sec.*, 2015, **10**, (4), pp. 833–848

[3] Ferrara, M., Franco, A., Maltoni, D.: 'The magic passport'. IEEE Int. Joint Conf. Biometrics (IJCB), 2014, pp. 1–7

[4] Makrushin, A., Neubert, T., Dittmann, J.: 'Automatic generation and detection of visually faultless facial morphs'. Proc. of the 12th Int. Joint Conf. Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6: VISAPP, (VISIGRAPP 2017), INSTICC, 2017, pp. 39–50

[5] Kraetzer, C., Makrushin, A., Neubert, T., *et al.*: 'Modeling attacks on photo-id documents and applying media forensics for the detection of facial morphing'. Proc. of the 5th ACM Workshop on Information Hiding and Multimedia Security, IHMMSec '17, New York, NY, USA, 2017, pp. 21–32

[6] Ferrara, M., Franco, A., Maltoni, D.: 'On the effects of image alterations on face recognition accuracy'. Face Recognition Across the Imaging Spectrum, 2016, pp. 195–222

[7] Wolberg, G.: 'Image morphing: a survey'. *Vis. Comput.*, 1998, **14**, (8), pp. 360–372

[8] Mallick, S.: 'Face morph using opencv – c++/python'. Available at http://www.learnopencv.com/face-morph-using-opencv-cpp-python/, last accessed 7 July 2017

[9] Raghavendra, R., Raja, K.B., Busch, C.: 'Detecting morphed face images'. 8th IEEE Int. Conf. on Biometrics Theory, Applications and Systems, BTAS 2016, Niagara Falls, NY, USA, 6–9 September 2016, pp. 1–7

[10] Chen, W., Shi, Y.Q., Su, W.: 'Image splicing detection using 2-d phase congruency and statistical moments of characteristic function', 2007, **6505**, p. 65050R–8

[11] Wolf, A.: 'Portrait quality (reference facial images for mrtd)'. Version: 0.06. Published by authority of the Secretary General, 2016

[12] Robertson, D.J., Kramer, R.S., Burton, A.M.: 'Fraudulent id using face morphs: experiments on human and automatic recognition', *PloS One*, 2017, **12**, (3), pp. 1–12

[13] Maio, D., Maltoni, D., Cappelli, R., *et al.*: 'Fvc-ongoing – benchmark area: face morphing challenge', 2018. Available at https://biolab.csr.unibo.it/FVCOnGoing/UI/Form/BenchmarkAreas

[14] Hancock, P.: 'Psychological image collection at stirling (pics) – 2d face sets – Utrecht ecvp'. Available at http://pics.psych.stir.ac.uk/, last accessed 14 July 2017

[15] Perez, P., Gangnet, M., Blake, A.: 'Poisson image editing', *ACM Trans. Graph.*, 2003, **22**, (3), pp. 313–318

[16] Tanaka, M., Kamio, R., Okutomi, M.: 'Seamless image cloning by a closed form solution of a modified Poisson problem'. Proc. of 5th ACM SIGGRAPH Conf. and Exhibition on Computer Graphics and Interactive Techniques in Asia (SIGGRAPH Asia 2012), p.c08-0164-1-1, 2012

[17] Luxand: 'Luxand – detect and recognize faces and facial features with luxand facesdk 2016'. Available at https://www.luxand.com/facesdk/, last accessed 7 July 2017

[18] Huang, G.B., Ramesh, M., Berg, T., *et al.*: 'Labeled faces in the wild: a database for studying face recognition in unconstrained environments'. Technical Report 07-49, University of Massachusetts, Amherst, 2007

[19] Parkhi, O.M., Vedaldi, A., Zisserman, A.: 'Deep face recognition'. British Machine Vision Conf., 2015

[20] Milborrow, S., Nicolls, F.: 'Active shape models with sift descriptors and Mars'. Int. Conf. Computer Vision Theory and Applications (VISAPP), 2014

[21] Frontex: 'Best practice technical guidelines for automated border control (abc) systems'. Tech. Rep., 2015

[22] Perez-Gonzalez, F., Heileman, G.L., Abdallah, C.: 'Benford's law in image processing'. IEEE Int. Conf. Image Processing, September 2007, vol. 1, pp. I-405–I-408

[23] Lowe, D.G.: 'Object recognition from local scale-invariant features'. Proc. of the Int. Conf. Computer Vision-Volume 2 – Volume 2, ICCV '99, Washington, DC, USA, 1999, p. 1150

[24] Bay, H., Ess, A., Tuytelaars, T., *et al.*: 'Speeded-up robust features (surf)', *Comput. Vis. Image Underst.*, 2008, **110**, (3), pp. 346–359

[25] Rublee, E., Rabaud, V., Konolige, K., *et al.*: 'Orb: An efficient alternative to sift or surf'. Proc. of the 2011 Int. Conf. Computer Vision, ICCV '11, Washington, DC, USA, 2011, pp. 2564–2571

[26] Rosten, E., Drummond, T.: 'Fusing points and lines for high performance tracking'. Proc. of the Tenth IEEE Int. Conf. Computer Vision – Volume 2, ICCV '05, Washington, DC, USA, 2005, pp. 1508–1515

[27] Mair, E., Hager, G.D., Burschka, D., *et al.*: 'Adaptive and generic corner detection based on the accelerated segment test'. Proc. of the 11th European Conf. Computer Vision: Part II, ECCV'10, Berlin, Heidelberg, 2010, pp. 183–196

[28] Canny, J.: 'A computational approach to edge detection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1986, **8**, (6), pp. 679–698

[29] Kanopoulos, N., Vasanthavada, N., Baker, R.: 'Design of an image edge detection filter using the sobel operator', *IEEE J. Solid-State Circuits*, 1988, **23**, (2), pp. 358–367, doi: 10.1109/4.996

[30] Ross, A., Nandakumar, K., Jain, A.K.: '*Introduction to multibiometrics*' (Springer US, Boston, MA, 2008), pp. 271–292

[31] Hall, M.: 'The weka data mining software: An update'. SIGKDD Explorations, 2009

[32] Neumann, C., Champod, C., Puch-Solis, R., *et al.*: 'Computation of likelihood ratios in fingerprint identification for configurations of three minutiae', *J. Forensic Sci.*, 2006, **51**, (6), pp. 1255–1266