Abstract—Face morphing attack detection is a challenging task. Automatic classification methods and manual inspection are realised in automatic border control gates to detect morphing attacks. Understanding how a machine learning system can detect morphed faces and the most relevant facial areas is crucial. Those relevant areas contain texture signals that allow us to separate the bona fide and the morph images. Also, it helps in the manual examination to detect a passport generated with morphed images. This paper explores features extracted from intensity, shape, texture, and proposes a feature selection stage based on the Mutual Information filter to select the most relevant and less redundant features. This selection allows us to reduce the workload and know the exact localisation of such areas to understand the morphing impact and create a robust classifier. The best results were obtained for the method based on Conditional Mutual Information and Shape features using only 500 features for FERET images and 800 features for FRGCv2 images from 1,048 features available. The eyes and nose are identified as the most critical areas to be analysed.

Index Terms—morphing, differential morphing attack detection, feature selection.

I. INTRODUCTION

In recent years, ID verification systems have been exposed to variations of presentation attacks. For instance, they compare the user selfie with a picture of the photo ID extracted from the user ID card or passport, where the critical challenge becomes ensuring whether or not the ID card image has been tampered with in the digital or physical domain. Image tampering is a significant issue for such scenarios and biometric systems at large [1].

One of these approaches is related to the passports, and the Morphing attack on face recognition systems based on the enrolment of a morphed face image, which is averaged from two-parent images and allowing both contributing subjects to travel with the passport [1], [2], [3]. Morphing attack detection is a new topic aimed to detect unauthorised individuals who want to gain access to a “valid” identity in other countries. Morphing can be understood as a technique to combine two or more look-alike facial images from one subject and an accomplice, who could apply for a valid passport exploiting the accomplice’s identity. Morphing takes place in the enrolment process stage. The threat of morphing attacks is known for border crossing or identification control scenarios. A morphing attack’s success depends on the decision of human observers, especially a passport identification expert. The real-life application for a border police expert who compares the passport reference image of the traveller (digital extracted from the embedded chip) with the facial appearance of the traveller [4] is too hard because of the improvements of the morphing tools and because of the difficulty for the human expert to localise facial areas, in which morphing artefacts are present.

This work proposes to add an extra stage of feature selection after feature extraction based on Mutual Information $MI$ to estimate and keep the most relevant and remove the most redundant features from the face images to separate bona fide and morphed images. The high redundancy between features confuses the classifier.

The contributions of this work are described as follows:
a) Identify the most relevant and less redundant features from faces that allow us to separate bona fide from morphed images. b) Localise the position of the most relevant areas on the images. c) Visualise the areas that contain morphing artefacts d) Reduce the algorithm’s complexity, sending fewer features to the classifier. e) Analysis of the feature level fusion, the intensity, shape, and texture information. All these contributions may help to guide the manual inspection of morphed images.

This paper is organised as follows: a summary background in features selection and $MI$ is presented in section III-B. The related work is describe in Section II. The methods are described in Section III. The database are described in section IV and the experiments and results are presented in section V and conclusion are presented in section VII.

II. RELATED WORK

Face morphing attack has captured the interest of the research community and government agencies in Europe. For instance the EU founded the iMARS project [1], developing new techniques of manipulation and detection of morphed images.

Ferrer et al. [1] was the first to investigate the face recognition system’s vulnerability with regards to morphing attacks. He has evaluated the feasibility of creating deceiving morphed face images and analysed the robustness of commercial face recognition systems in the presence of morphing.

Juan Tapia and Christoph Busch, da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, e-mail: (juan.tapia-farias@h-da.de, christoph.busch@h-da.de).

Manuscript received xxx; revised xx.
Scherlag et al. [2] studied the literature and developed a survey about the impact of morphing images on face recognition systems. The same author [3] proposed a face representation from embedding vectors for differential morphing attack detection, creating a more realistic database, different scenarios, and constraints with four automatic morphed tools. He also reported detection performances for several texture descriptors in conjunction with machine learning techniques.

Indeed, the NIST FRVT MORPH [5] evaluates and reports the performances of different morph detection algorithms organised in three tiers according to the morph images quality. Tier 1 evaluates low-quality morph images; Tier 2 considers automatic morph images; and Tier 3 for high-quality images. Further, the NIST report is organised w.r.t local (crop faces) and global (passport-photos) morphing algorithms. This fact confirms and shows that morphing images is a problem considering many scenarios.

Most of the state-of-the-art approaches are using machine learning and deep learning to detect and classify morph images. Also, they are utilising embedding vectors from deep learning approaches to detect and classify the images. However, those approaches did not analyse the most relevant features and their localisation on the original images. An efficient feature selection method may help to improve this limitation.

Regarding feature selection, in image understanding, raw input data often has very high dimensionality and a limited number of samples. In this area, feature selection plays an important role in improving accuracy, efficiency and scalability of the object identification process. Since relevant features are often unknown a priori in the real world, irrelevant and redundant features may be introduced to represent the domain. However, using more features implies increasing computational cost in the feature extraction process, slowing down the classification process and also increasing the time needed for training and validation, which may lead to classification over-fitting. As is the case in most image analysis problems, with a limited amount of sample data, irrelevant features may obscure the distributions of the small set of relevant features and confuse the classifiers.

Peng et al. [6] develop a general framework to analyse the interaction between the redundancy and the relevance of the features in a machine learning method to look at the most valuable features based on $MI$.

Guyon et al.[7] proposed the Conditional Mutual Information Maximisation (CMIM) to estimate the relationship of the relevance of the features among three pairs of features.

Vergara et al. [8] proposed an improvement for CMIM [7] approach based on the selection of the first relevant feature. The traditional method maximised the conditional mutual information to select relevant features. This author proposes the average of the $MI$ to reduce the difference among chosen features.

Tapia et al. [9], [10] used the measures of $MI$ to guide the selection of bits from the iris code to be used as features in gender prediction. Also, in [10] used complementary information to create clusters of the most relevant features based on information theory to classify gender from faces.

According to those previous works, we believed that $MI$ is suitable for detecting morphed images to localised and detect the artefact present in morphed images using an efficient number of features.

### III. Methods

Figure 1 shows the proposed framework used in this paper, where a feature selection stage is added after traditional feature extraction approaches.

#### A. Feature extraction

Three different features were extracted from the morphing face images: Intensity, Texture and Shape.

1) **Intensity:** For raw data the intensity of the values in grayscale were used and normalised between 0 and 1.

2) **Uniform Local Binary Pattern:** For texture, the histogram of uniform local binary pattern were used [11]. LBP is a gray-scale texture operator which characterises the spatial structure of the local image texture. Given a central pixel in the image, a binary pattern number is computed by comparing its value with those of its neighbours. The original operator used a $3 \times 3$ windows size. LBP features were computed from relative pixels intensities in a neighbourhood, as is show in the following equation:

$$LBP_{p,R}(x,y) = \bigcup_{(x',y') \in N(x,y)} h(I(x,y), I(x',y'))$$

where $N(x,y)$ is vicinity around $(x,y)$, $\cup$ is the concatenation operator, $P$ is number of neighbours and $R$ is the radius of the neighbourhood.

The uniform Local Binary Pattern (uLBP) was used as texture information. The uLBP was introduced, extending the original LBP operator to a circular neighbourhood with a different radius size and a small subset of LBP patterns selected. In this work we use, ‘U2’ which refers to a uniform pattern. LBP is called uniform when it contains at most 2 transitions from 0 to 1 or 1 to 0, which is considered to be a circular code. Thus, the number of patterns is reduced from 256 to 59 bins.

The reasons for omitting the non-uniform patterns are twofold. First, most of the LBP in natural images are uniform. It was noticed experimentally that uniform patterns account for a bit less than 90% of all patterns when using the (8,1) neighbourhood. In experiments with facial images, it was found that 90.6% of the patterns in the (8,1) neighbourhood and 85.2% of the patterns in the (8,2) neighbourhood are uniform [12]. The second reason for considering uniform patterns is the statistical robustness. Using uniform patterns instead of all the possible patterns has produced better recognition results in many applications. On one hand, there are indications that uniform patterns themselves are more stable, i.e. less prone to noise and on the other hand, considering only uniform patterns makes the number of possible LBP labels significantly lower and reliable estimation of their distribution requires fewer samples. See Figure 2.
3) **Inverse Histogram Oriented Gradient:** From Shape, the inverse Histogram of oriented gradients [13], [14] were used. The Histogram of oriented gradient was proposed by Dalal et al. [14]. The distribution directions of gradients (oriented gradients) are used as features. Gradients, $x$, and $y$ derivatives of an image are helpful because the magnitude of gradients is large around edges and corners (regions of abrupt intensity changes). We know that edges and corners contain more information about object shape than flat regions. However, this descriptor presents some problems. For instance, when we visualise the features for high-scoring false alarms in the object detection area, they are wrong in image space. They look very similar to true positives in feature space. To avoid this limitation that confuses the classifiers, we used the visualisation proposed by Vondrik et al. [13] to select the best parameters that allows us to visualise the artefacts contained in morphed images. This implementation used $10 \times 12$ blocks and $3 \times 3$ filter sizes. One example is shown in Figure 3.

**B. Feature selection**

Feature selection (FS) is the process in which groups of features derived from image areas and textures respectively pixels (in raw images) from facial images out of a dataset are selected based on some measure or the correlation such as F-statistic, Logistic regression or $MI$ between the features and the class of the labels. See Figure 4. It is closely related to feature extraction, a process in which feature vectors are created from the facial image. This takes place through domain transformation or manipulation of the data space and can be considered as selecting a subset of features.

Figure 4 shows a random morphed image with three different correlation metrics. The heat maps show the most correlated features in blue and the less correlated in red. All the features (relevant and redundant) are present in the image.

**Fig. 1:** Framework proposed with feature selection stage.

**Fig. 2:** Example of LBP images. Left: Grayscale image. Middle: traditional LBP (256 bins). Right: LBP with uniform pattern implementation (59 bins).

**Fig. 3:** Example images of inverse HOG. Left: Morphed images. Middle: Traditional HOG. Right: Inverse HOG.

**Fig. 4:** Example images whit different correlation metrics. Red pixels represent the less correlated features.

FS can be classified into three main groups: Filters, Wrappers, and Embedding methods [7]. A filter does not have a
dependency with classifiers when looking for the most relevant features as it. Filters estimates the correlation values according to the MI values. Conversely, wrappers search for the most relevant features according to the classifier. Therefore, if the classifier changes, then the relevant features vary. The embedding method is looking to estimate an optimisation function according to the data and the classifier.

For this work, we propose to use a filter methods based on MI as correlation metrics to estimate the most relevant features to classify bona fide versus morphed face images.

### C. Mutual Information

MI is defined as a measure of how much information is contained jointly in two variables or how much information of one variable determines the other variable [15]. MI is the foundation for information theoretic feature selection since it provides a function for computing the relevance of a variable with respect to the target class [7]. The MI between two variables, \( x \) and \( y \), is defined based on their joint probabilistic distribution \( p(x, y) \) and the respective marginal probabilities \( p(x) \) and \( p(y) \) as:

\[
MI(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \tag{2}
\]

A categorical MI is used in this paper, which can be estimated by tallying the samples of categorical variables in the data building adaptive histograms to compute the joint probability distribution \( p(x, y) \) and the marginal probabilities \( p(x) \) and \( p(y) \) based on the Fraser algorithm [16] for bona fide and morphing images. According to that, if more than two pairs of features reach the same value then, the information is redundant. Conversely, if a couple of features is not contained in any, other pair of features is considered relevant and therefore can help to disentangle and separate the two classes.

If a feature extracted from an image is randomly or uniformly distributed in different classes (bona fide or morph), then the MI between these classes is zero. If a feature is strongly differently expressed for other classes (morph), it should have a large MI. Thus, we use MI as a measure of the relevance of features presented in the images.

The following protocol was used:

- Each image of size \( M \times N \) was flattened to \( 1 \times M \times N \) for each class (bona fide and morphed).
- The matrix \( A \) is formed by \( K \) flattened images of size \( 1 \times M \times N \) features, and the class vector (c).
- MI for each pair of column of matrix \( A \) is estimated.
- The relevance (RI) and redundancy (Rd) are estimated from matrix \( A \).
- The trade-off between the relevance and redundancy (RI and Rd) matrices is estimated, sorted and indexes according to the MI values.
- A vector \( v \) with the index value of each column (feature) with the higher relevance and less redundant is formed.
- Only the \( N \) columns according to with index value are selected.

- A small matrix from \( A \) and element \( v \) is conformed in the step of 100 features up to 1,000 features to be evaluated for the classifier.

Different implementations have been proposed in state-of-the-art [7] to estimate the trade-off between relevance and redundancy. Estimate all the combinations \( 2^N \) to remove all the redundancy is not possible because of high dimensionality problem. Then, the following methods based on MI and Conditional MI have been used and are described as follows:

### D. minimum Redundancy Maximal Relevance (mRMR)

Two forms of combining relevance and redundancy operations are reported in [6]: MI difference (MID), and MI quotient (MIQ). Thus, the mRMR feature set is obtained by optimising MID and MIQ simultaneously. The trade-off both conditions requires to integrate them into a single criterion function [6] as follows:

\[
f_{mRMR}(X_i) = MI(c; f_i) - \frac{1}{S} \sum MI(f_i; f_s), \tag{3}
\]

where, \( MI(c; f_i) \) measures the relevance of the feature \( f_i \) to be added for the class \( c \), and the term \( \frac{1}{S} \sum_{f_s \in S} MI(f_i; f_s) \) estimates the redundancy of the \( f_{ih} \) feature with respect to the previously selected features \( f_s \) to belong to set \( S \).

### E. Normalized Mutual Information Feature Selection (NMIFS)

Estevez et al. [17] proposed with the Normalised Mutual Information (NMIFS) an improved version of mRMR based on the normalised feature of MI. The MI between two random variables is bounded above by the minimum of their entropies \( H \). As the entropy of a feature could vary greatly, this measure should be normalised before applying it to a global set of features as follows:

\[
f_{NMIFS}(X_i) = MI(c; f_i) - \frac{1}{S} \sum_{f_s \in S} MI_N(f_i; f_s) \tag{4}
\]

Where, \( MI_N \) is the normalised MI by the minimum entropy of both features, as defined in:

\[
MI_N(f_i; f_s) = \frac{MI(f_i; f_s)}{\min(H(f_i), H(f_s))} \tag{5}
\]

### F. Conditional Maximisation Mutual Information (CMIM)

The CMIM criterion is a tri-variate measure of the information associated with a single feature about the class, conditioned upon an already selected feature [18]. It loops over the chosen features and assigns each candidate to feature a score based upon the lowest Conditional Mutual Information (CMI) between the features selected, the candidate feature, and the class [7], [18]. Then, the selected feature is the one with the maximum score.

\[
CMIM = \begin{cases} 
\arg\max_{f_i \in F} \{MI(f_i; c)\text{ for } S = \emptyset \} & \text{for } S = \emptyset , \\
\arg\max_{f_i \in F \setminus S} \{\min_{f_j \in S} MI(f_i; c|f_j)\} & \text{for } S \neq \emptyset . 
\end{cases} \tag{6}
\]
G. **Conditional Maximisation Mutual Information-2 (CMIM2)**

The CMIM criterion selects relevant variables and avoids redundancy. However, it does not necessarily choose a variable that is complementary to the already chosen variables. A variable with high complementarity information (max) to the already selected variable will be had by a high (CMI). In general, in problems where the variables are highly complementary (or dependent) to predict c, the CMIM algorithm will fail to find that dependence among the variables. The $\text{CMIM} - 2$ was proposed in order to improve CMIM and changes the max function for the average function $(1/d)$. Then, the selected feature is the one with the average score.

$$MI(x, y) = 1/d \sum_{f_j \in \mathcal{S}} MI(f_i; c \mid f_j).$$

IV. DATABASES

The FERET and FRGCv2 databases were used to create the morph images based on the protocol described by [3]. A summary of the databases is presented in Table I. All the images were captured in a controlled scenario and include variations in pose and illumination. FRGCv2 presents images more compliant to the passport portrait photo requirements. The images contain illumination variation, different sharpness and changes in the background. The original images have the size of 720 × 960 pixels. For this paper, the faces were detected, and images were resized and reduced to $180 \times 240$ pixels. These images still fulfill the resolution requirement of the intra-eye distance of 90 pixels defined by ICAO-9303-p9-2015. The $\alpha$ value to define the contribution of each subject to morph image results was 0.5.

Figure 5 shows examples of the morphing portrait images and the different output qualities with the artefact in their background. For instances OpenCV implementation.

TABLE I: Number of images used for FERET and FRGCv2 Database. Column 1, show the software used to create morph images. The number of images is per dataset.

| Database   | Nº Subjects | Bona fide | Morphs | Probes |
|------------|-------------|-----------|--------|--------|
| FERET      | 533         | 984       | 964    | 1726   |
| FaceFusion | 533         | 984       | 964    | 1726   |
| FaceMorpher| 533         | 984       | 964    | 1726   |
| FaceOpenCV | 533         | 984       | 964    | 1726   |
| UBO-Morpher| 533         | 984       | 964    | 1726   |

The following algorithms were used to create morph images:

- **FaceFusion** is an open-source morphing algorithm, developing for IOS app. This algorithm to create high-quality morph images without visible artifact.
- **FaceMorpher** is an open-source algorithm to create morph images. This algorithm introduce also some artifacts in the background.
- **FaceOpenCV**, this algorithm is based on the OpenCV implementation. The images contain visible artifacts in the background and some areas of the face.

- **Face UBO-Morpher**. The University of Bologna developed this algorithm. The resulting images are of high quality without artifact in the background.

As we mentioned before, after creation of the morphed images, all the faces were cropped using a modified dlib face detector implementation. Figure 6 shows examples of the FERET cropped face database. We can observe that cropped images represent a more challenging scenario because all the background artefacts of the morphing process result were removed. However, some artefacts remain and can be observed in the images, for instances for the FaceMorpher and OpenCV implementations.

V. EXPERIMENTS AND RESULTS

This section presents the quantitative results of the proposed scheme based on feature selection for automated single-morph attack detection. In addition to the proposed system, we evaluated six different contemporary classifiers such as K-Nearest Neighbors (KNN), Logistic regression (LOGIT), Support Vector Machine (SVM), Decision Tree (DT), Random Forrest (RF), and Multilayer Perceptron (MLP). Overall, Random Forest and SVM reached the best results. See Figure 7. To compare and to estimate the baseline method, only the Random Forest classifier was used.

Fig. 7: DET Curves comparing the baseline classifiers using RF. RF and SVM reached the best results. KNN, LOGIT and MLP are not showed in the curve because of poor results.

The experiments, tested a leave-one-out (LOO) protocol and an RF classifier with 300 trees. These datasets allow subject-disjoint results to be computed; that is, no subject has an image in both the training and the testing subset.

The FERET and FRGCv2 databases were partitioned to have 60% training and 40% testing data for feature selection. The selection of features was made using only the training set. The output of the four methods delivers the index of each column of the matrix $A$ that represents the more relevant

2 www.wearemoment.com/FaceFusion/
3 github.com/alyssaq/face\_morpher
4 www.learnopencv.com/face-morph-using-opencv-cpp-python
5 https://www.pyimagesearch.com/2018/09/24/opencv-face-recognition/
features. The number of features were evaluated in steps of 100 features up to the end of the vector.

The performance of the detection algorithms is reported according to metrics defined in ISO/IEC 30107-3. The Attack Presentation Classification Error Rate (APCER) is defined as the proportion of attack presentations using the same attack instrument species incorrectly classified as bona fide in a specific scenario. The bona fide Presentation Classification Error Rate (BPCER) is defined as the proportion of bona fide images incorrectly classified as a morphing in the system. The D-EER is the operation point where APCER = BPCER is reported for the different morphing methods.

A. Experiment 1

Three different kinds of features were extracted from faces. Intensity, HOG, and uLBP. From raw images, we used the values of intensity of the pixels normalised between 0 and 1. For shape, we used the histogram of HOG. For texture, the histogram of the Uniform Local Binary Patterns (uLBP) was used. For the uLBP all radii values were explored from uLBP81 to uLBP88. The fusion of LBPs was also investigated, concatenating the LBP81 up to LBP88 (LBP_ALL). The vertical (uLBP_VERT) and horizontal (uLBP_HOR) concatenation of the image divided into 8 patches also was explored. After feature extraction, we fused that information at the feature level by concatenating the feature vectors from different sources (Intensity, HOG, and uLBP) into a single feature vector that becomes the input to the classifier (FUSION). The classifier was trained with each feature extraction method’s selected features and the fused chosen features.

Table II and III show the baseline results for the intensity, shape and texture feature extraction methods. This baseline was estimated using a leave-one-out protocol for all the morphing methods. The intensity (Raw) and HOG reached the higher D-EER (worst results). Most of the time, the (LBP_ALL) obtained the lower average D-EER results (Best results).

Table II shows the results on the left side for the FERET database were trained with FaceFusion and tested with FaceMorpher, OpenCV, and UBO-Morpher. Right side, trained with FaceMorpher and tested with FaceFusion, OpenCV, and UBO-Morpher.

Table III shows the results on the left side for FERET database were trained with OpenCV and tested with FaceFusion, FaceMorpher, and UBO-Morpher. Right side, trained with UBO-Morpher and tested with FaceFusion, FaceMorpher and OpenCV. The same protocol was applied to Tables IV and V with FRGCv2 database.

B. Experiment 2

This experiment explores the application of the proposed method based on feature selection. The four feature selection methods, mRMR, NMIFS, CMIM, and CMIM2, were applied
TABLE II: Baseline performance reported in % of D-EER for FERET LOO trained on FaceFusion and FaceMorpher.

| Method   | FACEFUSION | FaceMorpher | UBO-MORPHER | Average |
|----------|------------|-------------|-------------|---------|
| Train    | RAW        | 37.47       | 40.47       | 39.91   |
|          | HOR        | 40.6        | 40.04       | 39.91   |
|          | LBP1       | 27.35       | 32.33       | 32.72   |
|          | LBP2       | 24.01       | 27.46       | 29.59   |
|          | LBP3       | 24.88       | 26.92       | 29.61   |
|          | LBP4       | 22.25       | 26.24       | 27.67   |
|          | LBP5       | 25.45       | 26.25       | 28.83   |
|          | LBP6       | 25.79       | 26.98       | 30.57   |
|          | LBP7       | 27.78       | 28.37       | 32.19   |
|          | LBPHOR     | 26.59       | 28.66       | 30.98   |
| FUSION   | 43.64      | 44.56       | 46.88       | 45.03   |

TABLE III: Baseline performance reported in % of D-EER for FERET LOO trained on OpenCV and UBO-Morpher.

| Method   | FACEFUSION | FaceMorpher | UBO-MORPHER | Average |
|----------|------------|-------------|-------------|---------|
| Train    | RAW        | 47.45       | 20.21       | 48.82   |
|          | HOR        | 43.17       | 35.7        | 40.02   |
|          | LBPHOR     | 48.6        | 25.72       | 41.88   |
|          | UBP_VERT   | 37.66       | 23.95       | 39.52   |
|          | UBP_HOR    | 36.08       | 22.72       | 38.52   |
|          | FUSION     | 42.23       | 22.38       | 41.36   |

TABLE IV: Baseline performance reported in % of D-EER for FRGCv2 LOO trained on FaceMorpher and FaceFusion.

| Method   | FACEFUSION | FaceMorpher | UBO-MORPHER | Average |
|----------|------------|-------------|-------------|---------|
| Train    | RAW        | 26.4        | 27.02       | 29.89   |
|          | HOR        | 17.61       | 10.8        | 17.44   |
|          | LBPHOR     | 14.18       | 11.4        | 19.2    |
|          | LBPajor    | 10.58       | 13.36       | 21.67   |
|          | LBPminor   | 11.71       | 15.38       | 22.88   |
|          | LBP5       | 13.44       | 14.46       | 25.25   |
|          | LBP6       | 14.61       | 15.43       | 28.87   |
|          | LBP7       | 15.88       | 15.78       | 26.2    |
|          | LBP8       | 15.96       | 16.29       | 20.94   |
| FUSION   | 14.88      | 28.84       | 45.92       | 39.87   |

TABLE V: Baseline performance reported in % of D-EER for FRGC LOO trained on OpenCV and UBO-Morpher.

| Method   | FACEFUSION | FaceMorpher | UBO-MORPHER | Average |
|----------|------------|-------------|-------------|---------|
| Train    | RAW        | 40.3        | 15.07       | 41.3    |
|          | HOR        | 31.21       | 22.17       | 32.44   |
|          | LBP1       | 20.94       | 13.29       | 17.88   |
|          | LBP2       | 20.52       | 8.43        | 18.89   |
|          | LBP3       | 19.57       | 7.26        | 20.11   |
|          | LBP4       | 20.76       | 8.05        | 23.1    |
|          | LBP5       | 20.74       | 9.12        | 24.03   |
|          | LBP6       | 21.66       | 10.66       | 26.57   |
|          | LBP7       | 22.64       | 10.68       | 26.84   |
|          | LBP8       | 23.5        | 11.94       | 27.78   |
| FUSION   | 17.4       | 16.09       | 27.68       | 19.06   |
in order to reduce the size of the data and estimate the position of the relevant features before entering classifiers from Intensity, HOG, and uLBP. The best 5,000 from 43,200 features were extracted from the raw data (intensity). The best 1,000 from 1,048 features were extracted from HOG, and the best 400 features from 472 were selected from the fusion of uLBP (uLBP_ALL).

Table VI and VII show the results for FERET and FRGCv2 database for single morphed detection from the best feature selected from HOG applied to FaceFusion, FaceMorpher, OpenCV-Morph and UBO-Morpher. The results reported shown an improved in comparison to the baseline in Experiment 1 using the HOG features extracted from the images. The number of feature was reduced on average down to 10%. This reduction would enable the application in mobile devices hardware and also allow us to see the localisation of the most relevant features.

**TABLE VI: D-EER in % of HOG + Fea / FERET.**

| Method       | FaceFusion (bestFea) | FaceMorpher (bestFea) | OpenCV-Morph (bestFea) | UBO-Morpher (bestFea) |
|--------------|----------------------|-----------------------|------------------------|-----------------------|
| mRMR         | 11.15 (400)          | 7.07 (700)            | 6.15 (700)             | 15.68 (100)           |
| NMIFS        | 19.98 (300)          | 9.74 (300)            | 5.84 (800)             | 13.88 (300)           |
| CMIM         | 17.83 (300)          | 5.84 (600)            | 7.07 (500)             | 11.07 (400)           |
| CMIM2        | 8.12 (300)           | 4.92 (900)            | 6.15 (500)             | 13.52 (300)           |

**TABLE VII: D-EER in % of HOG + Fea / FRGCv2.**

| Method       | FaceFusion (bestFea) | FaceMorpher (bestFea) | OpenCV-Morph (bestFea) | UBO-Morpher (bestFea) |
|--------------|----------------------|-----------------------|------------------------|-----------------------|
| mRMR         | 15.06 (1000)         | 2.17 (1000)           | 4.99 (500)             | 4.99 (500)            |
| NMIFS        | 4.83 (600)           | 2.1 (900)             | 1.68 (900)             | 4.73 (700)            |
| CMIM         | 6.50 (700)           | 1.83 (400)            | 2.17 (1000)            | 3.83 (500)            |
| CMIM2        | 6.65 (900)           | 1.92 (900)            | 1.50 (1000)            | 4.02 (600)            |

Table VIII and IX show the results for FERET and FRGCv2 database for single morphed detection from the best feature selected from the fusion of uLBP (LBP8,1 up to LBP 8,8) applied to FaceFusion, FaceMorpher, OpenCV-Morph and UBO-Morpher. The results reported shown an improved in comparison to the Experiment 1 using all the features extracted from the images. The number of feature also is reduced on average down to 10% for texture features.

Figure 8 shows the accuracy obtained for the UBO-Morpher tool when features selected were applied from intensity features. The UBO-Morpher constitutes a high-quality morphing implementation and then is used and analysed on FERET and FRGCv2 databases. Conversely, FaceMorpher is the more straightforward method to be detected based on the artefacts present in the images. The mRMR and NMIFS methods based on MI obtained the lower results. The method based on conditional MI (CMIM and CMIM-2) reached the best results. These results show that the complementary information captures the relationship between the feature selected and the feature candidate in a better way. CMIM with only 400 features and CMIM-2 with 1,000 features reached higher accuracy and lower D-EER.

**TABLE VIII: D-EER in % of Fusion uLBP + Fea / FERET.**

| Method       | FaceFusion (bestFea) | FaceMorpher (bestFea) | OpenCV-Morph (bestFea) | UBO-Morpher (bestFea) |
|--------------|----------------------|-----------------------|------------------------|-----------------------|
| mRMR         | 22.7 (200)           | 12.92 (400)           | 13.84 (400)            | 23.37 (200)           |
| NMIFS        | 21.22 (100)          | 11.84 (300)           | 12.30 (400)            | 23.98 (200)           |
| CMIM         | 21.53 (100)          | 12.30 (200)           | 11.84 (400)            | 23.75 (200)           |
| CMIM2        | 18.45 (100)          | 10.45 (400)           | 11.68 (400)            | 12.11 (200)           |

**TABLE IX: D-EER in % of Fusion uLBP + Fea / FRGCv2.**

| Method       | FaceFusion (bestFea) | FaceMorpher (bestFea) | OpenCV-Morph (bestFea) | UBO-Morpher (bestFea) |
|--------------|----------------------|-----------------------|------------------------|-----------------------|
| mRMR         | 9.15 (400)           | 1.60 (200)            | 8.54 (300)             | 10.47 (300)           |
| NMIFS        | 9.65 (400)           | 1.49 (200)            | 4.50 (400)             | 8.99 (400)            |
| CMIM         | 8.52 (300)           | 1.09 (100)            | 4.17 (200)             | 8.70 (300)            |
| CMIM2        | 7.53 (300)           | 1.33 (200)            | 3.99 (400)             | 8.15 (400)            |

Figure 9 shows the accuracy obtained for the UBO-Morpher tool, when feature selected were applied from HOG features. Again, The method mRMR and NMIFS based on MI obtained the lower results. The method based on conditional MI (CMIM and CMIM-2) reached the best results with 500 and 600 features respectively.

Figure 10 shows the accuracy obtained for the UBO-Morpher tool, when feature selected were applied from the fusion of uLBP. This time NMIFS and CMIM reached the
Fig. 9: FRGCv2 Feature selection for HOG features. X axis represents the number of the best features. Y axis represents the Accuracy in %.

Fig. 10: FRGCv2 Feature selection for uLBP_ALL features. X axis represents the number of the best features. Y axis represents the Accuracy in %.

best results with 300 and 400 features respectively. Consolidating the Conditional $MI$ over traditional $MI$.

Table X shows the D-EER for HOG feature with the best method CMIM-2. Surprisingly, the shape feature (HOG) reached the best results with the lower D-EER using CMIM-2 and FRGCv2 database. FaceMorpher reached the lower D-EER with 1.3% with a BPCER10 of 0.3% and BPCER20 of 1.0%. Conversely, UBO-Morpher reached the higher D-EER of 9.4% with a BPCER10 of 2.9% and BPCER20 of 13.8%. The second column shows, the comparison (D-EER) between the uLBP_all results from baseline using only the fusion of uLBP features versus the proposed method with feature selected from uLBP.

Figure 11 show the DET curves obtained for the four feature selected method for the three feature selected (Intensity, Texture and Shape). The UBO-Morpher constitutes a high-quality morphing implementation and is applied on FERET and FRGCv2 databases. Conversely, FaceMorpher is the more straightforward method to be detected based on the artefacts present in the images. The features selection applied to intensities values reached the lower results. Even these results improve the baseline, the D-EER are not competitive with the literature. Conversely, uLBP and HOG improve a lot in comparison with the baseline and reached results competitive with the literature as is shown in Tables X and XI.

In order to compare and analysed which extracted feature delivers more useful information for the detection task, the Figures 12 and 13 shows a comparison of FERET and FRGCv2 for best results obtained by CMIM-2 from intensity, shape (HOG) and texture (uLBP). Both figures have shown that HOG reached a lower D-EER in both databases. This result shows that the shape algorithms also can detect morphing images as a complement of textures. The exploration parameters to find the most representative inverse HOG features and their visualisation allows us to improve the results. This is shown in Figure 3.

VI. VISUALISATION

Once we select the best features, it is possible to recover the coordinates of the features into the images. Then, we can visualise the attributes for each method. Figure 14 shows the localisation of the most relevant features for an FRGCv2 random image. The 5,000 features were divided into five equal parts and assigned to five different colours. The most relevant features from 1 to 1,000 are represented as red pixels. From 1,001 to 2,000 are pink. 2,001 to 3,000 are green. 3,001 to 4,000 are light green, and 4,001 to 5,000 are represented as blue. It is essential to highlight that the

Table X: D-EER in % for the best results reached by CMIM-2 using HOG.

| Method      | FRGCv2-HOG (D-EER) | BPCER10 | BPCER20 |
|-------------|---------------------|---------|---------|
| FaceFusion  | 27.7 / 5.8          | 3.7     | 7.7     |
| FaceMorpher | 29.1 / 1.8          | 0.3     | 1.0     |
| OpenCV-Morpher | 28.8 / 2.0   | 0.0     | 0.0     |
| UBO-Morpher | 26.4 / 4.0          | 2.0     | 4.4     |

Table XI: D-EER in % for the best results reached by CMIM-2 using All_LBP.

| Method      | FRGCv2-uLBP (D-EER) | BPCER10 | BPCER20 |
|-------------|---------------------|---------|---------|
| FaceFusion  | 14.2 / 9.2          | 7.4     | 20.0    |
| FaceMorpher | 17.2 / 1.3          | 0.3     | 1.0     |
| OpenCV-Morpher | 16.2 / 4.0   | 1.3     | 4.4     |
| UBO-Morpher | 11.7 / 9.4          | 2.9     | 13.8    |
Fig. 11: DET curves for FRGCv2 and FERET using Feature selection method. Top: RAW. Middle: HOG. Bottom: uLBP Fusion.

Fig. 12: D-EER for comparison of the features selected using CMIM from intensity, shape and texture for FERET database. R: represents RAW. H: represents HOG and L, represents uLBP.

Fig. 13: D-EER for comparison of the features selected using CMIM from intensity, shape and texture for FRGCv2 database. R: represents RAW. H: represents HOG and L, represents uLBP.

pixels in colours represent the best features selected, which means the most relevant less redundant from the four methods: mRMR, NMIFS, CMIM, and CMIM2, from 1,000 up to 5,000. The CMIM features are distributed in all the images and only concentrate in some areas. The CMIM-2 focalised the features in the most relevant areas. The eyes and the nose areas are selected as relevant to detect morphed images.

VII. CONCLUSION

After analysing all the results, we can conclude that morphing based on the FERET database is more challenging to detect than the FRGCv2 database. The leave-one-out protocol is essential to estimate the actual performance of the proposed method. In the literature, the test set typically contains images from the same morphing tools. The feature selection reduces the number of features used drastically to separate bona fide for morphed images and reduce the D-EER in all the cases. For the feature selected from HOG, the D-EER decreased from 26.4% (baseline) to 4.0% for UBO-Morpher, reached a BPCER10 of 2.0%. For the chosen feature from the fusing uLBP, the D-EER decrease from 11.7% (baseline) to 8.4% obtained a BPCER10 of 2.9%. These results are very competitive with the state of art. The localisation of the features enabled us to select the most relevant and less redundant features. The nose and eyes are identified as relevant areas in the face for manual analysis of morphed images. This tool may help the border police detect morphing images and address the areas to be analysed for the artefacts. In summary, the shape feature (HOG) results outperform the texture performance as is shown in Figures 12 and 13. In future work, we will apply this method to embedding features extracting from the face-recognition system in order to choose the best features.

ACKNOWLEDGMENT

This work is supported by the European Union’s Horizon 2020 research and innovation program under grant agreement
No 883356 and the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

DISCLAIMER

This text reflects only the author’s views, and the Commission is not liable for any use that may be made of the information contained therein.

REFERENCES

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

[2] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face recognition systems under morphing attacks: A survey,” IEEE Access, vol. 7, pp. 23012–23026, 2019.

[3] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, “Deep face representations for differential morphing attack detection,” IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3625–3639, 2020.

[4] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, “Face morphing attack generation & detection: A comprehensive survey,” 2020.

[5] M. Ngan, P. J. Grother, K. K. Hanooka, and J. Kao, “‘Face recognition vendor test (FRVT) part 4: MORPH - performance of automated face morph detection’ National Institute of Technology (NIST),” 2020.

[6] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.

[7] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Feature Extraction, Foundations and Applications, Studies in Fuzziness and Soft Computing, Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[8] J. Vergara and P. Estevez, “Cmim-2: An enhanced conditional mutual information maximization criterion for feature selection,” Journal of Applied Computer Science Methods, vol. 2, pp. 5–20, 01 2010.

[9] J. Tapia, C. Perez, and K. W. Bowyer, “Gender classification from the same iris code used for recognition,” IEEE Transactions on Information Forensics and Security, vol. 11, no. 8, pp. 1760–1770, 2016.

[10] J. Tapia and C. Perez, “Clusters of features using complementary information applied to gender classification from face images,” IEEE Access, vol. 7, pp. 79374–79387, 2019.

[11] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[12] J. E. Tapia, C. A. Perez, and K. W. Bowyer, “Gender classification from iris images using fusion of uniform local binary patterns,” in Computer Vision - ECCV 2014 Workshops, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham: Springer International Publishing, 2015, pp. 751–763.

[13] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, “Hoggles: Visualizing object detection features,” in 2013 IEEE International Conference on Computer Vision, 2013, pp. 1–8.

[14] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1, 2005, pp. 886–893 vol. 1.

[15] T. M. Cover and J. A. Thomas, Elements of information Theory, C. C. o. N. Y. Donald L. Schilling, Ed. Wiley Series in Telecommunications, 1991.

[16] A. M. Fraser and H. L. Swinney, “Independent coordinates for strange attractors from mutual information,” Physical Review A, vol. 33, no. 2, pp. 1134–1140, Feb. 1986.

[17] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189–201, 2009.

[18] F. Fleuret, “Fast binary feature selection with conditional mutual information,” Journal of Machine Learning Research, vol. 5, pp. 1531–1555, 11 2004.

Juan Tapia received a P.E. degree in Electronics Engineering from Universidad Mayor in 2004, a M.S. in Electrical Engineering from Universidad de Chile in 2012, and a Ph.D. from the Department of Electrical Engineering, Universidad de Chile in 2016. In addition, he spent one year of internship at University of Notre Dame (USA). In 2016, he received the award for best Ph.D. thesis. From 2016 to 2017, he was an Assistant Professor at Universidad Andres Bello. From 2018 to 2020, he was the R&D Director for the area of Electricity and Electronics at Universidad Tecnologica de Chile - INACAP. He is currently a Senior Researcher at Hochschule Darmstadt(HDA), and R&D Director of TOC Biometrics. His main research interests include pattern recognition and deep learning applied to iris/face biometrics, vulnerability analysis, morphing, feature fusion, and feature selection.
Christoph Busch is member of the Department of Information Security and Communication Technology (IIK) at the Norwegian University of Science and Technology (NTNU), Norway. He holds a joint appointment with the computer science faculty at Hochschule Darmstadt (HDA), Germany. Further he lectures the course Biometric Systems at Denmark’s DTU since 2007. On behalf of the German BSI he has been the coordinator for the project series BioIS, BioFace, BioFinger, BioKeyS Pilot-DB, KBEinweg and NFIQ2.0. In the European research program he was initiator of the Integrated Project 3D-Face, FIDELITY and iMARS. Further he was/is partner in the projects TURBINE, BEST Network, ORIGINS, INGRESS, PiDaaS, SOTAM, RESPECT and TReSPAsS. He is also principal investigator in the German National Research Center for Applied Cybersecurity (ATHENE). Moreover Christoph Busch is co-founder and member of board of the European Association for Biometrics (www.eab.org) that was established in 2011 and assembles in the meantime more than 200 institutional members. Christoph co-authored more than 500 technical papers and has been a speaker at international conferences. He is member of the editorial board of the IET journal.