Kraetzer et al. EURASIP Journal on Information Security (2021) 2021:9
https://doi.org/10.1186/s13635-021-00123-4

EURASIP Journal on Information Security

RESEARCH Open Access

Potential advantages and limitations of using information fusion in media forensics—a discussion on the example of detecting face morphing attacks

Christian Kraetzer*, Andrey Makrushin, Jana Dittmann and Mario Hildebrandt

Abstract

Information fusion, i.e., the combination of expert systems, has a huge potential to improve the accuracy of pattern recognition systems. During the last decades, various application fields started to use different fusion concepts extensively. The forensic sciences are still hesitant if it comes to blindly applying information fusion. Here, a potentially negative impact on the classification accuracy, if wrongly used or parameterized, as well as the increased complexity (and the inherently higher costs for plausibility validation) of fusion is in conflict with the fundamental requirements for forensics.

The goals of this paper are to explain the reasons for this reluctance to accept such a potentially very beneficial technique and to illustrate the practical issues arising when applying fusion. For those practical discussions the exemplary application scenario of morphing attack detection (MAD) is selected with the goal to facilitate the understanding between the media forensics community and forensic practitioners.

As general contributions, it is illustrated why the naive assumption that fusion would make the detection more reliable can fail in practice, i.e., why fusion behaves in a field application sometimes differently than in the lab. As a result, the constraints and limitations of the application of fusion are discussed and its impact to (media) forensics is reflected upon.

As technical contributions, the current state of the art of MAD is expanded by:

a) The introduction of the likelihood-based fusion and an fusion ensemble composition experiment to extend the set of methods (majority voting, sum-rule, and Dempster-Shafer Theory of evidence) used previously
b) The direct comparison of the two evaluation scenarios "MAD in document issuing" and "MAD in identity verification" using a realistic and some less restrictive evaluation setups
c) A thorough analysis and discussion of the detection performance issues and the reasons why fusion in a majority of the test cases discussed here leads to worse classification accuracy than the best individual classifier

Keywords: Information fusion, Media forensics, Face morphing attacks, Morph attack detection (MAD), Fusion methods, Fusion ensemble composition

* Correspondence: christian.kraetzer@iti.cs.uni-magdeburg.de
Otto-von-Guericke University Magdeburg, Magdeburg, Germany

1 Introduction

Information fusion has a long research history and its core concept, the combination of outputs of different expert systems, has been rigorously studied and applied for at least two decades in various application domains. The concept of fusion has been studied under many different terminologies, e.g., classifier ensembles [1], combining pattern classifiers [2], or cooperative agents [3]. As a result of the growing popularity of machine learning at that point of time and practical problems arising from ever increasing feature space complexities, in 2002 [4] stated that "instead of looking for the best set of features and the best classifier, now we look for the best set of classifiers and then the best combination method." This statement was rephrased by [5] into "the role of information fusion […] is to determine the best set of experts in a given problem domain and devise an appropriate function that can optimally combine the decisions rendered by the individual experts [...]." In [2], the following three different types of reasons why a classifier ensemble might be better than a single classifier are identified: Statistical (instead of picking a potentially inadequate single classifier, it would be a safer option to use a set of unrelated ones and consider all their outputs), computational (some training algorithms use hill-climbing or random methods, which might lead to different local optima when initialized differently) and representational (it is possible that the classifier space considered for a problem does not contain an optimal classifier). Whatever the exact reason for choosing a fusion approach instead of a single classifier, [2] explicitly warns that "an improvement on the single best classifier or on the group's average performance, for the general case, is not guaranteed. What is exposed here are only 'clever heuristics' [...]". In summary, by combining classifiers (or other expert systems), the applicants hope for a more accurate decision at the expense of increased complexity.

The huge potential for accuracy improvement gained by applying fusion has been well illustrated in many fields of applied pattern recognition. A good example is the field of biometric user authentication where, e.g., [5] shows various benefits that this field can draw from fusion at different steps of the pattern recognition pipeline. When it comes to blindly applying information fusion, among the disciplines that are currently still hesitant are the forensic sciences. Here, the potentially negative impact to classification accuracy as well as the increased complexity (and the inherently higher cost for plausibility validation) of fusion are in conflict with fundamental requirements for (media) forensics (as is discussed in more detail in section 2.1). The goals of this paper are to explain the reasons for this reluctance to accept a potentially very beneficial technique such as information fusion and to illustrate the practical problems of applying fusion. To this end, an exemplary application scenario from media forensics called face morphing attack detection (MAD) is selected. This scenario is currently a hot research topic due to the fact that this kind of attack imposes a recent and currently unsolved threat to face image based authentication scenarios such as border crossing using travel documents (i.e., passports), see section 2.3.

By facilitating the understanding of the reluctance to blindly use fusion in (media) forensics as well as the potential pitfalls of practically applied fusion techniques, it is the hope to facilitate acceptance both in the media forensics community as well as the community of forensic practitioners. To achieve this, the paper provides the following contributions:

a) As general contributions, it is illustrated why (even with a set of classifiers relevant to a specific problem) the naive assumption that fusion would make the detection more reliable can fail in practice, i.e., why fusion behaves in a field application sometimes differently than in the lab and often delivers lower detection performances than single detectors. As a result, the constraints and limitations of the application of fusion are discussed and its impact to (media) forensics is reflected upon. The two main aspects addressed in this discussion are the generalization power of classification models and the relationship between training and test data sets. In the evaluations, it is shown that both aspects, despite being similar in nature, have to be considered separately for applied information fusion.
b) As technical contributions for face morphing attack detection (MAD), the current state of the art is expanded by:

- Introduction of likelihood ratio (LR) based fusion for face morphing attack detection (MAD) to extend the set of methods (majority voting, sum-rule, and Dempster-Shafer Theory (DST) of evidence [6]) used in [7].
- Direct comparison of the two evaluation scenarios: "MAD in document issuing" vs. "MAD in identity verification."
- Analysis and discussion of detection performance issues found with the fusion based detectors (note: questions of feature or classifier selection are out of scope for this paper), the results show that:
- Fusion can fail even when a set of accurate individual classifiers is available. The results presented for fusion detectors are in the vast majority of the cases worse than the results of the best individual classifier used.

- Trained thresholding and weighting strategies as well as sophisticated (context adapted) fusion methods (especially DST and LR based) can under specific circumstances perform significantly worse than unweighted, simplistic fusion approaches like the sum-rule or majority voting.
- Different fusion ensemble composition strategies (i.e., using all available detectors vs. selecting a subset of those) have an influence on the decision error rates.
- For the two evaluation scenarios "MAD in document issuing" (*SC1*) vs. "MAD in identity verification" (*SC2*) different detection and fusion trends are observed, resulting from differences in the inherent characteristics of the application scenario (esp. the amount and type of data available for investigations).

The rest of the paper is structured as follows: section 2 performs a discussion of related work on requirements for media forensic methods, the current state of the art in face morphing attacks detection (MAD) and information fusion approaches in MAD. In section 3, the investigation concept from [7] is summarized and extended into the concept for fusion-based face morphing attack detection used in this paper. Section 4 defines the evaluation setup (incl. the two application scenarios "MAD in document issuing" vs. "MAD in identity verification"). Section 5 presents the evaluation results and their discussion, while in section 6 the conclusions are drawn from the presented results.

2 Related work

Technical capabilities (such as accuracy) are by far not the most significant characteristics of forensic methods. In general, those are usually rated by practitioners in criminal investigations by their maturity, i.e., by their scientific admissibility. Section 2.1 discusses some issues of scientific admissibility in European contexts (where, due to the very nature of the EU and its member states, it is currently much less well regulated as for example in the USA) to establish an understanding on the requirements and limitations for forensic methods originating from this field.

Section 2.2 briefly summarizes the media forensics application domain selected for this paper, the face morphing attack detection (MAD). More detailed overviews over the research activities in this field, which is very active since 2014, can be found in the two survey papers [8, 9].

Several studies have demonstrated that both manually and automatically generated high-quality morphs cannot be recognized as such neither by algorithms nor by human examiners [10–13], and even low-quality morphs pose a threat to the identity verification process if it is completely automated. This explains the urgent need for automated face morphing detectors. At the time of writing this paper, none of the existing research initiatives working on this specific image manipulation detection problem has been able to present detectors that achieve sufficient detection accuracy on a wide range of morphed images (see the ongoing NIST FRVT MORPH challenge [14]). As a logical consequence fusion approaches are used to combine the existing detectors and thereby improve the overall performance. The state of the art approaches in information fusion for MAD are briefly discussed in section 2.3.

2.1 Requirements for media forensic methods in terms of scientific admissibility

When working in media forensics, the question of determining the maturity of methods arises. In lab tests analyzing data for which ground truth information exists, an answer to that question is easy. In that case, the degree of agreement between ground truth label and detector response can simply be used to express the accuracy of the method.

In field applications of forensics, there usually exists no ground truth information for an object under investigation. In these cases, other means of establishing the maturity or suitability of a forensic method have to be used. In forensics, the whole field of work looking into this aspect is termed "scientific admissibility." It is a very complex topic on which Champod and Vuille state in [15]: "The scientific admissibility of evidence, while subject to fairly precise rules in United States law, [...], is seldom addressed in European legal writings, [...]. The question of scientific reliability is seen as intrinsically linked with the assessment of the actual evidence, that is with the determination of its probative value […]." Researchers in the fields of computer science and applied pattern recognition have to rely on the verdict of legal experts defining the hurdles media forensics approaches have to take before achieving the ultimate goal of court admissibility. Looking at [15], it can be stated that there is no EU wide regulation on scientific admissibility questions but that there are common principles that would have to be considered. In that in-depth analysis of the current legal situation in [15] a non-exhaustive list of such principles is presented, containing in its core the following aspects:

- Methods should be peer reviewed and accepted within the corresponding scientific community.
- Error rates associated with a method should be precisely known,
- Existence of standards for the application and maintenance of methods.

This list is very similar to the state-of-the-art criteria used by judges in the USA to address the questions of court admissibility for forensic (and other) methods, i.e., the so called Daubert and FRE702 criteria [15]. While pointing out the benefits of such selection principles, Champod and Vuille also provide some form of criticism into their application: for peer reviewed methods they point out that "this criterion does not indicate whether a technique accepted in scientific literature has been used properly in a given case" and regarding the issue of ascertaining the error rates of a test, they claim that those "can prove misleading if not all its complexities are understood" [15].

In the context of work presented in this paper, those statements imply two important things: First, that a very careful investigation of the precise constrains for the application of a method such as information fusion is required for any specific forensic application case. Second, that the associated complexities in practical application (such as the attempt to improve MAD detection used for illustration purposed within this paper) are clearly and openly discussed.

2.2 Face morphing attacks and their detection

Face images in documents are an established and well accepted means of identity verification. Current electronic machine readable travel documents (eMRTD) are equipped with digital portraits to automate the identity verification process. The automation saves manpower and enhances security due to switching from subjective (officers) to objective (automated face recognition systems) matching of faces. The benefit of automation is especially relevant in high-throughput applications like an airport border control. However, the automation entails the risk of face morphing attacks [16].

In publications such as [12, 16], it has been shown that the blending of face images (here called face morphing) of two or more persons can lead to a face image resembling the faces of all persons involved. Using such an image as a reference in a document is referred to as face morphing attack because it enables illicit document sharing among several users. Such morphing attacks have been shown to be effective in an automated border control (ABC) scenario giving a wanted criminal a chance to cross a border with a chosen (i.e., wrong) identity [10, 17, 18].

Document issuing procedures are different depending on the country and its national regulations. In many countries, the biometric face image can be (and often is) submitted as a hard copy. Here, the attack aims at fooling an officer at the document issuing office by submitting a morphed face image. As long as persons are allowed to submit images to the document issuing office during the document generation, face morphing attacks will remain a severe threat to photo-ID-based verification. Indeed, if an officer accepts a morphed face image, the issued document would pass all integrity checks, and if an automated face recognition (AFR) system matches a live face with a morphed document image, access will be granted to an impostor.

The risk of the morphing attack can be reduced by supporting both officers and AFR systems with a dedicated morph detector. The only way to completely remove the threat of such attacks would be to take the picture directly in the controlled environment of the issuing office and by ensuring that there is no malware-enabled morphing attack embedded into the digital part of the document issuing pipeline, too. The question whether to take the picture directly in place is a political issue, which has in the past lead to many controversial discussions (e.g., in France and Germany) between governmental regulation and the photo industry. But even if this problem would be solved for one country, there would still be the issues of legacy passports (which might still be valid for up to 10 years) as well as foreign documents.

Figure 1 depicts the document life-cycle of a document with a face morphing attack present. While publications such as [19] also discuss the role of forensics (and anti-forensics) in the quality assessment (QA) of the attacker during the morph generation process, in the scope of this paper, only the image forensic analysis of the images submitted into the document creation and the corresponding analysis in every document usage (e.g., in an ABC gate) are relevant. These two investigation points are representing the evaluation scenarios "MAD in document issuing" (*SC1*) and "MAD in identity verification" (*SC2*) considered in this paper. They are discussed in detail in section 4.

The face morphing attack detection (MAD) approaches are typically categorized into two groups regarding whether a trustworthy reference face image is presented or not. The first group is often referred to as single-image or no-reference MAD approaches. The second group is referred to as two-image differential or reference-based MAD approaches. Despite the fact that the reference-based MAD has more potential for robust operation, the non-reference MAD approaches are better represented in the literature.

Within the group of reference-based MAD approaches, as ponted out in [21] there are two subcategories: Reconstruction-based and reference-based MAD. The most prominent examples from the first subcategory try to reconstruct a likely original face (from the assumedly morphed face image provided) by making use of a trustworthy reference face image taken life from the person in front of a camera. This process is often referred to as de-morphing. The detection is done in this

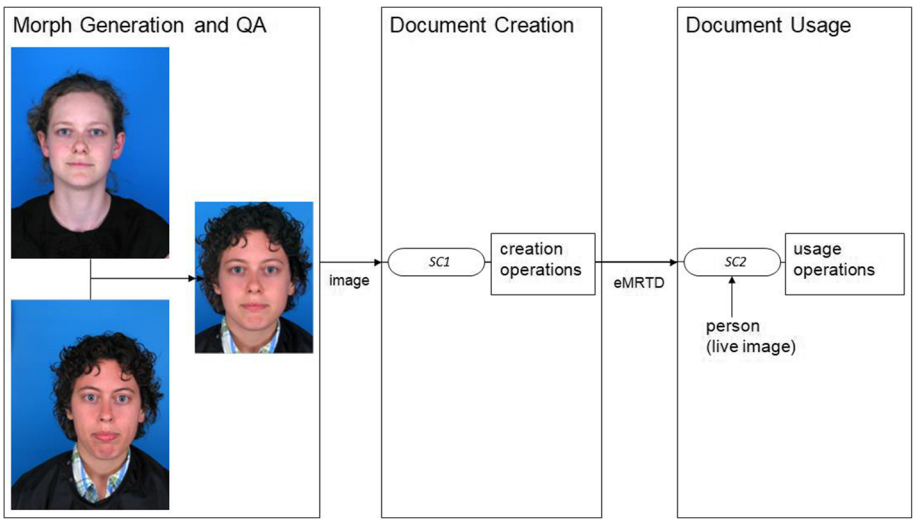

Fig. 1 Document life-cycle in case of a face morphing attack including the evaluation scenarios "MAD in document issuing" (*SC1*) and "MAD in identity verification" (*SC2*) (image derived from [19], combined morph generated based on [12] original face images taken from the ECVP face dataset [20]

case by comparing the reconstructed image and the reference one. The de-morphing is done either by inversion of the common morphing procedure [22] or by applying neural networks such as an autoencoder [23] or generative adversarial networks (GAN) [24]. Alternative approaches to implement reference-based MAD could also be relying on reference feature vectors instead of complete face images.

The approaches from the second subcategory extract features from both presented images (probe document image and trustworthy reference image) and either compare them to each other [13] or combine them for the further classification [25], or even train an additional classifier based on difference vectors [26]. The common problem of all single-image MAD approaches based on "hand-made" or "hand-crafted" features is that they do not detect morphing but rather traces of image manipulations. Since, there is a set of legitimate image manipulations such as in-plane rotation, cropping, scaling, and even some kinds of filtering the morphing characteristics can be easily simulated to prevent detection. The more sophisticated single-image MAD (like [27]) approaches make use of deep convolutional neural networks (DCNN) which are learned to automatically extract features characterizing morphing artifacts based on a large set of samples. If a training set is large and diverse enough covering all frequently used image manipulations, there is a chance that the network will learn not the characteristics of a special dataset, but actual characteristics of morphing. Training of different DCNN architectures for morphing detection was conducted in [17, 26, 28] applying transfer learning with pre-trained networks as well as learning from scratch. In [29], a feature-level fusion of two DCNNs (AlexNet and VGG19) trained by means of transfer learning is shown to outperform BSIF features.

The majority of the aforementioned detectors are learned with morphed face images created by the standard morphing approach which roughly includes three steps: alignment of faces, warping of face components given by polygons (usually triangles), and blending of color values [12, 17, 30]. However, the recent trend is the application of GAN to create realistic face images [31, 32]. The performance of MAD approaches to detect standard morphs and morphs produced by GAN are compared in [33, 34]. Several MAD approaches are compared within the framework of the ongoing NIST FRVT MORPH challenge [14].

2.3 Information fusion approaches in face morphing attack detection

Decision-making systems can be fused at four different levels [2]: data level, feature level, classifier level, and combination (or decision) level. The earlier the fusion is applied, the higher are implementation costs (esp. the computation power required), but also the higher accuracy is expected.

A huge number of different fusion approaches exist, ranging from simplistic methods, like the sum-rule (also known as average rule, meaning the linear combination of matching scores with equal weights) or majority voting to complex schemes like Dempster-Shafer Theory (DST) of evidence [35]. Since DST has a theoretical foundation for handling contradicting and missing decisions of expert systems, it has been successfully applied in a wide range of applications [36]. There, exist

different ways on how to exactly implement fusion based on DST. For details of our own realization, we refer to section 4.3 accordingly.

For the question which fusion method should be chosen, there exists, to the best of the authors' knowledge, no universally agreed upon theory to answer this question. Some experts put a strong focus on one specific method, e.g., Kittler et al. in [37], where the authors claimed that the sum-rule is not only simple, intuitive, remarkably robust, but also outperforms in their experiments all other aggregation operators tested. Other experts, like Ho [4] and Kuncheva [38], explicitly refrain to give any generalized recommendation. Acknowledging the fact that, even when a critical mass of single classification models has been accumulated in a field of application, there are still open questions regarding their combination and the interpretation of the combination output.

If, within media forensics, the field of image manipulation detection is considered (which also contains MAD as a research question) the same wide range of methods are used in research papers, ranging from the simple to complex. A good example in this domain would be the work of Fontani et al. in [39, 40]. In those papers, the authors apply with DST a very sophisticated approach to image manipulation detection task and additionally use its benefits to counter anti-forensics.

A face morphing attack detector is in its nature a binary pattern classifier. The methods for combining such pattern classifiers have been thoroughly studied for a long time, e.g., in [38]. The paper [7] summarizes the state of the art in information fusion for MAD and extends it by introducing DST to this field. The test results presented do show that the error rates with the DST-based fusion are significantly lower compared to those of individual detectors as well as some simplistic fusion approaches applied previously (majority voting and average rule). Here, the work from [7] is used as basis for this paper, taking its fusion framework and extending it even further by including likelihood-based fusion. The reason to do so is the prominent role that the forensic sciences currently attribute to the usage of likelihood ratios in expert testimony, see, e.g., [41] for the example of footwear marks (and underlying forensic analyses, see, e.g., [42]).

While many scientific publications address applying fusion under lab conditions, only very few publications address the question of generalization as well as the applicability for forensic procedures within the context of criminal investigations. In [43], classical probabilities are replaced by Shafer belief functions and an analogy of the Bayes' rule is introduced that is capable to overcome the traditional inability to distinguish between lack of belief and disbelief. Besides mathematical modeling, the consequences of applying the fusion theory for legal practice are discussed. They conclude that there is still a lot of room for explaining the advantages and limitations of using information fusion to forensic researchers as well as the actual practitioners in criminal investigations. Here, the discussion of the advantages and disadvantages of information fusion is continued and its limitations, if applied in real-life conditions, are empirically demonstrated.

3 The concept of fusion-based face morphing attack detection

In theory, a necessary and sufficient condition for a combination or fusion of classifiers to be more accurate than any of its members is that the individual classifiers are accurate and diverse. An accurate classifier has a classification performance better than random guessing and two diverse classifiers make errors on different data points [44]. In practice, experimental evidence has been provided that, for the case of classifiers with a low level of dependence, a consensual decision is likely to be more accurate than any of individual decisions [45]. It has been also shown that lowering correlation among classifiers increases the accuracy of combination [46].

Application of fusion to MAD approaches and especially of the Dempster Shafer Theory (DST) is initially discussed in [7]. In the experiments performed there, the fusion always outperforms individual classifiers in terms of lower error rates. The evaluation concept from this paper is considered here as a reference. It is expanded and it is demonstrated that under certain conditions the superiority of fusion is not always the case. In particular, it is illustrated why the assumption that fusion would make the detection more reliable can nevertheless fail in practice. This enables a discussion on the constraints and limitations of the application of fusion and reflects upon the impact of generalization power of single classifiers as well as fusion methods and the relationship between training and test data sets. Figure 2 roughly depicts the initial evaluation concept.

The concept consists of five major components:

1. The set D of individual morphing attack detectors. Each individual morphing detector is considered as a black box (i.e., they are used as pre-trained methods implying that we have no influence on the training of the classification model). An input for an individual detector is a face image and an output is a score between 0 and 1. High scores indicate morphs and low scores genuine samples.
2. The set of approaches for establishing weights for individual decisions in the fused one. In the case of DST, the mass (belief) functions are required. The process of deriving such parameters is referred to as training in Fig. 2.

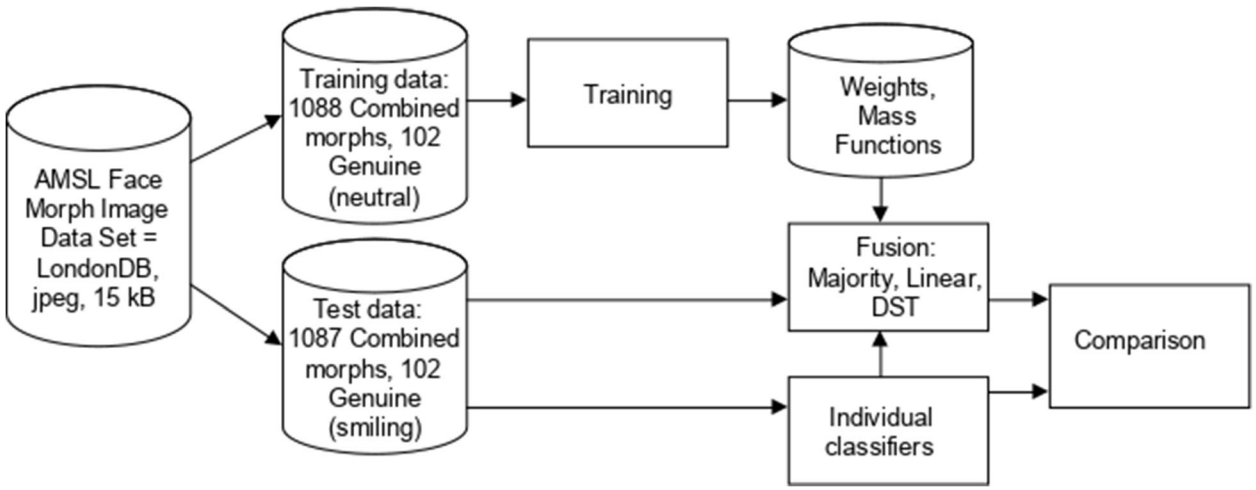

Fig. 2 Evaluation concept for MAD in identity verification from [7]

3. The set of fusion approaches *F*. A fusion approach gets a list of individual decisions and the "importance" of each decision and returns the consensual decision.
4. The evaluation data, which includes training data for establishing fusion parameters (e.g., weights or mass functions) and test data for estimation of error rates. The training and test datasets are created by splitting the AMSL Face Morph Image Data Set (made available via: https://omen.cs.uni-magdeburg.de/disclaimer/index.php). This dataset was initially created to simulate a border control scenario and includes cropped and JPEG-compressed face images which do not exceed 15 kByte and, therefore, fit onto a chip of an eMRTD. In the evaluation, this application scenario is referred to as "MAD in identity verification" (*SC2*). For creating morphed face images, the combined morphing approach from [30] is applied.
5. Comparison of individual detectors and fusion approaches. As a performance metric, we have chosen the error rates of classification approaches.

Here, this concept and its components are re-used and extended by the following: (1) providing a better separation between the training and test datasets by using completely different data sources, (2) adding a fusion approach based on forensic likelihood ratios, (3) adding two types of morphed face images: complete and splicing morphs [12], and (4) adding the application scenario "MAD in document issuing" (*SC1*).

For scientific rigor, it has been ensured in communication with the authors of the MAD approaches that the datasets used for training of the individual detectors do not overlap with the datasets used for training and testing of the fusion approaches.

4 Evaluation setup

Figure 3 depicts the evaluation concept for this paper. The components from [7] and the modifications and extensions summarized in section 3 are apparent in the comparison to Fig. 2.

The representation of the evaluation scenario is done by either using images in their native format and resolution (for application scenario "MAD in document issuing" *SC1*) or in the format specified for ICAO compliant eMRTD (for application scenario "MAD in identity verification" *SC2*). The evaluation scenarios are discussed in more detail in section 4.1. In section 4.2, the used single classifiers for MAD are discussed, while section 4.3 summarizes the fusion methods evaluated (including the strategies for determination of decision thresholds and score normalization). Section 4.4 introduces the performance metrics and 4.5 the databases that are used to create the evaluation data sets.

4.1 Detailed specification of two evaluation scenarios

So far, the evaluation of morphing attack detection (MAD) mechanisms has not been focused on the application scenario. The MAD approaches were rather classified in two groups regarding whether a trustworthy reference face image is presented or not (reference-based vs. single-image/no-reference approaches; see section 2.2). Here, two application scenarios "MAD in document issuing" (*SC1*) and "MAD in identity verification" (*SC2*), representing the two forensic checks required in the document life-cycle of a face image based identity document (see Fig. 1), are considered. Table 1 compares both application scenarios.

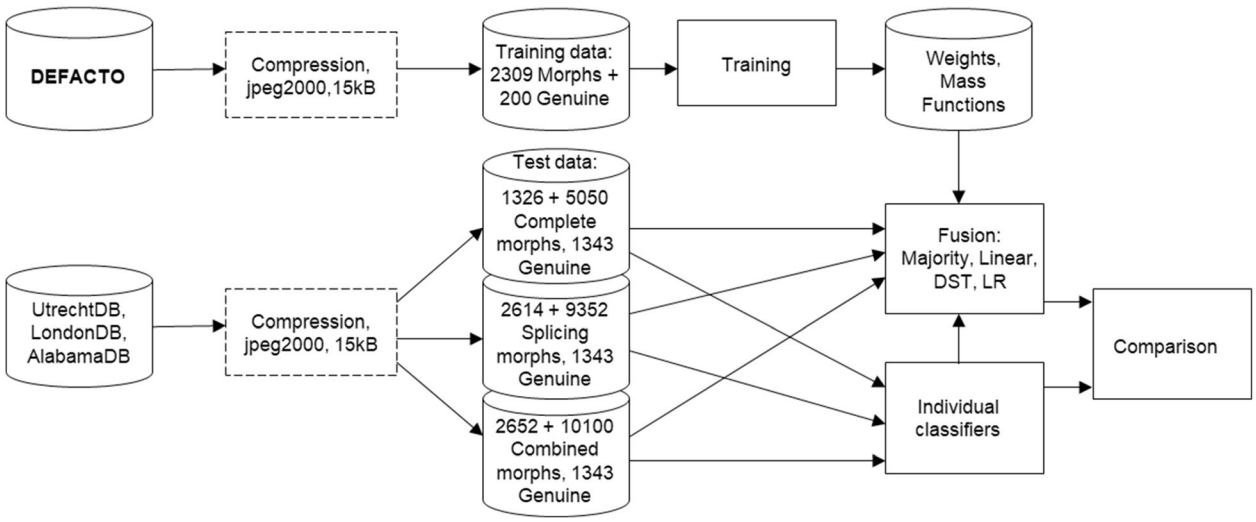

Fig. 3 Our enhanced evaluation concept—switching from "MAD in identity verification" to "MAD in document issuing" is done by deactivating the dashed boxes

The most intuitive mapping would be to link single-image MAD approaches to *SC1* and reference-based MAD approaches to *SC2*. In fact, both application scenarios can be tuned in the way that the reference image is presented. For *SC2*, taking a "live" face image is an inherent part of the procedure. Note that this image could be used solely for face recognition and ignored by the MAD module. For the document issuing in *SC1*, a webcam could be installed next to the officer at the issuing authority, providing a possibility for capturing "live" face images of an applicant.

No-reference MAD approaches are limited to the search for content-independent statistical anomalies or content-dependent visual artifacts caused by the morphing process. Such methods often apply techniques developed within the context of digital image forensics (see section 2.2). Reference-based MAD algorithms try to reconstruct the morphing process aiming at predicting the face of an "accomplice" and comparing this face to the trustworthy "live" image. Hence, the presence of a reference face image rather gives additional options for the choice of detection mechanisms, but does not determine the application scenario.

In contrast, the face image format in *SC2* is very closely defined by national and international regulations, especially by the International Civil Aviation Organization (ICAO) standardization of eMRTD. As a result, the limitations to the digital image that should be stored in an eMRTD are caused by antiquated physical storage limitations. For instance, the current generation of German (and other countries) passports limits the free space for a digital face to 15 kB. During the application for a new document, an applicant submits a printed face photograph of the size of 35 × 45 mm. These images are scanned with the resolution of 300 dpi and undergo lossy compression before they are stored in the passport. The submission of printed face images is in fact the main vulnerability spot making the face morphing attack easy to execute. The reason is that the printing process destroys almost all traces of image manipulation so that human examiners are highly prone to

Table 1 Comparison of the document issuing (*SC1*) and identity verification (*SC2*) scenarios

	Document issuing (*SC1*)	**Identity verification (*SC2*)**
Attack's target	Officer at the document issuing authority	Identity verification system
Time constraints	Up to several minutes	Few seconds (< 2 s)
Face image format	- Low-size printed document image - High-resolution digital image from a certified photo-kiosk	- Low-resolution compressed digital document image - Low-size re-printed document image partially occluded by watermarks
Currently used morphing detection mechanisms	- Naked eye, comparison to the person in front of the desk	- No explicit mechanisms - AFR systems may be set to rejecting at low similarity
Proposed morphing detection mechanisms	- Primarily non-reference (blind) detection - Could be extended by reference-based detection	- Reference-based detection - Demorphing - Could be extended by non-reference (blind) detection

errors when categorizing such images [12]. The straightforward way to reduce the danger of the morphing attack is a prescription to submit high-resolution digital face photographs of a decent quality. Having done this, the image resolution would not be an issue any more for at least a document issuing scenario. As described in section 2.2, taking the picture directly in the controlled environment of the issuing office would limit the threat by morphing attacks. This is not only a political issue but would also require the elimination of further attack vectors.

The file format used in this paper to implement *SC2* is a face image compliant with ICAO specifications for eMRTD: 531 × 413 pixels (inter-eye distance of at least 120 pixels), in JPEG2000 format, compressed to fit the 15 kB size constraint. The file format to implement *SC1* is not that narrowly defined; here, the original file format of the reference databases (see section 4.5) is used.

4.2 Morph attack detection approaches

In this paper, five morph attack detection (MAD) approaches are examined. The first one ($D_{keypoints}$) is based on localization and counting of keypoints [19]. The keypoint-based morphing detector indirectly quantifies the blending effect as an indispensable part of the morphing process. Blending leads to a reduction of face details and therefore to a reduction of "significant corners" and edge pixels. The detector counts the relative number of keypoints in the face region detected by different approaches as well as the relative number of edge pixels. For classification within $D_{keypoints}$, a linear support vector machine (SVM) was trained based on 24-dimensional feature vectors with a dataset of 2000 genuine and 2000 morphed high-resolution passport images. These morphs were created using the approaches from [12, 30].

The other four MAD approaches are based on Deep Convolutional Neural Networks (DCNN). Two of them designated as $D_{ArXivNaive}$ and $D_{ArXivMC}$ are described in [26]. The other two designated as $D_{BIOSIGNaive}$ and $D_{BIOSIGMC}$ are described in [17]. All four of these detectors are based on the VGG19 network. Transfer learning is applied to build a binary classifier from the classification model originally trained for the ILSVRC challenge. The training dataset is comprised of approximately 2000 genuine images and the same number of morphs. Genuine images were collected from several public face databases and scraped from the internet. The major difference between classifiers is in the approach for generation of morphed face images for training. While the $D_{ArXivNaive}$ is an older detector trained with lower quality morphs and $D_{ArXivMC}$ is the same detector with an updated data augmentation strategy in the training, the $D_{BIOSIGNaive}$ and $D_{BIOSIGMC}$ detectors applied for the creation of the training data sophisticated morphing with artificially added high-frequencies to compensate the blurring effect of the blending operation. The differences between the *Naive* training and the *MC* (multiclass/complex morphs) versions lie in the composition of the training data: For *Naive* 50% genuine images and 50% complete morphs are used. For *MC* 50% genuine images and a mix of complete and partial morphs are used, with the aim of forcing the network to take all available information for its decision-making into account (i.e., prevent it from focus on selected face regions like the eyes to detect morphing attacks). The details on the training concept for *Naive* and *MC* versions of the detectors used here can be found in [17].

4.3 Fusion approaches

Here, each MAD approach operates as a "black box" returning a matching score for an input sample. As a consequence of the evaluation concept, fusion on signal level is out of scope for this paper and fusion on feature level (see section 2.3) is not feasible. Hence, the detection accuracy gain from one fusion approach at the decision level (majority voting) and three fusion approaches at the matching score level (weighted linear combination, Dempster-Shafer Theory (DST) of evidence, and forensic likelihood ratios (LR)) is explored. Below, the fusion operators F are described in detail:

4.3.1 Majority voting (F_M)

The naive consensus pattern of simple majority [38] is used for opinion combination. If the number of votes for every alternative is equal, the majority rule returns "no decision."

4.3.2 Weighted linear combination (F_{WLC})

The sum-rule (or weighted linear combination) extends the average rule by assigning different weights to the output of the individual classifiers to be combined. For the case of the same weights, the fusion strategy is often referred to as average rule. Here, two different strategies are used: average rule as well as weighted linear combination with pre-determined weights (see section 5.1 for details on these two strategies).

4.3.3 Fusion based on Depster-Shafer Theory (F_{DST})

The Depster-Shafer Theory (DST) is based on two concepts: belief functions representing degrees of belief for one question from subjective probabilities for a related question and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

In our case, the frame of discernment is defined as $\Theta = \{mor, gen\}$, with m(mor)/m(gen) representing the basic beliefs that the face is morphed/genuine respectively, and $m(\Theta)$ is a mass of uncertainty. A degree of belief (mass) is assigned to each subset. As proposed in [7], we construct mass functions as cumulative distribution functions of matching scores obtained from an experiment. Let $p_{mor}(s)$ and $p_{gen}(s)$ be the approximations of probability density functions of scores for verification

attempts with morphed and genuine images respectively. For a detector outcome s* ranging from 0 to 1, we define the mass m(mor) as an area under $p_{mor}(s)$ between 0 and s* and m(gen) as an area under $p_{gen}(s)$ between s* and 1, and the mass of uncertainty as a complement to the sum of both masses:

$$m(mor) = \int_{s=0}^{s^*} p_{mor}(s)ds, m(gen) = \int_{s=s^*}^{1} p_{gen}(s)ds \quad (1)$$

$$m(\Theta) = 1-(m(mor) + m(gen)) \quad (2)$$

Note that we interpret the detector outcome s* (also called matching score) as a decision confidence with 1 for 100% confidence that the image is morphed and 0 for 100% confidence that the image is genuine.

Technically, the three masses are calculated for each morphing detector based on the matching scores of training samples and stored as a parameter of our fusion engine. At the time of decision-making, for each outcome s_i^* of the i_{th} detector, we obtain the values $m_i(mor)$, $m_i(gen)$, and $m_i(\Theta)$ as the nearest points on the corresponding discrete mass curves.

Dempster's rule of combination for two beliefs from independent sources is given by:

$$m(A \neq O) = \frac{1}{K} \sum_{A=A_1 \cap A_2} (m_1(A_1) \cdot m_2(A_2)) \quad (3)$$

$$K = 1- \sum_{A_1 \cap A_2 = 0} (m_1(A_1) \cdot m_2(A_2)) \quad (4)$$

where m(A) represents the combined mass on A (a given member of the power set), m_1 and m_2 represent the masses of first and second items of evidence respectively, and K represents the normalization constant. The second term in K describes the conflict between two items of evidence. If it is equal to 1 then K is equal to 0 implying that these two items contradict each other and cannot be combined by applying Dempster's rule.

The efficient application of the Dempster's rule for computation of combined belief can be found in [6]:

$$m(mor) = 1-\frac{1}{K}\prod_{i=1}^{n}(1-m_i(mor)) \quad (5)$$

$$m(gen) = 1-\frac{1}{K}\prod_{i=1}^{n}(1-m_i(gen)) \quad (6)$$

$$m(\Theta) = \frac{1}{K}\prod_{i=1}^{n} m_i(\Theta) \quad (7)$$

$$K = \prod_{i=1}^{n}(1-m_i(mor)) + \prod_{i=1}^{n}(1-m_i(gen)) - \prod_{i=1}^{n} m_i(\Theta) \quad (8)$$

4.3.4 Fusion using likelihood ratios (F_{LR})

Likelihood ratios (LR) are used in forensics in order to express uncertainty [47]. The basic concept relies on the quotient of the probabilities of the correctness of two hypotheses with respect to an observation within binary decisions which are common in forensics. Semantically, the LR describe how much more probable one of the hypotheses is in comparison to a complementary one when specific observations can be made.

Within the scope of a forensic comparison of face images, LR are discussed, e.g., in [42] and is already used in some countries in the forensic practice as well, as shown, e.g., in [41] for a case involving footwear marks in the UK. Sometimes the observed LR are mapped to particular levels regarding the confidence in the hypothesis in order to make the result more accessible to forensic laymen as the requirements for particular LR differ between forensic domains, see, e.g., [48]. Generally, a likelihood ratio close to 1 indicates a weak decision as the probabilities for the two hypotheses are almost identical.

With the availability of multiple detection algorithms, a fusion using LR is also possible as suggested, e.g., in [49] for multiple biometric matchers. For each detection algorithm, a quality value needs to be determined as a weight in the fusion algorithm.

In our experiments, the LR for a single detector D providing confidence levels c in a two-class problem is determined by the quotient of the detectors confidence for a sample s toward a genuine sample—$c_D(gen)$—divided by the confidence toward a morphed sample—$c_D(mor)$:

$$LR(s, D) = \frac{c_D(gen)}{c_D(mor)} \quad (9)$$

Note that the inverse of the LRs is used in the experiments performed here, in order to achieve a defined value of zero for a confident decision. Usually the tested hypothesis—in this case whether an image is a morph—would be used as the numerator. As a result, the F_{LR} shows the same behavior. In addition to that, it is possible to normalize F_{LR} using the number of detectors (in this paper 5). Otherwise, this number would have to be taken into account during the interpretation of fusion operator.

The LR-based fusion score F_{LR} of a sample image in question for the k = 5 detectors D = {$D_{keypoints}$,

$D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGNaive}$, $D_{BIOSIGMC}\}$ is determined as the quotient of weighted sum of LRs toward a genuine sample (LR_g) divided by the LRs toward a morph (LR_m) with $LR_g(s,D) = \frac{1}{LR_m(s,D)} = \frac{c_D(mor)}{c_D(gen)}$:

$$F_{LR}(s) = \frac{\sum_{i=1}^{k} LR_g(s,D_i) * w_i}{\sum_{j=1}^{k} LR_m(s,D_j) * w_j} \quad (10)$$

The factor w_i/w_j represents here the weighting factor for the LR fusion as described in section 5.1. A quotient $F_{LR}(s)$ closer to zero indicates a larger confidence of the decision toward a morph.

4.3.5 Normalization

In order to perform a reasonable fusion, the matching scores of the individual classifiers should be brought into the same range. The detectors $D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGNaive}$, and $D_{BIOSIGMC}$ return negative values for genuine faces and positive values for the morphed faces. The default decision threshold is 0. In contrast, the detector $D_{keypoints}$ returns values between 0 and 1. Lower values are for genuine faces and higher values for morphed faces. The default decision threshold is 0.5. Within the training phase performed in this paper using the DEFACTO dataset (see section 4.5), we perform min-max normalization of the matching scores and adapt the default decision thresholds. As a result, the normalized matching scores of all detectors range then from 0 to 1 and the new default decision threshold can be found in Table 3 (column τ_{fixed}). For each classifier, the MIN and MAX values of matching scores are stored to perform the min-max score normalization at the evaluation phase. The aforementioned decision thresholds are also stored as parameters of the fusion and are used in the evaluations in *SC1* and *SC2*.

4.4 Performance metrics

Morphing detection is a standard two class problem with two possible outcomes: "passport image is morphed" or "passport image is not morphed" and two types of errors: morphed image is recognized as non-morphed and vice versa. Driven by the idea that the morphing attack can be seen as a special case of the presentation attack, the detection performance metrics from the presentation attack detection testing standard [50] are adopted. Attack Presentation Classification Error Rate (APCER) describes the proportion of morphed face images incorrectly classified as genuine (bona fide) and Bona Fide Classification Error Rate (BPCER) describes the proportion of genuine (bona fide) face images incorrectly classified as morphed. MAD approaches are typically designed to report two values: a binary decision on whether the image is morphed or not and a confidence score for this decision from the interval [0; 1]. Higher values indicate higher confidence that the image is morphed. In fact, the binary decision is derived from the confidence score by comparing it to an algorithm-dependent predefined decision threshold. Hence, APCER and BPCER are the reciprocal functions of decision threshold. Formally, the BPCER is computed as the proportion of bona fide images over the threshold and the APCER as the proportion of morphed images below the threshold. At the stage of development, when an algorithm can be evaluated with different decision thresholds, the more informative way to compare algorithms is drawing the detection error trade-off (DET) curves (respectively the area under curve (AUC)) on the same plot. Traditionally, BPCER is seen as a convenience measure while APCER as a security measure. The DET curve represents BPCER as a function of APCE R. Here, also the half total error rate (HTER) is used as an average of BPCER and APCER with the fixed decision threshold to compare performances in an easier way.

4.5 Evaluation datasets

There are four databases used in the experiments in this paper: The DEFACTO database [51] containing morphs and genuine face images is used for the training of the fusion methods (see Fig. 2). This database is chosen as a neutral dataset for training because it ensured by the authors that it was not used in the creation (i.e., training) of any of the five used "black box" individual detectors and its used morphing method being unknown. By this choice, a realistic evaluation setup can be ensured, with training data (DEFACTO material) having an unknown similarity to test data (for *SC1* and *SC2*; see Fig. 1), reflecting the constraints that will be encountered in field application. The following datasets (and subsets) are used:

- The DEFACTO dataset contains 200 genuine face images and 39980 morphs. Since using the whole dataset would represent an extremely strong bias toward morphs, only a subset of 2309 randomly selected morphed images is used.
- Three other databases are used to simulate the evaluations conducted within the comparison between single classifiers and fusion methods performances:
  - For two of them (the ECVP (aka Utrecht) [20] and London Set [52] databases) morphed images are generated using the approaches from [12, 30]. The subsets of morphed images are

denoted as *complete*, *splicing*, and *combined* according to the generation method used.
- Additionally, as a source for further genuine face images, mugshots from the Alabama News Network [53] are taken.

Using the original sized images (and morphs based on those), the experiments simulate the passport issuing scenario (*SC1*). In order to simulate the verification scenario (*SC2*), the images are down-scaled (to 413 × 531 pixels) and compressed using the JPEG2000 format in a way that the image size does not exceed 15 kilobyte (kB) as described in section 4.1. Figure 3 shows the exact evaluation concept and Table 2 summarizes the information about the image (sub-)sets used in our experiments.

5 Evaluation results and discussion

This chapter contains a large number of results from different empirical evaluations as well as their interpretation. It is structured as follows:

- Section 5.1 summarizes the DEFACTO experiments, which serve as a baseline as well as an estimator for fusion weights (or mass functions).
- Section 5.2 evaluates the individual detectors and fusion methods (using the full ensemble of detectors) for the two simulated application scenarios *SC1* and *SC2*.
- Section 5.3 discusses the impact of the performed fusion to the field of MAD.
- Section 5.4 determines the impact of using smaller ensembles (i.e., subsets of the available detectors) for fusion.
- Section 5.5 determines the impact of less restrictive assumptions in the evaluation setup composition on the error rates achieved in fusion.
- Section 5.6 provides a final summary and generalization on the obtained results.

5.1 DEFACTO training and baseline experiments

The experiments with the DEFACTO dataset have two objectives:

1. Fair comparison of the MAD approaches to each other regarding their error rates with a disjunctive dataset. In fact, face images in the DEFACTO dataset do not overlap with those used for the training of MAD approaches. Moreover, the morphing procedure with the DEFACTO significantly differs from those with the individual MAD approaches.

2. Training of the fusion parameters including fusion weights and decision thresholds of the individual MAD approaches as well as mass curves for the DST-based fusion. An importance (or in other words a credibility) of one or another detector in the fusion is given by the fusion weight. Here, we consider two thresholding strategies "*fixed*" and "*adaptive*" to define at the same time the decision thresholds and weights (the latter only for F_{WLC} and F_{LR}):

For the "fixed" strategy, we rely on the default decision thresholds suggested by the developers of the MAD approaches and assign equal weights for fusion approaches that accept weights. This trivial strategy (which considers all available detectors as being equally important) is typically the only choice if no additional evaluation of classifiers can be performed, or if there is a suspicion that the evaluation dataset does not fit to the in-field data.

For the "adaptive" strategy, we set a new decision threshold at the point at which the EER of a MAD approach is reached. Additionally, we calculate the fusion weights for F_{WLC} and F_{LR} based on the EER values. To be more precise, the inverse of the EER values are used as weights of the individual MAD approaches in the fusion. Since the possible EER values for a binary

Table 2 Evaluation data sets

Database	Number of images	*SC1* (document issuing)	*SC2* (identity verification)
DEFACTO morphs	2309	tiff, 500 × 652	15kB, jpeg2000, 413 × 531
DEFACTO genuine	200	jpg, 500 × 652	15kB, jpeg2000, 413 × 531
ECVP complete	1326	png, 900 × 1200	15kB, jpeg2000, 413 × 531
London complete	5050	png, 1350 × 1350	15kB, jpeg2000, 413 × 531
ECVP splicing	2614	png, 900 × 1200	15kB, jpeg2000, 413 × 531
London splicing	9352	png, 1350 × 1350	15kB, jpeg2000, 413 × 531
EVCP combined	2652	png, 900 × 1200	15kB, jpeg2000, 413 × 531
Alabama genuine	1343	jpg, image resolution varies	15kB, jpeg2000, 413 × 531

classifier range from 0 (for a perfect classifier) to 0.5 (for a random guess) and the weight should spread over the interval [0, 1], an EER value is multiplied by 2, see Equation (11).

$$w_i = \max(0, 1-2 \cdot EER_i) \tag{11}$$

with i representing one of the five MAD approaches.

Figure 4 shows the DET curves of the five addressed MAD approaches on the original-sized DEFACTO images. Crossings with the dashed black line represent the EER of the detectors. Regarding the EER, three detectors $D_{ArXivNaive}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$ demonstrate comparable performances, with $D_{BIOSIGMC}$ achieving the best performance by a small fraction. The $D_{ArXivMC}$ demonstrates slightly worse performance and the $D_{keypoints}$ is by far the worst detector.

Table 3 demonstrates the EER values of the individual MAD approaches, the decision thresholds τ at which the EER are reached, and the weights assigned to the approaches for fusion for both strategies "fixed" and "adaptive." If the fusion is done at the decision level, the decision thresholds are used to derive decisions from matching scores.

The mass functions for the DST fusion are demonstrated in Fig. 5. The mass curves for the "genuine" and "morphed" matching scores reproduce the classic error curves so that the crossing point indicates the EER.

What can be observed from the results in Table 3 is that $D_{BIOSIGMC}$ outperforms the other four detectors by presenting the smallest EER (resp. the highest AUC). As a result, it is assigned the highest weight for the fusion operations. The results for $D_{keypoints}$ confirm what was already indicated in Fig. 4: Despite its good performance on other image sets, this detector is here performing significantly worse than the other four. As a result, it gets with 0.42 the lowest weight assigned for the fusion.

If the EER locations (the projection of the EER onto the x-axis) and the uncertainty curves shown in Fig. 5 are analyzed, it can be seen that four of the five curves (resp. EER locations) are shifted from the center to the left (indicating a bias toward morphed images) and only $D_{keypoints}$ is shifted to the right with a strong bias toward genuine images. The amount of the shift correlates with the ranking of the detectors: $D_{BIOSIGMC}$ shows the smallest shift (a nearly centered uncertainty curve with a very small skew) while the other four show an increase in the shift (and skew) with their higher EER.

5.2 Experiments with individual detectors and fusion methods

The sections 5.2.1 and 5.2.2 summarize the results on the performance of the individual detectors and fusion methods evaluated with the two simulated application

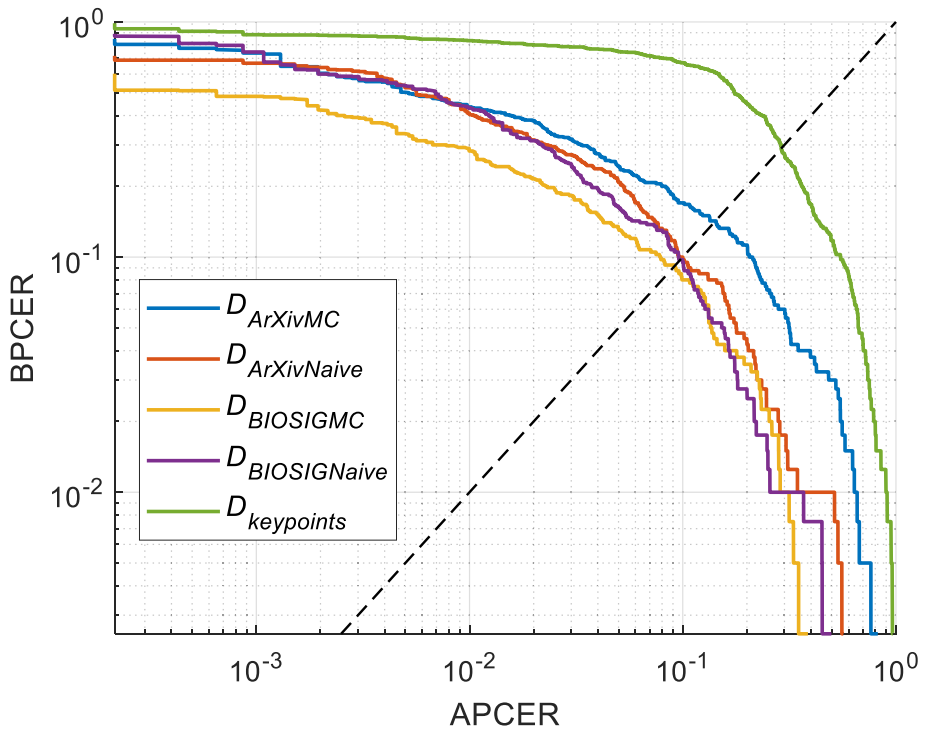

Fig. 4 DET curves of the individual detectors with the DEFACTO dataset (original-sized images)

Table 3 Evaluation of detectors with the DEFACTO dataset and associated weights

Detector	AUC	EER	$\tau_{adaptive}$	$w_{adaptive}$	τ_{fixed}	w_{fixed}
$D_{ArXivMC}$	0.94	0.14	0.35	0.72	0.47	1.00
$D_{ArXivNaive}$	0.97	0.10	0.40	0.80	0.59	1.00
$D_{BIOSIGMC}$	0.98	0.09	0.48	0.82	0.53	1.00
$D_{BIOSIGNaive}$	0.97	0.10	0.36	0.81	0.52	1.00
$D_{keypoints}$	0.77	0.29	0.87	0.42	0.50	1.00

scenarios *SC1* and *SC2*. All these tests use as data the combined images from the ECVP, London and Alabama datasets (see section 4.5). For *SC1* the original-sized images are used and for *SC2* the 15 kB versions.

5.2.1 Scenario SC1 ("MAD in document issuing")

Figure 6 shows the DET curves for the tests on complete, splicing, and combined morphs in *SC1*. The individual classifier performance is displayed by solid lines (with the same color coding as in Fig. 4), and the performance of the fusion methods is given as dashed lines (where a continuous space of operation points is possible) or symbols (in case only one operation point, either the "fixed" setting or the "adaptive," is possible).

For all three morphing types, the individual classifier $D_{arXivNaive}$ achieves the best performance for *SC1*, followed by the weighted linear combination (F_{WLC}). The three single classifiers $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, and $D_{keypoints}$ show the lowest performance. F_M with "fixed" and "adaptive" thresholding strategy achieve the lowest performance of the fusion methods. The more sophisticated fusion operators (F_{DST} and F_{LR}) perform better than F_M, in some cases F_{DST} even outperforms F_{WLC}, but both show a significant bias toward morphed images. Especially for F_{DST}, this is apparent with an APCER close to 0 at a BPCER of roughly 0.2.

5.2.2 Scenario SC2 ("MAD in identity verification")

Figure 7 shows the DET curves for the tests on complete, splicing, and combined morphs in *SC2*. The same color coding and symbols are used as in Figs. 4 and 6.

The general performances of the individual and fusion based detectors in *SC2* are very similar to the *SC1* results shown in Fig. 6. A slight decrease in the detection performances can be observed for all tested methods. This decrease can be attributed to the fact that the 15 kB image format that is used in *SC2* leaves generally less room for media forensic investigations on image manipulation. What is remarkable in the results is that the results of the more sophisticated fusion operators (F_{DST} and F_{LR}), while also showing some performance decrease, loose some of their bias toward morphed images. Especially for the splicing morphs, it can be observed in Fig. 7 that F_{DST} shows an APCER larger than 0, even slightly outperforming at the corresponding APCER values all other detectors.

5.3 Discussion of the impact of fusion to face morphing attack detection

Tables 4, 5, and 6 summarize the results. Table 4 demonstrates a baseline using only the individual classifiers,

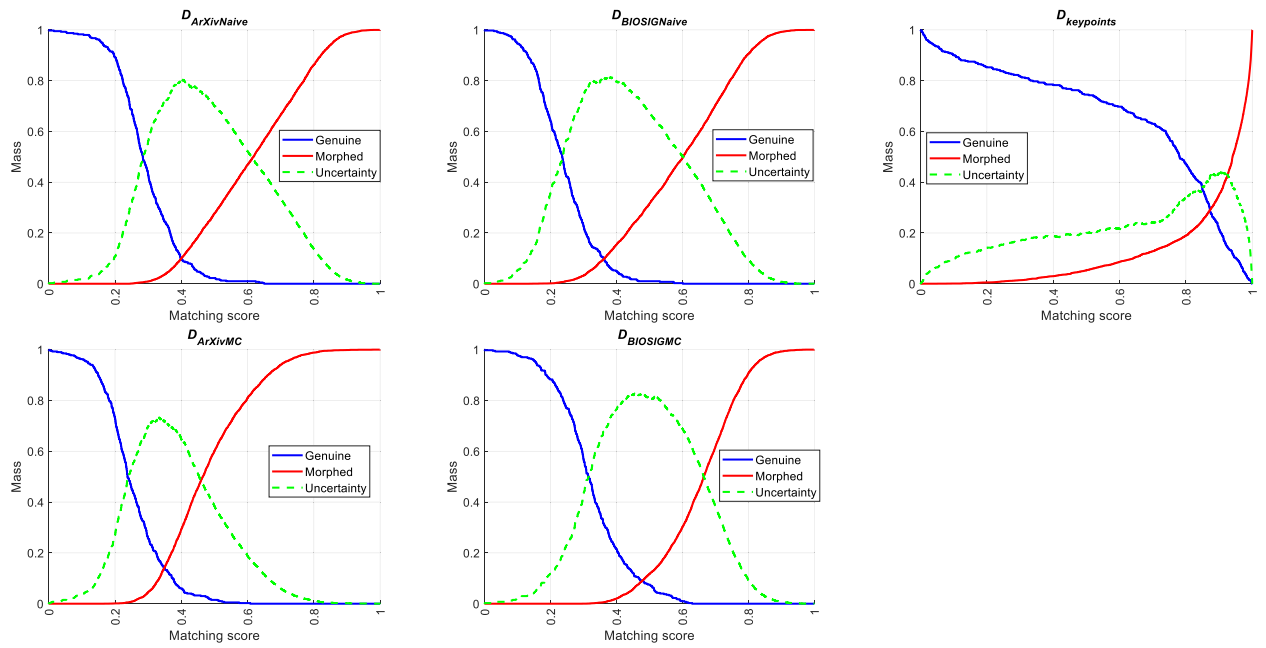

Fig. 5 Mass functions for the DST fusion resulting from the evaluation with the DEFACTO dataset (original-sized images)

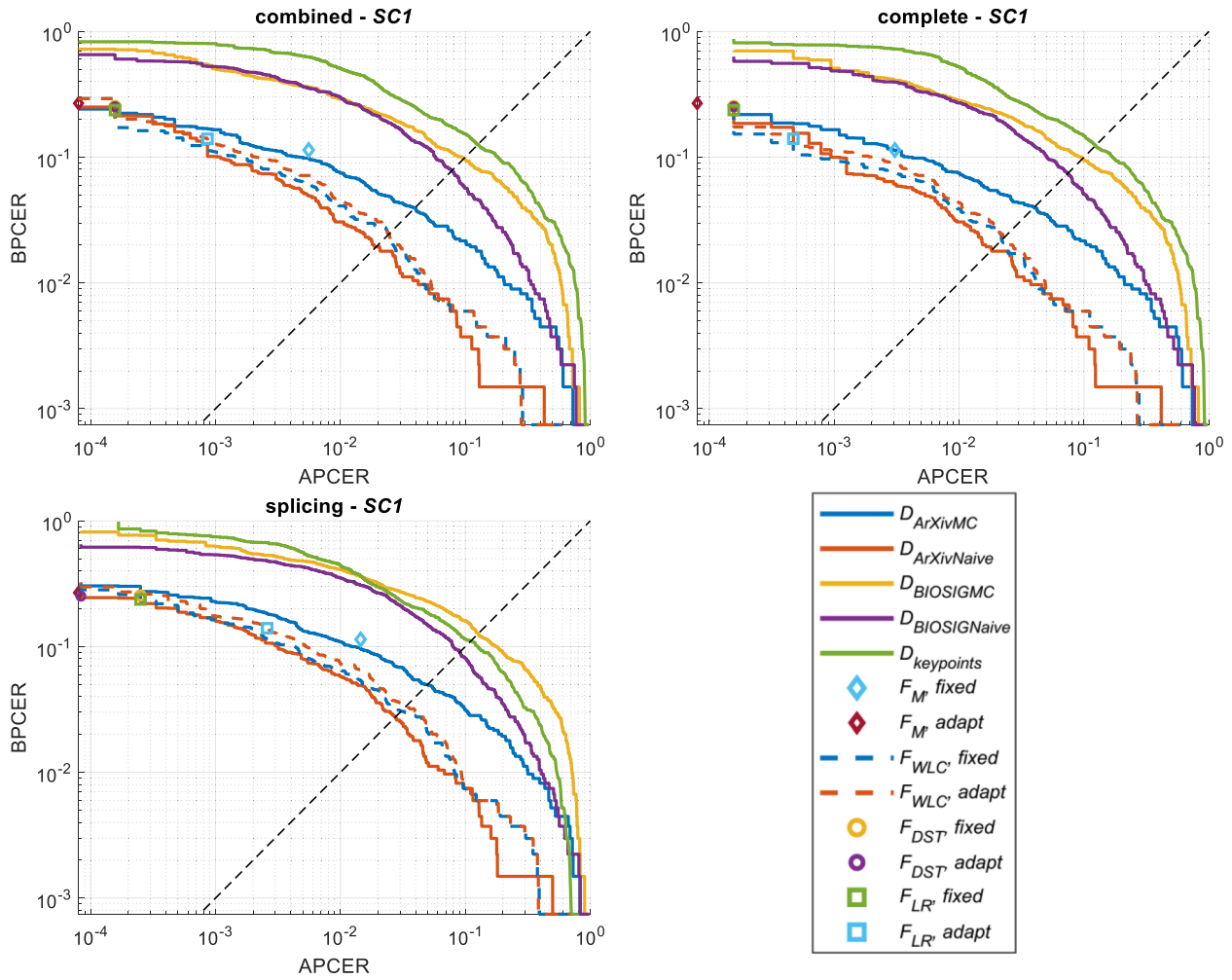

Fig. 6 DET curves for the tests on complete, splicing, and combined morphs in *SC1*

showing that $D_{ArXivNaive}$ performs best in testing in both application scenarios *SC1* and *SC2* on all three morph types.

Tables 5 and 6 present the single classifier and fusion results in the "fixed" (Table 5) and "adaptive" (Table 6) thresholding strategies. The difference lies in the basic assumption for the similarity of training data (here DEFACTO) and the material encountered in field application (here, the mix of ECVP, London, and Alabama material, either in original (for *SC1*) or the 15 kB version (*SC2*)). While the "adaptive" setting is the setting encountered in most lab experiments, the "fixed" one (which assumes a much lower similarity between training and test data) is a more realistic assumption, leading to more trustworthy error estimates in this media forensic analysis.

When focussing on the single classifier results obtained for both thresholding strategies ("fixed" decision threshold and fusion weights vs. "adaptive" decision threshold and fusion weights), it can be seen that $D_{BIOSIGMC}$, which performed best on the DEFACTO dataset (see Fig. 4 in section 5.1) demonstrates in the evaluations significantly worse performance in both application scenarios *SC1* and *SC2*. In Fig. 4, in two of the six tests (the two evaluations run on splicing morphs), it actually shows the lowest performance (i.e., highest HTER). When looking at Tables 5 and 6, these results are confirmed. For both thresholding strategies and all three different morphing types, $D_{BIOSIGMC}$ achieves the second lowest detection performances, followed only by $D_{keypoints}$. The best performance for a single classifier is in all cases achieved by $D_{arXivNaive}$ with the "fixed" decision threshold.

When comparing the single classifier and fusion results in Tables 5 and 6, the general picture established in section 5.2 is confirmed: In nearly all cases for *SC1* as well as *SC2*, the fusion approaches fail to outperform the best individual detector. Neither for selected morphing

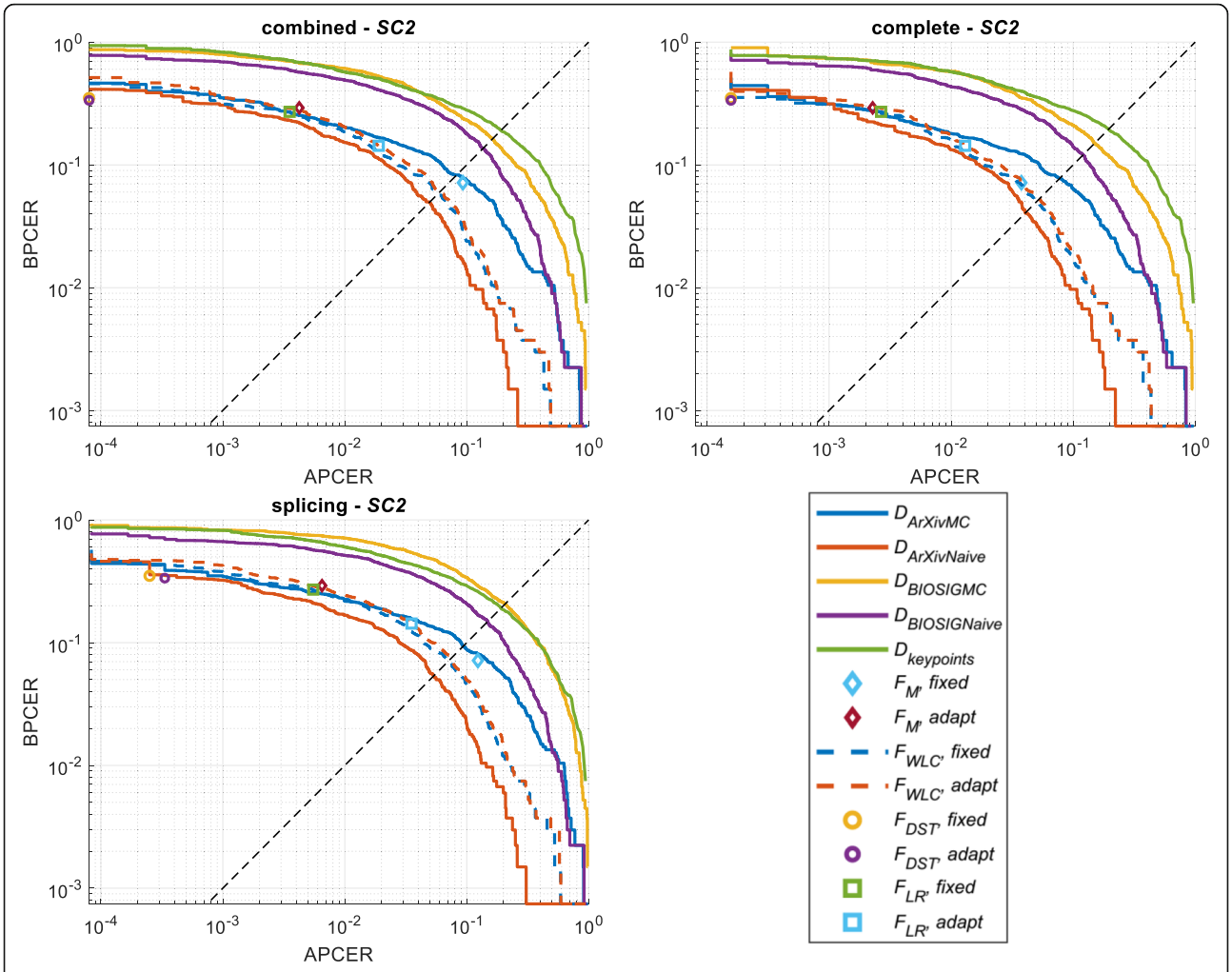

Fig. 7 DET curves for the tests on *complete*, *splicing*, and *combined* morphs in *SC2*

approaches nor for one of the two thresholding strategies, the fusion generally outperforms the best single classifier, even though in one case for *SC2* and splicing morphs it is close (best single is $D_{arXivNaive}$ with "fixed" at an HTER of 8.5% and the best fusion is F_{LR} with "adaptive" and an HTER of 8.92%). Most interestingly, the DST-based fusion, which is the most sophisticated fusion strategy and which is highly regarded in many other application fields, leads here in all cases to low performances.

For the thresholding strategies, it can be summarized that for the four classifiers $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$, there is a tendency that the best results are obtained with the "fixed" decision threshold while for $D_{keypoints}$ in the majority of the cases better results are obtained with the adaptive decision threshold. For the fusion, no clear tendency which thresholding strategy leads to better results can be observed.

When considering the differences in the detection performance for the three tested morph types (*combined*, *complete*, and *splicing*), it can be summarized that all detection approached discussed here yield very similar detection performances (both in *SC1* as well as *SC2*).

5.4 Variation of the fusion ensemble

During the review phase for this journal paper, the reviewers raised the question why it is assumed that a fusion using all five single classifiers is the optimal choice at hand. Alternative fusion ensembles using three or four classifiers might be capable to outperform the whole set of five, especially when removing the weakest candidate ($D_{keypoints}$). To address this issue, Table 7 compares the results of three different sets of fusion ensembles for the "fixed" decision thresholds. The results shown are for the complete set of 5 detectors as baseline, the best performing ensemble of 4 (here $D_{BIOSIGNaive}$, $D_{BIOSIGMC}$,

Table 4 Theoretical performance of the individual detectors with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1		*SC2*	
Detector	**Morph type**	**EER**	**$\tau_{adaptive}$**	**EER**	**$\tau_{adaptive}$**
$D_{ArXivMC}$	Combined	3.95%	0.528241	8.27%	0.417467
$D_{ArXivNaive}$		**1.94%**	0.594687	**4.96%**	0.499938
$D_{BIOSIGMC}$		9.75%	0.617098	16.31%	0.561516
$D_{BIOSIGNaive}$		7.74%	0.558175	13.56%	0.478364
$D_{keypoints}$		12.65%	0.971509	19.08%	0.990942
$D_{ArXivMC}$	Complete	4.00%	0.526729	7.75%	0.424468
$D_{ArXivNaive}$		**1.82%**	0.600357	**4.06%**	0.507648
$D_{BIOSIGMC}$		9.75%	0.616997	14.98%	0.565297
$D_{BIOSIGNaive}$		7.45%	0.563476	12.07%	0.496052
keypoints		12.43%	0.972011	19.53%	0.990758
$D_{ArXivMC}$	Splicing	4.99%	0.501098	9.37%	0.406828
$D_{ArXivNaive}$		**2.76%**	0.566983	**5.37%**	0.492199
$D_{BIOSIGMC}$		12.67%	0.594189	20.64%	0.541497
$D_{BIOSIGNaive}$		9.08%	0.54235	14.84%	0.470685
$D_{keypoints}$		11.09%	0.976876	19.08%	0.990933

$D_{arXivNaive}$, and $D_{arXivMC}$; the evaluations performed in this case were a complete leave one out sequence but only the most relevant result is presented here) and the ensemble of three with the most disparate characteristics ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$; i.e., selection by limiting redundancy). The results show an apparent decrease of the HTER for *SC1* and *SC2* if switching from an ensemble of 5 (denoted as "5 det" in Table 7) to an ensemble of (the most suitable) 4 detectors (denoted as "4 det" in Table 7). When compared to the single detector performance reported in Table 5 above, it can be seen that the best ensemble of 4 also seems to outperform the individual detectors. Some of the figures presented have to be considered very carefully since they are hiding a problem in the scheme: This is absolutely no problem for cases where the individual weighting makes deadlocks neigh to impossible (e.g., in case of the F_{WLC}) but is especially relevant for the majority vote where significant numbers of "undecided" events occurred (e.g., cases where 2 detectors predicted one class and the other 2 the other) that are not reported in the table. These "undecided" events amount over the various tested ensembles to up to 10% of all majority vote cases.

In case of the chosen ensemble of 3 detectors (denoted as "3 det" in Table 7) all HTER values increased significantly, showing that this ensemble (which more strongly relies on the opinion of the rather weak $D_{keypoints}$) is outperformed by the bigger ensembles.

Similar to Table 7, Table 8 performs the same ensemble tests for the "adaptive" thresholding strategy. Here, the results also show better results for the best ensemble of 4 detectors when compared to the complete ensemble of 5. In contrast to the "fixed" thresholding strategy discussed above, the performance increase obtained by leaving $D_{keypoints}$ out seems smaller but also the number of "undecided" events is way smaller (less than 3%) so that here the gain has to be considered higher. This performance gain is also evident in the comparison to the single detector results discussed in Table 6.

Like in the case of the "fixed" thresholding strategy, the tested cases of 3 detector ensembles showed significantly worse results, increasing the HTER to 18% or even higher.

Summarizing the results on these detector ensemble selection experiments, it has to be said that the best performing set of 4 detectors outperformed for both thresholding strategies ("fixed" and "adaptive") and *SC1* as well as *SC2* the complete ensemble of 5. For fusion methods that are prone to deadlock or "undecided" situations (esp. the majority vote), the even number of detectors in this cased caused a small issue, generating in the worst case up to 10% deadlock results that would have to be handled in application. All results for the chosen ensemble of the 3 most dissimilar detectors proved near fatal for the system performance since the HTER was significantly increased in all these cases.

5.5 Discussion on alternative evaluation setups

Another issue, raised during the review phase for this journal, is the choice of a realistic but rather challenging experimental scenario where the dataset used for training is disjoint from the ones used for testing. The question was how an overlap between training and testing set (i.e., more favorable conditions for the individual detectors) would influence the outcome of the experiments. To address this question, two different sets of less realistic experimental setups are discussed below: first, a tenfold stratified cross-validation with disjoint sets of genuine samples and morphs, and second an even less realistic (i.e., more lab-condition) test with a static percentage split on one a set containing genuine and morphs that are derived directly from these genuine images.

For the first of these alternative setups, additional tests are performed here to show how a deviation from rigorous evaluation routines reflects in the error rates obtained. Table 9 summarizes the results for the "fixed" as well as the "adaptive" thresholding strategy. If comparing the results in Table 9 to the results in Tables 4 and 5, then the single detector performances in the "fixed" thresholding remain nearly unchanged while the HTER values in case of the fusions decrease (e.g., from 11.85% to 2.6% in case of F_{LR} in *SC1* for *combined* morphs of from 13.70% to 5.9% in case of F_{LR} in *SC2* for *combined*

Table 5 Realistic performance of the individual detectors and fusion approaches with the fixed decision thresholds and equal fusion weights with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
$D_{ArXivMC}$	Combined	7.45%	1.00%	4.22%	3.87%	18.56%	11.22%
$D_{ArXivNaive}$		2.01%	1.82%	**1.91%**	0.97%	13.32%	**7.14%**
$D_{BIOSIGMC}$		25.76%	1.35%	13.56%	23.10%	10.12%	16.61%
$D_{BIOSIGNaive}$		11.47%	5.25%	8.36%	9.54%	18.05%	13.80%
$D_{keypoints}$		87.86%	0.00%	43.93%	96.94%	0.00%	48.47%
F_M		11.39%	0.56%	5.97%	7.15%	9.30%	8.23%
F_{WLC}		18.09%	0.02%	9.05%	19.90%	0.84%	10.37%
F_{DST}		25.47%	0.02%	12.74%	35.02%	0.01%	17.52%
F_{LR}		23.68%	0.02%	11.85%	27.05%	0.35%	13.70%
$D_{ArXivMC}$	Complete	7.45%	1.00%	4.22%	3.87%	15.73%	9.80%
$D_{ArXivNaive}$		2.01%	1.60%	**1.81%**	0.97%	10.57%	5.77%
$D_{BIOSIGMC}$		25.76%	1.38%	13.57%	23.10%	8.38%	15.74%
$D_{BIOSIGNaive}$		11.47%	4.47%	7.97%	9.54%	14.31%	11.92%
$D_{keypoints}$		87.86%	0.00%	43.93%	96.94%	0.00%	48.47%
F_M		11.39%	0.31%	5.85%	7.15%	3.76%	**5.46%**
F_{WLC}		18.09%	0.02%	9.05%	19.90%	0.60%	10.25%
F_{DST}		25.47%	0.02%	12.74%	35.02%	0.02%	17.52%
F_{LR}		23.68%	0.02%	11.85%	27.05%	0.27%	13.66%
$D_{ArXivMC}$	Splicing	7.45%	2.57%	5.01%	3.87%	24.81%	14.34%
$D_{ArXivNaive}$		2.01%	3.54%	**2.77%**	0.97%	16.04%	**8.50%**
$D_{BIOSIGMC}$		25.76%	3.54%	14.65%	23.10%	17.39%	20.25%
$D_{BIOSIGNaive}$		11.47%	7.39%	9.43%	9.54%	21.22%	15.38%
$D_{keypoints}$		87.86%	0.02%	43.94%	96.94%	0.00%	48.47%
F_M		11.39%	1.45%	6.42%	7.15%	12.35%	9.75%
F_{WLC}		18.09%	0.07%	9.08%	19.90%	1.42%	10.66%
F_{DST}		25.47%	0.03%	12.75%	35.02%	0.03%	17.52%
F_{LR}		23.68%	0.03%	11.85%	27.05%	0.55%	13.80%

morphs). For the "adaptive" thresholding, the single detector HTER values reported significantly improve (e.g., from 9.62 to 2.2% for $D_{ArXivNaive}$ in *SC1* for *combined* morphs). In some cases, they are getting really close to the EER values for the corresponding experiment, which represents the best value that could be achieved in this test. The fusion results for this thresholding strategy see an even more significant drop in the HTER values presented (e.g., 13.41% to 2.8% for F_M in *SC1* for *combined* morphs).

For the second, an even less realistic (i.e., more lab-condition) test no additional test has to be performed here. Instead results from an earlier publication on fusion in face morph attack detection are re-used here. As authors of [7], we used a static percentage split (50%: 50%) on one a set containing genuine (originating from exactly one public database) and morphs that are derived directly from these genuine images to perform initial tests with DST in this field. The results presented were astonishing HTER values of less that 1%. While the results did indicate the potential benefit of using fusion in MAD, the observed lack of realism in the setup made us question the actual extend of the performance increase we could realistically hope for. This realization motivated the research work on the empirical limitations of using information fusion and the constraints for its application that lead to this journal paper.

Summarizing the results obtained on alternative (i.e., less realistic) evaluation setups, it has to be said that the error rates obtained achieved when drawing training and test data from the same parent population are obviously lower than in a setup with disjoint populations used. In the experiments discussed above, the fusion approaches benefit more from the unrealistic lab-condition like

Table 6 Realistic performance of the individual detectors and fusion approaches with the adaptive decision thresholds and fusion weights based on the estimated EER with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
$D_{ArXivMC}$	Combined	30.08%	0.01%	15.04%	20.79%	0.89%	10.84%
$D_{ArXivNaive}$		19.21%	0.03%	9.62%	20.34%	0.50%	10.42%
$D_{BIOSIGMC}$		39.76%	0.35%	20.05%	37.03%	4.51%	20.77%
$D_{BIOSIGNaive}$		34.18%	0.65%	17.41%	33.76%	3.29%	18.52%
$D_{keypoints}$		47.95%	1.15%	24.55%	73.17%	0.26%	36.72%
F_M		26.81%	0.01%	13.41%	29.14%	0.42%	14.78%
F_{WLC}		0.60%	10.87%	**5.73%**	0.30%	46.48%	23.39%
F_{DST}		25.17%	0.02%	12.59%	33.53%	0.01%	16.77%
F_{LR}		14.00%	0.09%	7.04%	14.31%	1.90%	**8.10%**
$D_{ArXivMC}$	Complete	30.08%	0.00%	15.04%	20.79%	0.63%	10.71%
$D_{ArXivNaive}$		19.21%	0.02%	9.61%	20.34%	0.35%	10.34%
$D_{BIOSIGMC}$		39.76%	0.38%	20.07%	37.03%	3.45%	20.24%
$D_{BIOSIGNaive}$		34.18%	0.44%	17.31%	33.76%	2.46%	18.11%
$D_{keypoints}$		47.95%	1.13%	24.54%	73.17%	0.09%	36.63%
F_M		26.81%	0.01%	13.41%	29.14%	0.23%	14.68%
F_{WLC}		0.60%	10.30%	**5.45%**	0.30%	41.15%	20.72%
F_{DST}		25.17%	0.02%	12.59%	33.53%	0.02%	16.77%
F_{LR}		14.00%	0.05%	7.02%	14.31%	1.30%	**7.80%**
$D_{ArXivMC}$	Splicing	30.08%	0.03%	15.05%	20.79%	1.38%	11.08%
$D_{ArXivNaive}$		19.21%	0.05%	9.63%	20.34%	0.64%	10.49%
$D_{BIOSIGMC}$		39.76%	1.11%	20.44%	37.03%	8.60%	22.82%
$D_{BIOSIGNaive}$		34.18%	1.07%	17.62%	33.76%	4.35%	19.05%
$D_{keypoints}$		47.95%	0.78%	24.36%	73.17%	0.25%	36.71%
F_M		26.81%	0.01%	13.41%	29.14%	0.65%	14.89%
F_{WLC}		0.60%	17.18%	8.89%	0.30%	56.84%	28.57%
F_{DST}		25.17%	0.01%	12.59%	33.53%	0.03%	16.78%
F_{LR}		14.00%	0.26%	**7.13%**	14.31%	3.53%	**8.92%**

evaluation setups than the single detectors and the "adaptive" thresholding strategy benefits more than the "fixed" one.

5.6 Summary on the fusion experiments results

There are three main reasons why fusion fails to outperform the best individual classifier in the results discussed in section 5.3:

1. *Lack of diversity of the individual detectors.* The detectors $D_{ArXivNaive}$, $D_{arXivMC}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$ are developed by the same research group and rely on training of DCNN with similar data sets but strong variances in data augmentation. Hence, it is very likely that these detectors make in field application mistakes on the same samples. Only the $D_{keypoints}$ detector relies on entirely different morphing detection clues and is developed by another research group using a different data set for training. In theory, an assumed clustering of four apparently very similar detectors might prove a strong prejudice in fusion that should be avoided at any cost. In practice, our experiment on different ensembles of classifiers showed a better performance if only those four detectors are used instead of all five.

2. *Lack of performance in individual detectors.* It can be seen from the evaluation with the DEFACTO dataset, that $D_{keypoints}$ lacks generalization power.

Table 7 Comparing fusion ensembles consisting of all five, one set of four ($D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$), and one set of three ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$) detectors with the fixed decision thresholds and equal fusion weights with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type and ensemble size marked in bold)

		SC1			*SC2*		
Fusion	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
F_M (5 det)	Combined	11.39%	0.56%	**5.97%**	7.15%	9.30%	**8.23%**
F_{WLC} (5 det)		18.09%	0.02%	9.05%	19.90%	0.84%	10.37%
F_{DST} (5 det)		25.47%	0.02%	12.74%	35.02%	0.01%	17.51%
F_{LR} (5 det)		23.68%	0.02%	11.85%	27.05%	0.35%	13.70%
F_M (4 det)		2.98%	0.56%	**1.77%**	1.56%	9.30%	**5.43%**
F_{WLC} (4 det)		5.29%	1.07%	3.18%	2.31%	12.17%	7.24%
F_{DST} (4 det)		22.34%	0.02%	11.18%	19.75%	0.49%	10.12%
F_{LR} (4 det)		7.67%	0.61%	4.14%	4.47%	8.53%	6.50%
F_M (3 det)		26.14%	0.19%	13.16%	23.25%	3.85%	**13.55%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.01%	**12.96%**	44.86%	0.01%	22.43%
F_{LR} (3 det)		60.46%	0.01%	30.23%	77.35%	0.01%	38.68%
F_M (5 det)	Complete	11.39%	0.31%	**5.85%**	7.15%	3.76%	**5.46%**
F_{WLC} (5 det)		18.09%	0.02%	9.05%	19.90%	0.60%	10.25%
F_{DST} (5 det)		25.47%	0.02%	12.74%	35.02%	0.02%	17.52%
F_{LR} (5 det)		23.68%	0.02%	11.85%	27.05%	0.27%	13.66%
F_M (4 det)		2.98%	0.30%	**1.64%**	1.56%	3.76%	**2.66%**
F_{WLC} (4 det)		5.29%	0.97%	3.13%	2.31%	9.57%	5.94%
F_{DST} (4 det)		22.34%	0.02%	11.18%	19.75%	0.64%	10.19%
F_{LR} (4 det)		7.67%	0.58%	4.12%	4.47%	6.51%	5.49%
F_M (3 det)		26.14%	0.09%	13.11%	23.25%	1.36%	**12.30%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.02%	**12.96%**	44.86%	0.00%	22.43%
F_{LR} (3 det)		60.46%	0.00%	30.23%	77.35%	0.00%	38.67%
F_M (5 det)	Splicing	11.39%	1.45%	**6.42%**	7.15%	12.35%	**9.75%**
F_{WLC} (5 det)		18.09%	0.07%	9.08%	19.90%	1.42%	10.66%
F_{DST} (5 det)		25.47%	0.03%	12.74%	35.02%	0.03%	17.52%
F_{LR} (5 det)		23.68%	0.03%	11.85%	27.05%	0.55%	13.80%
F_M (4 det)		2.98%	1.45%	**2.21%**	1.56%	12.35%	**6.96%**
F_{WLC} (4 det)		5.29%	2.33%	3.81%	2.31%	16.71%	9.51%
F_{DST} (4 det)		22.34%	0.05%	11.19%	19.75%	0.84%	10.29%
F_{LR} (4 det)		7.67%	1.42%	4.55%	4.47%	11.66%	8.06%
F_M (3 det)		26.14%	0.48%	13.31%	23.25%	6.01%	**14.63%**
F_{WLC} (3 det)		88.38%	0.00%	44.19%	98.06%	0.00%	49.03%
F_{DST} (3 det)		25.91%	0.03%	**12.97%**	44.86%	0.03%	22.44%
F_{LR} (3 det)		60.46%	0.01%	30.24%	77.35%	0.00%	38.67%

The default decision threshold of 0.5 is far away from the sub-optimal (i.e., containing an offset due to training data vs. test data mismatch) threshold of 0.87252 obtained from its evaluation. Even higher are the sub-optimal decision thresholds with the mixed test data set (London, ECVP , and Alabama images). The values of approximately 0.97 for the *SC1* and 0.99 for the *SC2* indicate a large discrepancy between the data used for the training of the classifier and for evaluation/testing. As a consequence, the APCER and BPCER values are imbalanced, both are on the margins of the [0, 1] interval

Table 8 Comparing fusion ensembles consisting of all 5, 4 ($D_{BIOSIGNaive}$, $D_{BIOSIGMC}$, $D_{arXivNaive}$, and $D_{arXivMC}$), and 3 ($D_{arXivNaive}$, $D_{BIOSIGMC}$, $D_{keypoints}$) detectors with the adaptive decision thresholds and fusion weights based on the estimated EER with the combined LondonDB/UtrechtDB/Alabama datasets (best result per morph type and ensemble size marked in bold)

		SC1			*SC2*		
Detector	**Morph type**	**BPCER**	**APCER**	**HTER**	**BPCER**	**APCER**	**HTER**
F_M (5 det)	Combined	26.81%	0.01%	13.41%	29.14%	0.42%	14.78%
F_{WLC} (5 det)		0.60%	10.87%	**5.73%**	0.30%	46.48%	23.39%
F_{DST} (5 det)		25.17%	0.02%	12.59%	33.53%	0.00%	16.77%
F_{LR} (5 det)		14.00%	0.09%	7.04%	14.31%	1.90%	**8.10%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.42%	8.26%
F_{WLC} (4 det)		6.40%	0.86%	3.63%	3.06%	10.31%	6.68%
F_{DST} (4 det)		23.90%	0.01%	11.95%	21.68%	0.68%	11.18%
F_{LR} (4 det)		7.89%	0.62%	**4.26%**	4.77%	8.36%	**6.56%**
F_M (3 det)		29.41%	0.01%	14.71%	36.36%	0.19%	**18.28%**
F_{WLC} (3 det)		0.00%	64.59%	32.29%	0.00%	93.25%	46.62%
F_{DST} (3 det)		25.69%	0.01%	**12.85%**	43.59%	0.00%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.03%	21.10%
F_M (5 det)	Complete	26.81%	0.01%	13.41%	29.14%	0.23%	14.68%
F_{WLC} (5 det)		0.60%	10.30%	**5.45%**	0.30%	41.15%	20.72%
F_{DST} (5 det)		25.17%	0.02%	12.59%	33.53%	0.02%	16.77%
F_{LR} (5 det)		14.00%	0.05%	7.02%	14.31%	1.30%	**7.80%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.23%	8.16%
F_{WLC} (4 det)		6.40%	0.78%	**3.59%**	3.06%	8.03%	5.54%
F_{DST} (4 det)		23.90%	0.02%	11.96%	21.68%	0.77%	11.23%
F_{LR} (4 det)		7.89%	0.61%	4.25%	4.77%	6.29%	**5.53%**
F_M (3 det)		29.41%	0.01%	14.71%	36.36%	0.04%	**18.20%**
F_{WLC} (3 det)		0.00%	64.16%	32.08%	0.00%	91.78%	45.89%
F_{DST} (3 det)		25.69%	0.02%	**12.85%**	43.59%	0.00%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.02%	21.10%
F_M (5 det)	Splicing	26.81%	0.00%	13.40%	29.14%	0.65%	14.89%
F_{WLC} (5 det)		0.60%	17.18%	8.89%	0.30%	56.84%	28.57%
F_{DST} (5 det)		25.17%	0.01%	12.59%	33.53%	0.03%	16.78%
F_{LR} (5 det)		14.00%	0.26%	**7.13%**	14.31%	3.53%	**8.92%**
F_M (4 det)		17.20%	0.00%	8.60%	16.10%	0.65%	8.37%
F_{WLC} (4 det)		6.40%	2.01%	**4.21%**	3.06%	14.28%	8.67%
F_{DST} (4 det)		23.90%	0.05%	11.98%	21.68%	1.17%	11.43%
F_{LR} (4 det)		7.89%	1.45%	4.67%	4.77%	11.32%	**8.05%**
F_M (3 det)		29.41%	0.00%	14.71%	36.36%	0.24%	**18.30%**
F_{WLC} (3 det)		0.00%	75.02%	37.51%	0.00%	97.08%	48.54%
F_{DST} (3 det)		25.69%	0.01%	**12.85%**	43.59%	0.02%	21.80%
F_{LR} (3 det)		33.88%	0.00%	16.94%	42.18%	0.07%	21.12%

and the HTER values are close to 43% in *SC1* and 48% in *SC2* for the "fixed" thresholding strategy. If the decision threshold for $D_{keypoints}$ is readjusted, based on the training set (DEFACTO), the HTER values in testing become significantly lower, approximately 24% in *SC1* and 36% in *SC2*. However, the APCER and BPCER values are still imbalanced. The impact of one bad detector on the overall fusion is shown very well in the experiment on different ensembles of classifiers showed where a better performance was achieved when only an ensemble of four (all except $D_{keypoints}$) is used.

Table 9 Fusion under laboratory conditions: tenfold stratified cross-validation with 90% training/10% test split; genuine samples from the Alabama dataset [53]; morphs from LondonDB and UtrechtDB (best result per morph type and application scenario marked in bold)

	Combined				Complete				Splicing			
	SC1		*SC2*		*SC1*		*SC2*		*SC1*		*SC2*	
	EER	HTER	EER	HTER	EER	HTER	EER	HTER	EER	HTER	EER	HTER
Fixed												
$D_{ArXivMC}$	3.8%	4.3%	8.2%	11.2%	3.9%	4.2%	7.2%	9.8%	4.8%	5.0%	9.2%	14.3%
$D_{ArXivNaive}$	1.5%	**1.9%**	3.9%	7.1%	1.3%	**1.8%**	3.4%	**5.8%**	1.8%	**2.7%**	4.4%	8.5%
$D_{BIOSIGMC}$	9.3%	13.6%	15.7%	16.6%	9.3%	13.6%	14.7%	15.7%	12.8%	14.7%	20.2%	20.2%
$D_{BIOSIGNaive}$	7.1%	8.4%	13.4%	13.7%	7.0%	7.9%	11.8%	11.9%	8.4%	9.4%	14.2%	15.4%
$D_{keypoints}$	12.3%	43.9%	18.6%	48.8%	12.2%	43.9%	19.3%	48.8%	8.9%	43.9%	18.3%	48.8%
F_M		6.0%		8.2%		6.2%		5.9%		6.4%		9.7%
F_{WLC}		9.6%		10.9%		9.2%		10.6%		9.2%		10.6%
F_{DST}		2.6%		**5.9%**		3.0%		6.7%		2.9%		**7.3%**
F_{LR}		2.6%		**5.9%**		3.0%		6.7%		2.9%		**7.3%**
Adaptive												
$D_{ArXivMC}$	3.8%	3.9%	8.2%	8.3%	3.9%	4.0%	7.2%	7.9%	4.8%	4.9%	9.2%	9.4%
$D_{ArXivNaive}$	1.5%	**2.2%**	3.9%	**5.0%**	1.3%	**2.1%**	3.4%	**4.4%**	1.8%	**3.0%**	4.4%	**5.6%**
$D_{BIOSIGMC}$	9.3%	9.8%	15.7%	16.4%	9.3%	9.8%	14.7%	15.0%	12.8%	12.7%	20.2%	20.7%
$D_{BIOSIGNaive}$	7.1%	7.9%	13.4%	13.6%	7.0%	7.6%	11.8%	12.1%	8.4%	9.1%	14.2%	15.0%
$D_{keypoints}$	12.3%	12.7%	18.6%	19.0%	12.2%	12.5%	19.3%	19.3%	8.9%	11.4%	18.3%	19.2%
F_M		2.8%		6.0%		2.2%		4.5%		3.3%		6.8%
F_{WLC}		15.2%		39.2%		14.3%		35.5%		17.7%		45.8%
F_{DST}		2.8%		5.8%		3.3%		6.6%		3.1%		7.3%
F_{LR}		2.8%		5.8%		3.3%		6.6%		3.1%		7.3%

3. *Lack of similarity between the training and test data.* Different proprietary data sets are used for training individual classifiers, which is a very common case, but the datasets for adjusting fusion parameters (evaluation data set) and for actual testing are also very different from each other and the training data set. One can say that it makes absolutely no sense to use different data sources for adjusting fusion parameters and for testing, but this is the real-life situation. In practice, it is very difficult to precisely foresee and provide significant in-field data at the stage of system development or parameter adjustment. Moreover, there is no guarantee that the in-field data that will be obtained in the future is even similar to the presented training data.

The case study performed in this paper clearly demonstrates that if the training, evaluation, and test datasets lack similarity, the adaptation of the classifier parameters such as a decision threshold may lead to performance degradation. This can be well explained on the example of the classifier $D_{ArXivNaive}$ which in the tests performed shows the best generalization power. The classifier is well trained with the default decision threshold of 0.59072. An attempt to adapt the decision threshold based on the DEFACTO data set actually fails with shifting it to 0.39958, resulting in an EER of 10%. As a consequence, the APCER and BPCER values are imbalanced in the test leading to the HTER values of approximately 9.5% in *SC1* and 10.5% in *SC2* (see Table 6). However, if there is no adaptation of the decision threshold, the sub-optimal (i.e., offset) thresholds of 0.594687, 0.600357, and 0.566983 are close to the default one and the APCE R and BPCER values are well balanced in *SC1* leading to HTER values of 1.91%, 1.81%, and 2.77% for combined, complete, and splicing morphs respectively (see Table 5). In contrary, the sub-optimal thresholds in the *SC2* would be 0.499938, 0.507648, and 0.492199 for combined, complete, and splicing morphs respectively which are far away from the default value of 0.59072. Hence, in the test within *SC2* the APCER and BPCER values are imbalanced leading to the HTER values of 7.14%, 5.77%, and 8.50% for combined, complete, and splicing morphs respectively. The same situation can be observed with the detectors $D_{arXivMC}$, $D_{BIOSIGMC}$, and $D_{BIOSIGNaive}$.

Considering the results of different fusion strategies, it can be said that in almost all cases, the APCER and BPCER values are imbalanced in the case when training, evaluation, and test datasets lack similarity. This results in the conclusion that pre-determining the proper decision thresholds (as well as the fusion weights) in real-life conditions (where the training, evaluation, and in-field data might be dramatically different) is hardly possible.

When considering alternative (less strict) evaluation setups, where training and test data show and artificial similarity due to the fact that they have been drawn from the same parent distribution, we see in section 5.5 significantly lower HTER values not only for fusion results but in some cases also for the individual detectors.

The results presented more clear indicators that the similarity between the training and test data is the dominating factor for the error rates achieved. If this similarity is an artificial one (e.g., in an unrealistic setup where training, parameterization, and test data are drawn from the same parent population) instead of a natural one (i.e., the fusion as well as the individual detectors are suitably well trained) the low error rates obtained are meaningless.

The practical consequence of these three issues is that one of the individual detectors (obviously accurate but far from perfect in its performance) in all evaluations outperforms four different fusion approaches, ranging from simplistic to very sophisticated, in different parameterizations in the tests performed in 5.3 but becomes marginalized by fusion approaches as soon as either the ensemble of detectors used in the fusion is optimized (as done by removing one disturbing detector in section 5.4) or the similarity between training and test data is increased (as in section 5.5).

6 Conclusions

The results presented in the empirical evaluations in this paper demonstrate that fusion can fail even with a set of relevant individual classifiers. This can be seen in both application scenarios ("MAD in document issuing" and "MAD in identity verification") evaluated in this paper. Here, the three reasons for this phenomenon discussed above are (a) low diversity of the detectors, (b) lack of performance in individual detectors, and (c) lack of similarity between the training and test data.

Summarizing the lessons learned from the approach of using fusion for MAD detection as done in this paper and drawing some generalization toward other media forensics classification or decision problems, the following has to be said: The requirements for (media) forensic methods in terms of scientific admissibility (or Daubert compliance) are obviously important! Methods should indeed be published upon and peer reviewed, their error rates should be precisely known and standards for the application of methods should be known. But the threat that Champod and Vuille identify as a problem of ascertaining the error rates of a test "can prove misleading if not all its complexities are understood" [15] plays a very significant role as demonstrated in the evaluations performed here.

Besides the requirements for individual expert systems to be used in forensic investigations (including its accurateness), if it comes to information fusion, additional constraints have to be observed. These are, at least:

- The diversity of the detectors, which has to be ascertained either by knowledge about the precise means of decision generation and the diversity of those means or empirically.
- An independent and thorough benchmarking of detectors to establish also an idea on the generalization power of performance claims made by their creators.
- Considerations on the similarity/correlation between training data available (during training of the individual classifiers and the training of the fusion methods) and the data to be expected in field application are very important. If very precise assumptions are possible on the application data, weighting might be applicable in fusion. Else-wise, only unweighted fusion strategies like majority voting or the sum-rule should be employed, if any fusion is used in those cases at all.

The diversity issue becomes very problematic if features (as the means to represent a decision problem in a feature space) are not hand crafted by experts but learned, e.g., by DCNN. In this paper, the diversity problem of the detectors used here as "black boxes" has been established in direct contact with the developers of those methods, which is hardly an option in most field applications.

Also, the recent trend to generate synthetic data sets for the training of pattern recognition methods (either traditional or neural network based) introduces another degree of freedom into the characteristics of datasets. In publications such as [54], this approach is used to avoid tedious data collection tasks while creating sufficiently sized data sets for modern day data-greedy classifiers. The problem here is the influence of the synthesis process on its output (i.e., the synthesis-specific artifacts) that will become part of the model trained by each classifier. It is related to the questions of source characteristics imposing themselves into trained models but carries a different degree of relevance for forensic application scenarios.

The general problem with training- and test data being mismatched in practice is hardly new. It hardly ever occurs in scientific papers on applied pattern recognition, because it can easily be prevented in lab tests. Nevertheless, it is a very good argument why media forensic methods should undergo rigorous testing and benchmarking by third parties, like it is done in the field of MAD in the NIST FRVT MORPH challenge. Only such joint efforts can lead to methods that might become mature enough to aim at court admissibility.

Abbreviations

ABC: Automated border control; AFR: Automated face recognition; APCER: Attack Presentation Classification Error Rate (from [ISO/IEC 30107-3:2017]); AUC: Area under curve; BPCER: Bona Fide Classification Error Rate (from [ISO/IEC 30107-3:2017]); BSIF: Binarized Statistical Image Features; DB: Database; DCNN: Deep convolutional neural networks; DET: Detection error trade-off; dpi: Dots per inch; DST: Dempster-Shafer Theory; ECVP: European Conference on Visual Perception; EER: Equal error rate; HTER: Half total error rate; IEC: International Electrotechnical Commission; eMRTD: Electronic machine readable travel documents; EU: European Union; FRE: Federal Rules of Evidence; FRVT: (NIST) Face Recognition Vendor Test; GAN: Generative adversarial networks; HOG: Histogram of oriented gradients; ICAO: International Civil Aviation Organization; ID: Identity document; ILSVRC: ImageNet Large Scale Visual Recognition Challenge; ISO: International Organization for Standardization; JPEG: Joint Photographic Experts Group image file format; LBP: Local binary patterns; LR: Likelihood ratio; MAD: Morphing attack detection; MAX: Maximum; MIN: Minimum; NIST: National Institute of Standards and Technology; PNG: Portable Network Graphics; QA: Quality assessment; SC: Scenario; SIFT: Scale-invariant feature transform; SURF: Speeded up robust features; SVM: Support vector machine; TIFF: Tagged image file format; WLC: Weighted linear combination

Acknowledgements

The authors wish to thank Clemens Seibold (Fraunhofer HHI, Berlin, Germany) for collecting the version of the Alabama image set used here and for running our data through his four detectors as well as the important discussions about the training data used to generate those "black box" detectors. The authors wish to thank Tom Neubert for providing the keypoints detector used.
The experiments in sections 5.4 and 5.5 were inspired by the reviewers in the first round of reviews for this journal paper. We wish to express our gratitude to those experts unknown to us because we feel that their recommendations significantly helped to improve this paper.

Authors' contributions

CK works on the media forensic perspectives and fusion theory parts as well as the interpretation of results. AM work on the biometric perspective, the dataset creation, the conduction of the experiments (classifier selection and fusion operator implementation), and the interpretation of the experimental results. JD initial structuring of the work and definition of focus and scope of the work (incl. suggesting the two application scenarios as well as the usage of DST and LR based fusion). MH theoretical and practical work on likelihood based fusion and the interpretation of its results. The authors read and approved the final manuscript.

Funding

The work in this paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the research program under the contract no. FKZ: 16KIS0509K (research project ANANAS). The work in this paper has been funded in part by the Deutsche Forschungsgemeinschaft (DFG) under contract no.: 421860277 (research project GENSYNTH). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The empirical work in this paper is based on the following publicly available datasets:

- The AMSL Face Morph Image Data Set (made available via: https://omen.cs.uni-magdeburg.de/disclaimer/index.php; last accessed Sept. 10, 2020)
- The Utrecht/ECVP as part of Psychological Image Collection at Stirling (PICS), (available at: http://pics.stir.ac.uk/2D_face_sets.htm; last accessed Sept. 10, 2020)
- The London DB has been made available by *L. DeBruine and B. Jones as: Face Research Lab London Set:* https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666 *(last accessed Sept. 10th, 2020)*
- The DEFACTO dataset (including the face morphing subset used in this paper) introduced in [51] is available at: https://defactodataset.github.io/ *(last accessed Sept. 10, 2020)*
- The used Alabama database is the collection of mugshots of the Alabama News Network (available at: https://www.alabamanews.net/mugshots/; last accessed Sept. 10, 2020)

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 21 September 2020 Accepted: 17 June 2021
Published online: 29 July 2021

References

1. H. Drucker, C. Cortes, L.D. Jackel, Y. LeCun, V. Vapnik, Boosting and other ensemble methods. Neural Comput. **6**(6), 1289–1301 (1994)
2. L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms* (Wiley-Interscience, 2004)
3. M. Tan, *Multi-agent reinforcement learning: independent vs. cooperative agents. Readings in agents* (Morgan Kaufmann Publishers Inc., San Francisco, 1997), pp. 487–494
4. T.K. Ho, in *Hybrid Methods in Pattern Recognition*, ed. by A. Kandel, H. Bunke. Multiple classifier combination: lessons and the next steps (World Scientific Publishing, 2002), pp. 171–198
5. A. Ross, K. Nandakumar, A.K. Jain, *Handbook of Multibiometrics* (Springer, 2006)
6. R.P. Srivastava, Alternative Form of Dempster's Rule for Binary Variables. Int. J. Intell. Syst. **20**(8), 789–797 (2005)
7. A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, P. Eisert, in *Proc. 27th European Signal Processing Conference (EUSIPCO)*. Dempster-Shafer Theory for Fusing Face Morphing Detectors (A Coruna, 2019), pp. 1–5
8. A. Makrushin, A. Wolf, in *Proc. 26th European Signal Processing Conference (EUSIPCO)*. An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack (2018)
9. U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, C. Busch, Face Recognition Systems Under Morphing Attacks: A Survey. IEEE Access **7**(2019), 23012–23026 (2019)
10. M. Ferrara, A. Franco, D. Maltoni, in *Face Recognition Across the Electromagnetic Spectrum*, ed. by T. Bourlai. On the effects of image alterations on face recognition accuracy (Springer, Cham, 2016), pp. 195–222
11. R.S.S. Kramer, M.O. Mireku, T.R. Flack, K.L. Ritchie, Face morphing attacks: investigating detection with humans and computers. Cogn. Res. Princ. Implications **4**, 28 (2019)
12. A. Makrushin, T. Neubert, J. Dittmann, in *Proc. 12th Int. Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6*. Automatic generation and detection of visually faultless facial morphs (VISAPP, 2017), pp. 39–50
13. U. Scherhag, R. Raghavendra, K.B. Raja, M. Gomez-Barrero, C. Rathgeb, C. Busch, in *Proc. 5th International Workshop on Biometrics and Forensics (IWBF)*. On the Vulnerability of Face Recognition Systems: Towards Morphed Face Attacks (2017)
14. National Institute of Standards and Technology (NIST) FRVT MORPH, https://pages.nist.gov/frvt/html/frvt_morph.html
15. C. Champod, J. Vuille, in *International Commentary on Evidence. Vol. 9, Issue 1*. Scientific evidence in Europe – admissibility, evaluation and equality of arms (2011) Available at: https://core.ac.uk/reader/85212846 (Last accessed: 26 Aug 2020)
16. M. Ferrara, A. Franco, D. Maltoni, in *Proc. Int. Joint Conf. on Biometrics (IJCB)*. The magic passport (2014), pp. 1–7

17. C. Seibold, W. Samek, A. Hilsmann, P. Eisert, in Proc. Int. Workshop Digital Watermarking (IWDW2017). Detection of Face Morphing Attacks by Deep Learning (Springer, Berlin, 2017)
18. U. Scherhag, C. Rathgeb, C. Busch, in *Proc. 13th IAPR Workshop on Document Analysis Systems (DAS'18)*. Towards detection of morphed face images in electronic travel documents (2018)
19. C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, J. Dittmann, in *Proc. 5th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec'17)*. Modeling attacks on photo-ID documents and applying media forensics for the detection of facial morphing (ACM, New York, 2017), pp. 21–32
20. Utrecht ECVP as part of Psychological Image Collection at Stirling (PICS), http://pics.stir.ac.uk/2D_face_sets.htm, last accessed: 31 Aug 2020.
21. R. Raghavendra, S. Venkatesh, K. Raja, C. Busch, in *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. Towards making morphing attack detection robust using hybrid scale-space colour texture features (2019), pp. 1–8
22. M. Ferrara, A. Franco, D. Maltoni, Face demorphing. Trans. Inf. Forensics Secur. **13**(4), 1008–1017 (2018)
23. D.O. del Campo, C. Conde, D. Palacios-Alonso, E. Cabello, Border control morphing attack detection with a convolutional neural network de-morphing approach. IEEE Access **8**, 92301–92313 (2020)
24. F. Peng, L. Zhang, M. Long, FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplice's Facial Image. IEEE Access **7**, 75122–75131 (2019)
25. U. Scherhag, D. Budhrani, M. Gomez-Barrero, C. Busch, in *International Conference on Image and Signal Processing (ICISP 2018)*. Detecting Morphed Face Images Using Facial Landmarks (2018), pp. 444–452
26. C. Seibold, W. Samek, A. Hilsmann, P. Eisert, Accurate and robust neural networks for security related applications exampled by face morphing attacks. Arxiv/CoRR abs/1806.04265 (2018)
27. U. Scherhag, C. Rathgeb, J. Merkle, C. Busch, in *IEEE Transactions on Information Forensics and Security (TIFS)*. Deep Face Representations for Differential Morphing Attack Detection (2020)
28. L. Wandzik, G. Kaeding, R.V. Garcia, in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO), Sep. 2018*. Morphing detection using a general- purpose face recognition system (2018), pp. 1012–1016
29. R. Raghavendra, K. Raja, S. Venkatesh, C. Busch, in *Proc. 30th Int. Conf. on Computer Vision and Pattern Recognition Workshop*. Transferable Deep-CNN features for detecting digital and print-scanned morphed face images (2017)
30. T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, J. Dittmann, Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images. IET Biometrics **7**(4), 325–332 (2018)
31. T. Karras, S. Laine, T. Aila, in *IEEE Conference on Computer Vision and Pattern Recognition*. A style-based generator architecture for generative adversarial networks (2019), pp. 4401–4410
32. T. Karras, T. Aila, S. Laine, J. Lehtinen, in *International Conference on Learning Representations*. Progressive growing of GANs for improved quality, stability, and variation (2018)
33. N. Damer, A.M. Saladie, A. Braun, A. Kuijper, in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS), Oct. 2018*. MorGAN: recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network (2018), pp. 1–10
34. S. Venkatesh, H. Zhang, R. Raghavendra, K. Raja, N. Damer, C. Busch, *Can GAN generated morphs threaten face recognition systems equally as landmark based morphs? -vulnerability and detection* (International Workshop on Biometrics and Forensics (IWBF), 2020)
35. G. Shafer, *A Mathematical Theory of Evidence* (Princeton University Press, 1976)
36. P. Smets, in *Proc. 15th Conf. On Uncertainty in Artificial Intelligence*. Practical uses of belief functions, vol 99 (1999), pp. 612–621
37. J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **20**(3), 226–239 (1998)
38. L. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd edn. (Wiley, New York, 2014)
39. M. Fontani, A. Bonchi, A. Piva, M. Barni, in *Proc. Media Watermarking, Security, and Forensics 2014, San Francisco, CA, USA, February 2, 2014*, ed. by A. M. Alattar, N. D. Memon, C. Heitzenrater. Countering anti-forensics by means of data fusion, vol 9028 (SPIE Proceedings, 2014), p. 90280Z SPIE
40. M. Fontani, T. Bianchi, A. De Rosa, A. Piva, M. Barni, A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence. IEEE Trans. Inf. Forensics Secur. **8**(4), 593–607 (2013)
41. Royal Courts of Justice, "R v T", [2010] EWCA Crim 2439, Redacted Judgment, 2011, Available at: http://www.bailii.org/ew/cases/EWCA/Crim/2010/2439.pdf (last accessed: 10 Mar 2021)
42. Y. Peng, L.J. Spreeuwers, R.N.J. Veldhuis, in *Proceedings of the 3rd International Workshop on Biometrics and Forensics, IWBF 2015*. Likelihood Ratio Based Mixed Resolution Facial Comparison (IEEE Computer Society, USA, 2015), pp. 1–5
43. T. Kerkvliet, R. Meester, Assessing forensic evidence by computing belief functions. Law Probability Risk **15**(2), 127–153 (2016)
44. T.G. Dietterich, in *Multiple classifier systems*. Ensemble methods in machine learning (Springer LNCS 1857, 2000), pp. 1–15
45. B. Quost, M.-H. Masson, T. Denœux, Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules. Int. J. Approx. Reason. **52**(3), 353–374 (2011)
46. K. Tumer, J. Ghosh, Error correlation and error reduction in ensemble classifiers. Connect. Sci. **3–4**(8), 385–404 (1996)
47. S.P. Lund, H. Iyer, Likelihood ratio as weight of forensic evidence: a closer look. J. Res. Nat. Instit. Stand. Technol. **122**, 122.027 (2017) 2017
48. A. Nordgaard, R. Ansell, W. Drotz, L. Jaeger, Scale of conclusions for the value of evidence. Law Probability Risk **11**(1), 1–24 (2012)
49. K. Nandakumar, Y. Chen, S.C. Dass, A. Jain, Likelihood Ratio-Based Biometric Score Fusion. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 342–347 (2008)
50. ISO/IEC JTC1 SC37 Biometrics, ISO/IEC 30107-3:2017 Information technology- biometric presentation attack detection - Part3: Testing & reporting. ISO, 2017.
51. G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. L. Dugelay, M. Pic: DEFACTO: Image and Face Manipulation Dataset. 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5, (dataset: https://defactodataset.github.io/), 2019.
52. L. DeBruine, B. Jones: Face Research Lab London Set. May 30th, 2017. Available at: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666 (last accessed 10 Sept 2020).
53. Alabama News Network Mugshot database, online: https://www.alabamanews.net/mugshots/ (last accessed: 9 Sept 2020).
54. D. Maltoni, D. Maio, A.K. Jain, S. Prabhakar, in *Synthetic Fingerprint Generation*. Handbook of Fingerprint Recognition (Springer, London, 2009), pp. 271–302

Publisher's Note