Dempster-Shafer Theory for Fusing Face Morphing Detectors

Andrey Makrushin, Christian Kraetzer, Jana Dittmann
University of Magdeburg
Magdeburg, Germany
andrey.makrushin@ovgu.de, {kraetzer, jana.dittman}@iti.cs.uni-magdeburg.de

Clemens Seibold, Anna Hilsmann
Fraunhofer HHI
Berlin, Germany
{clemens.seibold, anna.hilsmann}@hhi.fraunhofer.de

Peter Eisert
Fraunhofer HHI & Humboldt University
Berlin, Germany
eisert@informatik.hu-berlin.de

***Abstract*—Revealing that a human face on a biometric image is a mixture of two or more faces is of immense importance for document issuing authorities and document checking services. If not done, several persons can use the same photo-ID document for identity verification without being condemned. The development of automated face morphing detectors is currently in its early phase. The detectors reported so far are not mature for the market which is reflected in high error rates when tested with "unseen" data. Here, we demonstrate that fusion of several by far non-optimal detectors may lead to significant improvement of detection accuracy compared to that of individual detectors. Among the examined fusion approaches, Dempster's Rule of combination has the best accuracy allowing for coherent decision making even with contradicting decisions of individual detectors.**

Keywords— face morphing attack, morphing detection, fusion

I. Introduction

Morphing is formally a process of gradual transformation of one object (source) to another (target). Facial morphs would result from morphing one face image to another if the process stopped in between. Usage of facial morphs as biometric portraits intended for an identity document application is malicious, because, if successful, it compromises further identity verification by means of the issued document, namely the document can be shared among persons who provided their face images for morphing.

Several studies have demonstrated that manually generated high-quality morphs cannot be recognized as such neither by algorithms nor by human examiners [1], [2] 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.

Since the development of dedicated morphing detectors is in its early phase, many current solutions have too high error rates for practical use. Possessing several detectors, the straight-forward and inexpensive way to improve the overall detection performance is to combine the decisions of individual detectors into a consensual one keeping an eye on the fact that individual detectors may make contradicting decisions or provide a low degree of confidence.

In theory, a necessary and sufficient condition for a combination of classifiers to be more accurate than any of its members is that the classifiers are *accurate* and *diverse*. An accurate classifier has an error rate better than random guessing and two diverse classifiers make errors on different data points [3]. 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 [4]. It has been also shown that lowering correlation among classifiers increases the accuracy of combination [5]. In our case, the higher detection accuracy is expected because individual morphing detectors may rely on different morphing artifacts.

The research question of this study is whether the *recently* proposed face morphing detectors possess sufficient degree of diversity enabling for a detection accuracy gain through fusion and which fusion strategy is superior.

Generally, decision-making systems can be fused at three different levels: feature level, matching score level and decision level. The earlier the data is fused, the higher implementation coasts are, but the higher accuracy is expected. For the case of "black box" detectors returning a matching score for an input sample, the feature level fusion is not feasible. Hence, we empirically evaluate the detection accuracy gain from fusing morphing detectors at decision and matching score levels by exploring several fusion techniques: majority voting, sum-rule and Dempster-Shafer Theory (DST) of evidence [6].

Our main contribution is in demonstrating how DST can be adopted for fusion of morphing detectors, or more specifically, we propose a novel technique on how to incorporate the uncertainty to induce belief functions from distributions of matching scores and show an efficient way to apply Dempster's Rule of combination. Furthermore, we explore several strategies of assigning degrees of reliability to detectors which are based on experimental error rate estimates.

An empirical evaluation of our concept is made with four face morphing detectors introduced in [7], [8], [9] and [10]. The experimental results indicate that our proposed fusion technique leads to lower error rates than any of individual detectors and even any of the reference fusion techniques. Surprisingly, the best detection performance is reached when equal degrees of reliability are assigned to all detectors.

Hereafter, we review studies making an effort towards face morphing detectors fusion in Section II. Our fusion concept is introduced in Section III in detail. The experiments are reported in Section IV. Section V concludes the paper with a brief summary and future work.

978-9-0827-9703-9/19/$31.00

II. Related works

A face morphing detector is in its nature a binary pattern classifier and the methods for combining pattern classifiers have been thoroughly studied. A recent comprehensive overview of methods to combine pattern classifiers can be found in [11].

To the best of our knowledge, the first effort to combine two face morphing detectors is made in [12]. The authors combined keypoints-based detector [7] and Benford detector [13]. The former is based on localization and counting of SIFT, SURF etc. keypoints found in the face region while the latter relies on Benford-features extracted from DCT coefficients of JPEG-compressed images. The fusion is done *at the feature level*. The new combined feature vector is a result of concatenation of vectors comprised of keypoint features and Benford features. This kind of fusion brought no significant gain in detection performance compared to individual detectors. The probable reason is that both feature spaces are designed to formalize the blurring effect emerging in morphed face images and therefore do not possess a sufficient degree of diversity. Another reason could be that one set of features completely dominate another with the data points considered.

Another study on fusing face morphing detectors *at the matching-score level* is conducted in [14]. There are four different sets of feature extraction algorithms addressed: texture descriptors (LBP, BSIF), keypoint extractors (SIFT, SURF), gradient estimators (Sharp, HOG), and deep convolutional neural networks (DCNN). Feature vectors produced by the extractors are separately classified with the support vector machine (SVM) and the normalized matching scores (confidences of SVM decisions) are combined by the sum-rule. The experiments have shown that the more detectors are combined the better detection accuracy may be achieved provided that feature sets "fit" to each other. The equal error rate (*EER*) drops from 5.5% for the best individual detector to 3.1% for the best combination of two, to 3.1% for the best combination of three, and to 2.8% for the best combination of four detectors. Based on this result, the authors claim that the features are complementary in regard to morphing artifacts.

For the case of the document checking scenario where a live face image is taken for the automated matching with the document image, face morphing detection can be done not only blindly, but also in the presence of a reference image [15]. In [16], the aforementioned feature extraction algorithms are applied to document images as well as to the images representing the difference between a document and a live image and the matching scores are calculated for both. Further, the sum-rule fusion as applied *at the matching-score level*. In almost all cases, fusion leads to the significant improvement of detection *EER*.

Although the sum-rule is simple, intuitive, remarkably robust, and outperforms in experiments all other aggregation operators [17], we claim that adoption of DST for combining belief functions derived from decision confidences of individual classifiers has potential to outperform the sum-rule. Note that the sum-rule is just another name for the *average rule* meaning the linear combination of matching scores with equal weights.

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 [18]. In biometrics, DST is used, for instance, for multi-biometric fusion [19], or to fuse fingerprint verification algorithms based on different feature-levels [20]. In forensics, the advantages of reasoning using belief functions for legal practice are discussed in [21]. A framework for applying of DST in digital image forensics is proposed in [22].

III. Our Concept of Applying DST to the Fusion of Face Morphing Detectors

The DST is based on two concepts:

- Belief functions representing degrees of belief for one question from subjective probabilities for a related question;

- Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

A. Degrees of belief

Let $\Theta = \{A_1, A_2, ..., A_k\}$ be a finite set of k mutually exclusive hypotheses, referred to as the frame of discernment. The power set 2^Θ is the set of all subsets of Θ including itself and the null set O. In DST a degree of belief (mass) is assigned to each subset in the power set. In contrast, in probability theory the degree of belief is assigned only to each individual hypothesis. Formally, a basic belief assignment (BBA) is a function m, that assigns a value in the range of [0,1] to each subset A and satisfies the following conditions:

$$m(\mathrm{O}) = 0, \quad and \quad \sum_{A \subseteq 2^\Theta} m(A) = 1 \tag{1}$$

For a subset A, there exist two functions: belief (*Bel*) and plausibility (*Pl*). These can be seen as the lower and upper bounds of the interval containing the precise probability of A.

$$Bel(A) = \sum_{A_i : A_i \subseteq A} m(A_i), \quad and \quad Pl(A) = \sum_{A_i : A_i \cap A \neq \mathrm{O}} m(A_i) \tag{2}$$

Dealing with a binary problem (k=2), a questioned face image is either morphed or genuine, 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.

We propose to build masses as cumulative distribution functions of matching scores obtained from an experiment. Let $p_{mor}(s)$ and $p_{\mathrm{gen}}(s)$ 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 $p_{\mathrm{gen}}(s)$ between s^* and 1, and the mass of uncertainty as a complement to the sum of both masses:

$$m(mor) = \int_{s=0}^{s^*} p_{mor}(s)ds, \quad m(gen) = \int_{s=s^*}^{1} p_{gen}(s)ds \tag{3}$$

$$m(\Theta) = 1 - \big(m(mor) + m(gen)\big) \tag{4}$$

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. For discrete matching scores, the functions $p_{mor}(s)$ and $p_{gen}(s)$ are the corresponding histograms.

Technically, the three masses are calculated for each item of evidence (morphing detector) based on the training samples as functions of a decision threshold and stored as a parameter of our fusion engine. Since a particular sequence of decision thresholds is selected [0, 0.0001, 0.0002, ..., 1], the mass functions are discrete. 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. The mass-curves of the four morphing detectors used in our experiments are shown in Fig. 1.

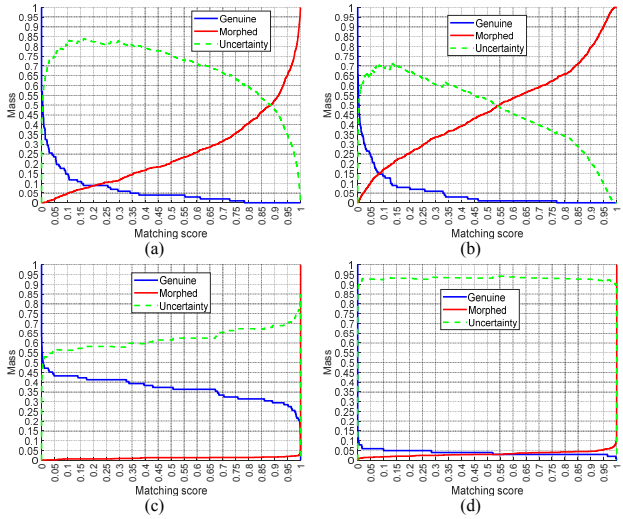

Fig. 1. Belief functions (masses) of the morphing detectors: (a) Keypoint-based detector, (b) High-Dim LBP detector, (c) GoogLeNet-based detector, and (d) VGG19-based detector.

B. Dempster's rule of combination

According to [6], Dempster's rule of combination for two beliefs from independent sources is given by:

$$m(A \neq \mathrm{O}) = \frac{1}{K} \sum_{A=A_1 \cap A_2} \left(m_1(A_1) \cdot m_2(A_2)\right) \tag{5}$$

$$K = 1 - \sum_{A_1 \cap A_2 = 0} \left(m_1(A_1) \cdot m_2(A_2)\right) \tag{6}$$

where $m(A)$ represents the combined mass on A, 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.

Having two detectors, first with $m_1(mor)$, $m_1(gen)$, $m_1(\Theta)$ and the second with $m_2(mor)$, $m_2(gen)$, $m_2(\Theta)$. Dempster's rule combining these two beliefs can be written as:

$$m(mor) = \frac{1}{K}\left(m_1(mor)m_2(mor) + m_1(mor)m_2(\Theta) + m_1(\Theta)m_2(mor)\right) \tag{7}$$

$$m(gen) = \frac{1}{K}\left(m_1(gen)m_2(gen) + m_1(gen)m_2(\Theta) + m_1(\Theta)m_2(gen)\right) \tag{8}$$

$$m(\Theta) = \frac{1}{K} m_1(\Theta)m_2(\Theta) \tag{9}$$

$$K = 1 - \left(m_1(mor)m_2(gen) + m_1(gen)m_2(mor)\right) \tag{10}$$

The general form of Dempster's rule for combining n items of evidence for a binary variable can be found in [23]. The authors also propose alternative equations which allow for efficient computation of combined belief and plausibility:

$$m(mor) = 1 - \frac{1}{K} \prod_{i=1}^{n} \left(1 - m_i(mor)\right) \tag{11}$$

$$m(gen) = 1 - \frac{1}{K} \prod_{i=1}^{n} \left(1 - m_i(gen)\right) \tag{12}$$

$$m(\Theta) = \frac{1}{K} \prod_{i=1}^{n} m_i(\Theta) \tag{13}$$

$$K = \prod_{i=1}^{n} \left(1 - m_i(mor)\right) + \prod_{i=1}^{n} \left(1 - m_i(gen)\right) - \prod_{i=1}^{n} m_i(\Theta) \tag{14}$$

$$Pl(mor) = 1 - m(gen) \quad , \quad Pl(gen) = 1 - m(mor) \tag{15}$$

In our experiment, we use equations (11)-(14) to obtain a final fusion score which is given by a combined belief. Since a combined plausibility is a complement to a combined belief in an opposite hypothesis, it does not bear additional information and therefore is not included in our fusion engine.

C. Reliability of individual detectors

Another important issue is the general reliability of the information sources or in our case morphing detectors. We propose to derive the degrees of reliability empirically based on the detection performance evaluation with a training dataset. Three strategies are explored. In the first one, all sources are absolutely reliable. In the second one, the degree of reliability is given by the area under the ROC curve (*AUC*). For the i^{th} detector $w_i = AUC_i$. In the third one, the degree of reliability is the inverted *EER*, namely $w_i = 1\text{-}EER_i$.

Note that we use the same training samples to build mass-curves and to obtain the degrees of reliability. If the training dataset is substantially different from real-life data, there is a risk that the mass-functions and degrees of reliability do not fit to the "unseen" data and the combined decision may be less accurate than individual decisions. Hence, choosing the training dataset can be seen as an additional source of ambiguity. In order to avoid the reduction of the generalization ability of a decision making system, if there is a risk to back the wrong horse when choosing the training dataset, we recommend considering all sources equally reliable.

IV. Experiments

A. Individual morphing detectors

There are four face morphing detectors in our experiments that are seen as *black boxes*. They produce matching scores in the range between 0 and 1 with 1 for a morphed and 0 for a genuine image. The default decision boundary is 0.5.

The **keypoint-based** morphing detector [7] relies on the assumption that blending as a part of the morphing process causes reduction of face details so that the amount of significant corners and edges becomes lower in face images after morphing. Five keypoint detectors (SIFT, SURF, FAST, ORB, AGAST) and two edge detectors (Canny, Sobel) are used to quantify the detail reduction. Linear SVM is utilized for classification. The detector is trained based on a proprietary dataset of 2000 genuine and 2000 morphed high-quality face images in an eMRTD-compatible format. Morphed faces are generated based on approaches from [12] and [13].

The **High-Dim LBP** morphing detector [8] exploits the ability of Local Binary Patterns (LBP), as texture descriptor, to grasp the change of textural skin characteristics after morphing. The detector includes the following steps: normalization of an input image based on five facial landmarks (eyes, nose, and mouth corners); building an image pyramid and extraction of fixed-size image patches centered around each landmark at each scale of the pyramid; dividing each patch into a grid of 4×4 cells and encoding each cell by an LBP descriptor; and finally concatenating the LBP descriptors to a 99120-dim feature vector. The classification is done with the linear SVM using face images from the Multi-PIE dataset. Morphed faces are generated based on the approach described in [24].

The next two morphing detectors are based on GoogLeNet and VGG19 DCNN models trained for the ImageNet Large Scale Visual Recognition Challenge. Using transfer learning, the **GoogLeNet-based** morphing detector [9] and the **VGG19-based** morphing detector [10] were trained as binary classifiers with two output neurons. Prior to feeding the face images into the networks these are cropped to the smallest bounding box that includes the eyebrows and mouth, and rescaled to the size of 224x224 pixels. The training is done based on images from proprietary and public face datasets using equal numbers of morphed and genuine images. Morphed faces are generated based on the approach described in [9]. The former detector is trained with about 700 images of each type and the latter detector with about 1500 images of each type. The latter detector is referred to as "naive" in the original paper.

TABLE I. DETECTION PERFORMANCE OF INDIVIDUAL MORPHING DETECTORS AND FUSION APPROACHES; *FPR*, *FNR* AND *HTER* ARE IN %.

Detection approach	***AUC***	***EER***	***T_{EER}***	***FPR***	***FNR***	***HTER***
Individual morphing detectors						
Keypoints [7]	97.44	8.82	0.20605	7.84	23.00	15.42
High-Dim LBP [8]	93.37	14.89	0.081681	1.96	47.01	24.49
GoogLeNet [9]	97.72	7.80	0.9999	31.37	0.83	16.10
VGG19 [10]	**99.48**	**2.94**	**0.54295**	**4.90**	**3.04**	**3.97**
Fusion approaches						
Majority voting	-	-	-	0.00	7.27	3.63
Average rule	-	-	-	1.96	1.93	1.95
Linear comb. (*AUC*)	-	-	-	0.00	2.21	1.10
Linear comb. (*EER*)	-	-	-	0.00	2.58	1.29
DST	-	-	-	**0.98**	**0.37**	**0.67**
DST (*AUC*)	-	-	-	0.98	0.55	0.77
DST (*EER*)	-	-	-	0.98	0.55	0.77

B. Fusion strategies

As reference fusion approaches we take the *majority voting* as a trivial example of fusion at decision level and the *average rule* as a trivial example of fusion at matching score level, and compare these with DST-based fusion. In case of equal number of votes, the majority rule decides "genuine". Moreover, the average rule is extended to a *linear combination* with the weights assignment according to the degrees of reliability (see Section III.*C*). Note that the weights undergo no normalization. The degrees of reliability are also exploited in the DST-based fusion. To do so, the values $m_i(mor)$, $m_i(gen)$, $m_i(\Theta)$ i=1..n are multiplied by the corresponding weight w_i prior to applying Dempster's rule of combination.

C. Evaluation Data

We evaluate the performance of the individual detectors as well as of their combination with the AMSL Face Morph Image Data Set provided for the ACM IH&MMSec'19 Special Session Media Forensics - Fake or Real? (https://www.ihmmsec.org/cms/special-session/index.html). The face images were generated in a way to comply with the technical requirements of the ICAO portrait quality standard for eMRTD [25] and to fit on a chip of an eMRTD. The dataset is comprised of 102 neutral, 102 smiling and 2175 morphed face images. Morphed faces were created based on neutral faces. We split morphs into two equally large non-overlapping subsets. Our training set containing neutral genuine faces (102) and the first subset of morphs (1088) is used for building mass functions and calculating reliabilities of individual detectors. Our test set containing smiling genuine faces (102) and the second subset of morphs (1087) is used for performance evaluation.

D. Evaluation metrics

We consider morphing detection to be a standard detection problem with morphed images as positive examples. The standard performance metrics for detection systems are the False Positive Rate (*FPR*) giving a ratio of falsely detected genuine images (false alarms) and the False Negative Rate (*FNR*) giving a ratio of falsely missed morphed images. With the training dataset, we locate the decision threshold T_{EER} where both errors are equal and determine the *EER* and *AUC*. With the test dataset, when the decision thresholds are fixed by detectors, we calculate *FPR* and *FNR* as well as the Half Total Error Rate (*HTER*) as an average of both.

E. Results

The results of our experiments are summarized in Table 1. For the detectors [7], [8] and [9], the default decision threshold of 0.5 is by far not optimal that can be read from the large difference between the *EER* with the training set and the *HTER* with the test set as well as the strong deviation of the decision threshold where the *EER* is reached (T_{EER}) from 0.5. In contrast, the morphing detector from [10] demonstrates the optimal performance close to the default decision threshold and achieves the best *HTER* of 3.97%. The balance between *FPR* and *FNR* indicates the robustness of the detector.

The majority voting only slightly improves the *HTER* over the best performing morphing detector (3.63% vs. 3.97%), but

lead to loosing the balance between *FPR* and *FNR*. The average rule reduces the *HTER* by half from 3.97% to 1.95% and even improves the balance making *FPR* and *FNR* almost equal. The DST-based fusion performs the best with the *HTER* of 0.67% which is approx. a third part of that with the average rule. Note that for only 6 samples the individual decisions were not combinable with Dempster's rule. The distributions of matching scores for the average rule and DST-based fusion are demonstrated in Fig. 2.

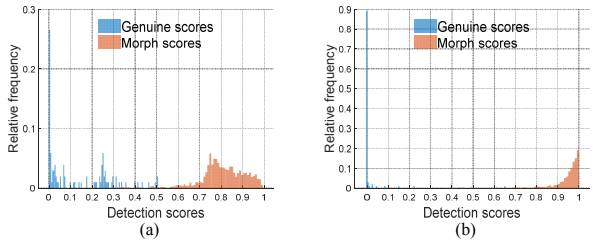

Fig. 2. Histograms of matching scores after fusion: (a) Average rule, (b) DST-based fusion

Noteworthy, assigning degrees of reliability to individual detectors and using these degrees as weights when combining scores to a linear combination (instead of the average rule) may significantly improve *HTER*, which is demonstrated in our experiment. When defining the weights, *AUC* seems to be preferable to inverted *EER*. In contrast, different degrees of reliability assigned to individual detectors only trigger the drop of *HTER* with the DST-based fusion. However, the observed tiny difference in *HTER* values might have been arisen due to quantization errors.

V. Conclusion

In this paper, we discussed an application of Dempster-Shafer Theory (DST) to the fusion of face morphing detectors. We proposed an approach to induce belief function based on cumulative distributions of matching scores in an independent experiment and a computationally effective way to combine the beliefs. We empirically demonstrated that the error rates with the DST-based fusion are significantly lower compared to those of individual detectors as well as of the reference fusion approaches: the majority voting and the average rule. Assigning degrees of reliability (also referred to as weights) to individual detectors improves the detection performance of the average rule transforming that to a linear combination, but does not help to reduce the error rates of the DST-based fusion.

Acknowledgment

The work in this paper has been funded in part by the German Federal Ministry of Education and Science (BMBF) through the research programme under the contract no. FKZ: 16KIS0509K and 16KIS0511.

References

[1] M. Ferrara, A. Franco, and D. Maltoni, "On the Effects of Image Alterations on Face Recognition Accuracy," in Face Recognition Across the Electromagnetic Spectrum, T. Bourlai, Ed. Springer: Cham, 2016, pp. 195–222.

[2] U. Scherhag, R. Raghavendra, K.B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the Vulnerability of Face Recognition Systems: Towards Morphed Face Attacks," Proc. IWBF, 2017.

[3] T. G. Dietterich, "Ensemble methods in machine learning, in: Multiple classifier systems," LNCS 1857, Springer-Verlag, 2000, pp. 1–15.

[4] B. Quost, M.-H. Masson, T. Denœux, "Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules," Int. J. of Approx. Reasoning, Vol. 52, No. 3, pp. 353–374, 2011.

[5] K. Tumer, J. Ghosh, "Error correlation and error reduction in ensemble classifiers," Connection Science, Vol. 3–4, No. 8, pp. 385–404, 1996.

[6] G. Shafer, A Mathematical Theory of Evidence. Princeton University Press, 1976.

[7] C. Kraetzer et al., "Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing," Proc. IH&MMSec, pp. 21–32, 2017.

[8] L. Wandzik, G. Kaeding, and R. Vicente-Garcia, "Morphing Detection Using a General-Purpose Face Recognition System," Proc. EUSIPCO, pp. 1012–1016, 2018.

[9] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Detection of Face Morphing Attacks by Deep Learning," Proc. IWDW, pp. 107–120, 2017.

[10] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Accurate and Robust Neural Networks for Security Related Applications Exampled by Face Morphing Attacks," CoRR abs/1806.04265, 2018.

[11] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, 2nd Ed., John Wiley and Sons, New York, 2014.

[12] T. Neubert et al., "Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images," IET Biometrics, Vol. 7, No. 4, pp. 325–332, 2018.

[13] A. Makrushin, T. Neubert and J. Dittmann. "Automatic generation and detection of visually faultless facial morphs," Proc. 12th Int. Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6: VISAPP, pp. 39–50, 2017.

[14] U. Scherhag et al., "Morph detection from single face images - a multi-algorithm fusion approach," Proc. ICBEA 2018.

[15] A. Makrushin and A. Wolf, "An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack," EUSIPCO 2018.

[16] U. Scherhag, C. Rathgeb, C. Busch, "Towards detection of morphed face images in electronic travel documents," Proc. 13th IAPR Int. Workshop on Document Analysis Systems (DAS), pp. 187–192, 2018.

[17] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On Combining Classifiers," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, pp. 226–239, 1998.

[18] P. Smets, "Practical uses of belief functions," Proc. 15th Conf. on Uncertainty in Artificial Intelligence, Vol. 99, pp. 612–621, 1999.

[19] K. Nguyen, S. Denman, S. Sridharan and C. Fookes, "Score-Level Multibiometric Fusion Based on Dempster–Shafer Theory Incorporating Uncertainty Factors," IEEE Trans. on Human-Machine Systems, Vol. 45, No. 1, pp. 132–140, 2015.

[20] R. Singh, M. Vatsa, A. Noore, S.K. Singh, "Dempster-Shafer Theory based Classifier Fusion for Improved Fingerprint Verification Performance," Computer Vision, Graphics and Image Processing, LNCS 4338, pp. 941–949, 2006.

[21] T. Kerkvliet and R. Meester, "Assessing forensic evidence by computing belief functions," Law, Probability and Risk, Vol. 15, No. 2, pp. 127–153, 2016.

[22] 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. on Information Forensics and Security, Vol. 8, No. 4, pp. 593–607, 2013.

[23] R. P. Srivastava, "Alternative Form of Dempster's Rule for Binary Variables," Int. J. of Intelligent Syst., Vol. 20, No. 8, pp. 789–797, 2005.

[24] L. Wandzik, R. V. Garcia, G. Kaeding, and X. Chen, "CNNs under Attack: On the Vulnerability of Deep Neural Networks Based Face Recognition to Image Morphing," Proc. 16th Int. Workshop on Digital Forensics and Watermarking (IWDW), pp. 121–135, 2017.

[25] A. Wolf, ICAO: Portrait Quality (Reference Facial Images for MRTD), Version 1.0. Standard. International Civil Aviation Organization, 2018.