SHORT-PAPER

Conditional Face Image Manipulation Detection: Combining Algorithm and Human Examiner Decisions

MATHIAS IBSEN, Darmstadt University of Applied Sciences, Darmstadt, Hessen, Germany

ROBERT NICHOLS, Darmstadt University of Applied Sciences, Darmstadt, Hessen, Germany

CHRISTIAN RATHGEB, Darmstadt University of Applied Sciences, Darmstadt, Hessen, Germany

DAVID J ROBERTSON, University of Strathclyde, Glasgow, Scotland, U.K.

JOSH P DAVIS, University of Greenwich, London, U.K.

FRØY LØVÅSDAL

View all

Published: 24 June 2024

IH&MMSEC '24: ACM Workshop on Information Hiding and Multimedia Security
June 24 - 26, 2024
Baiona, Spain

Conference Sponsors:
SIGMM

Conditional Face Image Manipulation Detection: Combining Algorithm and Human Examiner Decisions

Mathias Ibsen
mathias.ibsen@h-da.de
Hochschule Darmstadt
Darmstadt, Germany

Robert Nichols
Hochschule Darmstadt
Darmstadt, Germany

Christian Rathgeb
Hochschule Darmstadt
Darmstadt, Germany

David J. Robertson
University of Strathclyde
Glasgow, United Kingdom

Josh P. Davis
University of Greenwich
London, United Kingdom

Frøy Løvåsdal
National Police Directorate
Oslo, Norway

Kiran Raja
Norwegian University of Science and Technology
Gjøvik, Norway

Ryan E. Jenkins
University of Greenwich
London, United Kingdom

Christoph Busch
Hochschule Darmstadt
Darmstadt, Germany

ABSTRACT

It has been shown that digitally manipulated face images can pose a security threat to automated authentication systems (e.g., when such systems are used for border control). In such scenarios, a malicious actor can, in many countries, apply for an identity document using a manipulated face image, which can then be used to gain fraudulent access to a system. Research has shown that humans and algorithms struggle to detect digitally manipulated face images, especially when the type of manipulation is unknown or when evaluated across multiple types of manipulations. In this work, we consider the detection performance of algorithms and humans on datasets consisting of retouched, face swapped and morphed images. Specifically, we investigate the joint performance of algorithms and humans in a differential detection scenario where both a trusted and suspected image are presented simultaneously. To this end, we propose a conditional face image manipulation detection approach where the human decision is only considered when the algorithm is unsure about the decision outcome. The results show that the automated algorithm performs better than the human detectors and that combining the decisions of algorithms and humans, in general, leads to an increased detection performance. To our knowledge, this is the first study to explore the joint detection performance of algorithms and humans in a differential face manipulation detection scenario and when using a variety of face image manipulations.

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**; • **Computing methodologies** → **Biometrics**.

IH&MMSec '24, June 24–26, 2024, Baiona, Spain

ACM ISBN 979-8-4007-0637-0/24/06
https://doi.org/10.1145/3658664.3659649

KEYWORDS

Digital forensics; face manipulation detection; human performance; algorithm performance; information fusion

ACM Reference Format:
Mathias Ibsen, Robert Nichols, Christian Rathgeb, David J. Robertson, Josh P. Davis, Frøy Løvåsdal, Kiran Raja, Ryan E. Jenkins, and Christoph Busch. 2024. Conditional Face Image Manipulation Detection: Combining Algorithm and Human Examiner Decisions. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '24), June 24–26, 2024, Baiona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3658664.3659649

1 INTRODUCTION

Due to recent breakthroughs in artificial intelligence, and the accessibility to advanced image alteration software, the proliferation of digitally manipulated content has become a serious societal problem, with significant implications in identity and security related contexts [10, 22]. It is becoming progressively more challenging to know when digital content is authentic or fake (i.e., bona fide or manipulated) [13, 21]. For example, research has demonstrated that digitally manipulated facial images can present a security risk in authentication systems using automated face recognition, such as those used in border control systems [10]. Common digital manipulations include retouching [16], face morphing [5] and face swapping [22] (also called deepfakes). To address these issues, algorithms have been developed for detecting different types of digitally manipulated face images. Most approaches, consider only a single or a few related digital manipulations although some works have proposed algorithms for detecting different types of digital manipulations (e.g., using multi-task learning [4] or anomaly detection [9]). In numerous relevant application scenarios (e.g., when applying for a travel document in many countries), human examiners manually have to verify the authenticity of face images which have been submitted. Therefore, it is also important to explore how well humans perform at this task. Some works have investigated the performance of humans for detecting different types of digitally manipulated face images but only a few works (e.g., [2, 6]) have explored the performance in a differential detection scenario where both a trusted

and suspected image are available during detection. This scenario is relevant in the real world (e.g., when used in a biometric system), as both suspected (reference) and trusted (probe) images are available during authentication [10]. To our knowledge, none of the existing studies consider the joint performance of humans and algorithms in a differential detection scenario.

In this work, we consider the data acquired in the Darmstadt Face Manipulation Detection (DFMD) test [2] and compare the performance of a differential anomaly detection approach with the performance of humans across different experiments. Furthermore, we investigate and propose a framework for conditionally combining the decisions of humans and algorithms to obtain improved detection performance (see Figure 1). An application-oriented approach is taken in this work, where the effect and degree of human involvement in decision-making is considered and explored across different scenarios. In summary, this work makes the following contributions:

- A comprehensive analysis of the capabilities of algorithms and humans in detecting digitally manipulated face images given a differential detection scenario.
- A framework for combining the decisions of algorithms and humans for detecting manipulated face images.
- A detailed evaluation of the joint detection performance of algorithms and humans across multiple test conditions and multiple types of face manipulations.

2 RELATED WORK

Some existing works have investigated the detection performance of humans on different types of manipulated face images. In [13], the authors showed that face images which have been entirely synthetically generated using deep learning-based methods are indistinguishable from pristine images to humans. Some works (e.g., [8, 11]) have investigated the performance of humans and algorithms at detecting so-called deepfake videos. In [11], using a relatively small number of subjects per deepfake video, the authors found that deepfake videos can fool both humans and algorithms. In [8], the authors perform a more extensive study with more than 15,000 participants and show that human observers and state-of-the-art deepfake detection algorithms achieve similar performance, although making different mistakes. Furthermore, they show that participants who know the prediction of an algorithm are more accurate than either humans or algorithms alone but also that these participants seem to be more biased towards responding the same as the algorithm. In [12], the authors performed a psychophysical study of human performance in detecting three kinds of digital manipulations (i.e., retouched, morphed and face swapped images) using two test procedures: two-alternative forced-choice (2AFC) and ABX paradigms. The authors obtained an average accuracy of 75.43% for the 2AFC trials and 62.92% in the ABX trials when evaluated on 227 participants. In [17], the authors extended these experiments and found that combining multiple human examiner decisions can lead to an increased overall detection accuracy especially when factoring in the confidence of the examiners' decisions. In [7], the authors investigate the ability of humans for no-reference and differential face morphing attack detection considering both governmental employees with relevant professional expertise (e.g., border guards and document examiners) and a control group. The authors find that morphed face images cannot always be reliably detected by humans but that some experts (e.g., face comparison experts), in general, perform better for face morphing attack detection on the used tests. Furthermore, the results show that the best automated algorithms in general outperform the human observers for both no-reference and differential face morphing attack detection.

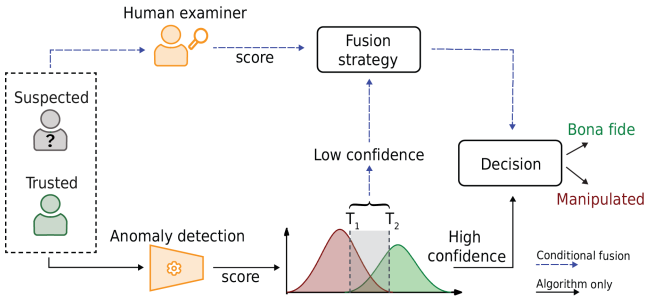


Figure 1: Overview of the proposed framework for combining human and algorithm decisions.

In [6], the authors investigated the capabilities of human examiners in a differential detection scenario considering three manipulation types (i.e., beautification, geometric distortion and morphing). The experiments involved 235 participants and consisted of 30 image pairs where approximately 33.33% of the suspected images was bona fide, 20% was beautification, 26.67% was geometric distortion and the remaining 20% was morphing. The authors found that some manipulations, especially morphing, can be difficult for humans to detect. In contrast to [6], this study considers the DFMD 1 and DFMD 2 tests which involve more participants (787) and image pairs (120) and the work focuses on the fusion of humans and algorithm decisions. A more detailed overview on relevant studies that consider human performance in detecting manipulated face images can be found in [12].

3 EXPERIMENTAL SETUP

The experiments are based on the DFMD 1 and DFMD 2 tests [2], which consist of data from 787 participants. Information about the participants is given in Table 1, which shows the gender and age distributions of the participants, whether a participant was using face analysis capabilities in a professional setting, and the order in which they took DFMD 1 and DFMD 2. As further described in [2], the experiments involve participants who have shown various face processing skills, including individuals with proven exceptional face processing skills, also called super-recognisers [1, 3, 18, 19]. The research presented in [2] finds a strong correlation between the ability to recognize faces exceptionally well and the proficiency in detecting manipulated face images.

3.1 Procedures

The procedures employed in this paper extend the experimental procedures of [12] to a differential detection scenario [2]. Specifically, the FRGCv2 [14] database and the same manipulation techniques are used. The experimental procedures are adjusted to consist of a

Table 1: Distribution of DFMD participant information based on self-reported data and the order of procedures.

Category	Subcategory	Distribution
Gender	Female	67.85%
	Male	31.26%
	Non-Binary	0.38%
	N/A	0.51%
Age	18-30	10.42%
	31-40	28.46%
	41-50	28.59%
	51-60	24.40%
	61+	8.13%
Professional Expertise	No	95.68%
	Yes	4.32%
Procedure Order	1	16.65%
	1,2	35.32%
	2	18.04%
	2,1	29.99%

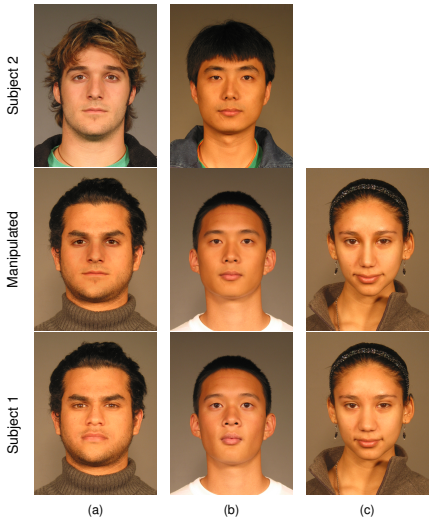


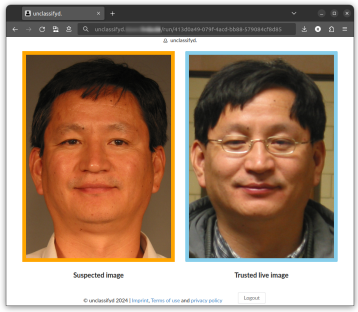
Figure 2: Examples of manipulated images generated from FRGC. (a) face swap, (b) morphing, and (c) retouching.

set of trials where both a trusted (bona fide) and suspected image are presented in each trial. The participant's task is then to determine if the suspected image is manipulated by comparing the image to the trusted image of the considered subject. The manipulated images consist of retouched [16], swapped [10] and morphed [20] images and have been selected to ensure that no apparent artefacts are present in the images. Examples of the different manipulation types are given in Figure 2.

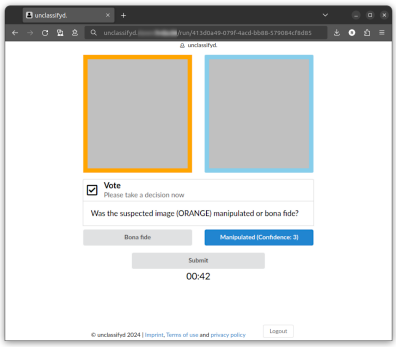
The experiment consists of two procedures, referred to as DFMD 1 and DFMD 2, employing different types of stimuli where DFMD 1 has a high prevalence of cases where the suspected image is manipulated (50%) and DFMD 2 has a lower prevalence (25%). An overview of the two procedures is given in Table 2. For each trial, the participants are asked to (1) select whether the suspected image is bona fide or manipulated and (2) provide a confidence value reflecting how sure they are in their decision ranging from 1 (unsure) to 5 (very sure). Each stimulus is displayed for 15 seconds, after which the participants have, at most, 90 seconds to make a decision. An example of a trial stimulus is shown in Figure 3a whereas Figure 3b shows the voting procedure. Two participants were removed during the evaluation as they did not provide a confidence value within the allocated time for, at least, one trial.

Table 2: Number of times the suspected image is bona fide or manipulated per procedure. Each procedure has 60 trials.

Procedure	Bona fide	Morphing	Retouching	Face swap
DFMD 1	30	15	10	5
DFMD 2	45	7	6	2



(a) Trial stimulus



(b) Trial voting phase

Figure 3: Examples from the online test. (a) Trial stimulus containing a suspected and trusted image and (b) the trial voting procedure.

3.2 Automated Detection

We utilize the differential anomaly detection method proposed in [9] for the automated detection procedure. During evaluation, the model produces an anomaly detection score. In [9], the best model obtained an average detection equal error rate (D-EER) of 4.23% when evaluated across multiple types of digital face manipulations and physical face impersonation attacks. The algorithm has been trained on only bona fide data by considering the natural changes between two bona fide images of the same subject. This approach is

used in this work because it has been trained on only bona fide data, and as such, like the human participants, it has received no prior training on the specific manipulations used for the experiments. Additionally, to our knowledge, it is the only algorithm proposed for detecting multiple types of digital manipulations in a differential detection scenario. The model used in this study corresponds to the best model proposed in [9] using a Variational Autoencoder model.

3.3 Score Normalization

To best compare and combine the decisions of humans and algorithms one must obtain a score range indicating how confident participants and the algorithm are when making a decision:

Algorithm We normalize the algorithmic scores to a range of [0,1], where a score close to 1 suggests a high likelihood of the suspected image being bona fide, while a score approaching 0 indicates a higher probability of image manipulation. This is achieved using min-max normalization, where the min and max scores are found by running the algorithm over bona fide and manipulated images from the FERET dataset [15]. As this dataset differs from the one used for the experiments, it ensures that the normalization is not unrealistic, analogous to a real-world scenario where an algorithm's true min and max values might not be known. The normalized algorithm scores are used during the experiments. Furthermore, when employed as an independent system, a threshold of 0.5 is used.

Human As described in section 3.1, humans assign a confidence value from 1 to 5 to each decision. Consider a specific stimulus s, a participant p with choice class c (i.e., bona fide or manipulated) on s and let v be the confidence value of p for s. We can then define a range for p on s as follows:

$$\mathrm{HS}(c, v) = \begin{cases} \frac{v+5}{10} & \text{if } c = \text{"bona fide"} \\ \frac{6-v}{10} & \text{otherwise} \end{cases} \tag{1}$$

In this case, a threshold > 0.5 would indicate that a participant believes that the suspected image of stimulus s contains a bona fide image.

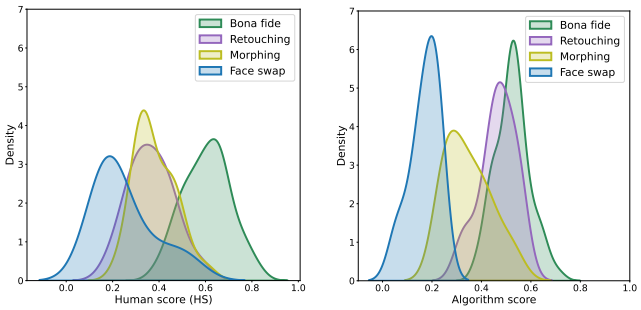
Normalized scores obtained for the bona fide and manipulated images on DFMD 1 and DFMD 2 are given in Figure 4.

3.4 Fusion strategies

In this work, we combine the decisions of humans and algorithms by using score fusion. We consider the normalized human confidence scores and the normalized algorithm score as the basis for the fusion. The following fusion strategies are explored:

Average In this fusion strategy, we perform a simple average fusion over the normalized human and algorithm scores, specifically for each participant on each trial, the normalized score (see Equation 1) is computed and fused with the algorithm score on the same trial by taking the average score. A threshold of 0.5 is used during the experiments.

SVM Another way to perform the fusion is to learn how to combine the confidence scores of humans and algorithms using machine learning. To accomplish this, a support vector machine (SVM) approach is explored. The SVM is trained using a leave-one-out protocol where the SVM is always trained on the algorithm and participants scores of the opposite test. Hence, when used during the DFMD 1 evaluation, the SVM has been trained on the DFMD 2 data and vice versa. The SVM utilizes a polynomial kernel with a degree of 3 and has been trained on all participant and algorithm scores obtained on either the DFMD 1 or DFMD 2 tests.



(a) Human scores **(b) Algorithm scores**

Figure 4: Normalized score distributions jointly on DFMD 1 and DFMD 2 for humans (a) and algorithms (b). Human scores are averaged per trial.

Conditional When automatic detection is used in collaboration with human examiners for detecting manipulated face images, it can make sense to consider human involvement only in cases where the algorithm is unsure about its decision, as this can reduce the load on the available human resources. To this end, two thresholds, T_1 and T_2, $T_1 < T_2$, can be defined, such that score fusion is only applied in case the algorithm score on a specific trial falls within the range of T_1 and T_2. Specifically, given the normalised human score s_h, the algorithm score s_a, and a fusion function $f(s_h, s_a)$, the conditional fusion strategy is given in Equation 2 and illustrated in Figure 1.

$$\mathrm{CF}(s_h, s_a) = \begin{cases} f(s_h, s_a) & \text{if } T_1 < s_a < T_2 \\ s_a & \text{otherwise} \end{cases} \tag{2}$$

The accuracy on DFMD 1 and DFMD 2 for different options of T_1 and T_2 is illustrated in Figure 5. Note, that to avoid undefined ranges as $T_1 < T_2$, the plot only shows the accuracy for the ranges where $T_1 \leq 0.5$ and $T_2 \geq 0.5$. Using a step size of 0.05, the accuracy does not increase beyond what is shown on the plot by changing the range of T_1 or T_2 to include the entire range of values from 0 to 1. Therefore, as seen, the best performance is obtained when the upper threshold, T_2 is 0.5 which given a normalized range from [0,1] would be a natural threshold to use. When choosing the optimal T_1, it is essential to consider not only its performance but also the number of images that human examiners need to assess. Therefore, during the experiments, we fix T_2 to 0.5 and find T_1 such that only 20% of the images are considered by the humans when applied on the opposite test. Hence, the thresholds are found on the opposite test than what is being evaluated in order to mitigate overfitting

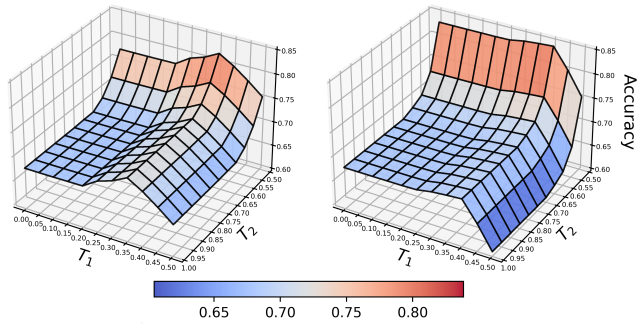


Figure 5: 3D surface grids visualising the accuracy on DFMD 1 (left) and DFMD 2 (right) when only applying the average fusion strategy in cases where the algorithm score lies in the range $[T_1, T_2]$.

to the evaluation set. During the experiments, two versions of conditional fusion are applied to correspond to the cases where the fusion function in Equation 2 is replaced by the average and SVM fusion strategies as explained above; we refer to these as conditional (average) and conditional (SVM), respectively, during the experiments.

4 EVALUATION

Table 3 shows different relevant performance measures achieved for the humans, algorithms and the different fusion strategies. Evidently, fusing the human performance with algorithm performance, under our test conditions, leads to improved detection performance. Specifically, on DFMD 1 an improvement in accuracy of more than 13.5 percent points can be observed using the conditional average fusion scheme and more than 19 percent points on DFMD 2 using a SVM fusion scheme. For both tests, it can be observed that the F1 score of both the human examiners and algorithm can be improved by fusing the scores.

Another aspect to be explored is if selecting only a subset of participants based on their self-reported perceptual face processing skills or detection performance on another test procedure, can further enhance the overall detection accuracy. To this end, participants were asked to indicate their face processing skill on a level from 0 to 100 and ranked according to their self-reported face processing skill. Additionally, by calculating the Pearson correlation coefficient (PCC) it was found that there was a moderate linear relationship (PCC≈0.33) between the accuracy of a participant across DFMD 1 and DFMD 2. Therefore, when evaluating DFMD 1, the participants can be ranked based on their performance on DFMD 2 and vice versa. The human accuracy results for the top 5% participants according to the different ranking criteria are given in Table 4. As seen, ranking the individuals based on their performance on a related test yields the best improvements where the top 5% best ranked participants always achieve better performance than when considering all the participants. For DFMD 2, an improvement in accuracy of more than 8.5 percent points can be observed. Figure 6 visualizes the accuracy in more detail when using DFMD 1 and DFMD 2 as the ranking criteria for selecting the best participants. Furthermore, the Figure shows the influence of combining only the selected participants with the algorithm scores

Table 3: Performance measures obtained on DFMD 1 and DFMD 2. Positive class is bona fide.

(a) DFMD 1

Scenario	Accuracy	Precision	Recall	F1 Score
Human	0.6550	0.6854	0.5729	0.6241
Algorithm	0.7500	0.8571	0.6000	0.7059
Average fusion	0.6807	0.7155	0.5998	0.6526
SVM fusion	0.7899	0.7208	0.9465	0.8184
Conditional (average)	0.7908	0.8137	0.7544	0.7829
Conditional (SVM)	0.7616	0.7310	0.8278	0.7764

(b) DFMD 2

Scenario	Accuracy	Precision	Recall	F1 Score
Human	0.6531	0.9233	0.5862	0.7171
Algorithm	0.7500	0.8947	0.7556	0.8193
Average fusion	0.6819	0.9296	0.6230	0.7460
SVM fusion	0.8478	0.8938	0.9045	0.8991
Conditional (average)	0.7977	0.8966	0.8254	0.8595
Conditional (SVM)	0.8087	0.8919	0.8478	0.8692

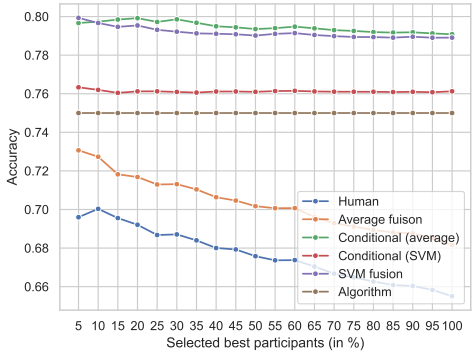
using the different fusion strategies. As can be observed, then in general selecting only a subset of best participants can lead to an improved detection accuracy although for some fusion strategies this improvement is only minor. Based on the results, and factoring in how often humans have to be involved in decision-making, the findings indicate that it is viable to consider a conditional fusion scheme where human involvement is considered for cases where the algorithm is ambivalent about its decision. Despite the accuracy being reduced with the conditional SVM fusion approach compared to when it is applied without a threshold it arguably represents an operational improvement in that human involvement is significantly reduced with only moderate differences in accuracy. This reduction in accuracy is expected, as the SVM was trained to classify optimally based on human and algorithm decisions, however requires input from both on every decision. The results in this study do not explicitly aim at optimizing the selected thresholds for the used algorithms and fusion schemes but rather suggest an intuitive and pragmatic approach. Furthermore, it would be advantageous to investigate whether automation bias affects the proficiency of human examiners when employed in conjunction with automated algorithms in a scenario involving differential face manipulation detection.

5 CONCLUSION

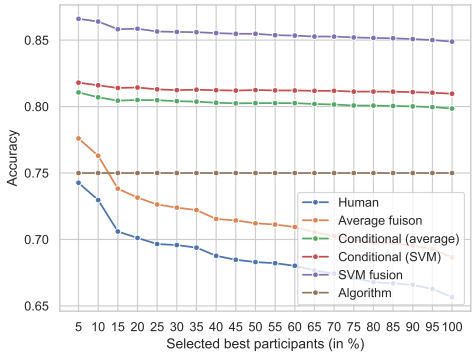
This work explored the performance of humans and algorithms for detecting three types of digitally manipulated face images in a differential detection scenario. Different fusion schemes for combining the algorithm and human examiner decisions were proposed. To this end, and to minimize the required involvement by human examiners, we proposed a conditional fusion scheme where human examiner decisions are only considered in case the algorithm

Table 4: Human accuracy results when selecting the participants based on different ranking criteria. When selecting participants based on perceptual skills all participants from a respective test are chosen. When selecting based on a specific test, only participants who completed both tests are selected.

Ranking criteria	Test	Acc. top 5%	Acc. all
DFMD 2	DFMD 1	0.6960	0.6550
DFMD 1	DFMD 2	0.7427	0.6565
Perceptual skills	DFMD 1	0.6781	0.6550
Perceptual skills	DFMD 2	0.6948	0.6531



(a) DFMD 1



(b) DFMD 2

Figure 6: Performance on DFMD 1 and DFMD 2 when selecting a subset of participants according to their performance on the other test.

lacks confidence about the decision. Lastly, it was explored if selecting only a subset of the participants, ranked based on their self-perceived face processing skills or performance on a similar detection test, could lead to further improvements. The results of this study show that under the given test conditions, combining the decisions of algorithms and humans yields improved detection accuracy. Furthermore, only considering the human decisions in cases where the algorithm is unsure about its decision can further improve the performance of the baseline systems where only the human or algorithm decisions are considered. Such a scenario could potentially minimize the number of images requiring assessment by a human examiner.

ACKNOWLEDGMENTS

The work has been partially funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

REFERENCES

[1] S. Bate, E. Portch, and N. Mestry. 2021. When two fields collide: Identifying "super-recognisers" for neuropsychological and forensic face recognition research. *Quarterly Journal of Experimental Psychology* 74, 12 (2021), 2154–2164.
[2] J. P. Davis, R. Nichols, D. J. Robertson, M. Ibsen, et al. 2024. The super-recogniser advantage extends to the detection of digitally manipulated faces. osf.io/preprints/psyarxiv/ye7ph
[3] J. P. Davis and D. J. Robertson. 2020. Capitalizing on the super-recognition advantage: a powerful, but underutilized, tool for policing and national security agencies. *The Journal of The United States Homeland Defence and Security Information Analysis Center (HDIAC)* 7, 1 (2020), 20–25.
[4] D. Debayan, X. Liu, and A. K. Jain. 2023. Unified Detection of Digital and Physical Face Attacks. In *Intl. Conf. on Automatic Face and Gesture Recognition (FG)*. 1.–8.
[5] M. Ferrara, A. Franco, and D. Maltoni. 2014. The magic passport. In *IEEE Intl. Joint Conf. on Biometrics (IJCB)*. 1–7.
[6] A. Franco, F. Løvåsdal, and D. Maltoni. 2023. On the human ability in detecting digitally manipulated face images. In *Proceedings Intl. Conf. on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*.
[7] S. R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, et al. 2023. Analyzing Human Observer Ability in Morphing Attack Detection—Where Do We Stand? *IEEE Transactions on Technology and Society* 4, 2 (2023), 125–145.
[8] M. Groh, Z. Epstein, C. Firestone, and R. Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119.
[9] M. Ibsen, L. J. Gonzalez-Soler, C. Rathgeb, P. Drozdowski, et al. 2021. Differential Anomaly Detection for Facial Images. In *IEEE Intl. Workshop on Information Forensics and Security (WIFS)*. 1–6.
[10] M. Ibsen, C. Rathgeb, D. Fischer, P. Drozdowski, and C. Busch. 2022. Digital Face Manipulation in Biometric Systems. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer Verlag, 27–43.
[11] P. Korshunov and S. Marcel. 2020. Deepfake detection: Humans vs. machines. *arXiv e-prints*, Article arXiv:2009.03155 (2020).
[12] R. Nichols, C. Rathgeb, P. Drozdowski, and C. Busch. 2022. Psychophysical Evaluation of Human Performance in Detecting Digital Face Image Manipulations. *IEEE Access* 10 (January 2022), 31359–31376.
[13] S. J. Nightingale and H. Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119.
[14] P. J. Phillips, P. J. Flynn, T. Scruggs, K.W. Bowyer, et al. 2005. Overview of the Face Recognition Grand Challenge. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. IEEE, 947–954.
[15] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing* 16, 5 (1998), 295–306.
[16] C. Rathgeb, A. Botaljov, F. Stockhardt, S. Isadskiy, et al. 2020. PRNU-based Detection of Facial Retouching. *IET Biometrics* 9, 4 (2020), 154–164.
[17] C. Rathgeb, R. Nichols, M. Ibsen, P. Drozdowski, and C. Busch. 2022. Crowd-powered Face Manipulation Detection: Fusing Human Examiner Decisions. In *Intl. Conf. on Image Processing (ICIP)*. IEEE, 181–185.
[18] D. J. Robertson, J. Black, B. Chamberlain, A. M. Megreya, and J. P. Davis. 2020. Super-recognisers show an advantage for other race face identification. *Applied Cognitive Psychology* 34, 1 (2020), 205–216.
[19] R. Russell, B. Duchaine, and K. Nakayama. 2009. Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review* 16 (2009), 252–257.
[20] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl. 2019. Detection of Face Morphing Attacks based on PRNU Analysis. *Trans. on Biometrics, Behavior, and Identity Science (TBIOM)* (2019).
[21] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131 – 148.
[22] L. Verdoliva. 2020. Media Forensics and DeepFakes: an overview. *IEEE Journal of Selected Topics in Signal Processing* (2020), 910–932.