On the Human Ability in Detecting Digitally Manipulated Face Images

On the human ability in detecting digitally manipulated face images

Annalisa Franco
Dept. of Computer Science and Engineering
University of Bologna
Cesena, Italy
annalisa.franco@unibo.it

Frøy Løvåsdal
Policing Department
National Police Directorate
Oslo, Norway
froy.lovasdal@politiet.no

Davide Maltoni
Dept. of Computer Science and Engineering
University of Bologna
Cesena, Italy
davide.maltoni@unibo.it

***Abstract*— Facial image manipulation in the context of electronic ID documents is a concrete security threat, confirmed by several real cases reported by the authorities of different countries. Such manipulations have a negative impact on the automated face recognition accuracy and should therefore be identified before the altered image is included in the document. This study reports and discusses the result of a test carried out on human examiners to evaluate their ability in detecting digital manipulations of facial images.**

***Keywords*— electronic Machine-Readable Travel Documents (eMRTD), face recognition, face image manipulation, face beautification, morphing attacks, presentation attacks.**

I. Introduction

Automated face recognition systems (FRSs) have reached impressive levels of accuracy, especially when applied in controlled application scenarios such as identity verification in electronic identity documents (e.g. biometric passports and identity cards). Face images in this case have to fulfill strict quality requirements [1] (e.g. frontal pose, neutral expression, natural skin color), which have been specifically designed to limit the possible errors in the verification process. However, in many countries, the document issuance process does not involve live enrollment for face, and citizens are allowed to bring their own ID photo printed on photographic paper. Unfortunately, such images might have been somehow digitally altered or manipulated, for instance to improve facial appearance (e.g. removing small skin defects), or could present some geometric distortions introduced by acquisition devices (for instance acquiring images at a too small subject-camera distance) or due to an uncareful printing process, or might finally have been intentionally manipulated with criminal intent. In recent years, morphing attacks has emerged as a serious security threat, as confirmed by several studies in the literature [2] [3]. Face morphing attack is a face manipulation attack that consists of mixing the faces of two subjects through a morphing process, i.e. the digital transformation of a visual representation of one subject into another. If a morphed image is included in a valid identity document, two different subjects might share it and use it for instance to cross borders at the airport ABC (Automated Border Control) gates. It is worth noting that this kind of attack is very insidious and consists in deceiving the officer who analyses the ID photo to include it in the document; if the picture is sufficiently similar to the applicant, the officer may not notice the manipulation and accept the image. Also, the above-mentioned types of manipulations could pass unnoticed, thus reducing the document utility for identity verification purposes [4]. It is therefore extremely important to spot them at the enrollment stage, before the image is stored in the document.

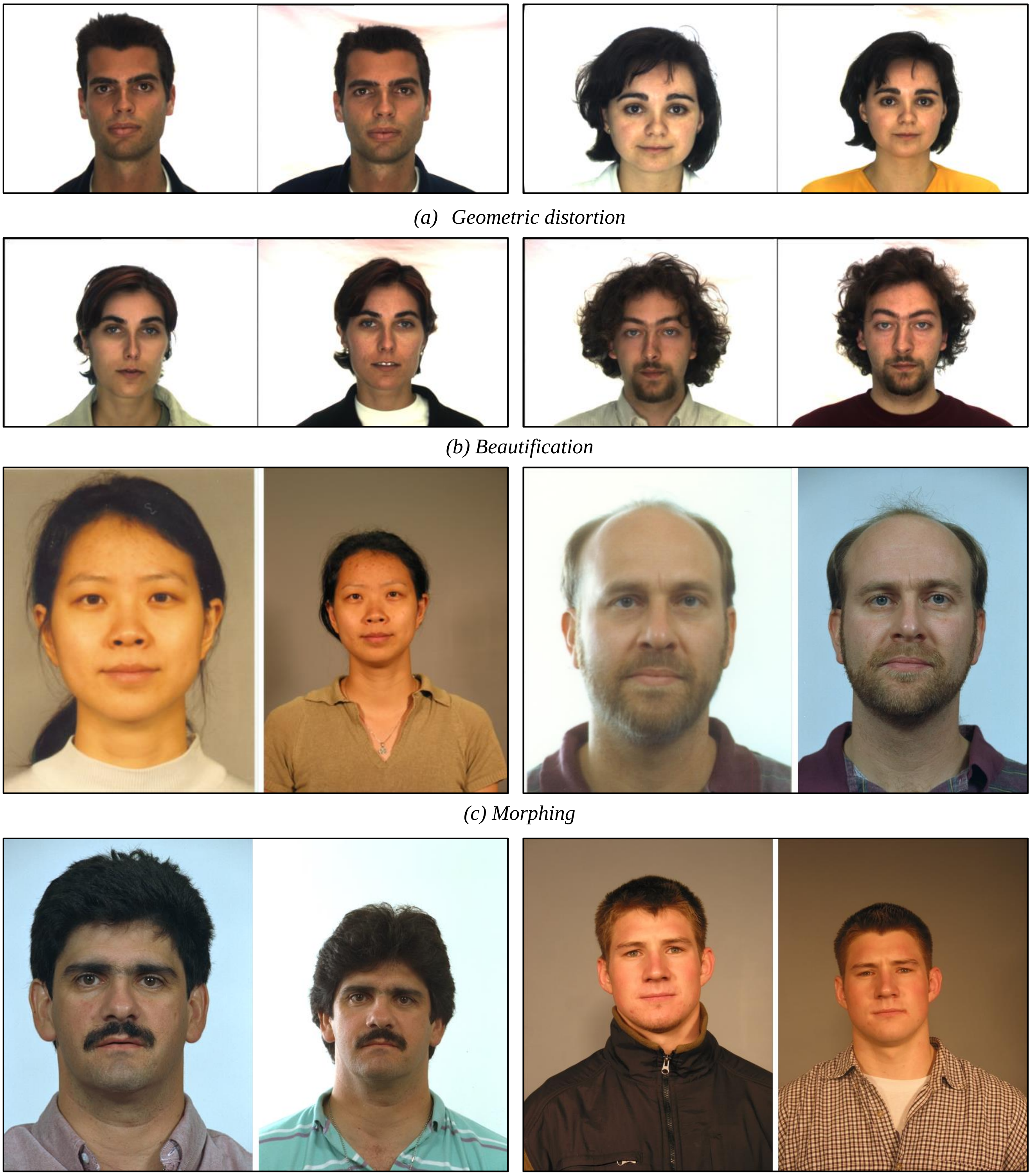
In order to assess the real extent of this potential issue, we organized a test for human examiners to evaluate their ability to detect digital manipulations in face images. Participants have been invited to evaluate the hypothetical citizen's photo to decide if it can be accepted and included in the document. A brief introduction is initially displayed to explain the context and to describe the task; participants were informed that about the manipulations considered in the test, i.e. geometric distortion, beautification and face morphing.

Different categories of examiners/observers have been involved, including border guards, case handlers (visas, residence permits, passports, etc.), document examiners, and face examiners. We also collected some information about the training attended by the examiners in order to analyze possible correlations between the accuracy level achieved and the prior training and experience.

This paper will analyze the results of this test and the result will be a valuable support to:

- assess human examiner ability to detect digital image manipulations;
- analyze possible correlations between human examiners' accuracy and the kind/duration of previous training attended (e.g. in face or document examination);
- identify the categories of manipulations more difficult to detect or for which a higher degree of uncertainty is observed in the decisions taken by examiners.

The literature reports the results of some studies, carried out to analyze the human capabilities in detecting morphed images. After the very preliminary study reported in [4], some other more extensive experiments have been conducted [5] [6] [7] [8] [9]; the main outcome of these studies is that face morphing detection is a complex task for humans, especially if they are not specifically trained on this kind of image manipulations. A few other works in the literature report experiments carried out with human examiners to assess their ability in manipulation detection [10] [11] [12] but, to the best of our knowledge, this is the first analysis including some kinds of manipulation. Moreover, a real application scenario is simulated here: the test is based on a differential approach (decision is taken based on the comparison between two images) and only experts working into different lines of work are involved in the experiment.



(a) Geometric distortion

(b) Beautification

(c) Morphing

(d) Non-manipulated

Fig. 1 Some examples of the image pairs submitted to the experts for their evaluation; experts know that the right image is always original (non-manipulated or altered) and have to take a decision on the left one. The examples reported refer to geometric distortions (a), digital beautification (b), face morphing (c), and finally (d) shows two non-manipulated samples.

II. Questionnaire structure

The test has been designed on an image-pair based approach. Several image pairs are shown to the examiner; in each pair one image (the right one) is always an original photo (not altered or manipulated) and, using this image as a reference, the examiner is asked to decide if the other image (on the left) is manipulated or not, and to quantify his/her confidence in the decision taken. Some examples of the image pairs in the test are given in Fig. 1. The test includes 30 image pairs, partitioned as follows:

- 10 image pairs include bona fide images, i.e. non-manipulated (both in digital and printed/scanned format):

- 6 image pairs include digital beautification of variable intensity;
- 8 image pairs include two kinds of geometric distortion, barrel and pincushion, also applied to different extents;
- 6 image pairs include face morphing, obtained by digitally mixing two different subjects with the specific intent of producing an image that can be matched by an FRS to both subjects, but is visually very similar to one of them (the document applicant) to increase the chance of deceiving the examiner.

The questionnaire has been developed on the EUSurvey platform[1] and the link to participate was published in the Norwegian ID Centre's website[2]. Invitations to participate were also distributed through professional networks

The participation was anonymous and no time limits where set to complete the test in order to allow participants to carefully analyze the images. At the end of the test, the obtained score was displayed to the participant, but no specific feedback was provided on the single questions to avoid possible biases.

The analysis of the results is based mainly on the accuracy, i.e. on the percentage of correct answers, and the confidence declared by each participant for each question in the test.

III. Results Analysis

This section presents and discusses the main outcomes that can be derived by analyzing the results of our experiment.

A. *Participants, line of work and training*

Overall, 235 participants took part in the experiment. Each participant has been requested to provide some information about his/her age, the line of work (each participant can even be assigned to multiple lines of work) and about possible specific training attended (face examination, document examination or other).

The age distribution of participants is given in Fig. 2, together with a box plot representing the accuracy observed for the different age groups. Participants' age ranges between 20 and 65 years, with a good balance between the different groups, especially between 25 and 50 years. Although the maximum accuracy is roughly comparable for the different age groups, some visible differences are observable in terms of variance. In fact, it's easy to observe that for some age groups the accuracy significantly varies and is quite low in some cases; in the age range 35-50 the variance is much lower, meaning that a constantly higher accuracy is achieved (neglecting a few outlier cases).

The possible impact of the line of work can be analyzed in the graphs of Fig. 3, which reports in *(a)* the percentage of participants working in the different lines of work and in *(b)* a boxplot representing the accuracy distribution across the lines of work. In particular, the following lines of work were considered: border guard 1st line, Border guard 2nd line, Case handler (Visas, Residence Permits, Asylum, Passports, Identity cards, etc.), Document examiner 2nd line, Document expert examiner, Face examiner 1st line, Face examiner 2nd line, Face expert examiner and Other.

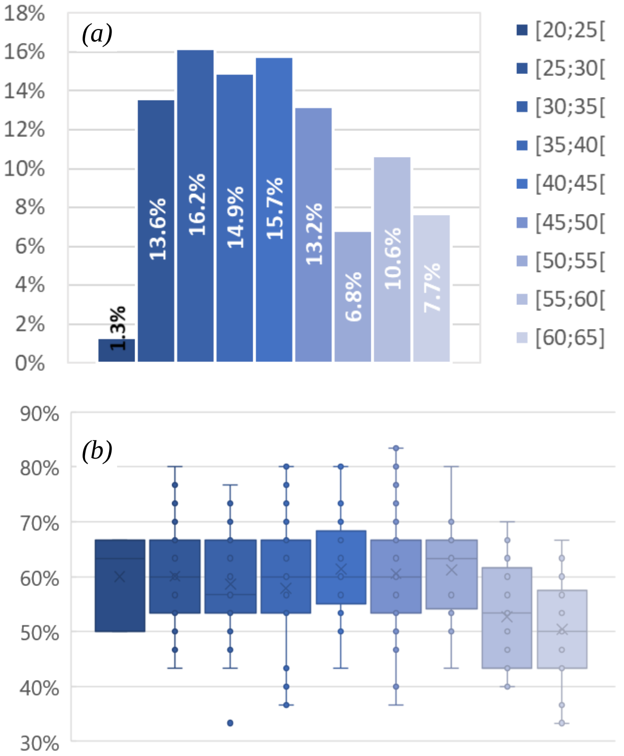


Fig. 2 Statistics on the age distribution of participants *(a)* and the accuracy achieved by the different age groups *(b)*.

Each participant can be associated to more than one line of work. Most of the participants who selected "Other" were immigration police officers or criminal investigators.

The results clearly show that the test is quite difficult, as confirmed by the overall average accuracy over all the participants of 58.6%. No marked differences can be appreciated between the accuracy observed in the different lines of work, even if some categories such as face examiners perform slightly better than others. Despite of the overall limited results, it is worth noting that some participants achieved very good results in detecting altered images, with an accuracy around 80%.

The training attended by participants does not seem to have a strong impact on the accuracy, as reported in Table 1; a bit counterintuitively, training on document examination seems to bring to overall slightly higher results. This result can be better understood if we consider that, even if this aspect varies from country to country and agency to agency, document examiners typically have some training also in face examination being the picture one of many security elements in the documents; some document examiners have therefore considerable training in face examination. A number of participants received both kinds of training, and in that case the average accuracy is 58.1%, in line with the one previously observed. Moreover, the training duration has no direct relation with the test accuracy. These results seem to confirm that human facial examination skills are to some extent innate and can only be limitedly influenced by specific training.

[1] https://ec.europa.eu/eusurvey/runner/FIMDSurvey

[2] https://www.nidsenter.no/en/subjects/face/testing/imars-project/

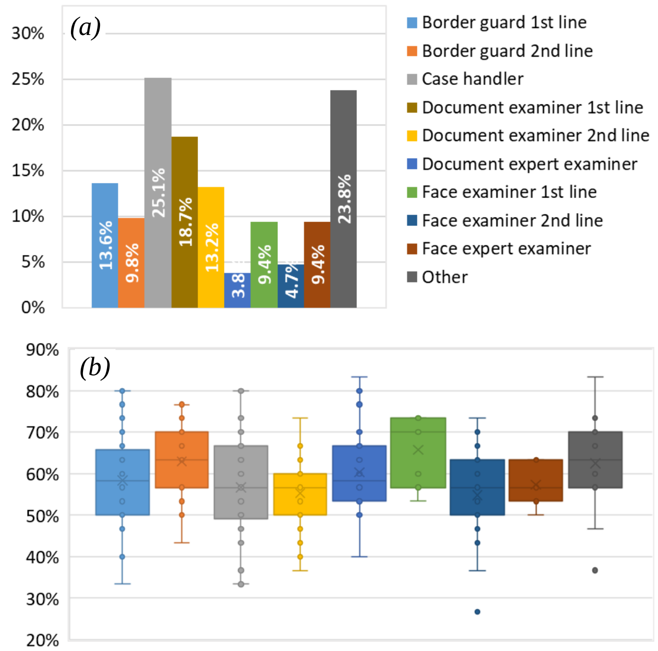


Fig. 3 Statistics on the line of work distribution of participants *(a)* and the accuracy achieved by the different age groups *(b)*.

TABLE 1 AVERAGE ACCURACY MEASURED AS A FUNCTION OF THE KIND OF TRAINING ATTENDED (DOCUMENT EXAMINATION OR FACE EXAMINATION) AND THE TRAINING DURATION.

Training period	**Document examination**	**Face Examination**
Less than 1 week	59.6%	60.3%
2-4 weeks	62.6%	60.9%
1-2 months	58.2%	55.8%
3-6 months	60.0%	60.8%
6 months 1 year	64.4%	58.9%
1 year-2 years	56.0%	58.2%
More than 2 years	56.3%	53.0%
Average	***59.4%***	***57.7%***

Anyway, the average accuracy of participants who did not receive any kind of training is 55.2, suggesting that training has a positive effect on the image examination capabilities, even if to a limited extent. It is, however, worth nothing that participants did not receive any specific training on face morphing or other kinds of image manipulation, so we are confident that the accuracy in alteration detection could be noticeably improved if properly trained.

B. *Alteration type*

Further statistics have been computed to analyze the participants' accuracy and confidence with respect to the different kinds of manipulations included in the experiment. The results are reported in Fig. 4 and Fig. 5, and clearly show interesting differences between the different kinds of alterations. The easiest manipulation to detect is face beautification, and the reason is quite easy to identify; digital beautification has a big impact on skin texture which is made smoother by this process and even unnatural when an aggressive beautification is carried out. Moreover, the facial traits are somehow visibly modified too. On the contrary, geometric distortions are very difficult to spot; the modifications introduced are not so evident and could be easily confused with slight pose variations. Morphing represents a manipulation including both geometric distortion (image warping) and texture alteration (blending), but we have to consider that in this case the morphing process has been applied with the explicit intent of obtaining an image that included facial characteristics of both subjects, but visually very similar to one of them. Some manual post-processing has been applied to the morphed images, in order to remove any visible artifact deriving from the morphing process. This kind of attack, directed to the face image examiner, confirms to be very insidious and leads to the lowest accuracy among the different alteration categories. A discrete level of accuracy is observed on bona fide images (no manipulations); most of the mistakes are reported on image pairs where the face appears at different scales, suggesting that a proper alignment of the two images to compare could ease the differential analysis.

The results in terms of confidence in the decision taken on the test image pairs is illustrated in 5 for the bona fide images as the manipulated ones; the confidence is given separately for correct and wrong answers. The general confidence level is quite low, confirming the complexity of the task. As to the different categories of images, the results confirm that facial image beautification can be detected with a higher degree of confidence, while the decisions taken on bona fide images, and on images with geometric distortion or morphing are very unsure.

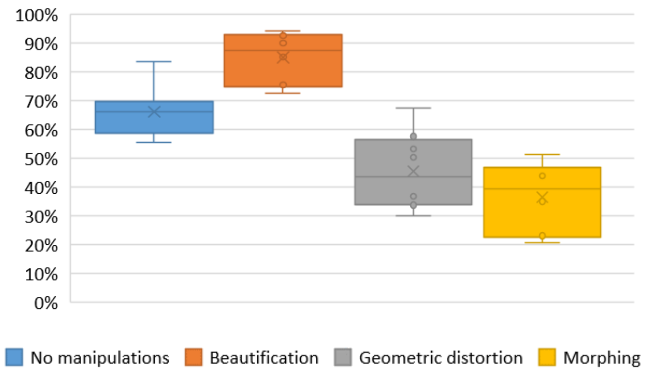


Fig. 4 Accuracy distribution for the bona fide images and the images manipulated with different kinds of alterations.

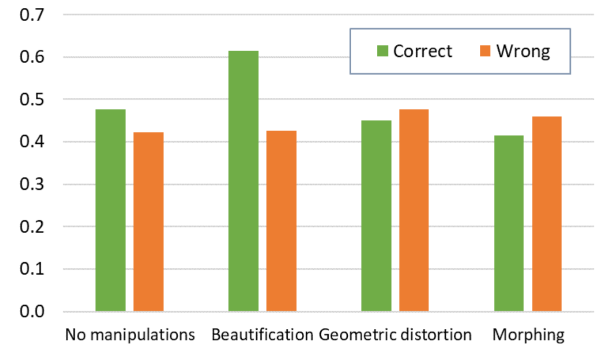


Fig. 5 Average participants' confidence in the decision taken, computed for bona fide images and images manipulated with different alterations. The average confidence is reported separately for correct and wrong answers.

To further analyze the results, the two images classified with the lowest accuracy for each category are given in Fig. 6, which also reports the percentage of correct answers. The minimum value observed is very low for morphing and geometric distortions, while non-manipulated images and beautification were easier to classify for the examiners involved in the test. An analysis of the most difficult images allows to identify some possible factors influencing the accuracy:

- Printing and scanning; this process, typical of the document enrollment pipeline in many countries, strongly impacts the image texture which becomes smoother, thus making it more difficult to identify some kinds of manipulation.
- Scale changes; when the two images are taken at different scales, the comparison seems to be more difficult;
- Illumination and pose changes in the live image, which have an impact mainly on the detection of geometric distortions.

IV. Conclusions

In this work, we analyze the capabilities of human examiners in detecting face image manipulations in a differential approach. The test, which involved a significant number of participants with different working and training experiences, clearly show that some categories of manipulations are far more difficult to detect then others.

In light of these results, our future research activity will be devoted to the development of a software tool able to support human examiners in the analysis of the face images to be included in electronic ID documents. The tool might facilitate the comparison, by highlighting, for instance, facial proportions or specific measures, which are difficult to analyze by the naked eye.

The main issues highlighted in the analysis provide some useful hints about the tool design; moreover we believe that the document [13], released by the Facial Identification Scientific Working Group, will represent a valuable reference to design effective measures covering the different facial components.

Acknowledgment

This project received funding from the European Union’s Horizon 2020 research and innovation program under Grant Agreement No. 883356. This text reflects only the author's views, and the commission is not liable for any use that may be made of the information contained therein.

V. References

[1] ISO/IEC, «39794-5:2019 Information technology — Extensible biometric data interchange formats — Part 5: Face image data,» 2019.

[2] M. Ferrara, A. Franco and D. Maltoni, «The magic passport,» in *International Joint Conference on Biometrics (IJCB)*, 2014.

[3] M. Ferrara and A. Franco, «Morph Creation and Vulnerability of Face Recognition Systems to Morphing,» in *Handbook of Digital Face Manipulation and Detection*, Springer, 2022.

[4] M. Ferrara, A. Franco and D. Maltoni, «On the Effects of Image Alterations on Face Recognition Accuracy,» in *Face Recognition Across the Electromagnetic Spectrum*, Springer, 2016, p. 195–222.

[5] D. J. Robertson, R. S. Kramer and A. M. Burton, «Fraudolent id using face morphs: Experiments on human and automatic recognition,» *PLoS One*, vol. 12, n. 3, 2017.

[6] R. S. Kramer, M. O. Mireku, R. T. Flack and K. L. Ritchie, «Face moprhing attacks: Investigating detection with humans and computers,» *Cognitive research: principles and implications*, vol. 4, n. 1, pp. 1-15, 2019.

[7] A. Makrushin, D. Siegel and J. Dittmann, «Simulation of border control in an ongoing web-based experiment for estimating morphing detection performance of humans,» in *ACM Workshop on Information Hiding and Multimedia Security*, 2020.

[8] S. J. Nightingale, S. Agarwal and H. Farid, «Perceptual and computational detection of face morphing,» *Journal of Vision*, vol. 21, n. 3, pp. 4-4, 2021.

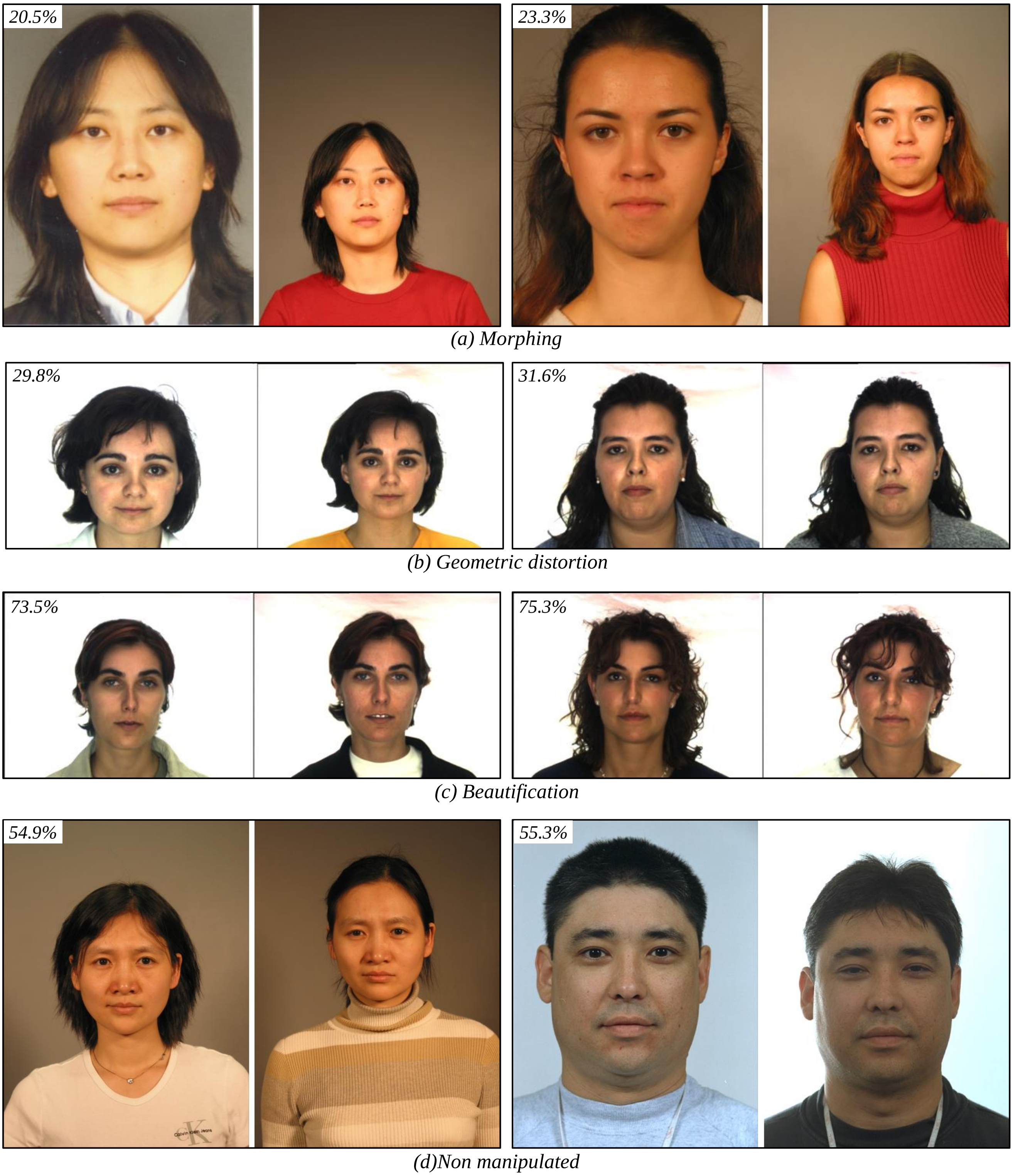
[9] S. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra and C. Busch, «Analyzing human observer ability in morphing attack detection-where do we stand?,» *IEEE Transactions on Technology and Society*, 2022.

[10] R. Nichols, C. Rathgeb, P. Drozdowski and C. Busch, «Psychophysical Evaluation of Human Performance in Detecting Digital Face Image Manipulations,» *IEEE Access*, vol. 10, pp. 31359-31376, 2022.

[11] Groh, M. Groh, Z. Epstein, C. Firestone and R. Picard, «Deepfake detection by human crowds, machines, and machine-informed crowds,» *PNAS*, vol. 119, n. 1, p. e2110013119, 2022.

[12] C. Rathgeb, R. Nichols, M. Ibsen, P. Drozdowski and C. Busch, «Crowd-Powered Face Manipulation Detection: Fusing Human Examiner Decisions,» in *International Conference on Image Processing* , Bordeaux, France, 2022.

[13] F. I. S. W. G. FISWG, «Facial Image Comparison Feature List for Morphological Analysis,» FISWG, 2018.



(a) Morphing

(b) Geometric distortion

(c) Beautification

(d)Non manipulated

Fig. 6 The most difficult images for each category: (a) Morphing, (b) Geometric distortion, (c) Beautification and (d) Non manipulated. For each pair, the percentage of correct answers is given in the top-left corner.