Building robust morphing attacks for face recognition systems

Roberto Gallardo-Cava, David Ortega-delCampo, Daniel Palacios-Alonso*, Cristina Conde, Javier M. Moguerza, Enrique Cabello
King Juan Carlos University
Madrid, Spain
daniel.palacios@urjc.es

Abstract— **In this paper, a method to build a robust morphing attack to a face verification system will be presented. The proposed method has been developed to investigate the robustness and the impact of morphing attacks to face recognition systems. In this kind of attack, an impostor accesses a face recognition system (FRS) which compares its real-time image with a stored morphed image, built with the impostor and a legitimate user. The attack succeeds when the FRS accepts the impostor and accesses the system. The current approach offers a method to build a robust attack to the FRS, in the sense that the morphed image will be closest to the decision threshold. Morphing attacks are usually evaluated only with images in which both subjects contributed in the same way to the morphing images. The image database considered, FRAV Database, was made up of 200 images. Likewise, two stages were carried out. The first stage was designed to build a baseline reference: an FRS system (trained only with legitimate users) was tested with morphed images. One contribution of this paper is that this test, which usually only considers a 50% fusion between two images has been enriched and some fusion contributions has been considered. Tests have been conducted with 20%, 40%, 50%, 60% and 80% contribution of each image to the morphed image. The comparison of the equal error rate (EER) achieved will show which contribution defines the best plausible attack. Notice that the attack that achieves the best rates with minimum disturbance of the images. The second stage consisted of the reinforcement of the FRS, training it with the contribution set defined in the previous stage. The outcomes obtained achieved improved by 3% of the EER scores.**

Keywords- morphing attack-detection, face recognition

I. INTRODUCTION

The use of facial recognition has taken an incredible relevance and is in our day to day [1]. Despite needing specific conditions capturing the biometric trait, such as scene lighting, the high accuracy and low intrusion to the subjects make the face recognition suitable for security tasks [2]. In recent years, commercial face recognition systems have been widely used with high performance and accurate results. These systems are exposed to different types of attacks already described, for instance, in [3]. The sensor attack uses a fake biometric trait to spoof a legitimate user. This kind of attacks requires that the attacker uses an artifact (3D mask or a printed image) to cheat the sensor. The second class of attack is the morphing attack in which criminal will be presented without any artifact. Morphing is a facial fusion method in which images of two subjects are merged into one [4].

This paper focuses specifically on the morphing attack due to this attack allows to easily cheat the facial verification system, impersonating the identity of the users. Indeed, the impostor uses a photo–ID in which the image is morphed with attacker and a legitimate user image. Morphing process allows a fluid transformation of an image to another [5]. Likewise, this process started being handcrafted but quickly changed due to the publication of automatic algorithms in the literature and their emergence in third–party applications. Morphing attack is one of the riskiest attacks because it is difficult to detect by attack detection systems. Unlike traditional approaches [6], the present study case does not use pre–defined 50% morphing images.

The current research can be included in the current research line of explainable artificial intelligence (XAI) aims to make predictive models making their decisions easier for humans to interpret. XAI could be categorized based on their origin, scope, and applicability, and it is extensively studied in [7]. An effective type of XAI is the counterfactual framework. Given a particular individual, this approach seeks the closest individual that alters the outcome of the prediction. Counterfactuals are used to understand the performance of the prediction system so that human beings do not require thorough knowledge of the internal logic of the model implemented by the decision system [8]. Indeed, they have been used in various fields such as computer vision [7] or adversarial risks [9], [10]. According to the current literature, counterfactuals have not been applied in the biometric area, yet.

II. PREVIOUS WORKS

In this research work, morphing and counterfactual techniques had been used. To this day, morphing images can be created through the Delaunay–Voronoi triangulation [11], [12] or generating these pictures with a generative adversarial network (GAN) [13] like [14]. Both models of morphing image generation are compared in detail in [15]. Counterfactuals provide the knowledge to obtain the target outcome by disturbing a feature of the income [16]. Counterfactual building can be categorized based on the optimality, the minimum distance to achieve the guarantees

978-1-6654-6828-2/22/$31.00 ©2022 IEEE

2022 11th Mediterranean Conference on Embedded Computing (MECO) | 978-1-6654-6828-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/MECO55406.2022.9797174

namely; the applicability, and the plausibility, the ability to generate counterfactuals that are likely under the dataset distribution [17]–[21]. There are studies about the counterfactuals in traditional machine learning approaches like tree ensembles [22], random forest [23] or support vector machines (SVM) [24]. This kind of studies mainly analyse the features to generate the separation planes, or hyperplanes, and build the counterfactuals.

III. DATABASE DESCRIPTION

The database has been acquired in a real environment of a biometric system provided by the FRAV research group [6]. The chosen database is composed of 200 subjects. The percentage of females is 59% whereas the percentage of males is 41%. Notice that 75% of the subjects are between 18 and 40 years old and the remaining are over 50 years old. There are two images for each subject with a temporal separation of up to seven years. The newest image to build morphings and the older to verify the identity of the subjects. Likewise, older images have been taken with different cameras, but always respecting acceptable conditions such as lighting and distance. The database is split into two sets of a hundred subjects. The first dataset is used for training and validation. The other dataset is used, exclusively, for testing tasks. All–by–all morphing building is performed, generating a total combination up to 100*100-100 in the process (9.900 images). Different morphing processes have been carried out changing the percentage of participation of the subjects. The following relations are explained as follows: 80%–20%, 60%–40%, 50%–50%, 40%– 60% and 20%–80%. Therefore, a total of 49.500 morphing images are generated per set (99.000 synthetic images in total).

IV. MODEL DESCRIPTION

This chapter is divided into two subsections. The first subsection is devoted to presenting the background of the morphing method. The second subsection is focused on counterfactuals.

A. Morphing construction

Given two images, Γ_A and Γ_B, and a contribution parameter denominated $\alpha \in [0...1]$, a morphing process $\Omega(\Gamma_A,\Gamma_B;\alpha)$ is an image fusion technique whose output is an image Γ_Ω, that is:

$$\Gamma_\Omega = \Gamma_{A,B} = \Omega(\Gamma_A,\Gamma_B;\alpha). \quad (1)$$

The morphing process Ω builds Γ_{AB} as a mixture of the initial pictures. This mixture contains different proportions of the input images Γ_A and Γ_B. Such proportions are established by the morphing contribution parameter α. It should be highlighted that the morphing process Ω is not that trivial. Morphing techniques based on a linear combination pixel by pixel generate ghost artifacts that are easily detected. Likewise, it includes a wrapping stage that guarantees a smooth fit of the two input images Γ_A and Γ_B, cutting down undesired artifacts such as hair or shadows between two faces/pictures. These research works created the morphing images using the Voronoi–Delaunay algorithm.

B. Counterfactuals

The internal state of modern machine learning algorithms consists of millions of variables, generating a long chain of dependency behaviours [28]. The explanation of an algorithm seeks to clarify its internal logic making decisions. Counterfactuals allow to understand dependencies on the external facts that lead to those decisions by finding similar individuals with different classification outcomes. It should be noticed that the key point, in this discussion, is the ability to obtain the desirable outcome doing the smallest change possible to a variable or a set of variables [8].

Consider a binary classification problem with classes +1 and -1, a classification function φ, a face recognition system ν and a threshold ρ. Let us define φ as:

$$\varphi(x) = \begin{cases} +1, & \text{if } \nu(x) \geq \rho, \\ -1, & \text{if } \nu(x) < \rho. \end{cases} \quad (2)$$

Given a particular individual x_f, denominated as factual, such that $\varphi(x_f) = +1$, the counterfactual of x_f, denoted as x_c, is defined as its closest individual such that $\varphi(x_c) = -1$. Up to our knowledge, there are not many research works studying the calculation of counterfactuals in a face recognition area (see [27]).

V. RESULTS AND DISCUSSION.

This chapter is divided into two subsections. Firstly, the construction of the baseline. The following point is the improvement of the system with the first counterfactual approach.

A. Construction of the baseline

FaceNet is a face recognition system (FRS) widely known, used, and referenced in the current literature. Unfortunately, this data set is not open-source, and the present work will use the implementation published in [28]. FaceNet has a good performance, tested using several databases and is a common reference in FRS literature. We will use it to test whether the morphed image is similar or not to initial images. Most part of the database images are used for training (70% of the total images) and a separate image set is used for testing (30% of the total images) [25-26].

The raw FRS obtains an equal error rate (EER) of 0.163. To obtain best results the common procedure is to train the raw FRS with images of the database considered, making FaceNet familiar with the kind of images in the problem. In our case, this reinforcement train achieves an EER of 0.131. Verification task improves around 3%. The reference system in the following study case is the FRS with reinforcement train due to this upgrade.

The system is analysed under gradual morphing attacks in the verification process (see Fig. 1). The FRS gets an EER between 0.174 and 0.420 depending on the level of morphing fusion. The higher level of fusion, the higher EER has the

FRS. Notice that the more EER means, the more errors in the system (both false acceptation rate and false rejection rate) and, therefore, the system assigns incorrect labels to morphing images.

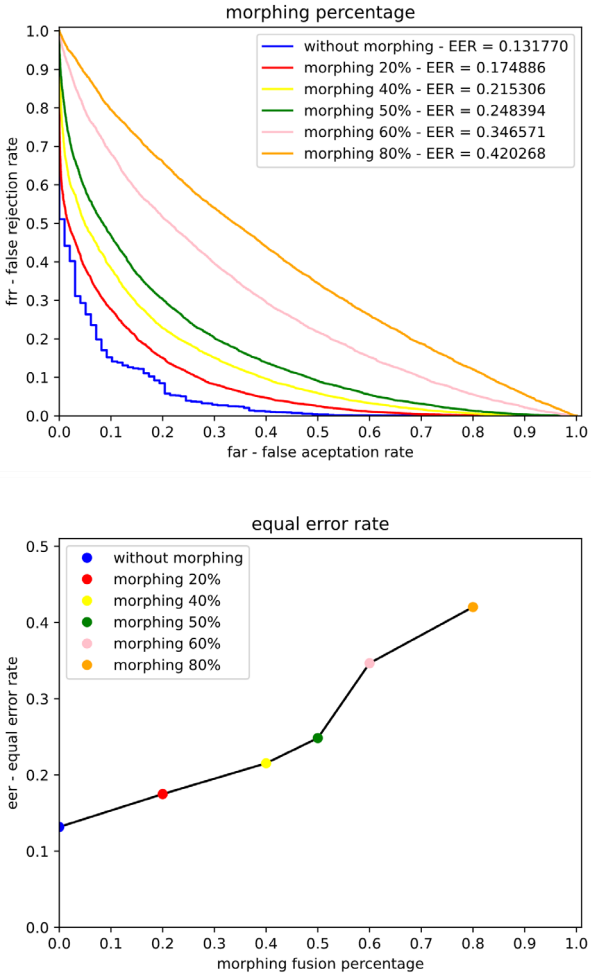

Fig. 1. Analysis of the FRS trained with the database. (top) FRS ROC trained with the database and (bottom) FRS EER trained with the database

B. *Improving the system with global counterfactuals*

In the previous experiment, the biggest leap in the EER is the stripe between 50–60% morphing fusion. For this purpose, we consider in this section that the decision boundary is in this region. It seems to be that the morphing images with 60% fusion can act as counterfactuals, because these pictures are the closest over the region boundary, setting $\alpha = 0.6$ and $\rho \leq \nu(\Omega(\Gamma_A,\Gamma_B;0.6))$. The FRS is trained with the counterfactuals and analyzed under the previous gradual morphing attack. The scores generated by the FRS get EER values between 0.146 and 0.385 depending on the contribution of the morphing images (see Fig. 2 (top image)). This model gives an EER of 0.109 in verification without attacks. The FRS trained with counterfactuals improves around 3% on all results, specially, in the morphing images with a bigger percentage of mixture. This system makes a special effort flattening the 50–60% stripe (see Fig. 2 (bottom image)) which is the critical point observed in the previous model. Furthermore, the FRS trained with counterfactuals gets better results in the verification process without morphing.

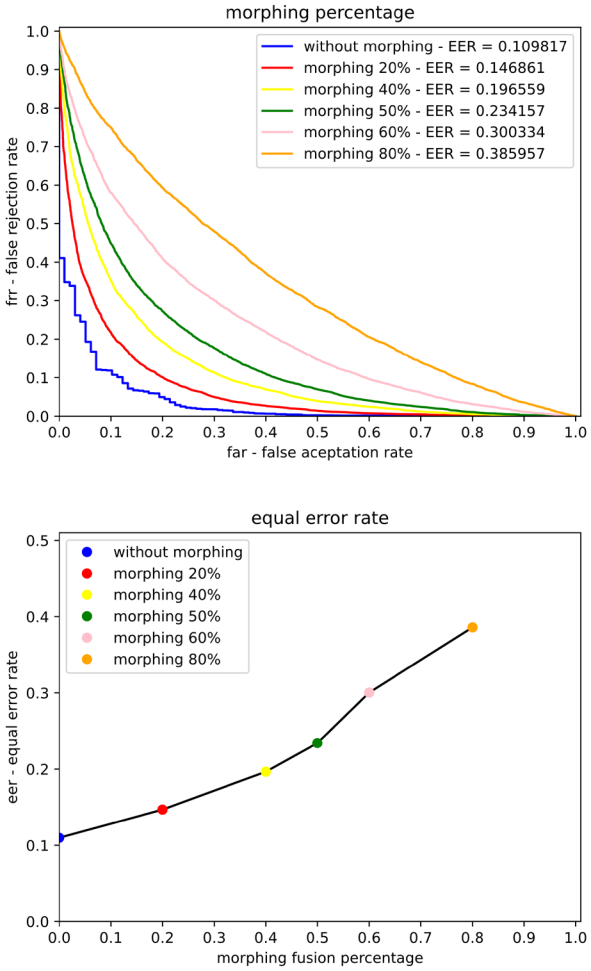

Fig. 2. Analysis of the FRS trained with α =0.6. (top) FRS ROC trained with α =0.6 and (bottom) FRS EER trained with α =0.6.

VI. CONCLUSIONS

This research work presents one of the first attempts to develop a counterfactual construction approach to increase the robustness of biometric face verification systems to morphing attacks. A novel counterfactual morphing bisection method (CMBM) is addressed, and its improvement is evaluated for a complete morphing face database.

This study explains how to improve the training of a facial biometric system through counterfactuals. It includes a substantial and mathematical morphing construction and the explanation about the selection of which one's act as counterfactuals. This research works created the morphing using the Voronoi–Delaunay algorithm, but it could be

modified to any other morphing construction method, such as generative adversary networks. Also, this investigation uses FaceNet as FRS, but any other system could be adapted instead.

A new database has been developed with a large amount of gradual morphing images (20%, 40%, 50%, 60% and 80%). These images are verified by the FRS generating a score table. The hardest morphings to be verified by the system can be located analysing the EER of the scores. This analysis creates a big stripe between 50–60% on morphing contribution scores (see Fig. 1). Being this band the boundary region, and the morphing images with 60% fusion the counterfactuals because they are the closest over the area division. Then, the system is trained with the counterfactuals. Training the FRS with counterfactuals improves the verification without attacks around 2%.

REFERENCES

[1] D. O. Gorodnichy, "Evolution and evaluation of biometric systems," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*. IEEE, 2009, pp. 1–8.

[2] D. N. Parmar and B. B. Mehta, "Face recognition methods & applications," *arXiv preprint arXiv:1403.0485*, 2014.

[3] C. Roberts, "Biometric attack vectors and defences," *Computers & Security*, vol. 26, no. 1, pp. 14–25, 2007.

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

[5] G. Wolberg, "Image morphing: a survey," *The visual computer*, vol. 14, no. 8, pp. 360–372, 1998.

[6] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, "Border control morphing attack detection with a convolutional neural network de-morphing approach," *IEEE Access*, vol. 8, pp. 92301–92313, 2020.

[7] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[8] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[9] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.

[10] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 895–905.

[11] L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch, "Prnu variance analysis for morphed face image detection," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–9.

[12] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Detection of face morphing attacks by deep learning," in *International Workshop on Digital Watermarking*. Springer, 2017, pp. 107–120.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, "Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10.

[15] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020, pp. 1–6.

[16] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617.

[17] R. R. Fernandez, I. M. de Diego, V. Ace´ na, A. Fern˜ andez-Isabel, and´ J. M. Moguerza, "Random forest explainability using counterfactual sets," *Information Fusion*, vol. 63, pp. 196–207, 2020.

[18] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[19] A. Adhikari, "Example and feature importance-based explanations for black-box machine learning models," 2018.

[20] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "Inverse classification for comparison-based interpretability in machine learning," *arXiv preprint arXiv:1712.08443*, 2017.

[21] S. Liu, B. Kailkhura, D. Loveland, and Y. Han, "Generative counterfactual introspection for explainable deep learning," *arXiv preprint arXiv:1907.03077*, 2019.

[22] H. Deng, "Interpreting tree ensembles with intrees," *International Journal of Data Science and Analytics*, vol. 7, no. 4, pp. 277–287, 2019.

[23] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, "Interpretable predictions of tree-based ensembles via actionable feature tweaking," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 465–474.

[24] N. H. Barakat and A. P. Bradley, "Rule extraction from support vector machines: A sequential covering approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 6, pp. 729–741, 2007.

[25] M. A. Shahin, H. R. Maier, and M. B. Jaksa, "Data division for developing neural networks applied to geotechnical engineering," *Journal of Computing in Civil Engineering*, vol. 18, no. 2, pp. 105–114, 2004.

[26] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation," 2018.

[27] A. Barredo-Arrieta and J. Del Ser, "Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[28] D. Sandberg, "Face recognition using tensorflow," https://github.com/davidsandberg/facenet accessed: 2021-04-28.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.