Explainable Visualization for Morphing Attack Detection

Henning Myhrvold[1], Haoyu Zhang[1], Juan Tapia[2], Raghavendra Ramachandra[1], and Christoph Busch[1,2]

[1] Norwegian University of Science and Technology, Department of Information Security and Communication Technology, NTNU Gjøvik Teknologiveien 22, 2815 Gjøvik, Norway. * haoyu.zhang@ntnu.no

[2] Hochschule Darmstadt, Schöfferstraße 3, 64295 Darmstadt, Germany

Abstract. Detecting morphed face images has become critical for maintaining trust in automated facial biometric verification systems. It is well demonstrated that better biometric performance of the Face Recognition System (FRS) results in higher vulnerability to face morphing attacks. Morphing can be understood as a technique to combine two or more look-alike facial images corresponding to the attacker and an accomplice, who could apply for a valid passport by exploiting the accomplice's identity. Morphing Attack Detection (MAD), with the help of Convolutional Neural Networks (CNN), has demonstrated good performance in detecting morphed images. However, they lack transparency, and it is unclear how they differentiate between bona fide and morphed facial images. As a result, this phenomenon needs careful consideration for safety and security-related applications. This paper will explore Layer-wise Relevance Propagation (LRP) to determine the most relevant features. We fine-tune a VGG pre-trained network for face morphing attack detection and LRP is then used to investigate the decision-making processes to understand what input pixels take part in the attack detection. This paper shows that CNN considers only a small part of the image, usually around the eyes, nose, and mouth.

Keywords: Face morphing · Morphing Attack Detection · Layer-wise Relevance Propagation · VGG19 · Deep Neural Network · Convolutional Neural Network

1 Introduction

Face recognition systems (FRS) is a technology that enables an individual to be recognized based on their unique biological traits, which are often represented in an identity document (e.g. passport) by means of a facial image [15]. Given the strong verification performance of such systems, an adversary can execute targeted attacks against FRS that use morphed face images, as presented by Ferrara

* This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356.

et al. [2]. Face morphing is the smooth transformation of two facial pictures into one. Morphing attacks have developed as a severe threat to enrolment in recent years, undermining facial recognition systems' capabilities. The most common scenario of face morphing attacks is automatic border control. In many contries, the enrolment of passport photos is not conducted by live-capturing [7] and hence the enrolled images may be manipulated by the attackers. To effectively address the face morphing attack problem, researchers have devised several face Morphing Attack Detection (MAD) algorithms based on both hand-crafted and deep learning techniques [15].

With the advancement of deep learning algorithms, biometric-based identification and verification have become a commonly utilized methodology for a variety of secure access control applications [15]. Classification of images has become a critical component of a wide variety of computer vision applications, with nonlinear methods such as convolutional neural networks (CNNs) serving as the gold standard [4]. While approaches based on learned features can reach a high level of accuracy, they act like black boxes [11]. As neural networks become more widely used, the topic of how these models' conclusions may be interpreted becomes increasingly important [8]. While precision is necessary for network performance, generality and robustness are equally critical. One aspect of neural networks is that they frequently employ only the data necessary to perform their task and reject additional helpful information. Worst case, a neural network learns to make correct decisions for the wrong reasons [13].

The discipline of explainable artificial intelligence has seen the development of a plethora of methodologies. Bach et al. [1] proposed the concept of layer-wise relevance propagation, which has established itself as a notable method for enhancing the interpretability of CNNs. This explanatory approach generally examines the model's interpretability from a black-box perspective and will be utilized in this paper concerning MAD networks.

2 Background

Morphing can be understood as a technique to combine two or more look-alike facial images from one subject and an accomplice, who could apply for a valid passport by exploiting the accomplice's identity [15]. This technique can be used to construct manipulated biometric samples that represent biometric information from both contributing subjects, as seen in Figure 2. Such face morphing attacks have implications on identity verification procedures, like those conducted at country borders [11].

Since the morphing process alters the pixel positions, some mismatched pixels may result in noise-generating artefacts and ghost-like pictures, giving the photos an unrealisitic appearance [15]. After creating the morphed face image, it can be further processed and manipulated to remove or minimize these unnatural aspects. Automatically created morphs may introduce artefacts, which can be avoided if the attacker creates a single high-quality morph and manually optimizes the final image [9]. In general, it is anticipated that mechanically cre-

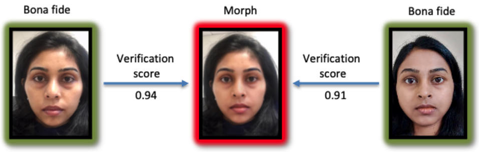

Fig. 1. Face morph illustrated in the middle getting a high similarity score against two bona fide samples from different individuals. Illustration is adapted from [15]

ated databases of morphed face photos will have a lower quality than real-world attack scenarios [9].

2.1 Morphing Attack Detection

The MAD algorithms proposed thus far have been trained and evaluated on datasets with constrained distributions of image features and technological variety [10]. The recent NIST FRVT MORPH results [5] indicate that most MAD algorithms submitted lack resilience and performance when applied to unknown datasets [10].

Single Image Based MAD (S-MAD) approaches are designed to detect a face morphing attack effectively using a single image supplied to the algorithm. The morphed image might be digital or re-digitized [15]. S-MAD is a difficult task since it is supposed to be robust against differences in sample quality, multiple types of cameras, morph creation tools, and various print-scan procedures [15] [14].

Researchers have successfully used deep learning S-MAD algorithms to classify bona fide and morphed images. Most previously published work uses pre-trained networks and transfer learning to cope up with small datasets. Although deep CNNs outperform hand-crafted texture descriptor-based MAD algorithms on both digital and print-scan data, their generalizability and robustness are restricted [15].

2.2 Explainability of deep learning models

Since deep learning is doing an excellent job in detecting morphing attacks, it is desired to explain what information the algorithm uses in its decision-making. Visualization techniques are a common approach. Most approaches for face morph detection are trained and evaluated on a single database utilizing a single morph generation algorithm [9] [14]. As a result, the training data must have a high degree of variance to avoid overfitting on database-specific artefacts [10].

Layer-wise Relevance Propagation Introduced by Bach et al. [1], the Layer-wise Relevance Propagation (LRP) interpretability approach assigns significance to each pixel in the input image [11]. The LRP's mathematical foundation is built on a deep Taylor decomposition. It assigns relevance layer by layer, starting with a single selected neuron representing a single class and ending with the image via the CNN. Each layer's relevance is communicated backwards into the preceding one by a set of rules. We follow the rules currently regarded as best practice for LRP [12]. These are epsilon-decomposition for the fully connected layers. Alpha-beta - decomposition with $a = 2$, $b = -1$ and flat decomposition for the convolutions layer [12]. These criteria are intended to direct attention to the neurons in the prior layer that are required for each neuron in the current layer to fire [12].

LRP considers the CNN's overall structure, the classification component, and the convolutional layer activations and weights. The relevance is assigned so that regions that considerably contribute to the activation are given a positive value, illustrated in Figure 2 with red colour. In contrast, areas that significantly inhibit its activation are assigned a negative value, presented with blue colour. This enables the production of finer heatmaps and assigning a relevance score to each pixel, defining its ability to either contribute to or prevent activation [11].

3 Methodology

To study the explainability of deep-learning based S-MAD algorithms, we first train a VGG19 network to classify morphed and bona fide images and then use LRP to interpret what has been learned by the model to make the classification. Due to the effort needed for manual post-processing for high-quality morphs, the size of the morphing dataset is usually limited. Hence, the VGG19 network in this work is fine-tuned based on weights pre-trained on ImageNet-1k dataset.

For the explanation of the model, we employ LRP [3] to gain insight into the decision-making process to interpret the MAD accuracy and robustness [13]. More specifically, we use LRP to determine the input relevance in the bona fide and morphed images. Different kinds of explanatory photos are computed and visualized for discussions.

The dataset used in this paper is a subset of the database presented by Zhang et al. [16]. The data originates from the FRGC-v2 dataset and consists of 140 unique participants from the FRGC-v2 collection based on the high-quality facial photos, where images similar to passport pictures were chosen. 47 of the 140 data individuals are female, whereas 93 are male. Each subject has a sample size of 7 to 21 images. The images are cropped by utilizing MTCNN presented by Zhang et al. [17]. Finally, in total 9971 morphs are generated using landmark-based morphing algorithm [6] and 8176 bona fide images are included in the dataset.

4 Experiments & Results

Since MAD can be considered a binary classification problem, the following metrics are widely used to benchmark MAD algorithms. The performance of the detection algorithms is reported according to metrics defined in ISO/IEC 30107-3 [3]. The Attack Presentation Classification Error Rate (APCER) is defined as

[3] https://github.com/fhvilshoj/TorchLRP

the proportion of attack samples incorrectly classified as bona fide images [15]. The Bona fide Presentation Classification Error Rate (BPCER) is defined as the proportion of bona fide images incorrectly classified as morphed images in the system [14].

The fine-tuned neural network that yielded the best results got a training accuracy of 0.995 and a validation accuracy of 0.892. Based on the results of this fine-tuned MAD algorithm, we used LRP to visualize what the neural network had used in its decision-making process. The bona fide images are presented in Figure 2 top, and the morphed images are presented in Figure 2 Bottom. Table 1 contains the models performance metrics.

Table 1. Overview of APCER and BPCER and the values used to calculate them for the validation phase. Sorted by the 10 epochs with the highest accuracy score.

Epoch	TP	TN	FP	FN	APCER	BPCER	Accuracy
1	5288	4183	116	1034	0.027	0.164	0.892
2	5260	4109	144	1108	0.034	0.174	0.882
3	4400	4864	1004	353	0.171	0.074	0.872
4	5356	3903	48	1314	0.012	0.197	0.872
5	5192	4024	212	1193	0.050	0.187	0.868
6	4900	4198	504	1019	0.107	0.172	0.857
7	3988	4954	1416	263	0.222	0.062	0.842
8	4368	4541	1036	676	0.186	0.134	0.839
9	5400	3497	4	1720	0.001	0.242	0.838
10	5376	3509	28	1708	0.008	0.241	0.837

Figures 2 shows a small sub-sample of twelve bona fide and morphed pictures from the overall dataset. Using LRP, we see what went into the decision of the MAD algorithm with the highest accuracy when classifying the images as bona fide or morphs.

The visual results show that the neural network primarily focuses on the eyes, nose, and mouth. This is best illustrated in the patternnet explanations. The algorithm also takes into account some of the hair features, as well as the edges of the faces. The patterned explanation shows a pattern where most of the image negatively influences its decision, while some areas in the forehead and cheek positively influence its decision.

5 Discussion

It should be mentioned that reliable detection of face morphing attacks continues to be a challenge, and numerous open issues exist in the research field of MAD algorithms [9]. One of these issues is the absence of large-scale publicly available datasets with more individual variation and a technological variation to reflect the real world [15]. In addition, generating high-quality face morphs automatically continues to be complicated. The dataset used in this paper ar-

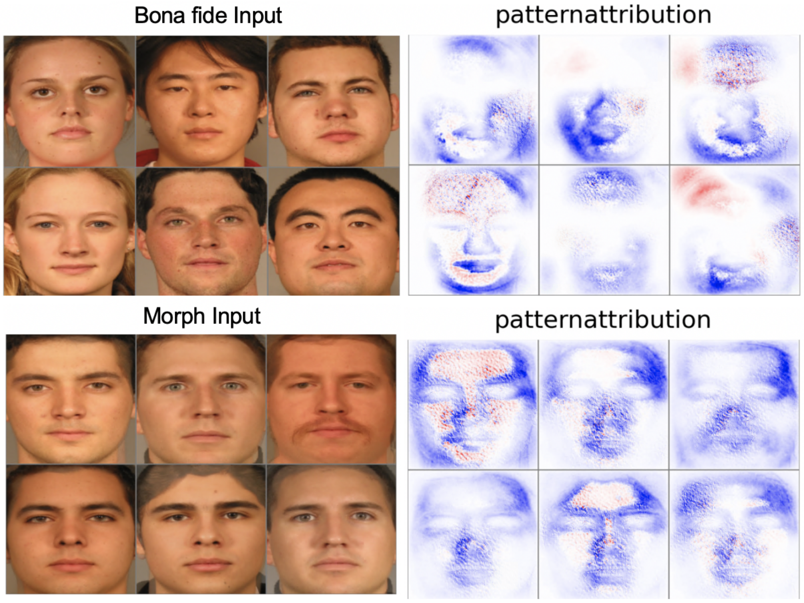

Fig. 2. Visualisation using the patternnet explanation. Top: bona fide images. Bottom: Morph images.

guably has more artefacts and ghosting in the morphed images than an attacker would achieve manually for a specific purpose.

From Table 1, the validation accuracy during the top ten epochs during fine-tuning is relatively high. However, the APCER and BPCER values show the algorithm is still inconsistent in its classification. The calculations indicate that the model cannot be used for any meaningful classification of bona fide or morphed images without more optimization and training.

We discovered using LRP that our fine-tuned neural network, with the highest accuracy, concentrates on the eyes, nose, and mouth areas while detecting morphed facial images. In most cases, the rest of the image is ignored. While focusing on these areas may be adequate to achieve relatively high accuracy, it has limitations. There is a high degree of inconsistency in what the algorithm deems relevant between the different ways of visualizing the input relevance. This could have unforeseen problems, which is especially serious for security-related applications. In critical systems, the algorithm should consider data from all image locations during the classification process to achieve robustness.

Due to the complexity of the behaviour of a CNN's fully connected layers, the relevance scores generated by LRP are not immediately interpretable, demanding additional research to comprehend the network's overall behaviour completely. Seibold et al. [12] found that LRP commonly assigns high relevance scores to artefact-free regions in morphed face shots, implying that these regions are critical for the decision to classify the images as morphs. This aligns with our results where we see that the neural network assigns relevance to areas without any visible artefacts and fails to detect other areas with clearly visible artefacts. In Figure 2 Bottom, this is visible where the hairline of multiple morphed images has artefacts that the algorithm does not detect very well.

Using LRP to visualize the decision-making process of convolutional neural networks fine-tuned for morph attack detection has been shown to be inconsistent. Problems with limited high-quality large datasets are an issue that makes the results of LRP challenging to assess and necessitates further research.

6 Conclusion

Given the strong generalization capabilities of face recognition systems, an adversary can execute targeted attacks that use morphed face images, as presented by Ferrara et al. [2]. Face morphing attacks are a significant security concern, given that several countries enable residents to provide a picture for passports or national identification cards. Without requiring specialist knowledge, these photos can be forged utilizing readily available tools or websites [13]. Advances in deep learning and machine learning approaches have enabled the development of relatively good-quality morphs through various novel techniques. Generalizing morphing attack detection is still a long way off, given the fundamental difficulty of gathering large-scale public databases with a variety of morph production strategies [15]. Robust MAD algorithms must account for the vast diversity of picture post-processing, printing, and scanning technologies. The observed accuracy for detecting face image morphing attacks does not yet represent generalization to datasets containing a range of real-world capture situations [9].

We discovered through LRP that a fine-tuned neural network focuses mainly on the eyes, nose, and mouth to detect morphed images. Though neural network analysis is still in its infancy, we demonstrated how methods such as LRP may be utilized to get valuable insights into a neural network's decision-making process. Future strategies for modifying the training process, particularly the training data, to increase resilience can be developed from this knowledge. Through the experiments, we got insights into how to train a network specifically for detecting face morphs and, more broadly, visual results on what the neural network used in its decision-making. High accuracy does not always imply robustness, implying the necessity for additional quality measures. We presented relatively high training and validation accuracy metrics for our MAD algorithm. LRP showed that most of the input images got ignored, implying a lack of robustness. In addition, our results show inconsistency in the algorithm's ability to detect artefacts and other features of morphed images.

Future work could improve the used dataset and codebase, trying to decrease the computational cost of training the neural network while still being able to calculate essential performance metrics for the MAD algorithm. Also, the applied experiments could also be extended for explaining the detection of other type of attacks against face recognition system. Meanwhile, it would be interesting to study on applying the explainability technique for monitoring CNN-based systems and detecting abnormal behaviours caused by general input attacks.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (July 2015). https://doi.org/10.1371/journal.pone.0130140
2. Ferrara, M., Franco, A., Maltoni, D.: The magic passport. In: IEEE International Joint Conference on Biometrics. p. pp. 4 (2014). https://doi.org/10.1109/BTAS.2014.6996240
3. ISO/IEC: Iso/iec 30107-3, information technology - biometric presentation attack detection - part 3: Testing and reporting (2017)
4. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: The lrp toolbox for artificial neural networks. The Journal of Machine Learning Research **17**(1), pp. 1–4 (2016)
5. Ngan, M., Grother, P.J., Hanaoka, K.K., Kuo, J.: "face recognition vendor test (FRVT): MORPH - performance of automated face morph detection" National Institute of Technology (NIST) (2022)
6. Raghavendra, R., Raja, K., Venkatesh, S., Busch, C.: Face morphing versus face averaging: Vulnerability and detection. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 555–563 (2017). https://doi.org/10.1109/BTAS.2017.8272742
7. Raja, K., Ferrara, M., Franco, A., Spreeuwers, L., Batskos, I., de Wit, F., Gomez-Barrero, M., Scherhag, U., Fischer, D., Venkatesh, S.K., et al.: Morphing attack detection-database, evaluation platform, and benchmarking. IEEE transactions on information forensics and security **16**, 4336–4351 (2020)
8. Raulf, A.P., Däubener, S., Hack, B., Mosig, A., Fischer, A.: Smoothlrp: Smoothing lrp by averaging over stochastic input variations. In: ESANN (2021)
9. Scherhag, U., Rathgeb, C., Merkle, J., Breithaupt, R., Busch, C.: Face recognition systems under morphing attacks: A survey. IEEE Access **7**, pp. 23012–23014, 23016, 23019–23026 (2019). https://doi.org/10.1109/ACCESS.2019.2899367
10. Scherhag, U., Rathgeb, C., Merkle, J., Busch, C.: Deep face representations for differential morphing attack detection. IEEE Transactions on Information Forensics and Security **15**, pp. 3625–3637 (2020). https://doi.org/10.1109/TIFS.2020.2994750
11. Seibold, C., Hilsmann, A., Eisert, P.: Feature focus: Towards explainable and transparent deep face morphing attack detectors. Computers **10**(9), pp. 1–7, 9, 12, 14 (2021). https://doi.org/10.3390/computers10090117, https://www.mdpi.com/2073-431X/10/9/117
12. Seibold, C., Hilsmann, A., Eisert, P.: Focused lrp: Explainable ai for face morphing attack detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. pp. 88–92, 95 (January 2021)
13. Seibold, C., Samek, W., Hilsmann, A., Eisert, P.: Accurate and robust neural networks for face morphing attack detection. Journal of Information Security and Applications **53**, pp. 1–6, 8, 10–11 (2020). https://doi.org/https://doi.org/10.1016/j.jisa.2020.102526, https://www.sciencedirect.com/science/article/pii/S2214212619302029
14. Tapia, J.E., Busch, C.: Single morphing attack detection using feature selection and visualization based on mutual information. IEEE Access **9**, pp. 1–2, 5–6, 10 (2021). https://doi.org/10.1109/ACCESS.2021.3136485

15. Venkatesh, S., Ramachandra, R., Raja, K., Busch, C.: Face morphing attack generation and detection: A comprehensive survey. IEEE Transactions on Technology and Society **2**(3), pp. 128–145 (2021). https://doi.org/10.1109/TTS.2021.3066254
16. Zhang, H., Venkatesh, S., Ramachandra, R., Raja, K., Damer, N., Busch, C.: Mipgan—generating strong and high quality morphing attacks using identity prior driven gan. IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(3), 365–383 (July 2021). https://doi.org/10.1109/TBIOM.2021.3072349
17. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342