Colour Processing in Adversarial Attacks on Face Liveness Systems

L. Abduh and I. Ivrissimtzis

Durham University, Department of Computer Science, UK

Abstract

*In the context of face recognition systems, liveness test is a binary classification task aiming at distinguishing between input images that come from real people's faces and input images that come from photos or videos of those faces, and presented to the system's camera by an attacker. In this paper, we train the state-of-the-art, general purpose deep neural network **ResNet** for liveness testing, and measure the effect on its performance of adversarial attacks based on the manipulation of the saturation component of the imposter images. Our findings suggest that higher saturation values in the imposter images lead to a decrease in the network's performance. Next, we study the relationship between the proposed adversarial attacks and corresponding direct presentation attacks. Initial results on a small dataset of processed images which are then printed on paper or displayed on an LCD or a mobile phone screen, show that higher saturation values lead to higher values in the network's loss function, indicating that these colour manipulation techniques can indeed be converted into enhanced presentation attacks.*

CCS Concepts

• ***Computing methodologies*** → ***Computer vision tasks;*** *Image manipulation;*

1. Introduction

Face recognition [ZCPR03] has long been established as the biometric method of choice for everyday applications, such as a mobile phone or PC login. However, its use in security critical applications is currently restricted to controlled environments, such as airport passport control, but not, for example, money withdrawal from street ATM machines.

The main reason behind this limitation is that face recognition is considered particularly vulnerable to *presentation attacks*, where one may gain access by presenting in front of the system's camera a photo or a video of the user they impersonate [GMF14]. Developed as countermeasures to such attacks, *liveness tests* are binary classifiers aiming at distinguishing between the genuine *client* and illegitimate *imposter* images or videos.

In this paper, we study *adversarial attacks* on liveness tests based on deep neural networks. In particular, we study the extent to which increasing the saturation of an imposter face image degrades the ability of the neural network to classify it correctly. Our approach was motivated by the observation that, generally, the client images have more vivid colours than the imposter ones.

We note that the study of adversarial attacks on machine learning classifiers is the focus of a large body of recent research and is considered one of the main methodologies for understanding and improving neural network performance. However, while deep neural networks are establishing themselves as the state-of-the-art in almost every classification task, to the best of our knowledge adversarial attacks on liveness tests have not been studied, with the exception of [OI16b, OI16a], where however adversarial attacks on traditional only machine learning methods are studied.

In the context of classifiers for liveness tests, beyond the issue of understanding and improving the classifier, another question we want to address is whether the proposed adversarial attack can be converted into a direct presentation attack. That is, we want to verify that the same performance degradation will be observed if instead of just manipulating the imposter images of the database, we create new imposter images by increasing the saturation of client images, printing them or displaying them on an electronic device and capture an image of them. In other words, we would like to verify that the all-digital adversarial attack on the test database of the classifier can be converted into a physical attack on a real liveness test system.

As the execution of that physical attack is a labour intensive process, we run a very limited experiment, which however gives a clear indication that the corresponding presentation attack is enhanced by the manipulation of the client images before presenting them to the system's camera. This was not an unexpected result, since we had already established that saturation increases lead to classifier performance degradation, and we naturally expect that by presenting to the camera a higher saturation image it will also result into a higher saturation image as the camera's output.

Contributions: We propose a colour manipulation adversarial attack to a face liveness system based on a deep neural network. To the best of our knowledge, it is the first study of adversarial attacks on deep neural networks in the context of face liveness detection.

In a second contribution, we conducted an experiment the result of which indicate that the proposed adversarial attack can be converted into a direct presentation attack.

Limitations: Due to the laboriousness of the task of creating a database with enhanced presentation attacks, the experiment corresponding to the second contribution was limited in scope and the creation of a sizeable database with enhanced presentation attacks is left as future work.

2. Related work

2.1. Liveness detection

Varghese and Matthew [VM15] classified liveness detection methods into intrusive and non-intrusive types, depending on their interference with the biometric data acquisition process. Alternatively, depending on the way the classification algorithm handles image features, liveness detection methods can be categorised into: traditional face anti-spoofing methods using hand-designed features and employing shallow machine learning, and deep learning methods.

Regarding traditional methods, Boulkenafet, at el. [BKH15] extracted local binary patterns (LBP) features from individual image channels in various colour spaces (RGB, HSV, YCbCr) and test on the CASIA database and Replay-Attack databases. LBPs are the most commonly used image feature for liveness detection, e.g. [CAM12a] used LBPs and shallow learning on three types of attacks: printed photographs, digital photos and videos. In addition, using several colour spaces such as RGB, HSV and YCbCr leads to more effective liveness detection algorithms [JXMA19].

Recent research has shown that in liveness detection tasks, deep learning methods could be more effective than those based on handcrafted features. The limitations of the latter become more apparent in a diversity of sensing environments; while they perform reasonably well within intra-dataset protocols, they are less suitable within multi-domain datasets as they cannot be easily adapted to new circumstances [BCV13]. Instead, the extraction of high-level (deep) features from a dataset, especially in systems that work on complex tasks, need multi-layered methods [WHJ15]. Convolutional Neural Networks (CNNs) in particular can achieve impressive results on image and video classification tasks.

One of the earliest attempts on liveness detection with CNNs is Yang et al. [YLL14], the results of which were improved by Atoum et al. [ALJL17] using a two-stream CNN-based face anti-spoofing method; the first stream extracts local and holistic features and the second is used to estimate a dense depth map. Their model achieved good performance in the intra-testing stage. Nagpal and Dubey [ND18] evaluated liveness detection algorithms on the Inception and ResNet architectures over the MSU database [PHJ16], covering several aspects of the architectures such as depth of the model, learning rate and random weight initialisation.

2.2. Adversarial Attacks

Adversarial attacks can be very simple in nature. In [NK16], the authors generated attacks by adding a small perturbation to a single pixel, or a small set of pixels. We also note that there are several open-source software tools for generating adversarial images, e.g. DeepFool [MDFF16]. However, in our context, the most relevant adversarial attacks are the black-box ones, where the attacker does not have access to the hidden layers of the network or more generally any information about the type and the parameters of the classification algorithm.

3. Experimental setup

The liveness detection classifier we use is based on ResNet, the winner of the classification task in the 2015 ImageNet Large-Scale Visual Recognition Challenge [RDS*15], reaching a 3.57 % error. Specifically, we used the ResNet50 variant, consisting of 50 layers. For training and testing, we used the Replay-Attack database [CAM12b]. While there are several other databases that are routinely used in liveness detection research, such as CASIA [ZYL*12] and NUAA [TLLJ10], we chose Replay-Attack which supports more types of presentation attacks: printed photos; video and photo playback on an iPhone; photos and videos displayed on an iPad screen.

3.1. Implementation and training

All code was written in Python, on the Pytorch deep learning platform, and the experiments ran on an Intel Core i7 CPU 64GB RAM PC. We used the pre-trained convolutional part of ResNet50 and trained with our images for 24 epochs, using the Adam optimizer with learning rate 0.0001, while the batch size was set to 20. The custom classifier contains a fully connected layer with ReLU activation and followed by a Dropout with 20% chance of dropping and a fully connected layer with log softmax output.

3.2. Validation

Using a within-subject validation protocol, we trained ResNet50 with 1279 images from 14 subjects. The test set consisted of 290 client and 310 imposter images, from all 14 subjects. On the clients, we obtained a True Negative Rate (TNR) of 99%, while on imposters a True Positive Rate (TPR) of 98% for a total accuracy rate of 98%.

The impressive performance of ResNet50 under a within-subject protocol masks the difficulty of the liveness classification task on images of previously unseen faces. In a next step, we validated the network under a cross-subject protocol, training ResNet50 on 1082 images from 12 subjects and testing it on 240 images (120 clients and 120 imposters) from 2 different subjects. This time we obtained a TPR of 88% and a TNR of 53% for a total accuracy rate of 70%. Indicative of the nature of the challenges in cross-subject liveness detection, we note that the very low TNR was almost exclusively due to very poor performance on one of the two subjects for which almost all client images were misclassified as imposters.

4. Results

4.1. Adversarial Attack

The adversarial attack was validated with the cross-subject validation protocol described in Section 3. After the images were converted to the HSV colour space, in a first experiment the saturation

α	0	0.25	0.5	0.75	1	1.25	1.5	1.75
T_p	95	89	99	89	88	75	75	77
L	.21	.33	.06	.27	.38	.57	.56	.53

Table 1: *TPR T_p and average loss L for various values of α. The grey shaded column corresponds to the original images.*

s	31	63	95	127	159	191	223	255
L	.20	.17	.15	.48	.77	.82	.98	.68
T_p	92	94	91	80	70	58	55	73

Table 2: *TPR T_p and loss L for various fixed values of s.*

value s was multiplied by a constant α and capped to 255. That is:

$$s \rightarrow \min\{\alpha \cdot s\ ,\ 255\}.$$

Figure 1 shows one imposter image from each subject undergoing that type of colour manipulation. Table 1 shows the obtained TPRs and the average loss. The main observation is that when $\alpha > 1$ the TPR is lower than that on the original images, which correspond to $\alpha = 1$, indicating a successfully adversarial attack. In contrast, when $\alpha < 1$, the TPR is higher, providing further evidence for the effectiveness of the attack. As expected, the average loss values exhibit the opposite pattern.So, we note the significant decrease in the TPR, which, for example, translates into 12% more imposter attacks being successful when the saturation value is multiplied by 1.25.Note that as we do not manipulate the client images, the TNR is the same as in Section 3.

In a second experiment, we put the saturation of all pixels of all images to a fixed value s. The corresponding example images are shown in Figure 2 and the corresponding TPR and average loss values are shown in Table 4.1. We notice that in the range 63-233 the TPR drops monotonically as the saturation increases, starting from a value as high as 94% for $s = 63$ and dropping to a low of 55% for $s = 223$. The extreme values of $s = 31$ at the left end of the table and $s = 255$ at the right end of the table exhibit different behaviour, in an interesting phenomenon that in the future we would like to study further.

4.2. Presentation Attack

The presentation attack was validated on imposter images created from the client images of the Replay-Attack. Three client images from each of the two subjects were:

i. printed on A4 paper
ii. displayed on a commodity laptop LCD screen
iii.displayed on an iPhone screen

and then captured with an iPhone camera. The acquired images were manually cropped and resized to 60×60. The whole process was repeated with the saturation of the client images put at a fixed value of 180. Figures 3-5 show one face image for each subject and each condition.

The corresponding TPR and loss are reported in Table 3. We note

	Original	Processed
paper	100 (0.05)	83 (0.46)
LCD	100 (0.03)	100 (0.07)
mobile	100 (0.18)	66 (0.42)

Table 3: *TPR and (loss) for all 6 types of presentation attacks.*

that, as expected, in all three forms of physical attacks, a high saturation value of $s = 180$ leads to a lower or equal TPR when compared to the corresponding imposter images that were produced by the same physical method from unprocessed client images. We note that while the test set is very small for the reported TPR to have significance, the reported loss values provide further evidence for the validity of the conclusions.

5. Conclusions

Our experiments demonstrated that the simple and intuitive adversarial attack of increasing the value of the saturation component of an image can be effective against neural network based face liveness systems. Initial results indicate that this adversarial attack can become the basis for an effective presentation attack, in which the imposter increases the saturation of a face image before printing it on paper or displaying on the screen of an electronic device.

In the future, we plan to create a database with imposter images corresponding to this type of presentation attack.

References

[ALJL17] ATOUM Y., LIU Y., JOURABLOO A., LIU X.: Face anti-spoofing using patch and depth-based cnns. *IJCB* (2017), 319–328. 2

[BCV13] BENGIO Y., COURVILLE A., VINCENT P.: Representation learning: A review and new perspectives. *IEEE PAMI 35*, 8 (2013), 1798–1828. 2

[BKH15] BOULKENAFET Z., KOMULAINEN J., HADID A.: face anti-spoofing based on color texture analysis. *CoRR abs/1511.06316* (2015). 2

[CAM12a] CHINGOVSKA I., ANJOS A., MARCEL S.: On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG* (2012), IEEE, pp. 1–7. 2

[CAM12b] CHINGOVSKA I., ANJOS A., MARCEL S.: On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG)* (2012), IEEE, pp. 1–7. 2

[GMF14] GALBALLY J., MARCEL S., FIERREZ J.: Biometric antispoofing methods: A survey in face recognition. *IEEE Access 2* (2014), 1530–1552. 1

[JXMA19] JAISWAL A., XIA S., MASI I., ABDALMAGEED W.: Ropad: Robust presentation attack detection through unsupervised adversarial invariance. *CoRR abs/1903.03691* (2019). URL: http://arxiv.org/abs/1903.03691. 2

[MDFF16] MOOSAVI-DEZFOOLI S.-M., FAWZI A., FROSSARD P.: Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR* (2016), pp. 2574–2582. 2

[ND18] NAGPAL C., DUBEY S. R.: A performance evaluation of convolutional neural networks for face anti spoofing. *arXiv preprint arXiv:1805.04176* (2018). 2

[NK16] NARODYTSKA N., KASIVISWANATHAN S. P.: Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299* (2016). 2

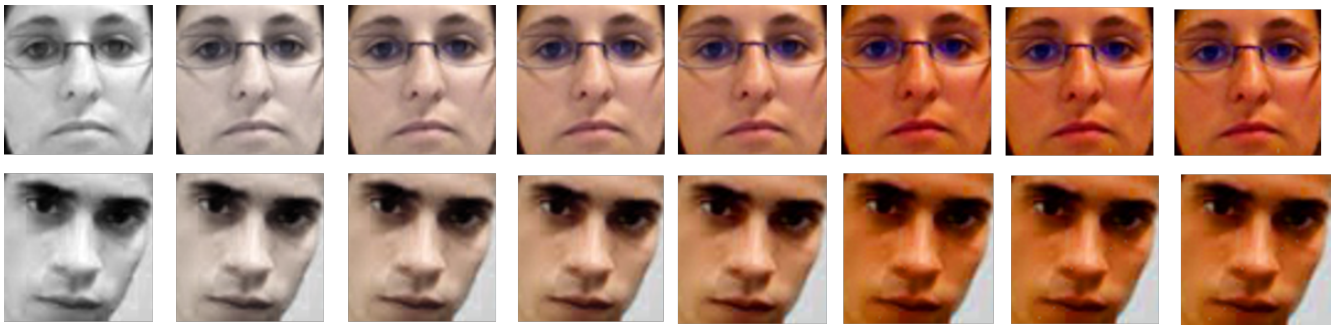

Figure 1: *Saturation linearly scaled by a constant* α *and capped to 255. From left to right:* $\alpha = 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75$.

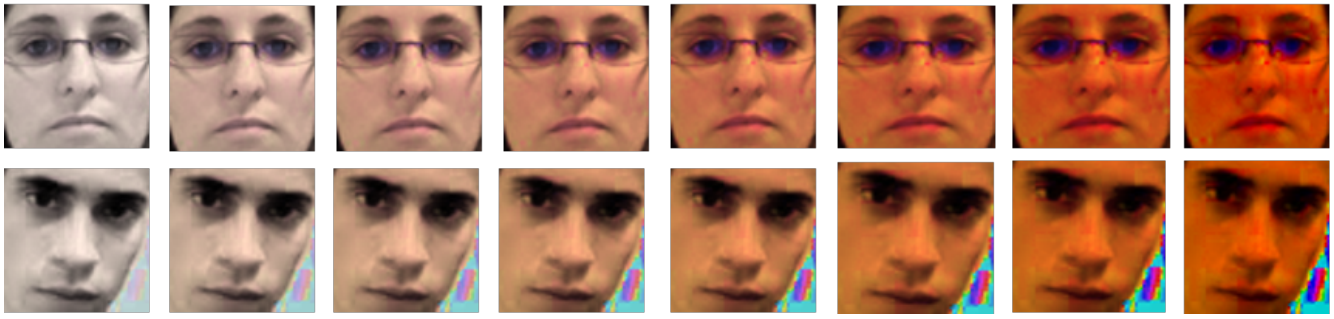

Figure 2: *Fixed saturation values s. From left to right:* $s = 31, 63, 95, 127, 159, 191, 223, 255$.

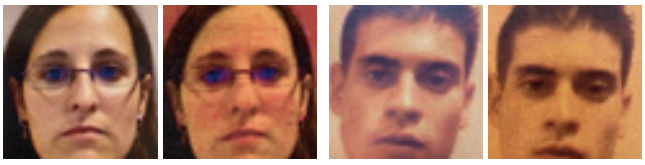

Figure 3: *Paper print: for each pair of images, the left is a photo of the original client and the right a photo of the processed client.*

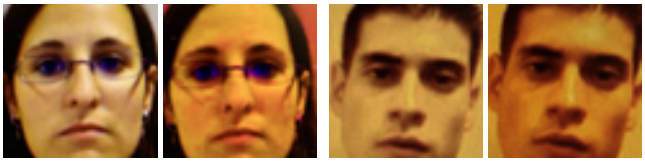

Figure 4: *LCD: for each pair of images, the left is a photo of the original client and the right a photo of the processed client.*

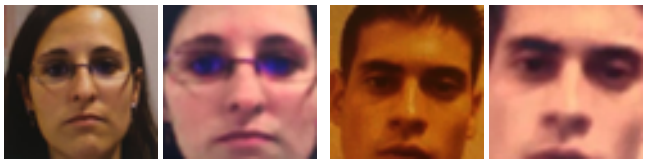

Figure 5: *iPhone: for each pair of images, the left is a photo of the original client and the right a photo of the processed client.*

[OI16a] OMAR L., IVRISSIMTZIS I.: Designing a facial spoofing database for processed image attacks. In *7th ICDP* (2016), p. 5 (6 .). 1

[OI16b] OMAR L., IVRISSIMTZIS I.: Resilience of luminance based liveness tests under attacks with processed imposter images. In *24th WSCG* (2016), pp. 79–82. 1

[PHJ16] PATEL K., HAN H., JAIN A.: Secure Face Unlock: Spoof Detection on Smartphones. *IEEE IFS 11*, 10 (2016), 2268–2283. 2

[RDS*15] RUSSAKOVSKY O., DENG J., SU H., KRAUSE J., SATHEESH S., MA S., HUANG Z., KARPATHY A., KHOSLA A., BERNSTEIN M., ET AL.: Imagenet large scale visual recognition challenge. *IJCV 115*, 3 (2015), 211–252. 2

[TLLJ10] TAN X., LI Y., LIU J., JIANG L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *ECCV* (2010), pp. 504–517. 2

[VM15] VARGHESE R. A., MATHEW J. S.: Face anti-spoofing methods. *IJSTE 2*, 4 (2015), 318–320. 2

[WHJ15] WEN D., HAN H., JAIN A. K.: Face spoof detection with image distortion analysis. *IEEE 10*, 4 (2015), 746–761. 2

[YLL14] YANG J., LEI Z., LI S. Z.: Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601* (2014). 2

[ZCPR03] ZHAO W., CHELLAPPA R., PHILLIPS P. J., ROSENFELD A.: Face recognition: A literature survey. *ACM computing surveys 35*, 4 (2003), 1–61. 1

[ZYL*12] ZHANG Z., YAN J., LIU S., LEI Z., YI D., LI S. Z.: A face antispoofing database with diverse attacks. *5th IAPR ICB* (2012), 26–31. 2