Detection of Attacks to Face Recognition Systems

Daniel Bento de Sousa
danielbsousa@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2018

Abstract

Face recognition systems are increasingly important in today's society, being mainly employed as a security measure. Everyday items, such as mobile phones and laptops, or more crucial security systems, such as airport access control, are good examples of its usage. Due to its popularity, these biometric systems are vulnerable to a wide range of attacks, which are becoming more and more complex. Therefore the development of effective counter-measures is necessary. The objective of this thesis is to develop a tool which detects intrusions at the sensor level, known as Presentation Attacks (PA). For this, state of the art contributions are reviewed in order to understand their main disadvantages and possibilities of improvement. An approach based on transfer learning using a pre-trained Convolutional Neural Network (CNN) model is presented. This network is then adapted to the problem and several steps are taken to optimise it accordingly. A novel approach is proposed by implementing a layer that performs video analysis for action classification, known as Long Short Term Memory (LSTM). The proposed solution achieves a half total error rate (HTER) of 1.09% in the Replay-Attack database. Finally, a conclusion is made about the detection of attacks to facial recognition systems and why is it still an open problem, even though state of the art methods show a high performance in such demanding databases.

Keywords: Face Recognition, Biometrics, Presentation Attacks, Transfer Learning, CNN, LSTM

1. Introduction

Face recognition together with fingerprint scanning are among the most used technologies for implementing security measures based on biometrics. However, face recognition systems are the most frequent targets of attacks [14], which proves that there exists a need to build a fail safe system that allows to securely "keep the prize" away from intruders.

This security measure is used worldwide not only because of the biometric passport (e-passport), which allows people to enter countries by simply comparing their face with the passport's picture, but also because of the increasingly presence of biometric applications for personal computers and mobile phones to an extent of opening/accessing your personal bank account using face identification [12]. Furthermore, the India Unique Identification Authority is implementing an ID system for every Indian resident based on facial recognition and fingerprint scanning.

Unfortunately, these biometric systems are all vulnerable to different kinds of attacks, being the focus of this work the Presentation Attack (PA) [1] at the sensor, where a person tries to masquerade as another by creating a fake biometric trait of the user and presenting it to the sensor. This work particularly on trying to determine if the presented trait originated from a real legitimate client or not.

Overall, these attacks can be divided in three types [14]: photo, video and mask attacks. **Photo attacks** are carried out by presenting to the sensor a photograph of the original user. They can be printed on paper (print attacks), where the eyes and mouth region can be cut-out so the imposter can wear it as a 2D mask, mimicking natural face behaviour such as eye-blinking. Or it can be displayed on a screen (digital-photo attacks); **Video attacks**, or replay attacks, consist in the attacker presenting to the sensor a video of the genuine user, where the dynamics and the movement of a specific user are presented; Finally, in **mask attacks**, a 3D mask of the genuine user is created and used to spoof the system. There are several types of 3D masks but they all aim to capture the 3D model of the original person in detail, with rich textures, so that the use of any depth, thermal or some feature analysis techniques may become ineffective. As expected, all previous discussed attacks depend on the resolution of the device, or on the type of support used to present the fake copy (handheld or fixed support). External variability, such as illumination or background conditions, can also heavily influence the outcome. The objective of this work is to propose a method robust enough to detect these attacks, fast and non-intrusive, being able to be deployed without needing additional hardware.

2. Related Work

Techniques that detect the previous mentioned attacks are known as Presentation Attack Detection (PAD) techniques [1]. The key goal of a PAD system is to automatically distinguish between a normal access, referred to as bona-fide, and any type of attack presented to the sensor from the imposter. As it can be seen in Figure 1, PAD methods can be separated in four dimensions: User Interaction Support, Imaging Sensor, Contextual Information and Feature Extraction [35]. The combination of such dimensions is often used in order to achieve a more complete PAD algorithm.

- **User Interaction Support** - This dimension can be applied when the user is willing to undergo a more

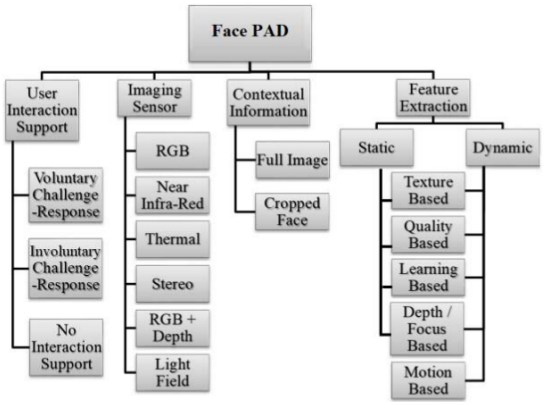

Figure 1: Classification of face presentation attack detection (PAD) algorithms. [35]

thorough identity check. In such scenarios, PAD systems usually employ the so called *challenge response* methods. Voluntary challenges would be portrayed as a pre-determined set of movements that the user has to fulfill. On the other hand, involuntary challenges consist in detecting natural responses of the human body, such as natural pupillary response to external stimuli using illumination [38]. Or blink detection, eye or mouth movement [27]. Unfortunately, these systems often fail against replay or mask attacks, need dedicated hardware, which increases the computational cost, and are often labelled too intrusive.

- **Imaging Sensor** - This type of face analysis is related to the selected sensor. Typically, 2D RGB cameras are the most commonly used sensor nowadays. Nevertheless, the recent availability of richer imaging sensors is opening new possibilities for PAD solutions. The Kinect [13], Lightfield Cameras [35], thermal imaging and near infra-red cameras [46] are recent sensors that have been employed to PAD. However, these are not so frequently used since they can be expensive and require additional hardware.

- **Contextual Information** - This dimension is related to the possibility of using background or scenic cues or motion to detect a presentation attack [43].

- **Feature Extraction** - This dimension does not require user cooperation, has relatively low cost and can be subdivided into two classes: dynamic, which takes advantage of the temporal information or motion, and static, which only takes into consideration the analysis of a single image. Textural based methods exploit the textural patterns presented by the used artefact, some examples are LBP, BSIF and LPQ [26, 17, 29]. Quality based methods explore the image quality characteristics in order to detect the bona-fide sample or an artefact used for the attack. Learning based methods derive features by modelling and learning relationships between images and/or sequences. Depth/focus based techniques explore different depth information between images captured from different sensors [18, 31, 19], as well as frequency analysis by using the Fourier Transform, Euler motion magnification [5] or Difference of Gaussian (DoG) [31]. Finally, motion based methods explore the movements of the user, such as eyes or mouth, as well as calculating the optical flow between consecutive frames [37], which provides all the required motion information in order to detect a PA.

When reviewing PAD methods, feature extraction techniques are the most used since they are practical to implement, requiring only the camera system to detect any attacks, as well as not requiring any user interaction. Nevertheless, as each method focuses on solving a specific problem, the overall system may lack the robustness to detect all types of attacks. Due to this, fusion between several techniques is possible, being referred to as multi-biometrics. However, blindly fusing methods without first determining if they are compatible just to increase the system's complexity may lead to worse results [7].

Furthermore, several PAD techniques depend on a hand-crafted feature extraction, usually simple features are extracted and analysed. However, some are processed and never connected. Nowadays, this task is eased by the uncovering potential of Deep Learning (DL). Such algorithms are able to extract several distinct features that are deemed important, and successfully find correlations between these in order to produce a successful outcome. Learnable feature based PAD techniques often use Convolutional Neural Networks (CNN)[22] that do not require fixed, hand crafted features. Instead, these CNN based algorithms learn what features are important by using input data as training.

The remarkable success of these CNNs in the ImageNet competition [33] has attracted a multitude of researchers in the computer vision community to investigate the full potential of these networks. Usually, early CNN based algorithms consist in using a CNN network as a simple feature extractor, using afterwards a Support Vector Machine (SVM) or other conventional classifier [44], or a simple CNN with low depth. Other works such as [25, 24, 32], simply propose a simple adapted neural network architecture to detect presentation attacks. However, more recent works tend to complement a CNN with other techniques. In [2] a nonlinear diffusion filter is first applied to the input data and in [44] background information is also provided, in order to obtain depth cues. [42] proposes a very basic implementation, without any optimization, of a LSTM layer to perform a temporal analysis from a sequence of frames and a fusion approach is proposed by [4], where a deep neural network learns rich patch-colour features while other outputs a depth map of an image, then the scores of these are fused together. In this work, a deep learning time analysis CNN network is proposed.

3. Proposed Architecture

The overall architecture of the proposed deep learning solution can be seen in Figure 2. Given an example video containing a face image access, the objective is to detect the facial region, pre-process and compile it into a sequence of frames, giving it as input to the CNN and, afterwards, introducing the video analysis layer, producing a decision. For this to be possible, the following processes have to be taken into account.

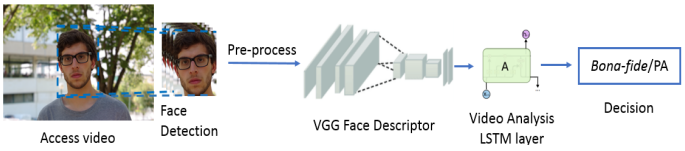

Figure 2: Architecture of the proposed solution.

3.1. Face Detection

The used architecture receives as input a facial image with a fixed size of 224x224 pixels. Therefore, facial regions in an image need to be first detected. When regarding the PAD problem, the user needs to be frontally facing the sensor, without any type of occlusion. Accordingly, a simple classifier, as the Viola-Jones [40], can be used.

The Viola-Jones [40] algorithm, introduced in the early 2000's, was a ground breaking technique since it was fast and computationally cheap, being the first face detection algorithm to run on simple cameras and mobile phones. In order to do so the algorithm is trained with positive images (images of faces) and negative images (images without faces). Haar-features [39] are extracted from each image by several haar-blocks. Each block is divided in a black and white segment that covers part of the image. Then, a single value per block is obtained by subtracting the sum of pixels that lay under the white from the sum of pixels that lay under the black, an example can be seen in Figure 3. During training, the best threshold which classifies the faces to positive or negative is found.

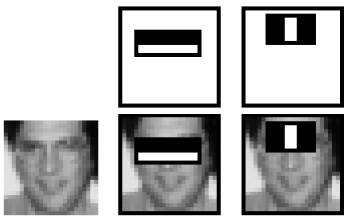

Figure 3: The top row are the haar-features which are then overlayed on a training face. [39]

3.2. VGG-Face Descriptor

The used CNN is the VGG-Face Descriptor [30] introduced by the Visual Geometry Group, which is a network created for identity recognition, being nominated as "very deep" as it comprises a 16 layer network, 13 convolutional blocks plus 3 fully connected layers, with over 140M of weights. However, this model can be adapted to the PAD problem by using the pre-trained model approach, a network that was previously trained for a similar problem, and can now be retrained, taking advantage of its original weights. This proves advantageous when both problems are similar and there is not enough robust training data available for the new problem, which is the case with most PAD methods. This model may be seen as a starting point for the transfer learning process, with the original model suffering several alterations in order to adapt the network to PAD.

In order to adapt the VGG network, illustrated in Figure 4, to the PAD problem several factors had to be taken into consideration. The result of the last layer

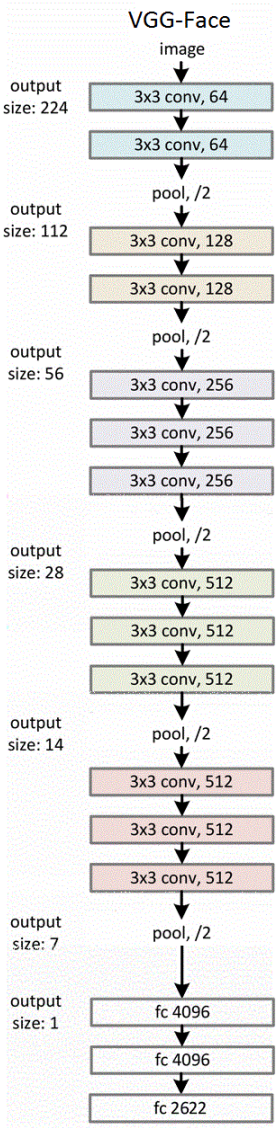

Figure 4: Architecture of the original VGG Face descriptor.

was changed to a binary output, either bona-fide or a presentation attack. The chosen loss function was the cross-entropy, or logarithmic loss, which increases as the predicted probability diverges from the actual label. The used optimizer was Adam, or the Adaptive Moment Estimation [20], which is based on a gradient descent having the particularity of computing individual adaptive rates for different weights of the network, while other optimizers usually have a unique adaptive rate. This adaptive rate, or learning rate, has the value of 10^{-5}. This low value is due to the application of transfer learning. As the original network also dealt with a similar problem, face analysis, weights need to be fine-tuned and not drastically changed, therefore weight updating needs to be done smoothly. With the idea of optimizing the converging efficiency during training, early stopping was applied. This concludes training earlier if the performance of the algorithm does not increase within a range of epochs, usually between 10 to 20, concluding that the algorithm already converged to the optimal solution.

3.3. Video Analysis

The analysis of a single image might not be enough to fully detect any type of fraud. Having this in consideration, a method is proposed where instead a small sequence of frames is analysed with the objective of action classification, bona-fide or presentation attack. For this to be possible, a type of Recurrent Neural Networks was utilized, the Long Term Short Memory (LSTM) layer [16]. LSTM networks are capable of learning long-term dependencies and were the chosen video analysis layer since they are able to remember information for long periods of iterations.

This layer was implemented in the proposed architecture by removing the last VGG layer, feeding its features directly to the LSTM which then will analyse their evolution over a period of time. This period, known as timestep, is a hyper parameter of the proposed architecture and it is a measure of how many steps does the LSTM need to keep in memory in order to make a decision. The ideal value would be the size of the video, however, a value too large introduces noise, as well as memory problems, deeming the LSTM network useless. There are also several possible architectures when using the LSTM layer, but only two were deemed important to experiment. A many-to-one architecture is when, during a timestep, each LSTM cell gives the output vector to the following layer and only the last layer produces the final output, given then to the softmax layer for classification. This architecture is often used when dealing with an action classification problem, which is the case. On the other hand, a many-to-many architecture is also possible, where, during a timestep, every LSTM cell produces an output which is combined at the end and given to the softmax layer, which is often used for video classification, where every frame produces a label. The algorithm's architecture is displayed in Figure 5.

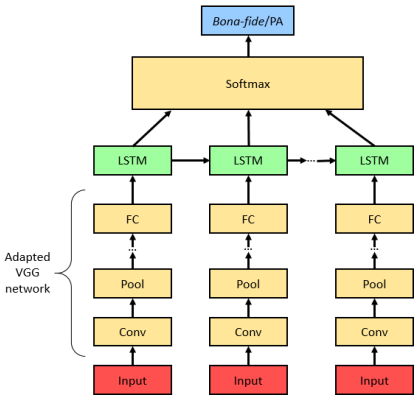

Figure 5: Architecture of the neural network used for Presentation Attack Detection. In red are displayed the inputs, blue the output, green the RNN layer while the rest is the adapted VGG network.

4. Experimental Setup

When assessing the effectiveness of PAD techniques, publicly available face spoofing databases are often used for benchmarking. Several of these have been collected and made available, using different capturing devices, conditions and operating scenarios. The used databases were the Replay-Attack [9] and the CASIA-FASD [45] which are composed by video recordings of genuine accesses and presentation attacks. Even though the used databases are not the most recently available, they were selected due to their difficulty and popularity amongst other state of the art algorithms, so a wide comparison can be made. The contents of each database consists in recordings of genuine user accesses and several types of presentation attack attempts.

4.1. Replay-Attack

The Replay-Attack database [9] consists in video accesses recorded by a built-in camera of a MacBook Air 13-inch laptop, with a resolution of 320x240 pixels. 50 users participated in this database acquisition, creating three disjoint datasets: train, validation and test, leading to 200 genuine accesses and 1000 attack samples. These were captured under two lighting conditions: controlled, with an illumination support in an homogeneous background; and adverse, with natural lighting and a more complex background, as seen in Figure 6. The presentation attacks are: high-resolution print attacks, where photos and videos were captured using a Canon PowerShot SX150 IS with 720p; mobile display attacks, using an iPhone 3GS with a resolution of 480x320; and high quality photos and videos displayed on an iPad screen, with 1024x768 resolution. These artefacts were held by the imposter or by a fixed stand.

Figure 6: Examples of accesses from the Replay-Attack database. From left to right: real accesses, printed photograph, display attack and tablet attack. The top row is from a controlled scenario whereas the bottom is from an adverse scenario. [9]

It is also important to note that this was the primarily used database to fine tune the neural network and test its alterations, since it is the most used database by state of the art techniques, as well as the most robust since it presents diverse types of attacks with different conditions among different people.

4.2. CASIA-FASD

The CASIA-FASD [45] database contains captured data from 50 subjects where the PAs were constructed from the high quality recordings of the genuine users. Several attacks were considered: warped photo attack, where the facial motion is simulated by bending a printed photograph; photographic masks; and replay attacks. Video accesses were acquired with three imaging qualities: low, with a long-time used USB camera that degrades image quality; normal, using a newly bought USB camera which keeps the original quality; and high, using a Sony NEX-5 camera for recording. These have resolutions of 640x480, 480x640 and 1280x720, respectively. Presentation attacks were shown using an iPad with a resolution

of 1280x720 and in prints the same Sony camera was used. The complete image set is illustrated in Figure 7. There are 12 videos per subject, 3 genuine and 9 fake, resulting in 600 clips.

Figure 7: Complete image set for a subject. The top left images represent the low quality videos, the bottom left normal quality videos and the right set are the high quality videos. For each set of photos from left to right: genuine, warped photo attack, cut photo attack and replay attack. [45]

5. Results

The experiments follow the protocols associated with each used database. For each, the training set is used to learn the CNN model, the validation set used to check the overall performance of the algorithm and convergence, and the testing set is used for evaluation purposes using the HTER [34]. For the image processing OpenCV [8] is used and for the CNN+LSTM optimization, training and testing Keras [3] is used. All of the used databases present a frame rate of 25 images/s, this leads to thousands of images with very low variability between consecutive frames. To decrease the amount of redundant data as well as to respect the memory usage of the system, the algorithm captures only a few frames per second, depending on the database's size usually being around 3 to 5 frames, as it is enough to perform a complete analysis of the video. Afterwards these images are resized according to the input of the neural network, with a fix size of 224x224 pixels and, according to the type of analysis, the colour space might be converted and frames are grouped into a timestep series. All images are also subtracted by their average corresponding facial image.

5.1. Neural Network Fine Tuning

Before applying the LSTM layer, experiments were done in order to optimize the neural network to the PAD problem as well as to check if either the novel video analysis layer has any impact as a presentation attack detection method. First off, when applying Transfer Learning, there exists a similarity between the original network's objective, face identification, and PAD, as in both cases the CNN still needs to identify discriminative features of the face. So, when adapting the VGG CNN to PAD, we know that initial convolutional layers capture low-level image features, edges for example, while deeper convolutional layers capture increasingly more complex details [22]. This way, an hypothesis can be made in which the proposed neural network adaptation should achieve better performance by only fine-tuning the last parts of the neural network instead of the whole model as the first layers do not need any weight tuning. In order to confirm this, a test was made in which the entirety of the network was frozen, meaning that the weights cannot be altered during training, keeping its original values and, subsequently, deeper layers were unfrozen and trained, testing the performance of the network.

Furthermore, performing an analysis in the RBG colour space can be quite limiting since there is a high correlation between the three channels, leading to a low quality feature extraction. Having this in consideration, other colour spaces that allow a better separation between luminance and chrominance were considered. By fully separating these channels, the images are richer in features, presenting an increase in textural information. This way, a HSV and YC_bC_r analysis was performed. These results can be seen in Table 1.

As it is demonstrated by the table, a higher performance can be achieved when using other colour spaces rather than RGB. Also, but only when working in the RGB colour space, performance increases by skipping training for the first two convolutional blocks, as they present introductory operations and don't need further tuning, confirming the proposed hypothesis. However, this does not happen in HSV or YC_bC_r. This is because the original network was trained using only RGB images therefore the weights are not adapted/trained for the colour space changes, thus needing tuning as well. So, in this latter case, best performance is achieved when all of the network is trained.

5.2. Video Analysis

When applying this layer only the model that achieved the best performance, presented in Table 1, is used, therefore meaning that the chosen model performs a HSV colour analysis. As it was previously discussed, when applying the LSTM layer, a tuning of the newly introduced hyper parameter, timestep, requires to be done. This parameter allows to determine for how many frames, or for how long, the algorithm keeps analysing the video images before making a decision, forgetting previous iterations. If the timestep value is too low then it would result in a premature decision, which may lead to errors, and if it is too big it introduces noise, underperforming as well. Secondly, it is also important to experiment with which architecture the best performance is achieved, many-to-one or many-to-may as previously explained. Results to all these experiments are displayed in Table 2 and the graphs extracted during the training of the neural network can be seen in Figure 8.

According to Table 2, it is possible to conclude that a many-to-one architecture is the most adequate architecture for the problem. This is mainly because this approach is better for action classification, which is the case. Better performance is achieved by outputting an answer after analysing the whole group of frames rather than deciding in each singular frame what action is taking place, since we do have more information after the process. Regarding the timestep analysis, the best value found is seven, this shows that by analysing groups of seven frames, which corresponds between a three to five seconds window in the presented videos, the best possible outcome is achieved.

5.3. Comparison with State of the Art

After fully developing and tuning the proposed system, a performance assessment and comparison against state of the art methods is presented. The algorithms considered for comparison are the ones reporting the best performance values until this date, which propose mostly

Table 1: Performance table of different colour spaces. Best values of each space highlighted in bold.

	RGB		HSV		YC_bC_r	
Trained	Accuracy (%)	HTER (%)	Accuracy (%)	HTER (%)	Accuracy (%)	HTER (%)
Fully Connected (FC)	76.9	63.5	93.2	14.6	87.4	27.7
FC + 1 conv. block*	81.7	44.0	96.2	12.4	89.1	22.8
FC + 2 conv. block*	83.0	39.5	95.0	10.5	93.6	13.3
FC + 3 conv. block*	**89.4**	**23.2**	96.7	7.0	94.4	11.7
Full Network	84.2	37.0	**96.9**	**6.5**	**95.7**	**8.5**

*Conv. block represents the group of 2/3 convolutional blocks before the max pooling operation as it was explained in the VGG architecture.

Table 2: Network's performance with various timesteps and different architecture. All values are in % and best highlighted in bold.

	Many-to-one				Many-to-many			
Timestep	Accuracy	FAR	FRR	HTER	Accuracy	FAR	FRR	HTER
5	98.5	1.6	7.8	4.7	97.8	4.5	6.1	5.3
6	98.8	2.3	4.1	3.2	98.1	4.3	3.9	4.1
7	**99.2**	0.6	1.6	**1.1**	98.4	4.0	2.8	3.4
8	97.9	4.3	2.4	3.4	97.7	6.5	3.1	4.8
9	97.1	9.0	3.0	6.0	96.0	11.0	4.4	7.7

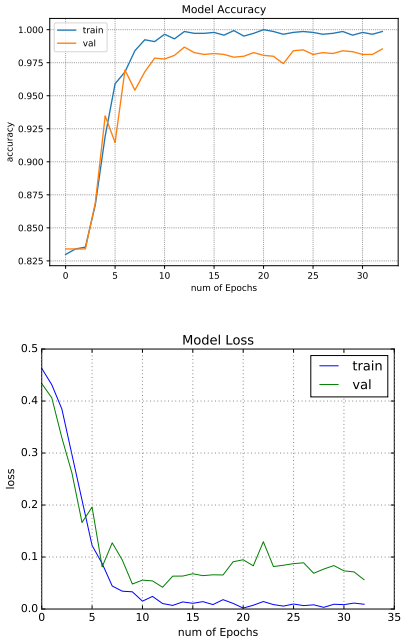

Figure 8: Network's performance when adding a LSTM layer with 7 as the chosen timestep with a many-to-one architecture.

textural and/or motion analysis solutions, or other deep learning algorithms.

Two assessment evaluations were made, the first one where each database was used for training and testing independent from one another, being named as intra-database test. On the other hand, to test the generalization ability of the proposed solution and others, in terms of testing with completely new data, an inter-database test was performed. In this latter assessment, one database is used to train the algorithm, but testing is performed with sets from the other database. This type of test allows assessing the full robustness of the systems in a more demanding cross-database scenario, since by changing the databases every aspect of acquisition is completely altered: conditions, image quality, people, capturing sensor and so on.

The results of the first test are summarized in Table 3. It shows that amongst all of the tested databases, the algorithm performs well with quite good results when compared to other techniques. By analysing with more detail the conditions where misclassifications happen, it is possible to conclude that the majority of errors happen when high quality presentation attacks are given to the sensor, specifically in the case photographic masks are presented to the sensor, in the CASIA dataset. In a high quality image, usually with a high resolution, around 1024x720 pixels, the face is detected and then it is resized by an image with 224x224 pixels. This greatly down-samples the target image. With this, rich features that can identify the attacks are lost together with quality and, by analysing the remainder features it leads to an attack being incorrectly identified as a genuine user. An example of this is again the case of photographic masks, which are prints with a high quality, where both the eyes and mouth are cut out in order to mimic normal face movements. By down-sampling these images the algorithm cannot detect the usual print attack features, such as the paper reflection and its low image texture when compared to a genuine face, but can detect movement which influences the system into accepting the sample as genuine.

As it is displayed by Table 3, most algorithms present a good performance throughout all of the chosen databases, which may raise doubts concerning the necessity of the still ongoing investigation about face presentation attack detection. If there are algorithms that achieve such low error rates then why is PAD an open problem. In spite of these methods achieving high performance rates when following each datasets' set of rules when performing an intra-database test, the same does not apply in the inter-database test, as illustrated in Ta-

Table 3: Performance and comparison between state of the art methods.

Methods	Replay Attack	CASIA FASD
	HTER (%)	
LBP[9], (2012)	13.80	18.20
LBP - TOP [11], (2012)	7.60	10.60
LBP - GLCM [41], (2013)	7.20	-
Motion [10], (2013)	11.70	26.00
Motion + LBP [21], (2013)	5.10	-
Motion Mag [6], (2014)	0.25	14.40
Deep Learning [28], (2015)	2.10	7.34
Fine-Tuned VGG [24], (2017)	4.30	-
DPCNN [24], (2017)	6.10	-
Nonlinear Diffusion CNN [2], (2017)	10.00	-
FASNet (CNN) [25], (2017)	1.20	-
Patch + Depth CNN [4], (2017)	0.72	2.27
(1)(HSV + YCbCr) LBP [23], (2018)	2.90	6.09
Proposed Method	1.09	10.32

ble 4. This latter table shows that unfortunately all of the state of the art methods, including the proposed solution, perform poorly when displayed with a more challenging robustness test. By completely changing all of the conditions in a database, all algorithms fail, which only shows that most of them would perform poorly in a real world situation.

6. Conclusions and Future Work

6.1. Conclusions

This works presents a novel approach, utilizing deep learning, to detect attacks to facial recognition systems. In order to create an acceptable algorithm, the challenges faced by the community were evaluated, analysed in detail as to what are the most common type of attacks. Moreover, an evaluation of the current state of the art framework was made, describing the several existing algorithms and its different approaches when addressing the problem. Techniques were subdivided depending on the type of analysis and a particular focus was given to methods that use machine learning. As a result an approach was presented, which adapts to different conditions and variabilities not focusing only on specific attacks but rather in its training, trying to create a well prepared model.

Throughout the process of developing the algorithm several milestones were achieved, Transfer Learning presented as being a good starting method when the problems are similar, allowing to reuse the original weights to the new task, achieving a high performance with a fast convergence. Either way, even if the problems are not similar or the training process differs, the architecture can be used as a starting point to a new network, as it was shown. Regarding colour analysis, and as it was demonstrated, the RGB colour space has a poor performance when regarding presentation attack detection since there is a high correlation between the three channels. Therefore, other colour spaces should be taken into consideration, preferably ones that take the luminance and chrominance of a picture into consideration as they are better for image recognition, such as YC_bC_r and HSV, where the latter achieved the best performance. Instead of performing the usual single frame analysis, the architecture of the network was altered so that a full video analysis could be performed. For this, the LSTM layer was used and it proved fruitful to the task, it allowed a spatio temporal analysis of a picture, analysing its feature evolution throughout the video. This layer has a high adaptability to different problems, achieving the best performance in a many to one architecture and with a timestep of 7, in this case.

The suggested method has acceptable results, presenting a higher performance, in most cases, when compared to state of the art methods, showing that deep learning can be a good solution to the problem. During the intra-database test, misclassifications happened mostly when high quality presentation attacks were captured by the sensor, mainly due to the high loss of texture features when downsampling the target image. In order to demonstrate why PAD is still an open problem and there is not an accepted solution by the community, an inter-database test was made. This assessment demonstrated that almost all proposed state of the art methods have a high error rate when presented with completely new capture conditions and/or unexpected types of attacks. In the particular case of the proposed method, the algorithm underperforms in the inter-database test as expected since the testing set differs drastically from its training, presenting a disadvantage of deep learning. Algorithms based on deep learning work well within the training set specifications and variations however, if new data outside the training range emerges, it proves unsuccessful. Presentation attack detection presents an open problem since it has many variations in every condition possible as well as its wide range of attacks.

6.2. Future Work

When regarding inter-database analysis and since the proposed solution fails most often when the target image is downsampled, losing rich features that can identify a PA, an analysis without changing the original image size should be made. In order for this to be possible, the input layer of the network should accept any image size, instead of the expected 224x224 pixel face image. This might be possible by using the Spatial Pyramid Pooling (SPP) layer [15] instead of the usual max pooling layer. A max pooling layer resizes the number of features in order to reduce their dimensionality, however both input and output size need to be previously determined values. By using a SPP layer, the target size is outputted regardless of the input size, having no input size regulation. Although in theory this seems possible, this layer is still in development showing few practical situations, therefore its viability needs to be verified before it can be applied. This approach would ideally increase the number of richer features in the image since there is no initial resizing needed, leaving the facial region of the image untouched. Other method that might present a solution to the downsampling problem would be to use patches of the target image instead of resizing the entirety of the image. This way, training data would increase and the original quality would be kept.

With respect to the robustness of the algorithm when

Table 4: Performance and comparison between state of the art methods in an inter-database test [23].

	Replay-Attack (trained on CASIA-FASD)		CASIA-FASD (trained on Replay-Attack)	
Test on: **Methods**	**Dev**	**Test**	**Train**	**Test**
	HTER (%)			
LBP[9], (2012)	44.9	47.0	57.3	57.9
LBP - TOP[11], (2012)	48.9	50.6	60.0	61.3
Motion[10], (2013)	50.2	50.2	47.7	48.2
Correlation[10], (2013)	47.7	48.3	50.2	50.2
Motion Mag[6], (2014)	50.0	50.2	43.8	50.3
Deep Learning[28], (2015)	48.2	48.8	45.7	45.4
(1) + SVM - RBF[23], (2018)	22.5	20.6	47.5	43.9
(1) + SVM - linear[23], (2018)	17.7	16.7	38.6	37.6
Proposed Method	50.5	49.2	44.5	45.3

using a test set that greatly differs from training, the performance decreases in all shown algorithms, presenting to be a deep learning disadvantage in this particular case. One possible solution may be using generative adversarial networks or adversarial neural networks [36], which generates an increasing dataset with several variations. In this scheme, there are two neural networks, one called generator, which generates new data instances based on the training set, while the other, the discriminator, evaluates them for authenticity, trying to assign the correct label to each new instance. So the generator creates new instances trying to fool the discriminator into accepting them as authentic while they are not, training the discriminator for new, different cases. When comparing to a usual neural network, which outputs a label when analysing the features, an adversarial network generates features taking the label into consideration, this allows to create different combinations of features that can represent a genuine sample or a presentation attack, being based, however, on the training set as well. Unfortunately, this type of networks cannot cover all the possible scenario variations, it may improve the algorithm but probably would not fully fix its miscalculations. These misclassifications are not so easily corrected since the algorithm mostly depends on its training. In order to seclude this, restricting capture conditions should be the most plausible approach, if some of the conditions could be manipulated or controlled, when possible, then the number of various scenarios would drastically decrease which, by consequence, would increase the algorithm's performance to disparate conditions.

Acknowledgements

I would like to thank Instituto de Telecomunicaes (IT), from IST, for funding this research under the name Research Grant - BIL/N25 - 19/04/2018 - UID/EEA/50008/2013 - SeLF-ICN.

References

[1] INTERNATIONAL STANDARD ISO / IEC Information technology Biometric presentation attack detection . 2016.

[2] A. Alotaibi and A. Mahmood. Deep face liveness detection based on nonlinear diffusion using convolution neural network. *SIVP*, 11(4):713–720, 2017.

[3] F. and others Chollet. Keras, 2015. Last Accessed on 2018-07-27.

[4] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu. Face Anti-Spoofing Using Patch and Depth-Based CNNs. *IEEE IJCB*, pages 319–328, 2017.

[5] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh. Computationally efficient face spoofing detection with motion magnification. *IEEE CSCCVP*, pages 105–110, 2013.

[6] S. Bharadwaj, S. Member, T. I. Dhamecha, and S. Member. Face Anti-spoofing via Motion Magnification and Multifeature Videolet Aggregation. *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (ii):1–12, 2014.

[7] B. Biggio, Z. Akthar, G. Fumera, G. L. Marcialis, and F. Roli. Robustness of multi-modal biometric verification systems under realistic spoofing attacks. *IJCB 2011*, 2011.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] I. Chingovska, A. Anjos, and E. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. *International Conference of the Biometrics Special Interest Group*, pages 1–7, 2012.

[10] T. De Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? *Proceedings - 2013 International Conference on Biometrics, ICB 2013*, 2013.

[11] T. De Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. LBP-TOP based countermeasure against face spoofing attacks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7728 LNCS(PART 1):121–132, 2013.

[12] E. Dunkley. Facial Recognition in Banking (Web), 2017. Last Accessed on 2018-03-30.

[13] N. Erdogmus and S. Marcel. Spoofing face recognition with 3D masks. *IEEE TIFS*, 9(7):1084–1097, 2014.

[14] J. Galbally, S. Marcel, and J. Fierrez. Biometric Antispoofing Methods: A Survey in Face Recognition - IEEE Journals & Magazine. 2:1–23, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[16] S. Hochreiter and J. J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1–32, 1997.

[17] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. *21st ICPR*, (Icpr):1363–1366, 2012.

[18] G. Kim, S. Eum, J. K. Suhr, D. I. Kim, K. R. Park, and J. Kim. Face liveness detection based on texture and frequency analyses. *2012 5th IAPR International Conference on Biometrics*, pages 67–72, 2012.

[19] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee. Face Liveness Detection Using Variable Focusing. *ICB*, pages 1–6, 2013.

[20] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. pages 1–15, 2014.

[21] J. Komulainen, A. Hadid, M. Pietikainen, A. Anjos, and S. Marcel. Complementary countermeasures for detecting scenic face spoofing attacks. *Proceedings - 2013 International Conference on Biometrics, ICB 2013*, 2013.

[22] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[23] L. Li, P. L. Correia, and A. Hadid. Face recognition under spoofing attacks: countermeasures and research directions. *IET Biometrics*, 7(1):3–14, 2018.

[24] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. *2016 6th International Conference on IPTA 2016*, (i), 2017.

[25] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo. Transfer Learning Using Convolutional Neural Networks for Face Anti-Spoofing. 10317(July), 2017.

[26] J. Maatta, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *Biometrics, IET*, 1(1):3–10, 2012.

[27] P. Majaranta and A. Bulling. Advances in Physiological Computing. 2014.

[28] D. Menotti, G. Chiachia, and A. Pinto. Deep representations for iris, face, and fingerprint spoofing detection. *International Journal of Professional AL Engineering Students*, VIII(3):138–142, 2017.

[29] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. *Lecture Notes in Computer Science*, 5099 LNCS:236–243, 2008.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep Face Recognition. *Procedings of the British Machine Vision Conference 2015*, (Section 3):41.1–41.12, 2015.

[31] B. Peixoto, C. Michelassi, and A. Rocha. Face liveness detection under bad illumination conditions. *Proceedings - ICIP*, pages 3557–3560, 2011.

[32] Y. Rehman, L. Po, and L. Mengyang. Deep learning for face anti-spoofing: An end-to-end approach. *IEEE Signal Processing Magazine*, pages 195–200, 2017.

[33] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[34] J. M. Samy Bengio. A Statistical Significance Test For Person Authentication. *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, (2), 2004.

[35] A. Sepas-Moghaddam, F. Pereira, and P. L. Correia. Light Field-Based Face Presentation Attack Detection: Reviewing, Benchmarking and One Step Further. *IEEE TIFS*, 13(7):1696–1709, 2018.

[36] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition. 2017.

[37] W. Shuigen, C. Zhen, and D. Hua. Motion Detection Based on Temporal Difference Method and Optical Flow field. *2009 Second International Symposium on ECS*, (May):85–88, 2009.

[38] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang. Face Flashing: a Secure Liveness Detection Protocol based on Light Reflections. (January), 2018.

[39] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE CVPR 2001*, 1:I–511–I–518, 2001.

[40] P. Viola and M. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[41] M. A. Waris, H. Zhang, I. Ahmad, S. Kiranyaz, and M. Gabbouj. Analysis of textural features for face biometric antispoofing. *European Signal Processing Conference*, pages 1–5, 2013.

[42] Z. Xu, S. Li, and W. Deng. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. *ACPR 2015*, pages 141–145, 2016.

[43] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Z. Li. Face liveness detection by exploring multiple scenic clues. *2012 12th ICARCV*, pages 188–193, 2012.

[44] J. Yang, Z. Lei, and S. Z. Li. Learn Convolutional Neural Network for Face Anti-Spoofing. 2014.

[45] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A Face Antispoofing Database with Diverse Attacks. *5th IAPR International Conference on Biometrics (ICB'12)*, pages 2–7, 2012.

[46] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li. Face liveness detection by learning multispectral reflectance distributions. *2011 IEEE International Conference on Automatic Face and Gesture Recognition, FG 2011*, pages 436–441, 2011.