Multi-Spectral Convolutional Neural Networks for Biometric Presentation Attack Detection

Marta Gomez-Barrero and Christoph Busch

da/sec Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
{marta.gomez-barrero,christoph.busch}@h-da.de

Abstract

Nowadays, reliable and automatic subject authentication has become of the utmost importance in multiple application scenarios. Over the last decades, biometric recognition has shown to be a good alternative to password based systems. In spite of their numerous advantages, biometric systems are vulnerable to presentation attacks (PAs), i.e., attempts to log into the system with a fake biometric characteristic or presentation attack instrument (PAI). These attacks pose a severe threat to the security of the authentication system: any person could eventually fabricate or order a gummy finger to impersonate someone else. Therefore, the development of accurate presentation attack detection (PAD) schemes is key to the wider deployment of secure biometric systems. In this paper, we present a novel approach for fingerprint PAD based on short wave infrared (SWIR) images and multi-spectral convolutional neural networks (CNNs). In particular, four samples are acquired at different SWIR wavelengths, which are subsequently fed to five different CNN models. These networks first pre-process the multi-spectral information to obtain images with three channels, and then apply regular CNN models. The approach is evaluated on a database comprising over 4700 samples, stemming from 562 different subjects and 35 different PAI species. The results show the soundness of the proposed approach with a detection equal error rate (D-EER) as low as 0.5%, outperforming the state-of-the-art D-EER of 1.4%. In addition, fusing the SWIR information with laser speckle contrast imaging (LSCI) sequences leads to an even lower D-EER of 0.2%.

1 Introduction

During the last decades, subject authentication has become a key task in a wide variety of applications. In contrast to traditional methods based on PINs, passwords, or tokens, biometrics makes use of the individuals’ biological (e.g., iris or fingerprint) or behavioural (e.g., voice or signature) characteristics to recognise them [16]. This way, biometrics provides a stronger link between the subject and the claimed identifier, and offers clear advantages over traditional authentication methods (e.g., you cannot lose or forget your finger). As a consequence, biometrics has emerged as the preferred authentication method for applications as diverse as crossing borders, unlocking smartphones, or in national-wide identification scenarios [9].

However, biometric systems are not free of vulnerabilities. Different attack points were listed in [25], including both inner modules of the system and communication channels. In particular, the biometric capture device is probably the most exposed attack point, since, in order to attack it, no further knowledge about the inner functioning of the system is required. Such attacks directed to the capture device are known in the literature as presentation attacks and defined within the ISO/IEC 30107 standard on biometric presentation attack detection [14] as the “presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system”. In other words, an attacker can present the capture device with a presentation attack instrument (PAI), such as a face mask, a gummy finger, or
multi-spectral CNNs for biometric presentation attack detection

M. Gomez-Barrero and C. Busch

a fingerprint overlay, in order to impersonate someone else (i.e., active impostor) or to avoid being recognised due to black-listing (i.e., identity concealer).

There is thus a clear need to prevent the aforementioned attacks. Presentation attack detection (PAD) methods refer to any technique developed to automatically distinguish between bona fide (i.e., real or live) presentations and access attempts carried out by means of PAIs [21]. Given the security risks posed by these attacks, this new area of research has attracted a considerable attention within the last decade, and different methods have been proposed for several biometric characteristics, including iris [4], fingerprint [20, 29], or face [5]. Moreover, several international projects, such as the European Tabula Rasa [32], BEAT [2], and RESPECT [26], as well as the US Odin research program [22], deal with these security concerns. In addition, the LivDet competition series on iris and fingerprint [6, 23] have been running since 2009.

For the particular case of fingerprint recognition systems, most PAD techniques are based on the output of traditional optical and capacitive sensors [20, 29]. However, it has been recently shown that images acquired within the short wave infrared (SWIR) spectrum can yield very accurate PAD approaches for face and fingerprint [30, 33]. This is due to the fact that all skin types according to the Fitzpatrick scale [3] present very similar remission curves for these wavelengths, and at the same time quite different from other materials commonly utilised for the fabrication of PAIs (e.g., silicone or paper) [30]. Therefore, the task of discriminating skin (i.e., bona fide presentations) from other materials (i.e., PAs) becomes easier in this part of the spectrum, in contrast to other wavelengths for which the skin types are very different among themselves and at the same time similar to, for instance, coloured silicone to name a challenging PAI.

In contrast to the first SWIR fingerprint PAD approach proposed by Gomez-Barrero et al. in [7], based on spectral signatures, Tolosana et al. carried out in [33] a thorough study on the soundness of using deep convolutional neural networks (CNNs) in combination with SWIR images. In particular, the sensor utilised in that work captures four grayscale images of the finger at different SWIR wavelengths with a 64 × 64 px. resolution. Given that most pre-trained CNN models expect RGB images (i.e., three channels), the authors defined a manual pre-processing of the samples to convert the four grayscale images into three channels. These RGB images were used as input to three different CNN models (i.e., VGG19 [28], MobileNet [11], and ResNet [31]). On the experimental evaluation, tested on a large dataset including 35 PAI species, the accuracy achieved with handcrafted features for SWIR data [7] was widely outperformed (i.e., the error rates were improved by 90%). In addition, the performance of the SWIR based CNNs was further improved with a score level fusion with handcrafted features extracted from laser contrast speckle images (LSCI) of the finger in [8]. In a parallel work, Hussein et al. [13] developed a patch-wise based CNN inspired on AlexNet [19] for SWIR and LSCI fingerprint samples, also with a score level fusion. Over a database comprising 778 samples and 17 PAI species, all samples were correctly classified for the fused system.

Building upon the work in [33], we propose an automatic pre-processing of the four grayscale images via an additional convolutional layer, integrated with the CNN model and trained together. This way, the four grayscale images can be regarded as a single four-channel image, and the network can learn the most discriminant features for the subsequent layers to process, thereby enhancing the overall detection performance. In addition to the three networks analysed in that previous work (i.e., a ResNet trained from scratch, and the pre-trained MobileNet and VGG19 models), we have studied i) the newer MobileNetV2 model [27], which includes residual connections in the form of inverted bottlenecks, and ii) the VGGFace network [24], pre-trained on facial images for recognition purposes. Since VGGFace has been trained on more skin data, this could be beneficial for the PAD task. Then, all PAD partial scores (i.e., one per
The proposed methodology is summarised in Fig. 1. First, a dedicated capture device (Sect. 2.1) acquires images of the finger at four different wavelengths within the SWIR spectrum. Then, those images are fed to five different CNN models, including an additional pre-processing layer at the beginning (Sect. 2.2.1). The models are described in detail in Sect. 2.2 and Fig. 3. Finally, the output of different models can be fused at score level, as presented in Sect. 2.3, in order to achieve a more robust PAD module.

2.1 Multi-Spectral Short Wave Infrared (SWIR) Sensor

The finger SWIR capture device used in the present work was developed within the BATL project [1] in cooperation with our project partners. In essence, the sensor is a closed box with a slot on the top for the finger (see Fig. 1, left), with the camera and lens placed inside the box. When the finger is placed over the slot, all ambient light is blocked and therefore only the desired wavelength is considered during the acquisition. In particular, a Hamamatsu G11097-0606S InGaAs area image sensor has been used, which captures 64 × 64 px. images.
Multi-Spectral CNNs for Biometric Presentation Attack Detection

M. Gomez-Barrero and C. Busch

with a 25 mm fixed focal length lens optimised for wavelengths within 900 – 1700 nm. Following
the findings of [30] for facial PAD, four wavelengths are captured, namely: 1200 nm, 1300 nm,
1450 nm, and 1550 nm. It is in these wavelengths that all skin types present similar remission
curves (i.e., the bona fide intra-class variability is reduced), thereby facilitating the PAD task of
discriminating skin vs. non-skin materials (i.e., presentation attacks).

It should also be noted that the camera captures the finger slot and the surrounding area
of the box. Since the finger is always placed over the fixed open slot, and the camera does not
move, the region of interest (ROI) can be extracted using a simple fixed size cropping. Thus,
the final ROI has a size of 18 × 58 px. The four ROIs for a bona fide, from now on referred to
simply as images or samples, are depicted in Fig. 1.

Furthermore, fingerprint verification can be carried out with contactless finger photos ac-
quired in the visible spectrum with an additional 1.3 MP camera and a 35 mm VIS-NIR lens,
which are placed next to the SWIR sensor within the closed box. As it has been shown in [18],
commercial off-the-shelf systems can extract minutiae correctly from these samples, in order to
allow compatibility with conventional fingerprint sensors.

2.2 Multi-Spectral Convolutional Neural Networks (CNNs)

A convolutional neural network (CNN) is a class of deep neural networks, most commonly
applied to image analysis. The name convolutional refers to the mathematical operation called
convolution, which is a particular type of linear operation and is the core of this type of networks.
In general, a convolution is defined as the integral of the product of two functions after one is
reversed and shifted. This way, the convolutional operation produces a third function expressing
how the shape of one is modified by the other. In the specific case of CNNs, a convolutional
layer consists of a set of learnable filters (or kernels). To train their corresponding parameters,
each filter is convolved across the width and height of the input volume (e.g., three for an input
RGB image), computing the dot product between the entries of the filter and the input. This
process results in a 2-dimensional activation map of that filter. This way, the network learns
filters that activate when some specific type of feature at some spatial position is detected in
the input (e.g., a simple vertical line, or something more complex such as the shape of the ear
of a cat). Stacking the activation maps for all filters along the depth dimension forms the full
output of the convolution layer.

In general, CNNs consist of an input and an output layer, as well as multiple hidden layers,
whose inputs and outputs are masked by the activation function. These hidden layers typically
consist of a series of the aforementioned convolutional layers. The activation function removes
negative values from an activation map by setting them to zero, thereby increasing the nonlinear
properties of the decision function and of the overall network without affecting the receptive
fields of the convolution layer. In addition, pooling and normalisation layers also increase the
convergence speed during training.

2.2.1 Input Pre-Processing

As mentioned above, most CNN models have been trained on, and thus expect, RGB images.
However, the SWIR sensor described in Sect. 2.1 outputs four different grayscale images acquired
at different wavelengths. Therefore, the PAD method proposed in [33] included a manual
preprocessing of the four SWIR samples to convert them to an RGB image, as depicted in
Fig. 2 (right).

In contrast to that manual preprocessing, which may not be optimised for the CNN model
at hand, we propose to let the network itself convert the four grayscale input channels intro
Multi-Spectral CNNs for Biometric Presentation Attack Detection

M. Gomez-Barrero and C. Busch

Figure 2: Left: The four SWIR images are automatically processed by the corresponding CNN model using a pre-processing convolutional layer with three filters of size $P \times P$ and a stride of 1. In addition, batch normalisation and a ReLu activation are used to facilitate convergence. The result is a 3-channel image. Right: the handcrafted conversion proposed in [33].

images (i.e., tensors) comprising three channels. This way, the network can apply different linear combinations to each region of the image and learn the most suitable features for the following layers. To that end, we include at the beginning of each CNN model the pre-processing module showed in Fig. 2 (left), as depicted in Fig. 3 in purple. This new convolutional layer has a four-dimensional tensor as input, a stride of one in order to preserve the image size, and a filter of size $P \times P$ px. The value of $P$ needs to be optimised ad hoc for each model. In addition, to facilitate convergence during training, batch normalisation and a ReLu activation function are added to the convolutional layer. The corresponding parameters will be trained together with the last layers of the pre-trained models, or the full network trained from scratch.

It should be noted that, as it will be shown in Sect. 4.1, each network will learn in this first layer different image representations, even though the same layer structure is used in all networks for pre-processing the four-dimensional input. This highlights the relevance of adequately pre-processing the data for the CNN model at hand, instead of using a fixed pre-processing method as in [33].

2.2.2 CNN Models

We consider five different CNN models, whose architectures are shown in Fig. 3. First, we study the three models analysed in [33], namely: i) a residual network trained from scratch, ii) a model based on the pre-trained MobileNet [12], and iii) a model based on the pre-trained VGG19 [28]. In addition, we also study two further CNN architectures: iv) a model based on the pre-trained MobileNetV2 [27], which is an improved version of MobileNet, and v) a model based on the pre-trained VGGFace (VGG16) [24], which has been trained on facial images, thus containing skin, instead of training it on the more general ImageNet database as the remaining pre-trained models. All strategies have been implemented under the Keras framework using Tensorflow as back-end, with a NVIDIA GeForce GTX 1080 GPU. Adam optimizer is considered with a learning rate value of 0.0001 and a loss function based on binary cross-entropy.

As already pointed out, the first approach is focused on training a residual CNN [10]
Figure 3: Proposed network architectures. From left to right: i) the residual CNN trained from scratch using only the SWIR fingerprint database (319,937 parameters); ii) the pre-trained MobileNet-based model (815,809 parameters); iii) the pre-trained MobileNetV2-based model (437,985 parameters, see Fig. 4 for details on the bottlenecks); iv) the pre-trained VGG19-based model (20,155,969 parameters); and v) the pre-trained VGGFace-based model (20,155,969 parameters). All pre-trained models are adapted using transfer learning techniques over the last white-background layers. Also, the first convolutional layer (purple) (i.e., “pre-processing layer”, see Fig. 2) is trained for all networks.
Figure 4: 3-layer structure of the bottleneck residual block of MobileNetV2, where \( t \) denotes the expansion factor, and \( c \) and \( s \) the number of filters and stride of the last convolutional layer.

After each convolution and before the ReLu activation.

The remaining CNN models considered are based on pre-trained models and transfer learning. In all cases, the fully-connected layers designed for the ImageNet [19] or facial classification tasks have been removed.

Both MobileNet [12] and MobileNetV2 [27] are based on depthwise separable convolutions. These layers perform a spatial convolution on each channel of their input, independently, before mixing output channels via a pointwise (i.e., \( 1 \times 1 \)) convolution. This is conceptually equivalent to separating the learning of spatial features and the learning of channel-wise features. This is justified by the correlations within neighboring pixels in the images, but the relative independence of a given pixel across channels, which is specially the case for the information captured at each SWIR wavelength. In addition, this type of convolutions require fewer parameters and computations, thereby allowing a speedy training using less data.

In both MobileNet networks, downsampling is directly applied by the convolutional layers that have a stride of 2 (represented by \( /2 \) in the convolutional layers of Fig. 3). This network architecture allows to reduce both model size and training/testing times, thus being a good solution for mobile and embedded vision applications. With respect to MobileNet, the main contribution of MobileNetV2 [27] is the use of residual connections and inverted bottlenecks (see Figs. 3 and 4). The rationale to implement such bottlenecks stems from the assumption of the low dimensionality of the manifold of interest on which the discriminative information extracted by the internal layers of the network lies. Therefore, linear bottleneck layers are introduced in the model and the residual connections are established between the aforementioned bottlenecks (i.e., in contrast to more common approaches where the residuals connect layers with a higher number of filters or output channels).

Finally, given the depth of both MobileNet models and the limited amount of data available, out of the 13 blocks of MobileNet, we retain only eight of the blocks, adapting the last one during training. Similarly, out of the 16 bottlenecks of MobileNetV2, 12 are used and the last two re-trained.

On the other hand, two different VGG based models have been studied, VGG19 [28] and VGGFace\(^1\) [24]. These networks are older and more simple than the MobileNets; however, due to its simplicity, VGG19 is still one of the most popular network architectures, providing very good results in a wide range of competitions. In fact, VGG19 showed a superior performance with respect to MobileNet for fingerprint PAD in [33].

Whereas VGG19 comprises 19 different layers, VGGFace is based on the smaller VGG16

\(^1\)Implementation available at https://github.com/rcmalli/keras-vggface
model, including 16 layers. In addition, the later has been trained on facial databases acquired in the wild (i.e., modelling realistic scenarios in opposition to controlled environments with frontal poses and fixed illumination). Therefore, VGG19 has been pre-trained on a multi-class task in contrast to the two class problem of face recognition for VGGFace. For our study, the last fully-connected layers have been replaced with 2 fully-connected layers with a final sigmoid activation function. In addition, the last three convolutional layers, depicted in white in Fig. 3, are re-trained in both models.

2.3 Fusion

As it was already observed in [33], different CNN models are more robust to specific PAI species than others. Therefore, the fusion of the final PAD scores output by several models yields a higher detection performance. In our case (see Sect. 4 for more details), we found that the optimal results are achieved fusing three different models: ResNet, MobileNetV2, and VGGFace. Therefore, we define the final PAD score as follows:

$$s = \alpha \cdot s_{vggF} + \beta \cdot s_{mob2} + (1 - \alpha - \beta) \cdot s_{res}$$

(1)

where $\alpha + \beta \leq 1$ are the weights assigned to VGGFace and MobileNetV2, respectively. Those weights are optimised over the validation set in order to minimise the detection error rates.

3 Experimental Setup

3.1 Database

The database considered in the experimental evaluation was acquired within the BATL research project [1], funded by the IARPA Odin program [22]. In collaboration with our project partners at USC, data from 562 subjects were collected in two different stages, and comprise both bona fide and PA samples. From each subject, five to six fingers were captured. It is important to highlight that people from different gender, ethnicity, and age were considered during the acquisition in order to model realistic conditions. In total, the database comprises 4,290 and 443 bona fide and PA samples.

For the PA samples, the selection of the PAI fabrication materials was based on the requirements of Odin program evaluation, covering the most challenging PAIs [29, 20]. There are a total of 35 different PAI species, which can be further categorised into eight main groups, namely: dragon skin, latex, overlay, playdoh, printed fingers, silicone, silly putty, and wax. It should be noted that all the materials used can be acquired in online shops, and the PAIs can be fabricated by non-experts in the area. This is why they pose a severe threat to biometric systems.

To carry out a fair evaluation, the database is split into non-overlapping training, validation, and test datasets, as summarised in Table 1. For the development of our proposed fingerprint PAD methods, both training and validation datasets are used in order to train the weights of the systems and select the optimal network architectures. It is important to highlight that we consider the same number of samples per class during the development of the systems in order to avoid bias towards one class. For the final evaluation, the test dataset comprises the remaining bona fide (4071) and PA (222) samples. Furthermore, the test dataset includes 5
Table 1: Partition of training, validation and test datasets.

<table>
<thead>
<tr>
<th></th>
<th># Samples</th>
<th># PA Samples</th>
<th># BF Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>260</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>Validation set</td>
<td>180</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>Test set</td>
<td>4293</td>
<td>222</td>
<td>4071</td>
</tr>
</tbody>
</table>

unknown PAI species, which were not considered during the development stage (i.e., they are not present either in the train or in the validation datasets). This way, we can also evaluate the robustness of our proposed methods to unknown attacks, thereby modelling realistic and more challenging scenarios.

3.2 Metrics

The performance of the PAD method is evaluated in compliance with the ISO/IEC IS 30107-3 on Biometric presentation attack detection - Part 3: Testing and Reporting [15]. To that end, the following metrics are used:

- **Attack Presentation Classification Error Rate (APCER):** percentage of attack presentations wrongly classified as bona fide presentations.
- **Bona Fide Presentation Classification Error Rate (BPCER):** percentage of bona fide presentations wrongly classified as presentation attacks.

The operating point at which APCER = BPCER is denoted as Detection Equal Error Rate (D-EER). In addition to the D-EER, the APCER at BPCER = 0.2\% (denoted as APCER\(_{0.2}\%)\) will be also reported to evaluate systems with a high user convenience.

3.3 Experimental Protocol

Two different sets of experiments are carried out:

- **Input pre-processing optimisation:** first, the optimal filter size \(P\) (see Sect. 2.2.1 and Fig. 2) needs to be determined for each model, in order to obtain the best detection performance. This is done individually for each CNN model described in Sect. 2.2.2.

- **Final fused system:** after determining the optimal filter size and the APCEs of each CNN model, the best fusion is carried out. In addition, the results are benchmarked with the state of the art reported in [8, 33].

4 Results

4.1 Input Pre-Processing Optimisation

We first analyse the detection performance obtained over the test set for different filter sizes \(P\) for each CNN model individually. The corresponding DET curves are presented in Fig. 5. It should be noted that, whereas ResNet and both VGG models are fed a ROI image resized to \(58\times58\) px, both MobileNets receive \(128\times128\) px. images as proposed in [33]. Therefore, the filter sizes analysed for the former are smaller than for the latter.

As we may observe in Fig. 5a, ResNet achieves the lowest D-EER (0.45\%) for a filter of size 11 (solid green). Even though this configuration presents a detection performance comparable
to its RGB version (dashed dark blue) for a BPCER $\geq 0.5$ (or APCER $\leq 0.5\%$), the most relevant operation points for PAD purposes are those with a lower BPCER (e.g., BPCER $= 0.2\%$). These yield a higher user convenience, which is for instance prioritized in the Odin Program [22], thereby achieving a higher user acceptance. For these lower BPCER values, $P = 11$ achieves the best detection performance, with an APCER$_{0.2\%} = 1.80\%$. In contrast, the RGB model reported an APCER$_{0.2\%} = 6.79\%$, thereby having achieved a 73\% relative improvement with the new input pre-processing module in the CNN. On the other hand, for an APCER $\leq 0.5\%$ (i.e., very high security operating point), the best configuration is achieved by the next filter size tested, with $P = 13$. Taking all these considerations into account, the best performing filter size for low BPCERs, $P = 11$, is considered for further experiments.

A similar behaviour to ResNet is observed in Fig. 5b for MobileNet and $P = 13$ (solid green), where the D-EER is reduced from 1.80\% to 0.90\%, and the APCER$_{0.2\%}$ from 19.91\% to 4.96\% (i.e., 75\% relative improvement).

Regarding MobileNetV2, Fig. 5c shows a somewhat different trend. Except for $P = 9$, all DET curves are close to each other. The main differences are observed for either low APCERs or low BPCERs. In particular, for low BPCERs all filter sizes $P \neq 17$ present horizontal asymptotes. In other words, the corresponding networks always classify a subset of bona fide samples as attacks and are thus unable to reach a BPCER $= 0\%$. In addition, $P = 17$ (solid green, APCER$_{0.2\%} = 4.96\%$) also shows the lowest BPCER for APCER $\leq 0.5\%$, thus making it the best configuration for this model.

For VGG19, two different RGB networks are analysed in Fig. 5d. First, Tolosana et al. [33] (dashed blue line) proposed re-training the last four layers of the network. However, with the addition of the pre-processing input layer, and the reduced amount of data available for adapting the network, convergence was not achieved. Therefore, we decided to adapt only the last three layers, as shown in Fig. 3. The corresponding DET curve is depicted in dashed orange. As a consequence, the D-EER increased from 1.36\% to 1.80\%. However, when all four images are fed to the multi-spectral network, we observe that the performance is improved both for $P = 7$ and 9. In particular, for lower BPCERs, we see that $P = 7$ (purple) manages to correct the main drawback of VGG19: it is able to achieve a BPCER $= 0\%$ instead of presenting an horizontal asymptote at BPCER around 1\%. Therefore, the performance of the original RGB CNN is considerably outperformed by the MS CNN for $P = 7$, which also achieves an APCER$_{0.2\%} = 3.60\%$.

Finally, the performance of the VGGFace based model is shown in Fig. 5e. In this case, we observe again a similar behaviour to VGG19, with horizontal asymptotes, which is reasonable since VGGFace is based on the very similar VGG16 model. However, in this case the asymptotes are shown for all APCER values instead only for APCER $\geq 0.5\%$. This is due to the nature of the database on which VGGFace was pre-trained: it contains only facial images. Therefore, the first layers of the network, which are not adapted for the PAD problem at hand, extract relevant information based on human skin and hair, relevant to face recognition, but have not seen other materials for the fabrication of PAIs. This leads to a poorer performance on average with respect to the other models, since a low BPCER of e.g. 0.2\% cannot be achieved. On the other hand, for low APCER values, the BPCER remains at 0.90\% for $P = 7$, thereby showing in this case the best performance for this limited range. As it will be shown in Sect. 4.2, this will be relevant for a fused approach.

To conclude this subsection, we analyse the output of the pre-processing layers: is it really different for each CNN model? In other words, is each model learning different information from the very beginning? The answer is yes, as it may be observed for a bona fide sample in Fig. 6a for the best filter sizes found before. In particular, both MobileNet and MobileNetV2 learn
Figure 5: DET curves for each individual CNN model and different filter sizes $P$. 
representations with information distributed across the whole image, with different weights per channel, similarly to the manual pre-processing carried out by [33]. However, the remaining models focus on particular regions of the image and present a bigger black area “empty” of information.

In addition, different PAI species are depicted in Fig. 6b for each model. For the “empty” representations of ResNet, VGG19, and VGGFace for the bona fide samples, we can see that the PAIs contain in contrast more information, and the CNNs can therefore discriminate bonafides from PAs. In contrast, for MobileNet and MobileNetV2, the difference between bonafides and PAs lies on the texture or the colour of the image.

Finally, it should be highlighted that all unknown attacks included in the test dataset were correctly detected by all the individual CNN models.

### 4.2 Final Fused System

In the second and final set of experiments, we tried different fusions of two and three CNN models in order to optimise the overall detection performance. To select which models would be fused, the APCES were analysed in detail, such that complementary models (i.e., making different errors) would be fused. The best configuration was the fusion of ResNet with a filter size of 11 (i.e., the best individual model) with MobileNetV2 with a filter size of 17, and VGGFace with a filter size of 7. Using different filter sizes for each network also implies that each model is focusing at different resolutions, thereby extracting different information. The score level fusion for all possible $\alpha$ and $\beta$ values (see Sect. 2.3) was analysed in terms of the APCER$_{0.2\%}$, and the best results (i.e., ACPER$_{0.2\%} = 1.35\%$) are presented in Fig. 7a. As it
Multi-Spectral CNNs for Biometric Presentation Attack Detection  M. Gomez-Barrero and C. Busch

<table>
<thead>
<tr>
<th>APCER (%)</th>
<th>BPCER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

## Figure 7: DET curves for (a) the best multi-spectral SWIR fusion, and (b) the fusion of the former MS SWIR with the LSCI approach proposed in [17], benchmarked with the results reported in [33] (RGB SWIR) and [8] (RGB Fusion).

<table>
<thead>
<tr>
<th>Fusion PAD Scores</th>
<th>Zoom: Fusion PAD Scores (&gt; 0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bona Fides</td>
</tr>
<tr>
<td></td>
<td>Presentation Attacks</td>
</tr>
</tbody>
</table>

---

Figure 8: Fused PAD score distributions: (a) both PA and bona fide scores for all samples, and (b) zoom of the bona fide PAD scores > 0.001.

---

SWIR and LSCI Fusion

<table>
<thead>
<tr>
<th>α * VGGFace + β * MobileNetV2 + (1 - α - β) * ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>α = 0.1, β = 0, D-EER = 0.45%</td>
</tr>
<tr>
<td>α = 0.3, β = 0.1, D-EER = 0.45%</td>
</tr>
<tr>
<td>α = 0.4, β = 0.1, D-EER = 0.45%</td>
</tr>
<tr>
<td>α = 0.7, β = 0, D-EER = 0.45%</td>
</tr>
</tbody>
</table>

---

Integration in the final PAD decision threshold of all scores may be observed, in all cases a remarkable D-EER = 0.45% is achieved.

Even if the detection performance obtained might be considered very good, we still can see that the BPCER remains constant for all APCER values under 0.5%. Let us thus analyse the PAD scores in Fig. 8 in detail. As we can see in Fig. 8a, all but one PA scores are close to one, in particular over δ = 0.9937. However, there is a single PAD score of 0.5933 (one conductive silicone overlay out of the 60 included in the test set). On the other hand, the distribution of the bona fide scores is wider, and comprises scores in the whole range. In particular, 3026 PAD scores out of the 4071 are lower than 0.0001 for the bona fides. The distribution of the remaining 1045 scores (i.e., 25.65% of the total number of scores) is shown in Fig. 8b.

Taking into account the analysis of the scores, we set the PAD decision threshold at δ = 0.9937. In this case, only 13 bona fide samples are misclassified; i.e., a BPCER = 0.32% is obtained. The corresponding BPCE SWIR samples at 1200 nm are shown in Fig. 9. Carefully analysing those images, we can observe that the high scores are mostly motivated by missposi-
tioning the finger on the slot, so that the ROI still includes black areas. In addition, one sample is badly focused and shows no texture, and the base of the finger is badly illuminated on the last images, most probably due to a non-horizontal position of the finger. That way, the base of the finger is under illuminated.

Finally, we focus on the fusion with $\alpha = 0.4$ and $\beta = 0.1$. Even if the performance is similar to the approach with $\alpha = 0.7$ and $\beta = 0.0$, including information from three instead of two models will make the approach more robust to yet unforeseen attacks. In Fig. 7b, we benchmark this approach (solid yellow) with the best RGB SWIR fusion in [33] (dashed green), and with the RGB SWIR and laser speckle contrast imaging (LSCI) fusion described in [8] (dashed orange). For completeness, we also show the individual LSCI results from [8, 17] (dashed purple). As it may be observed, the fused multi-spectral SWIR approach (solid blue) already outperforms the state of the art approaches presented in [33, 8]. This improvement also holds for BPCER $\leq 0.5\%$ even in the case of fusing the RGB SWIR CNNs with LSCI data. Furthermore, if we also fuse at score-level the multi-spectral SWIR CNNs with the same LSCI approach (solid yellow), we can further reduce the D-EER to 0.2%, and the ACPER$_{0.2\%}$ is as low as 0.90% (i.e., only two PA samples are not detected).

5 Conclusions

In this paper, we have analysed the use of multi-spectral CNN models with four input channels in combination with SWIR images for fingerprint PAD purposes. In contrast to the manual pre-processing method proposed in [33] to extract RGB images from the four grayscale SWIR samples, in our work we add a convolutional layer at the beginning of the CNN models with four channels as input and three channels as output. This additional layer is trained together with the full network or its last layers in the case of transfer learning approaches.

We thoroughly studied different filter sizes for the new pre-processing layer, and took into account five different and state of the art CNN models, both trained from scratch and previously trained on bigger databases acquired in the visible spectrum. The results, evaluated on a database comprising more than 4700 samples and 35 different PAI species, indicate a clear improvement with respect to the method proposed in [33]. For the best fusion configuration found in each work, the D-EER decreased from 1.35% to 0.45% (relative decrease of 66%).

In addition, a score-level fusion of the SWIR method with handcrafted features extracted from LSCI data following the example in [8] yields a very secure and very convenient PAD module. In more details, only two PA samples out of 222, stemming from 35 different PAI species, are misclassified for a high user convenience operating point with a BPCER = 0.2% (i.e., only 2 in 1,000 bona fide attempts are rejected).
Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under contract number 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References


