Domain Adaptation in Multi-Channel Autoencoder based Features for Robust Face Anti-Spoofing

Olegs Nikisins, Anjith George, Sébastien Marcel
Idiap Research Institute
Rue Marconi 19, CH - 1920, Martigny, Switzerland
{olegs.nikisins, anjith.george, sebastien.marcel}@idiap.ch

Abstract

While the performance of face recognition systems has improved significantly in the last decade, they are proved to be highly vulnerable to presentation attacks (spoofing). Most of the research in the field of face presentation attack detection (PAD), was focused on boosting the performance of the systems within a single database. Face PAD datasets are usually captured with RGB cameras, and have very limited number of both bona-fide samples and presentation attack instruments. Training face PAD systems on such data leads to poor performance, even in the closed-set scenario, especially when sophisticated attacks are involved. We explore two paths to boost the performance of the face PAD system against challenging attacks. First, by using multi-channel (RGB, Depth and NIR) data, which is still easily accessible in a number of mass production devices. Second, we develop a novel Autoencoders + MLP based face PAD algorithm. Moreover, instead of collecting more data for training of the proposed deep architecture, the domain adaptation technique is proposed, transferring the knowledge of facial appearance from RGB to multi-channel domain. We also demonstrate, that learning the features of individual facial regions, is more discriminative than the features learned from an entire face. The proposed system is tested on a very recent publicly available multi-channel PAD database with a wide variety of presentation attacks.

1. Introduction

Presentation Attacks Detection (PAD), also known as anti-spoofing, has gained a significant attention in the biometric society, since high accuracy state-of-the-art face recognition methods are known to be vulnerable to presentation attacks [4, 17]. This loophole in the security of recognition and verification systems is unacceptable in high security applications, such as border control or law enforcement, dictating the need of having a human-in-the-loop co-supervising the recognition process. Also, the fabrication of Presentation Attack Instruments (PAI) is getting more trivial, due to common availability of dozens of biometric samples in social networks, and improving fabrication technologies, such as 3D printers. Thus, the availability of *high accuracy* face PAD systems is the missing component in the wide deployment of face recognition technologies.

Presentation attacks in general are of unknown nature, meaning that it can be anything as long as it helps the attacker either *impersonate* or *obfuscate* the identity. Despite this fact, a vast majority of the research in face PAD, focuses on two types of attacks: photo and replay PAI. One explanation to this phenomena is a domination of publicly available databases containing just these two types of attacks, for example SiW [14], Replay-Mobile [8], OULU-NPU [6], MSU MFSD [24], or aggregations of such [18]. Also, the biometric samples in these databases are RGB only. Recent research [18, 2] states that RGB-based face PAD systems have relatively low performance in general, even for fore-mentioned two PAI types, and the situation is getting worse in the unseen-attacks tests. In our experiments we use a database containing a much wider range of PAIs, both 2D and 3D, as well as partial attacks, demonstrating that RGB-based PAD performs poorly even in the seen-attacks scenario, which coincides with [18, 2]. The solution boosting the performance of PAD system is to use multi-channel (**MC**) based approach, enhancing PAD system with specialized imaging sensors. Most of the effort with specialized sensors for facial biometrics was concentrated on face recognition (**FR**) [22]. However, recently some authors pointed out the applicability of this idea to face PAD. A preliminary study using NIR and LWIR and Depth (**D**) cameras for face PAD was introduced in [3]. In [3] authors argue that not just simple PAI, *e.g.* photo and replay PAI, but also advanced attacks, such as silicon masks, should be detectable much easier using MC PAD approach, however they don't introduce any practical algorithmical solution in the paper. In [23] a SWIR-camera based skin detection methodology is introduced, being potentially applicable to face PAD task, assuming that in the attack attempts the large fraction of the face is covered with synthetic materials. The potential challenge for skin-detection based PAD is partial attacks, covering just a small fraction of the face potentially important for FR methods, *e.g.* eye region. In [20] authors developed a PAD method using multi-spectral camera, capturing the samples in 7 spectral bands from 425 to 930 (nm). They proposed an algorithm based on hand-crafted, LBP and BSIF, features and SVM classifier, as well as various data fusion strategies at image and score levels. The reported results are promising, however the PAI types used in the experiments of [20] are limited to photo attacks only. To the best of our knowledge, our work is the first attempt introducing deep-learning based multi-channel face PAD system efficiently detecting a wide range of PAIs, such as 2D, 3D, and partial attacks.

The core of the *proposed* **MC face PAD** algorithm is a Convolutional Neural Network (CNN), which is decomposed into two components: a **set of MC encoders** extracted from pre-trained autoencoders (**AE**), and a **Multi-Layer Perceptron** (**MLP**) combining previous set of encoders. A task of the **set of MC encoders** is *feature extraction* from multi-channel input data, which in our case is a stack of gray-scale, NIR, and Depth (**BW-NIR-D**) facial images, or a stack of images of facial features, *e.g.* left-eye region. Is remarkable, that AE are trained using bona-fide samples only, learning the appearance of the real face. Moreover, instead of collecting more data for AE training, the domain adaptation technique is proposed, transferring the knowledge of facial appearance from RGB to BW-NIR-D domain. A task of **MLP** is *classification* categorizing the features of MC encoders as either bona-fide or an attack. Only MLP is trained using samples from both bona-fide and attack classes. To the best of our knowledge, both proposed Autoencoder+MLP based PAD algorithm, and domain adaptation (transfer learning) approach, are unique in the field of MC face PAD. Most of the research in the field of face PAD rely on the binary classifiers, *e.g.* SVM or LR, categorizing hand-crafted features, such as LBP or IQM [21]. In [20] similar strategy is applied to the task of MC face PAD, also testing different strategies for channels fusion. Relatively recent trend in face PAD is to use deep-learning based techniques. In [13] authors are using transfer learning ideas fine-tuning VGG-Face model [19], CNN originally designed for face recognition, to the face spoofing datasets. Then PCA is used reducing the dimensionality of the feature vectors, which are next classified with SVM. The experiments are done on CASIA-FA [27], Replay-Attack [7] containing photo and replay PAI only. Another CNN-based face PAD paper [10] is focusing on initialization procedure of CNN weights, arguing it improves the convergence of the training and overall performance of the system. Authors [10] first train a set of 9 patch-CNNs, learning features of a different facial regions. Afterwards, weights of patch-CNNs are substituted into a single CNN, which is then fine-tuned on the whole face. Again, a set of PAI in experimental section of [10] is limited to photo and video attacks. In [15] CNN-RNN model is proposed estimating two biometric traits from an input RGB video - depth pattern of the face, and rPPG signal. The results on SiW [14], and OULU-NPU [6] are promising, however from problem formulation, one can conclude that partial attacks can be problematic for this approach, since both rPPG and shape are preserved in this case. The are also attempts to use autoencoders for face PAD, in [25] authors combine hand-crafted LBP features and an autoencoder based outliers detection algorithm. It aims to detect unknown PAI better than other methods using hand-crafted features, however overall performance on OULU-NPU [6] is relatively low. In [5] authors suggest an interesting methodology on detecting anomalies in the face, *e.g.* partial attacks, however they don't discuss the application of the algorithm to the face PAD. To the best of our knowledge, there is no previous published research using deep-learning based approaches for MC face PAD, making our paper an initial contribution to this promising direction.

To summarize, the following **main contributions** are proposed in our paper. *First*, a novel deep-learning based MC face PAD algorithm is introduced, having a CNN-like structure composed of a *set of MC encoders* and an *MLP*. *Second*, instead of collecting more training data, the domain adaptation technique is proposed, transferring the knowledge of facial appearance from RGB to multi-channel domain. Domain adaptation is done via autoencoders, which are first pre-trained on a large publicly available RGB face database, and are then *partially* fine-tuned on the set of BW-NIR-D face images. *Third*, we demonstrate, that learning the features of individual facial regions, is more discriminative than the features learned from an entire face. Proposed MC face PAD method gives very promising results, significantly outperforming hand-crafted features based baseline, on the challenging database, namely WMCA, containing a wire rage of PAI - 3D, 2D, and partial attacks. *Finally*, the results reported in this work are fully reproducible: experimental database is publicly available, the evaluation protocols are strictly defined, and the source code for replicating experiments is published[*].

2. Proposed multi-channel face PAD approach

This section briefly introduces the proposed multi-channel face PAD system. The PAD algorithms is CNN based, where special CNN structure is used to categorize input MC data into bona-fide or attack classes, see Figure 1.

[*] Source code: https://gitlab.idiap.ch/bob/bob.paper.mcae.icb2019

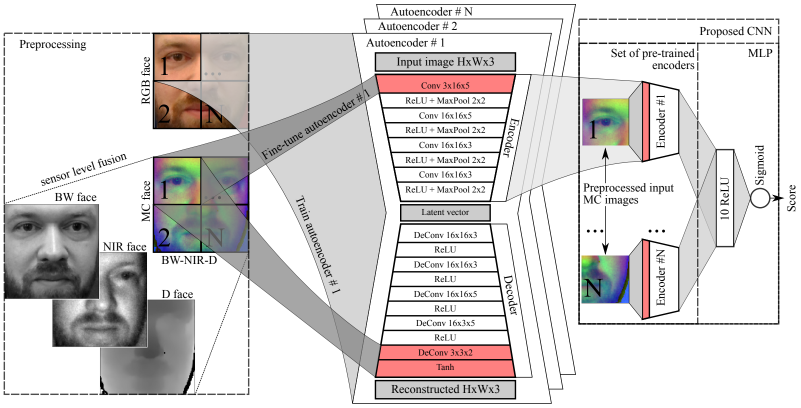

Figure 1. Schematic representation of the proposed MC face PAD approach: visualizing the preprocessing stage, internal structure of convolutional autoencoders, training and domain adaptation (fine-tuning) strategy, and the structure of the final CNN-based PAD system. The Conv/DeConv layers are parametrized as follows: number of input channels × number of output channels × size of the filter kernel.

While the network architecture can be represented as CNN, the proposed training technique, allowing domain adaptation, is unique, and is discussed later. An input MC data for the CNN is a *set of stacked BW-NIR-D* images corresponding to different facial regions, *e.g*. left-eye, see Figure 1. According to the widely accepted taxonomy [11], this type of biometric information integration from multiple cameras is called *sensor level fusion*. Sensor level fusion preserves maximum variance in the fused data, allowing the network to assess the hidden dependencies in the lowest level. However, this type of fusion requires special data preprocessing, making the streams to be blended compatible.

#	Input stream	Preprocessing
1.	RGB	Face, and landmarks detection in all frames.
2.	RGB	Conversion of all frames to BW format.
3.	NIR, D	Registration / alignment to BW channel.
4.	NIR, D	Normalization of dynamic range to $[0, 255]$.
5.	NIR, D	Data type-casting to 8-bit format.
6.	BW, NIR, D	Scaling, rotation, and cropping of facial regions.
7.	BW, NIR, D	Stacking of facial images into BW-NIR-D image.

Table 1. Preprocessing steps to generate BW-NIR-D facial images.

Preprocessing applied to all frames of input RGB, NIR, and D videos, allowing to generate BW-NIR-D facial images, is summarized in Table 1. An example of the preprocessing product is displayed in Figure 1, denoted as MC face, and RGB face. To generate RGB facial images, only steps 1 and 6 are applied to the input RGB stream. The face and landmark detection method is using MTCNN [26] deep neural network. The proposed *dynamic range normalization* technique of the NIR, and D streams is based on Median Absolute Deviation (MAD) measure. To clarify the MAD-based normalization, let **I** be a non-RGB image of the facial region. Given **I**, a vector **v** containing non-zero elements of **I** is obtained. Next, a MAD measure is computed as follows:

$$\mathbb{MAD} = median(|\mathbf{v} - median(\mathbf{v})|) \tag{1}$$

Given the $\mathbb{MAD}$ value, the input image **I** is normalized:

$$\hat{\mathbf{I}}_{i,j} = \frac{(\mathbf{I}_{i,j} - median(\mathbf{v}) + \sigma \cdot \mathbb{MAD})}{2 \cdot \sigma \cdot \mathbb{MAD}} \cdot (2^8 - 1), \tag{2}$$

where $\hat{\mathbf{I}}$ is normalized image, $i = 1, \dots, W$; $j = 1, \dots, H$, and W, H are the width and height of **I**. In our experiment: $\sigma_{NIR} = 3$, $\sigma_D = 6$. The size of the facial images is normalized to 128 × 128 pixels, and the images are rotated so that the eye-line is horizontal. Given preprocessed training data the subsequent step is CNN training.

#	Train step	Training data	DB, classes used
1.	Train N AEs	RGB face regions	CelebA [16], BF
2.	Fine-tune N AEs	MC face regions	WMCA, BF
3.	Train an MLP	MC latent encodings	WMCA, BF&PA

Table 2. Steps to train a CNN-based MC face PAD system. BF and PA stands for samples from *bona-fide* and *presentation attack* classes.

Training of the proposed CNN based multi-channel face PAD system is using both RGB and MC facial data, and is summarized in Table 2. An entire CNN is never trained, instead training steps are associated to two internal CNN components: a set of convolutional autoencoders, and an MLP, see Figure 1. First, a set of N convolutional autoencoders, reconstructing RGB facial regions, is pre-trained. Pre-training is done using images from publicly available database, namely CelebA [16], having significantly higher number of bona-fide samples and identities, than any PAD-specific database to date. AE pre-training helps to better position the network in the search space. The subsequent step is fine-tuning of AEs using MC face regions, extracted

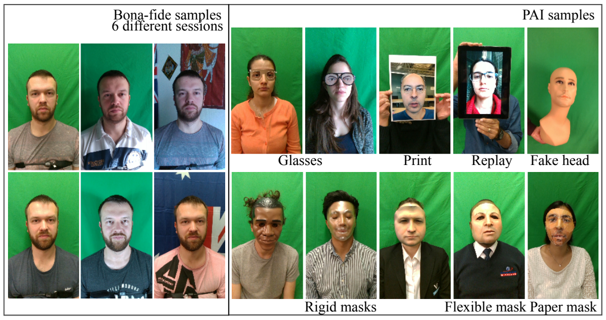

Figure 2. Examples of bona-fide data in 6 different sessions, and presentation attacks corresponding to 7 different categories.

from train set of WMCA. It is worth mentioning, that the dimensionality of RGB and MC training samples is identical. Here we propose to *fine-tune just the first layers of encoders* (partial fine-tuning), and last 2 layers of decoders (for symmetry), instead of fine-tuning the entire autoencoders. The intuition behind this proposition, is that only *low level features are domain dependent*, while deeper features are domain independent mostly preserving structural information of the face. In the experimental section we prove empirically, that proposed RGB-to-MC domain adaptation via *partial fine-tuning* is more efficient than full fine-tuning. The effectiveness of similar domain adaptation strategy has also been addressed in [9], focusing on the task of heterogeneous face recognition. Both pre-training and fine-tuning of AEs is using bona-fide samples only, and is taking 50 epochs each. The *latent vectors* of trained encoders are used as input features for the MLP, see Figure 1. An MLP is trained using latent features of both bona-fide and attack classes present in the training set of the WMCA database, Table 2. The best performing MLP model is selected cross-validating them on the development set of WMCA. Summarizing, the composition of trained convolutional autoencoders, and MLP, is forming the proposed MC face PAD system.

3. Experimental Database

The introduced face PAD system is evaluated on the *Wide Multi-Channel presentation Attack* (**WMCA**) database (DB), containing a total of **1679** video files of bona-fide and presentation attacks corresponding to **72** identities. The temporally synchronized video streams are recorded with two consumer capturing devices, Intel$^{\circledR}$ RealSense$^{\text{TM}}$SR300 capturing RGB-NIR-D streams, and Seek Thermal CompactPRO recording the thermal (T) stream. The data collection was split into **seven** sessions over an interval of **five** months. Bona-fide samples were recorded in 6 sessions. In each session, a bona-fide and at least two PA performed by the participant were captured.

The WMCA DB has more than **eighty** different presentation attack instruments, which can be groped into **seven** categories: glasses, fake head, print, replay, rigid mask, flexible/silicon mask, and paper mask. Examples corresponding to these PA categories, as well as the case of bona-fide data for 6 different sessions, are shown in Figure 2. More detailed information on the number of files for each PA category, and video data technicalities, are summarized in Table 3. Please refer to [12] for more details on the WMCA database.

Type	#Videos	Video Details
bona-fide	347 (72 IDs)	Length:
glasses	75	10 seconds
fake head	122	
print	200	#Frames:
replay	348	RGB-NIR-D: 300
rigid mask	137	T: 150
flexible mask	379	
paper mask	71	Format:
Total	1679 (5.1 TB)	Uncompressed

Table 3. Main statistics of the WMCA DB.

4. Experimental evaluation

This section covers the details on the evaluation protocols for the WMCA DB, following by the experimental results for baseline MC PAD algorithm, and the results for the proposed PAD setup with different settings. In all experiments, only **RGB-NIR-D** channels available in WMCA are used, making the proposed PAD system dependent on one capturing device only, specifically Intel$^{\circledR}$ RealSense$^{\text{TM}}$SR300.

The WMCA **evaluation protocol**, namely **grandtest-10**, follows the legacy evaluation strategy, in this protocol samples of all PAI categories, and bona-fide, are evenly distributed across all subsets: *training*, *development*, and *evaluation*. The identities of bona-fide samples are not intersecting across these subsets. The number **10**, in the protocol name, stands for the *number of frames* uniformly sampled from each video, thus the total number of biometric samples, using frame-level evaluation strategy, is $1679 \cdot 10$. Training set is used for training the PAD system. The threshold corresponding to the selected operation points is chosen on the development set, and the system performance is reported on the evaluation set given the threshold.

The **evaluation metrics** is adopted from the ISO/IEC 30107-3 standard [1], APCER (Attack Presentation Classification Error Rate), and BPCER (Bona-fide Presentation Classification Error Rate). We also adopt two measures, namely BPCER20, and BPCER100, which are the BPCER at APCER 5.0%, and 1.0%, respectively. Again, the thresholds corresponding to BPCER20, and BPCER100, are selected on the development set, then BPCER, and APCER

Channel	RGB	NIR	D	Fused
Method	IQM+LR	LBP+LR	LBP+LR	LR
BPCER20	77.7	9.9	13.8	**3.0**
APCER	13.2	7.1	9.6	10.4
BPCER100	94.6	35.6	57.5	**14.1**
APCER	8.5	1.9	2.0	2.9

Table 4. Baseline results for *evaluation* set of WMCA, **grandtest-10** protocol. Thresholds are computed on *development* set.

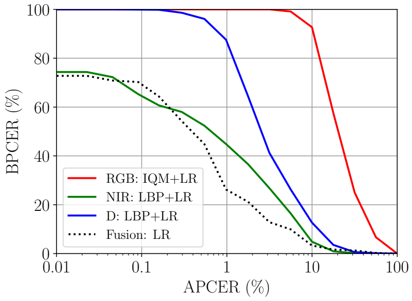

Figure 3. DET curves for baselines. For the *evaluation* set of WMCA, **grandtest-10** protocol.

values are reported on the evaluation set given threshold. In addition to numerical performance, the DET curves are given for all experiments.

4.1. Results: MC face PAD baselines

In this subsection a baseline RGB-NIR-D face PAD system is discussed and evaluated, being composed of successful, channel specific, hand-crafted features combined with two-class classifiers. The baseline PAD system is composed of 4 blocks: *preprocessing*, *feature extraction*, *classification*, and *fusion*. The **preprocessing** stage is similar to the one introduced in Section 2, excluding sensor level fusion. In the baseline setup each channel is handled individually, and preprocessed data examples are displayed in Figure 1, denoted as RGB, NIR, and D face. The cropped faces are normalized to 64×64 pixels, and the frames with faces smaller than 50×50 are discarded.

In the **feature extraction** stage, a grid search across popular hand-crafted features was adopted to identify the best performing ones. The considered features demonstrating reasonable performance in recent literature are Image Quality Measures (IQM) [18], and LBP/MCT histograms [18, 25]. An IQM is used for the RGB channel, producing feature vectors of 139 quality measures. As a result of a grid-search in the parameter space of spatially enhanced LBP/MCT histograms, an $MCT_{8,1}$ (8 sampling points on a radius of 1) features are chosen for NIR data, and $LBP_{8,1}$ spatially enhanced histograms computed over 2×2 regions are selected for D frames.

Method	**Single AE for entire MC face + MLP**		
AE training	CelebA	CelebA+WMCA all encoder layers	CelebA+WMCA 1 encoder layer
BPCER20	3.0	1.9	4.1
APCER	4.7	4.5	2.8
BPCER100	59.0	51.7	51.6
APCER	0.0	0.3	0.0
Method	**9 AE using MC face blocks + MLP**		
AE training	CelebA	CelebA+WMCA all encoder layers	CelebA+WMCA 1 encoder layer
BPCER20	1.1	5.0	1.5
APCER	5.9	3.8	3.1
BPCER100	11.9	12.3	**7.3**
APCER	1.1	0.7	0.8
Method	**16 AE using MC face blocks + MLP**		
AE training	CelebA	CelebA+WMCA all encoder layers	CelebA+WMCA 1 encoder layer
BPCER20	1.7	3.0	**1.0**
APCER	3.1	3.6	3.5
BPCER100	10.7	16.4	20.4
APCER	0.7	0.6	0.2

Table 5. BPCER20, BPCER100, and corresponding APCER in %, for the proposed MC face PAD system, reported for *evaluation* set of WMCA, **grandtest-10** protocol. Experiments for different regioning of facial images, and different AE training strategies. Thresholds are computed on *development* set.

The **classification** stage deploys a Logistic Regression (LR) for all channels. The features are normalized to zero-mean and unity standard deviation before the training, and the normalization parameters are computed using samples of bona-fide class only. In the prediction stage, a probability of a sample being a PA is computed given pre-trained LR model. A **fusion** of individual RGB-NIR-D channels is done in the score-level using the same LR-based approach. An LR for the fusion stage is trained using channel scores computed on the *development* set of WMCA.

The results for this sequence of experiments are introduced in Table 4, accumulating BPCER20, BPCER100 and corresponding APCER values on the *evaluation* set of WMCA. Additionally, DET curves are given in Figure 3. Both from table and DET, one can observe, that MC approach boosts the performance, and legacy RGB baseline operates poorly on the WMCA DB containing a wide range of challenging PAIs.

4.2. Results: proposed MC face PAD algorithm

In this subsection a proposed CNN-based BW-NIR-D face PAD system is evaluated under different training scenarios and parametrization. Training of the network incorporates both RGB data from CelebA, and BW-NIR-D data from WMCA databases. To be suitable for training the biometric recordings are first preprocessed, as discussed in Section 2, producing facial images of the size $128_{(pixels)} \times 128_{(pixels)} \times 3_{(chanels)}$ in both RGB and MC

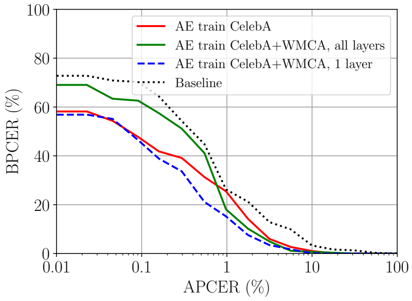

Figure 4. DET curves for PAD system using AE for the *entire* MC face, and MLP classifier. For the *evaluation* set of WMCA, **grandtest-10** protocol.

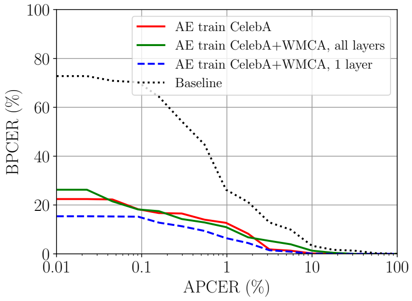

Figure 5. DET curves for PAD system using 9 AE for MC facial blocks, and MLP classifier. For the *evaluation* set of WMCA, **grandtest-10** protocol.

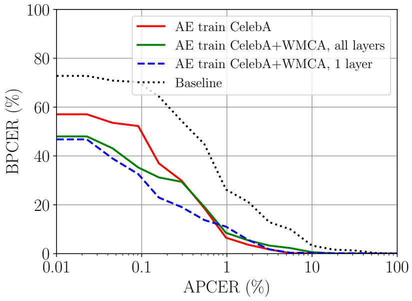

Figure 6. DET curves for PAD system using 16 AE for MC facial blocks, and MLP classifier. For the *evaluation* set of WMCA, **grandtest-10** protocol.

cases. The preprocessed RGB data from CelebA database undergoes additional quality assessment procedure, before being used for training. More specifically, an eye detection algorithm is applied to face images, assuring the deviation of eye coordinates from expected positions is not significant. The Haar-based eye detector from OpenCV is integrated for this purpose. This is done to exclude images with occlusions, *e.g.* sun-glasses, which can be considered as PAI, rather than bona-fide. The resulting amount of *bona-fide* RGB and MC training images is 42499, and 1240 respectively. An example of preprocessed/training data is displayed in Figure 1, denoted as RGB face and MC face.

Three training strategies are explored in current experiments: N autoencoders (AEs) are trained using only RGB CelebA facial regions (*no domain adaptation*); N AEs are trained using RGB CelebA facial regions, and fully fine-tuned using MC facial regions (RGB-to-MC domain adaptation, *full fine-tuning*); N AEs are trained using RGB CelebA facial regions, and only first and 2 last layers are fine-tuned using MC facial regions (RGB-to-MC domain adaptation, *partial fine-tuning*). As discussed in Section 2, training of the proposed network is associated to two components: a set of convolutional AEs, and an MLP. Here, training of an MLP remains the same, for all three strategies of autoencoders training. An MLP is trained using latent features (concatenation of latent vectors of all AEs) of both bona-fide and attack classes present in the training set of the WMCA database. In each experiment, an MLP is re-trained 10 times, initializing it differently. The best MLP model is selected cross-validating them on the *development* set of the WMCA. A BCE loss is used in both training and cross-validation of an MLP. AEs are trained with MSE loss in the unsupervised manner, running training and fine-tuning for 50 epoch each. The detailed parametrization of autoencoders and MLP is displayed in Figure 1.

Additionally, **three face regioning approaches** are observed for each training strategy discussed above. In the *first* sequence of experiments, an autoencoder is trained using entire face region, $N = 1$. *Second* and *third* approaches assume splitting of the facial region into $N = 9$, and $N = 16$ regions respectively, training an individual AE for each region. The dimensionality of RGB/MC facial blocks is $64_{(pixels)} \times 64_{(pixels)} \times 3_{(chanels)}$ for $N = 9$, and is $32_{(pixels)} \times 32_{(pixels)} \times 3_{(chanels)}$ for $N = 16$, with a patching stride of 32 pixels in both cases. The dimensionalities of latent feature spaces are 1296 for $N = 1$, 3600 for $N = 9$, and 2304 for $N = 16$.

The results for this sequence of experiments are introduced in Table 5, summarizing BPCER20, BPCER100 and corresponding APCER values on the *evaluation* set of WMCA. The DET curves are given in Figure 4 for $N = 1$, Figure 5 for $N = 9$, and Figure 6 for $N = 16$. From Table 5 one can see, that top BPCER20 and BPCER100 values are observed in the case of AEs training incorporating *partial fine-tuning*, and with *face regioning*. The same trend can be observed in the DET curves, with clearly **best performing PAD system** based on $N = 9$ autoencoders, which are pre-trained using RGB CelebA facial regions, followed by partial fine-tuning on MC data from WMCA, see Figure 5. Also, latent features learned combining bona-fide training

Channel	RGB	
Method	IQM+LR	9 AE trained on CelebA + MLP
BPCER20	77.7	**10.5**
APCER	13.2	17.3
BPCER100	94.6	**29.2**
APCER	8.5	7.8

Table 6. BPCER20, BPCER100, and corresponding APCER in %, using RGB channel only for *evaluation* set of WMCA, **grandtest-10** protocol. Thresholds are computed on *development* set.

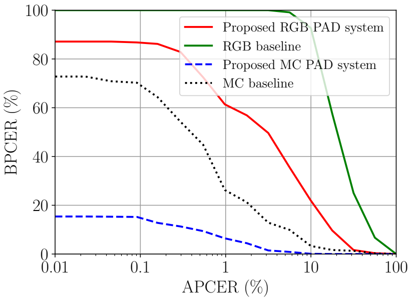

Figure 7. DET curves for proposed RGB PAD system (using 9 AE for RGB facial blocks, and MLP classifier) and RGB baseline, vs. proposed MC PAD approach and MC baseline. For the *evaluation* set of WMCA, **grandtest-10** protocol.

samples from RGB and MC domains, demonstrate the superior discriminative capacity, as opposed to the MC baselines using hand-crafted features, especially in the range of low APCER values, Figure 5.

Interestingly, that *full AEs fine-tuning* using MC data doesn't boost, or even degrades, the overall performance of the PAD system in all face regioning approaches, Figures 4 - 6. While the proposed *partial fine-tuning* has a stable positive impact, as opposed to the AEs trained using CelebA data only. It is worth mentioning, that we have also tested other RGB-to-MC domain adaptation strategies, fine-tuning more than just one layer of encoders, but it doesn't improve the performance further.

4.3. Results: proposed face PAD in RGB mode

In the final sequence of experiments, the introduced face PAD approach is tested in RGB mode, meaning that AEs are trained using RGB CelebA facial regions only, and an MLP is trained using latent encodings obtained from RGB channel of WMCA. No fine-tuning of AEs is involved. The input data in the experiments is RGB channel, present in the WMCA DB. Other parameters of the system, *e.g*. preprocessing and face regioning ($N = 9$), are the same as in the *best performing* system of subsection 4.2. Following the results summarized in Table 6 and Figure 7, offered RGB CNN-based system outperforms the IQM-based RGB baseline by a large margin. However, an auxiliary discriminative information introduced by MC approach seems to have a higher impact on the performance, rather than using advanced PAD methods in RGB domain only, see Figure 7.

5. Conclusion

Current paper suggests a multi-channel face anti-spoofing solution, being motivated by the need of high security PAD systems, capable of dealing with challenging attacks, such as 3D and partial PAIs. The introduced approach is using a consumer grade imaging sensors, capturing BW-NIR-D streams, in combination with a deep-learning based PAD algorithm. In the experimental section we demonstrate, that heavily studied RGB-based legacy systems are inefficient in detecting a wide range of challenging PAIs present in the WMCA database. In contrast, our proposed solution gives very promising results, especially in the range of low APCER values, which is critical for high security applications, see Figure 7.

Current paper introduces a number of novelties. *First*, a CNN-based MC face PAD algorithm, which is decomposed into a set of encoders, processing individual MC facial regions, and an MLP, categorizing latent encodings into real or attacks classes. *Second*, CNN decomposition allows us to introduce a special training procedure, transferring the knowledge of facial appearance from RGB to multi-channel domain. Domain adaptation is done via autoencoders, which are first pre-trained on a large set of RGB facial data from CelebA, and are then *partially fine-tuned* on the BW-NIR-D data from WMCA DB. Here partial fine-tuning means training of the first layers of encoders, and last 2 layers of decoders (for symmetry). The intuition behind this proposition, is that only low level features are domain dependent, while deeper features/layers of AEs are domain independent mostly preserving *structural information* of the face. *Full fine-tuning* of the AEs will try to learn this *structural information* from a training set of WMCA database, which has a significantly smaller number of *bona-fide* training samples, than CelebA DB, leading to over-fitting of the system. The effectiveness of the proposed RGB-to-MC domain adaptation, via partial fine-tuning, is proved experimentally. *Third*, we demonstrate, that learning the features of individual facial regions, is more discriminative than the features learned from an entire face.

To the best of our knowledge, this work is one of the first attempts applying deep-learning technologies, and domain adaptation, to the task of multi-channel face PAD, giving promising results and motivating to enhance the research in this direction.

Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the 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 thereon.

References

[1] *Information technology - Biometric presentation attack detection - Part 3: Testing and reporting*, 2017. 4

[2] S. R. Arashloo and J. Kittler. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 80–89, Oct 2017. 1

[3] S. Bhattacharjee and S. Marcel. What you can't see can help you - extended-range imaging for 3d-mask presentation attack detection. In *2017 International Conference of the Biometrics Special Interest Group*, pages 1–7, Sept 2017. 1

[4] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. In *Proceedings of BTAS2018*, Oct. 2018. 1

[5] A. Bhattad, J. Rock, and D. A. Forsyth. Detecting anomalous faces with 'no peeking' autoencoders. *CoRR*, abs/1802.05798, 2018. 2

[6] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. May 2017. 1, 2

[7] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. 2012. 2

[8] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *Proceedings of the International Conference on Biometrics Special Interests Group (BioSIG)*, Sept. 2016. 1

[9] T. de Freitas Pereira, A. Anjos, and S. Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, page 13, Feb. 2019. 4

[10] G. B. de Souza, J. P. Papa, and A. N. Marana. On the learning of deep local features for robust face spoofing detection. *CoRR*, abs/1806.07492, 2018. 2

[11] S. N. Garg, R. Vig, and S. Gupta. A survey on different levels of fusion in multimodal biometrics. *Indian Journal of Science and Technology*, 10(44), 2017. 3

[12] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security (under review)*, 2019. 4

[13] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, Dec 2016. 2

[14] Y. Liu*, A. Jourabloo*, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 2018. 1, 2

[15] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 389–398, 2018. 2

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3

[17] A. Mohammadi, S. Bhattacharjee, and S. Marcel. Deeply vulnerable – a study of the robustness of face recognition to presentation attacks. *IET (The Institution of Engineering and Technology) – Biometrics*, pages 1–13, 2017. Accepted on 29-Sept-2017. 1

[18] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *2018 International Conference on Biometrics (ICB)*, 2018. 1, 5

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*. 2

[20] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Extended multispectral face presentation attack detection: An approach based on fusing information from individual spectral bands. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–6, July 2017. 2

[21] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Comput. Surv.*, 50(1):8:1–8:37, Mar. 2017. 2

[22] M. O. Simn, C. Corneanu, K. Nasrollahi, O. Nikisins, S. Escalera, Y. Sun, H. Li, Z. Sun, T. B. Moeslund, and M. Greitans. Improved rgb-d-t based face recognition. *IET Biometrics*, 5(4):297–303, 2016. 1

[23] H. Steiner, S. Sporrer, A. Kolb, and N. Jung. Design of an active multispectral SWIR camera system for skin detection and face verification. *J. Sensors*, 2016:9682453:1–9682453:16, 2016. 1

[24] D. Wen, H. Han, and A. Jain. Face Spoof Detection with Image Distortion Analysis. *IEEE Trans. Information Forensic and Security*, 10(4):746–761, April 2015. 1

[25] F. Xiong and W. Abdalmageed. Unknown presentation attack detection with face rgb images. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2018. 2, 5

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 2016. 3

[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 26–31, March 2012. 2