Journal of Pattern Analysis & Applications – manuscript No.
(will be inserted by the editor)

Face Spoofing Detection via Ensemble of Classifiers towards Low-Power Devices

Rafael Henrique Vareto · William Robson Schwartz

Received: date / Accepted: date

Abstract Facial biometrics tend to be spontaneous, instinctive and less human-intrusive. It is regularly employed in the authentication of authorized users and personnel to protect data from violation attacks. A face spoofing attack usually comprises the illegal attempt to access valuable undisclosed information as a trespasser attempts to impersonate an individual holding desirable authentication clearance. In search of such violations, many investigators have devoted their efforts to studying either visual liveness detection or patterns generated during media recapture as predominant indicators to block spoofing violations. This work contemplates low-power devices through the aggregation of Fourier transforms, different classification methods, and handcrafted descriptors to estimate whether face samples correspond to falsification attacks. To the best of our knowledge, the proposed method consists of low-computational cost and is one of the few methods associating features derived from both spatial and frequency image domains. We conduct experiments on recent and well-known datasets under same and cross-database settings with Artificial Neural Networks, Support Vector Machines and Partial Least Squares ensembles. Results show that although our methodology is geared for resource-limited single-board computers, it can produce significant results, outperforming state-of-the-art approaches.

Keywords Face spoofing · Liveness detection · Fourier transform · Machine learning · Biometrics.

Rafael Henrique Vareto · William Robson Schwartz
Universidade Federal de Minas Gerais
Department of Computer Science
Smart Sense Laboratory
Belo Horizonte – Minas Gerais – Brazil
E-mail: {rafaelvareto,william}@dcc.ufmg.br
http://smartsenselab.dcc.ufmg.br/en/
Tel.: +55-31-3409-5854

1 Introduction

Biometrics is the science of automatically identifying individuals based on their physiological or behavioral characteristics, ranging from face and fingerprint to iris and voice. Despite the significant progress of biometric authentication techniques in the past years, experts declare that novel technologies are constantly exposed to malicious authentication attacks and can be susceptible to emerging high-quality fraudulent mechanisms [25].

The term *spoofing*, also known as copy and presentation attack, represents a serious threat to any biometric system. It eventuates when a criminal manipulates fraudulent data to circumvent the security procedure and gain unauthorized access. More precisely, the attack occurs when an interloper attempts to impersonate someone who carries a desirable authentication clearance. As a countermeasure to presentation attacks, several researchers dedicate their time and efforts inspecting patterns generated during media recapture as well as building new databases as an attempt to anticipate spoofing infringements and leverage upcoming investigations [8,12,21,23,26,36,2].

Human pictures can be effortlessly collected since a person's face is probably the most natural biometric model due to its nonintrusive and obtainable characteristics when compared to others, such as iris and fingerprint. With the propagation of surveillance cameras and the growing number of individuals distributing personal images/videos on social media and networks, it is practically impossible to keep track of a subject's face photos as they spread out [17]. The low-cost access to face images contributes to the increase of criminals designing presentation attacks to be validated as authentic users, turning face spoofing into a popular way of deceiving biometric applications.

The approach proposed herein is an extension of the work of Vareto *et al.* [34], which contemplates a low computational-cost algorithm based on Partial Least Squares (PLS) and Support Vector Machine (SVM) classification models, originally designed for limited-resource equipment, such as IoT devices. This study introduces the implementation of an ensemble of Multi-Layer Perceptron (MLP) networks, provides further details of the proposed algorithms and contains additional evaluations on different face spoofing databases. It also provides a cross-dataset evaluation in behalf of determining the ensemble's generalization capability under previously unexplored media types. Moreover, this paper presents an objective comparison with recent state-of-the-art works and explores how each learning algorithm performs on each benchmark.

The proposed spoofing detection approach associates an ensemble of classification algorithms with simple handcrafted features extracted from spatial and frequency domains. LBP [24] and HOG [9] descriptors extract spatial information from video frames whereas GLCM [15] obtains features derived from Fourier transforms. In addition, PLS [29], SVM [32] and MLP [13] classifiers act as bootstrap aggregating meta-algorithms to achieve competitive results on the five most prominent databases, MSU-MFSD [35], OULU-NPU [6] and SIW [20], to mention a few.

To the best of our knowledge, this is one of the first studies that associates features extracted from frequency and spatial domains towards the spoofing detection problem. The leading premise is that interpreting the relationship between spatial and frequency domains can be suitable to enhance the accuracy and robustness of face anti-spoofing applications. We assume that authentic and counterfeit biometric data enclose distinct noise signatures derived from the media acquisition. In fact, we believe that the combination of different feature descriptors contributes to achieving higher performance considering that they acquire distinctive characteristics, which can enrich the classifier's robustness and generalization potential.

The main contributions of this work are: *1)* combination of three different learning algorithms fitted on randomly generated subsets in a bootstrap aggregating mode; *2)* association of features extracted in spatial and temporal domains; *3)* efficient method for image and video-based copy attack receiving as input high-resolution videos; *4)* low complexity and computational cost algorithm, capable of being deployed in embedded systems and computers with small processing capabilities; *5)* clear study and experimental evaluation of the proposed approach considering fundamental feature descriptors, such as GLCM, HOG and LBP.

2 Related Works

There is a large number of works focusing on print and replay spoofing attacks. Many methods deal with the design of handcrafted descriptors and learning algorithms as others focus on the neural networks trend.

Many researchers have worked with handcrafted feature extraction and learning design: Pinto *et al.* [26] extract low-level feature descriptors gathering temporal and spectral information across biometric samples. Wen *et al.* [35] developed an algorithm based on the analysis of image distortion and low-level feature descriptors. The method consists of an ensemble of SVM classification algorithms evaluated on cross-dataset scenarios. Pinto *et al.* [27] investigate the spatial domain during the recapture process as it takes over the noise with Fourier transforms accompanied by visual rhythm algorithms and the extraction of gray-level co-occurrence matrices. Using color texture analysis and low-level descriptors, Boulkenafet *et al.* [4,5] detect copy attacks through the examination of luminance and chrominance information of each image color channel separately.

In recent years, Deep Neural Networks (DNN) have confirmed to be effective in a myriad of computer vision and biometric problems. Li *et al.* [18] use a hierarchical neural network with multiple inputs incorporating either *shearlet* or optical-flow-based features. Similarly, Feng *et al.* [11] extract deep features from a convolutional neural network to identify real and fake faces. Liu *et al.* [20] combine DNN and Recurrent Neural Networks (RNN) to estimate the depth of face images along with rPPG signals to boost the detection of unauthorized access. Valle *et al.* [33] present a transfer learning method using a pre-trained DNN model on static features to recognize photo, video and mask attacks.

Handcrafted features are usually faster and present lower memory usage than methods based on deep neural networks, especially when it comes to resource-limited equipment. Still, they are susceptible of being restricted to particular datasets domains. Most neural networks are not invariant to image rotation or scale, and may struggle to handle scenarios consisting of various capture devices, lighting conditions and shooting angles [3]. In addition, top performing DNNs tend to suffer from either low speed or being too large to fit into single-board computers, preventing their deployment on remote applications. On the contrary of deep neural networks, the conventional descriptors as well as straightforward classifiers employed in our approach do not require cloud processing services or powerful dedicated servers since embedded devices are capable of running the proposed low-cost standalone algorithm fast enough to be employed in real environments.

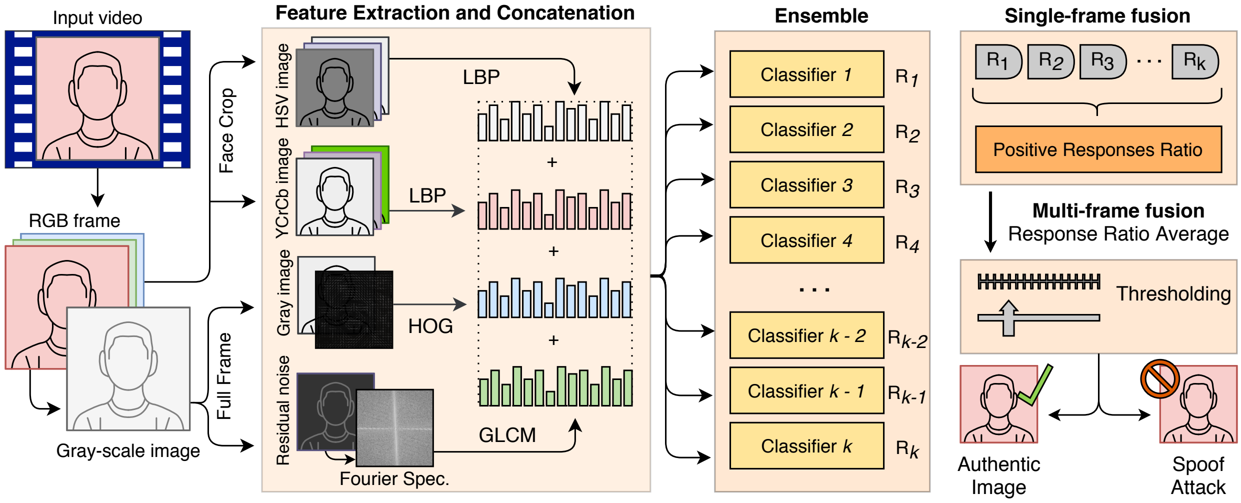

Fig. 1 Description of the suggested solution to face spoofing detection – *Training:* GLCM, HOG and LBP descriptors are extracted from the frames of the videos available for training. Such features are concatenated and used in an ensemble fashion to learn multiple classification models. Distinct models are learned containing different video samples in each subset. *Test:* The same features are extracted from the probe video frames and projected onto all binary classifiers. The procedure then performs a score fusion on the answers of the classifiers to determine if the probe video refers to an authentic presentation.

Along with the latest anti-spoofing methods, many databases have been designed in the last decade [37, 8,35]. Boulkenafet *et al.* [6] created one of the largest mobile-based benchmarks whereas Liu *et al.* [20] introduced a dataset covering a large range of expression, illumination and pose variations. A few masks-based attacks have also been proposed in the past 4 years [22,19, 2]. Despite the variety of benchmarks, many literature methods end up being restricted to specific datasets domains, especially when cameras do not have comparable capture quality. Therefore, there is room for improvement when it comes to achieving good results on cross-dataset evaluations.

3 Proposed Approach

This section describes an approach that captures visual noise signatures in both spatial and frequency domains. The method exploits GLCM [15], HOG [9] and LBP [24] descriptors to obtain low-level features. Then, an ensemble of classifiers is created as we group several identical classifiers to reinforce the method's overall efficacy [7]. Figure 1 illustrates the steps that compose the proposed approach, depicting the extraction of features, the learning of multiple classifiers and the aggregation of their response values to provide the final verdict.

Diversified feature descriptors make it possible to combine color, gradient magnitude and texture information. GLCM plays an important role examining artifacts added to biometric face samples during the recapturing process performed by the acquisition sensor [27]. HOG captures the structural shape of faces along with objects resembling still image frames and mobile borders, that is, regions of abrupt intensity changes around edges and corners [30]. LBP is capable of obtaining luminance and chrominance information, which are useful for telling real faces from fake ones, especially when applied to different color channels [4]. In summary, the proposed set of features combines multiple characteristics to distinguish between live and spoof media.

3.1 Extraction of Feature Descriptors

The process of extracting features inspects separate spatial colorspaces and the frequency domain in pursuance of discriminating spoofing patterns. The procedure starts converting every RGB colorspace video frame into HSV, YC_RC_B and gray-scale images. Contrastingly to the RGB color model, which holds high correlation between color components, HSV and YC_RC_B can isolate luminance from chrominance and are more robust to variations in illumination [28].

As the RGB video frame is transformed into HSV and YC_RC_B images, the method spots the area of interest, which is comprised by the subject's face. The approach extracts LBP descriptors from each HSV and YC_RC_B image color channel in an attempt to gather distinctive knowledge about color and texture. As a matter of fact, LBP computes local texture representation from all color bands comparing every pixel with its surrounding neighborhood of pixels. Both corresponding HSV and YC_RC_B feature descriptors derive from the integration of each channel's histogram, which accounts for the number of times every LBP pattern occurs [5].

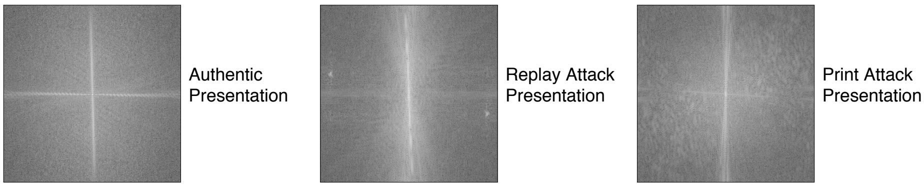

Fig. 2 Comparison among Fourier spectra extracted from different presentation images. Note that there are some artifacts spread throughout print and replay attacks.

Monochromatic video frames go through low-pass filtering (blurring) techniques for erasing artifacts, reducing noise and removing *Moire* patterns, resulting in images with lessened sharp transitions. Residual noises are acquired with the difference between a gray-scale image and its slightly blurred version [27]. A logarithmic-normalized Fourier transform function $\mathcal{F}_{log}(v,u)$ dissects each residual image $r(a,b)$ of size $M \times N$ into its sine and cosine constituents in which each pixel compounds a frequency from the spatial domain as

$$\mathcal{F}_{log}(v,u) = log(1 + |\sum_{a=0}^{M-1}\sum_{b=0}^{N-1} r(a,b)e^{-j2\pi[\frac{va}{M}+\frac{ub}{N}]}|).$$

Due to fast computation, the employed low-level feature descriptors provide great accuracy vs. speed trade-off. HOG and GLCM descriptors scan gray-scale images and their corresponding spectra, respectively, whereas LBP descriptor takes in HSV and $\mathrm{YC_RC_B}$ image color bands. GLCM calculates the residual image texture by producing co-occurring gray-scale values at a given offset while HOG provides shape information by counting gradient orientation occurrences using histograms. Figure 1 illustrates the steps required to build a robust feature descriptor by concatenating HOG and LBP features from the spatial-domain with GLCM information from the *log*-scaled Fourier spectrum.

3.2 Ensemble of Learning Algorithms

Instead of learning a single binary classifier, we learn a set of models as it reduces the risk of overfitting and seems to be more appropriate to handle contrasting chromatic distortions. The classification ensemble consists of collections of Multi-Layer Perceptrons [13], Support Vector Machines [32] or Partial Least Squares [29] learning algorithms.

Multi-Layer Perceptron (MLP) is a feed-forward artificial neural network characterized by several layers of input nodes connected as a directed graph between input and output layers. Besides being considered a universal approximator, shallow networks contain reduced number of parameters, analogous to PLS and SVM, well-suited for ensemble of classifiers [16,10].

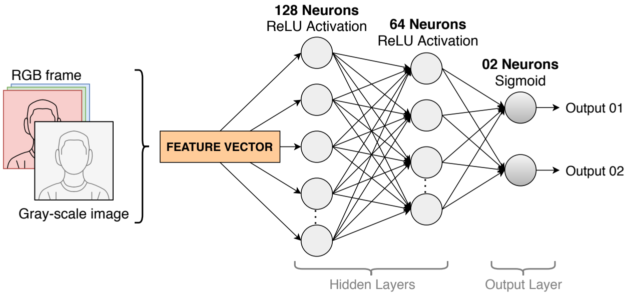

Fig. 3 Illustration of the proposed multi-layer perceptron classifier with three layers. Multiple MLP-based learning algorithms are trained and wrapped up in the bagging structure.

We propose a small network architecture with two hidden layers and an output layer. As depicted in Figure 3, each hidden layer is composed of neurons holding a nonlinear activation function (ReLU) that is connected to every neuron in the subsequent layer. The output layer consists of a sigmoid function, which is commonly employed in two-class logistic regression implementations. The network searches for a function f that relates observable variables $x \in X$ to output variables $y \in Y$ so that it satisfies $Y = f(X)$. The function f is optimized during training time in such a way that the network output for the observations in X is as close as possible to the target values in Y.

Support Vector Machine (SVM) searches for optimal separating hyperplanes among classes as it attempts to maximize the margins between classes' closest points. The adopted linear SVM considers a training set (X,Y) of n points of the form (x_1,y_1), $\cdots$, (x_n,y_n) where observation x_i is a p-dimensional feature vector, and y_i is either the -1 or $+1$ label that points out the class associated with observation x_i. The maximum-margin hyperplane is usually noted as a set of points $x \in X$ satisfying the equation $\mathbf{w} \cdot x - b = 0$, where $\mathbf{w}$ is the normal vector to the hyperplane. The distance between the separating hyperplane from the positive and negative support vectors is modeled by the equation $\frac{|\mathbf{w}\cdot x - b = 0|}{\|\mathbf{w}\|} = \frac{\pm 1}{\|\mathbf{w}\|}$. To that end, $\frac{2}{\|\mathbf{w}\|}$ is the total distance between the support vectors in such a way that maximizing the distance within the support vector hyperplanes is equivalent to minimizing $\|\mathbf{w}\|$.

Partial Least Squares (PLS) is a fast and effective regression technique based on covariance and dimensionality reduction. PLS usually works well when the number of explanatory variables is both high and likely to be correlated. Besides, it is robust to unbalanced classes and supports high-dimensional feature vectors.

PLS creates latent variables as a linear combination of the independent zero-mean variables X and Y. It searches for latent vectors that can be simultaneously decomposed into equations $X = TP^T + E$ and $Y = UQ^T + F$ in order to identify the maximum covariance between these variables. Variables E and F indicate residuals. Matrix $T_{n \times p}$ portrays latent variables from feature vectors and matrix $U_{n \times p}$ denotes latent variables from target values. Variables $P_{p \times d}$ and $Q_{1 \times d}$ can be compared to the loading matrices from principal component analysis. We select NIPALS algorithm [29] to compute the maximal covariance between latent variables T and U. NIPALS outputs a matrix of weight vectors $W_{d \times p}$ and estimates the regression coefficients vector β using least squares satisfying the following equation: $\beta = W(P^T W)^{-1} T^T Y$. The PLS regression response for an video frame's feature vector x is given by $\hat{y} = \bar{y} + \beta^T(x - \bar{x})$ where $\bar{y}$ is the sample mean of Y and $\bar{x}$ the average values of X.

3.3 Execution Pipeline

Figure 1 shows that video samples go through a feature extraction and concatenation procedure before being presented to the ensemble of classifiers C. This process takes place during training and testing phases and, therefore, are carefully detailed below.

Learning stage. During the training phase, the matrix of observable variables X (obtained from the feature generation step) and its corresponding vector of target values Y are set apart to feed the array of classifiers. The proposed approach produces several binary learning algorithms, fitted on random subsets of the feature training set (X, Y) to create the ensemble C. The positive class only contains authentic feature vectors whereas the negative class holds features extracted from copy attacks. In favor of preventing instability issues that may arise when dealing with unbalanced training data, the method guarantees a balanced division within each classification model since v genuine live and v presentation attack videos are randomly selected, with replacement, out of all video samples available for training. This process is repeated k times, where $k = |C|$ is a user-defined parameter that defines the number of classification models in the ensemble as well as the total of authentic and counterfeit sampled subsets.

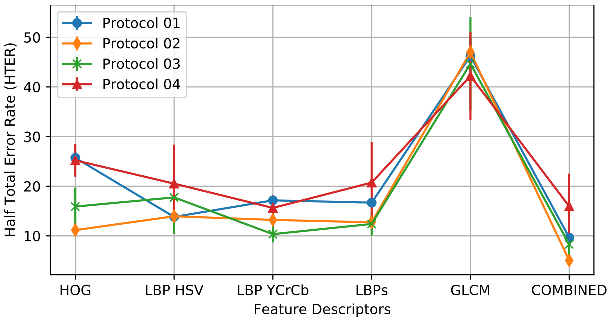

Fig. 4 Evaluation on all OULU-NPU protocols comprising standalone feature descriptors and their combination.

Prediction stage. Given a probe video V, consisting of a collection of frames, the proposed method loops through its image frames and extracts an equivalent set of visual descriptors. The approach sets up a feature vector for each probe video's frame and projects them onto all classification models. For each frame, the algorithm computes the ratio of the number of positive responses attained to the total number of classification models k. If most $c \in C$ classifiers return positive responses, it implies that the frame is likely to be a *bona fide* (authentic) sample. Otherwise, if they return negative responses, then the probe sample is likely to characterize a spoofing attack. As the approach examines all probe video frames, it keeps record of each frame's positive response ratio. Thereafter, it performs the multi-frame fusion through the computation of the numerical mean of all positive ratio scores for video V. A probe video is considered authentic if the averaged ratio score of all frames satisfies a threshold t. The appropriate value for t tends to be chosen according to a scenario's specifications and requirements.

4 Experimental Results

This section contains an objective evaluation of the proposed algorithm, namely, experiments regarding best parameters selection and literature comparison.

Evaluation Setup. We conduct experiments on Raspberry Pi, Nvidia Jetson Nano and on a Linux virtual machine to assess the performance of the proposed approach on different computer architectures. First, we analyzed the method on a CPU-based machine consisting of eight 2.0 GHZ-core processors and 16 GB RAM memory, but no more than 800 MB was required on test time. Then, we migrated to Raspberry Pi and Jetson devices, single-board microcomputers with a minimum specification of 1.2 GHZ Quad Core CPU and 1 GB RAM memory. Graphical processing units (GPU) can achieve higher frame rates; however, it would demand the acquisition of more advanced computer hardware.

Approach		Metric	50	100	200	300
Protocol 01	MLP	APCER	0.09 ± 0.17	0.08 ± 0.10	0.08 ± 0.17	0.11 ± 0.19
		BPCER	3.20 ± 3.91	2.22 ± 1.97	2.03 ± 1.33	1.89 ± 1.36
	PLS	APCER	0.67 ± 0.48	0.10 ± 0.13	0.78 ± 0.69	0.59 ± 0.46
		BPCER	2.91 ± 3.72	2.34 ± 1.56	1.51 ± 1.44	0.75 ± 0.50
	SVM	APCER	0.45 ± 0.29	0.07 ± 0.08	0.52 ± 0.42	0.39 ± 0.28
		BPCER	2.06 ± 2.33	1.65 ± 0.98	1.07 ± 0.90	0.53 ± 0.32
Protocol 02	MLP	APCER	3.00 ± 5.20	3.73 ± 4.75	0.00 ± 0.00	1.35 ± 1.54
		BPCER	6.68 ± 4.89	1.11 ± 0.75	4.33 ± 2.69	2.30 ± 1.36
	PLS	APCER	12.96 ± 11.06	9.45 ± 8.55	6.59 ± 6.96	6.57 ± 6.55
		BPCER	0.88 ± 0.84	1.31 ± 0.57	1.41 ± 1.17	1.16 ± 0.61
	SVM	APCER	13.83 ± 11.19	12.60 ± 11.41	8.79 ± 9.28	8.76 ± 1.39
		BPCER	0.97 ± 0.83	1.75 ± 0.76	1.88 ± 1.56	1.54 ± 0.81
Protocol 03	MLP	APCER	8.22 ± 6.24	7.02 ± 2.95	6.29 ± 0.70	1.77 ± 0.06
		BPCER	2.17 ± 1.28	0.97 ± 0.97	0.89 ± 0.84	2.44 ± 1.05
	PLS	APCER	24.09 ± 15.97	17.95 ± 12.45	11.21 ± 1.66	9.14 ± 2.66
		BPCER	1.01 ± 0.49	1.37 ± 0.50	1.62 ± 0.11	2.04 ± 0.84
	SVM	APCER	19.27 ± 13.63	14.36 ± 10.63	8.97 ± 1.41	7.31 ± 2.39
		BPCER	0.89 ± 0.45	1.21 ± 0.46	1.42 ± 0.10	1.80 ± 0.77

Table 1 Evaluation of all proposed approaches on different SIW protocols with an increasing number of classification models. Note that the method becomes more discriminative with the addition of classifiers.

Feature Descriptors. Three feature descriptors are employed in this work: GLCM [15], HOG [9] and LBP [24]. Each one of them searches for specific patterns and unique characteristics. The GLCM texture descriptor is computed with four directions $\theta \in \{0, 45, 90, 135\}$ degrees, 16 bins, a distance $d \in \{1, 2\}$, and six different texture properties: contrast, dissimilarity, homogeneity, energy, correlation, and angular second moment. The HOG shape descriptor is set with 96×96 cells and holding eight orientations. Lastly, the LBP texture descriptor comprises 256 bins, a radius equal to 1, and eight points arranged in a 3×3 matrix thresholded by its central point. Their low complexity and computational cost endorse our method so that it can be deployed to embedded systems.

Spoofing Datasets. For a complete experimental assessment, we adopt datasets with different evaluation protocols, medium characteristics and variable lighting conditions. Experiments are conducted on five benchmarks: CASIA-FASD [37], MSU-MFSD [35], OULU-NPU [6], REPLAY-ATTACK [8] and SIW [20]. Both OULU-NPU and SIW have been released in recent years and contain full high-definition videos of multi-ethnic individuals and featuring 30-FPS live and presentation attack videos.

CASIA-FASD, MSU-MFSD and REPLAY-ATTACK are traditional benchmark databases made up of genuine live recordings and distinct spoofing attack shots captured by distinct cameras in different scenarios. CASIA-FASD contains 50 individuals, and their analogous counterfeit faces are derived from the original ones. It consists of low, medium and high-quality images and three presentation attacks, including warped and trimmed photos, and video attacks. MSU-MFSD consists of 280 video clips of photo and video attack attempts to 35 persons, in which authentic and falsified face images were captured using smartphone, tablet and laptop cameras. REPLAY-ATTACK is composed of 50 individuals and 1,300 short recordings of photo and video attack attempts. Copy attacks comprise three different scenarios under controlled and adverse illumination conditions: good-quality videos, high-resolution photographs, and still and motion pictures taken with mobile devices.

OULU-NPU encompasses short videos of both real-access and attack attempts of 15 women and 40 men. 5,940 videos were recorded in three different illumination conditions using high-quality frontal cameras of six different mobile devices. The authors came up with four protocols that take into account unseen conditions not known during the training stage. The first protocol analyzes spoofing detection methods under previously unseen illumination and background scenes. The second one evaluates the effect of attacks created with different printers or displays mediums. Third protocol studies the influence input camera variations have on biometric systems whereas the fourth protocol simulates a scenario where all previous three variations are considered at the same time.

Spoof in the Wild (SIW) database contains 4,620 live and spoof videos from 165 subjects. The live videos are collected in four sessions with variations of distance, pose, illumination and expression. Spoofing recordings consist of printed paper and replay video attacks. The creators define three different protocols to study unusual attack properties: The first protocol evaluates anti-spoofing applications under different poses and expressions, where training video samples are restricted to their first 60 frames. Second protocol examines a system's generalization capability on the different types of replay attacks following a leave-one-out strategy. Protocol number three analyzes the performance on unknown presentation attacks, considering cross testing from print to replay attack and vice-versa.

Protocol	Method	APCER	BPCER	AVERAGE
1	Deep models [20]	3.58	3.58	3.58
	MLP approach	**0.08 ± 0.17**	2.03 ± 1.33	1.05 ± 0.75
	PLS approach	0.78 ± 0.69	1.51 ± 1.44	1.14 ± 1.06
	SVM approach	0.52 ± 0.42	**1.07 ± 0.90**	**0.79 ± 0.66**
2	Deep models [20]	0.57 ± 0.69	**0.57 ± 0.69**	**0.57 ± 0.69**
	MLP approach	**0.00 ± 0.00**	4.33 ± 2.69	2.16 ± 1.34
	PLS approach	6.59 ± 6.96	1.41 ± 1.17	4.01 ± 4.06
	SVM approach	8.79 ± 9.28	1.88 ± 1.56	5.33 ± 5.42
3	Deep models [20]	8.31 ± 3.81	8.31 ± 3.80	8.31 ± 3.80
	MLP approach	**6.29 ± 0.70**	**0.89 ± 0.84**	**3.59 ± 0.77**
	PLS approach	11.21 ± 1.66	1.62 ± 0.11	6.41 ± 0.88
	SVM approach	8.97 ± 1.41	1.42 ± 0.10	5.19 ± 0.75

Table 2 Literature comparison, presenting APCER and BPCER results (%) evaluated on the three protocols of SIW dataset.

Evaluation Metrics. For a consistent comparability among spoofing methods, it customary to engage the ISO/IEC 30107-3 metrics [1] denominated Bona Fide Presentation Classification Error Rate (BPCER) and Attack Presentation Classification Error Rate (APCER). The former, APCER, can be understood as the proportion of attacks under the same attack medium incorrectly classified as trustworthy. Contrarily, BPCER is defined as the fraction of genuine presentations incorrectly labeled as attack. Both metrics are described in the following equations:

$$\text{BPCER} = \frac{1}{V_{BF}} \sum_{i=1}^{V_{BF}} (Res_i)$$

$$\text{APCER} = \frac{1}{V_{PA}} \sum_{i=1}^{V_{PA}} (1 - Res_i)$$

Note that V_{PA} stands for the number of spoofing attacks whereas V_{BF} expresses the total number of authentic medium presentations. Res_i receives the value 1 when the i-th probe video presentation is categorized as an attack and 0 if labeled as *bona fide* presentation.

On cross-datasets evaluations, it is customary to employ Half Total Error Rate, $\text{HTER} = \frac{\text{FAR}+\text{FRR}}{2}$, which is half the sum of the False Rejection Rate (FRR) and the False Acceptance Rate (FAR) [20,31]. The reader must bear in mind that the closer APCER, BPCER and HTER values get to zero, the more accurate the described methods are. APCER and BPCER resemble False Acceptance and False Rejection Rates, traditionally employed in the literature when assessing binary classification methods. Nevertheless, APCER is estimated separately for each attack medium, such as print or replay, and the definite performance is made up of the highest APCER score – indicating the worst-case spoofing attack scenario.

Ablation Study. Figure 4 presents the proposed PLS approach's performance on all protocols of OULU-NPU dataset as we consider each one of the three adopted feature descriptors and their combination. This experiment aims at identifying the descriptors that best contribute to the generation of a robust set of features. With the ensemble size set to 100, the combination of HOG, LBP and GLCM is responsible for achieving at least an equivalent performance when compared to the best standalone feature descriptor in each protocol. Note that the benefits brought by each feature descriptor vary under different protocols, not being easy to choose the dominant descriptor. As a result, we employ the combination of all these feature descriptors in the experiments regarding literature comparison.

In the first parameter analysis, we conduct several experiments on SIW dataset to check how MLP, PLS and SVM-based methods respond to the addition of classification models within the ensemble. More precisely, this experiment considers homogeneous regression algorithms running in parallel, which are independent from each other and their outputs are combined in a deterministic averaging process. We search for the ensemble size k that provides the highest accuracy rate without compromising our concern for resource-limited single-board devices. Larger ensemble sizes are likely to require extra RAM memory to store additional working data as well as take more computational time to train and make predictions.

According to the results showed in Table 1, as the number of classifiers augments, the three designed approach become more discriminative. It can be explained by the fact that when the size of the ensemble increases, more prediction responses are taken into consideration in the positive response ratio step (majority voting). Using straightforward machine learning algorithms as building blocks for designing more complex and robust methods seem to have a positive impact on the method's performance. As we check the results, we notice a considerable performance gain when the ensemble size rises from 50 to 200, but the growth is not maintained when raised to 300 models. Therefore, we set the number of classification models in the ensemble to 200 in all remaining experiments.

Protocol	Method	APCER	BPCER	AVERAGE
1	Deep models [20]	1.60	**1.60**	**1.60**
	Gradiant [3]	**1.30**	12.50	6.90
	MLP approach	8.14 ± 1.53	13.3 ± 3.43	10.7 ± 2.48
	PLS approach	5.50 ± 2.11	9.79 ± 3.37	7.64 ± 2.74
	SVM approach	8.75 ± 3.76	16.3 ± 6.91	12.5 ± 5.33
2	Deep models [20]	2.70	2.70	**2.70**
	Gradiant [3]	6.90	**2.50**	4.70
	MLP approach	4.67 ± 0.74	4.61 ± 1.43	4.64 ± 1.08
	PLS approach	**2.13 ± 1.07**	3.61 ± 1.21	2.87 ± 1.14
	SVM approach	3.61 ± 1.54	13.3 ± 6.54	8.45 ± 4.04
3	Deep models [20]	2.70 ± 1.30	**3.10 ± 1.70**	**2.90 ± 1.50**
	Gradiant [3]	2.60 ± 3.90	5.00 ± 5.30	3.80 ± 2.40
	MLP approach	**1.33 ± 0.71**	5.81 ± 5.33	3.33 ± 3.26
	PLS approach	3.12 ± 2.58	8.51 ± 6.20	5.81 ± 4.39
	SVM approach	8.33 ± 7.00	11.4 ± 6.17	9.89 ± 6.59
4	Deep models [20]	9.31 ± 5.60	10.4 ± 6.00	**9.50 ± 6.00**
	Gradiant [3]	**5.00 ± 4.50**	15.0 ± 7.10	10.0 ± 5.01
	MLP approach	8.58 ± 5.79	19.2 ± 8.97	13.8 ± 7.38
	PLS approach	17.8 ± 9.83	**9.37 ± 4.31**	13.5 ± 7.07
	SVM approach	23.3 ± 13.4	10.8 ± 4.82	17.0 ± 9.11

Table 3 Literature comparison, presenting APCER and BPCER results (%) evaluated on the four protocols of OULU-NPU dataset.

Results Analysis. The evaluation of the methods proposed in Section 3 are assessed in consonance with the protocols available in the literature and following each dataset's guidelines. For those databases containing only training and test sets, like SIW, we set aside ten percent of all samples available in the training set for validation. Distinctively, OULU-NPU and REPLAY-ATTACK contain validation sets designed for parameter calibrations. We employ these validation sets to perform an automatic selection of threshold t through F1-SCORE [14]. In this work, F1-SCORE computes the harmonic mean of a test's accuracy and punishes extreme low values as it searches for the value t that optimizes the fusion of *precision* and *recall*.

Tables 2 and 3 present the results obtained on SIW and OULU-NPU datasets, respectively, along with other literature works. The proposed approach achieves state-of-the-art results on SIW's Protocols 1 and 3 and competitive results on Protocol 2. The ensemble holding MLP learning algorithms outperforms the ones with PLS and SVM classifiers. However, this hegemony does not repeat in most protocols of the OULU-NPU dataset. It seems that the adopted descriptors could not obtain features robust enough to provide a dominant generalization capability under the different illumination conditions and background scenes available.

A cross-database investigation gives an insight into the generalization power of countermeasure algorithms. As a result, Table 4 demonstrates the cross-testing HTER [1] performance metric for MLP, PLS, SVM approaches and other literature methods on long-established benchmarks. In this kind of evaluation scenario, an approach is trained and tuned in one among the available datasets and examined at the others. PLS-based method outperformed both MLP and SVM ones as it achieves a HTER of 34.44 ± 3.91 when trained on SIW and tested on OULU-NPU, and 17.55 ± 1.47 in the opposite way. Most datasets consistently carry some bias irrespective of their protocols due to the inherent and contextual information enclosed in their image and video data. Therefore, the combination of multiform data tends to culminate in a significant accuracy reduction in comparison to same-database evaluations.

Computational Cost Evaluation. The proposed approach is devised towards resource-limited single-board computers in order to reduce network communication. It contradicts most recent spoofing detection algorithms in the literature, in which deep neural networks routinely benefit from long training hours, "unlimited computational resources" and high-bandwidth video transmissions. GLCM, HOG and LBP feature descriptors seem to contain significant forensic signature information of image and video-based spoofing detection since results indicate that the association of spatial and frequency-based descriptors contributes to achieving both competitive and state-of-the art results.

Most researchers have neglected to work out biometric applications that can operate on low-power devices [11,18,33,20]. In fact, many commercial biometric systems require powerful dedicated servers or even cloud processing services. Table 5 demonstrates the methods' real-time frame frequency on multiple platforms. Note that the proposed algorithm is able to process up to 4.84 ± 0.07 frames per second (FPS) when considering the PLS-based ensemble and a Raspberry Pi 4 environment. For contrasting purposes, it executes at 26.02±0.58 FPS in a conventional computer enclosing a microprocessing unit (CPU), reached when the number of classifiers k is set to 100. Such frame rate, 4.842 FPS, enables tech developers to implement and run biometric IoT technologies in realistic environments.

Training Set	CASIA-FASD		MSU-MFSD		REPLAY-ATTACK	
Test Set	MSU-MFSD	REPLAY-ATTACK	CASIA-FASD	REPLAY-ATTACK	CASIA-FASD	MSU-MFSD
Color LBP [4]	36.6	47.0	49.6	42.0	39.6	35.2
Color Texture [5]	20.4	30.3	46.0	**33.9**	37.7	**34.1**
Spectral [26]	–	34.4	–	–	50.0	–
Deep Models [20]	–	**27.6**	–	–	**28.4**	–
MLP approach	**15.2 ± 1.7**	29.3 ± 2.8	34.6 ± 3.4	41.8 ± 2.8	41.3 ± 1.2	36.9 ± 1.6
PLS approach	19.2 ± 1.6	30.1 ± 0.7	**28.2 ± 2.7**	37.1 ± 3.2	35.6 ± 0.4	34.5 ± 2.3
SVM approach	17.3 ± 1.1	42.6 ± 2.5	34.8 ± 2.8	42.6 ± 1.7	38.3 ± 2.0	35.4 ± 1.9

Table 4 Cross-dataset evaluation (%) presenting HTER metric on CASIA-FASD, MSU-MFSD and REPLAY-ATTACK datasets.

Platform	**Method**		
	MLP approach	PLS approach	SVM approach
RaspberryPi 3	1.84 ± 0.44	**2.12 ± 0.04**	1.96 ± 0.07
RaspberryPi 4	4.31 ± 1.05	**4.84 ± 0.07**	4.69 ± 0.06
Jetson Nano	7.73 ± 2.12	**8.64 ± 0.14**	8.23 ± 0.14
Linux Machine	14.77 ± 2.01	**26.02 ± 0.58**	24.06 ± 1.04

Table 5 Real-time frame frequency performance presenting FPS metric on OULU-NPU dataset as executed on four different platforms.

Table 5 consists of results regarding a realistic evaluation on OULU-NPU dataset videos. To represent a surveillance scenario, the analysis follows *qHD* convention, the standard resolution for mobile devices and analog CCTV systems. All dataset videos are resized to 960×540 pixels, but keeping their original aspect ratio. Last, we provide the average price paid per FPS on the following microcomputer hardware[1]:

1. Raspberry Pi 3 Model B ($35.00), probably the cheapest model on sale;
2. Raspberry Pi 4 Model B ($70.00), including a faster processor than its predecessor;
3. Nvidia Jetson Nano ($100.00), designed for accelerating machine learning applications;
4. Intel i5 2.8 GHZ processor with 16 GB RAM ($450.00), similar to the Linux virtual machine tested;
5. Intel i7 3.2 GHZ CPU with 16 GB RAM and a GeForce GTX 1080Ti ($1600.00).

The first three specified microcomputers attained different frame rates and are identical to devices we have evaluated. The fourth specification is comparable to the Linux virtual machine most of the experiments have been conducted on. As a consequence, we assume a frame rate of 26.02 for both computers chiefly because most quality CCTV cameras record videos between 15 and 30 FPS. As a result, the lowest price paid per FPS on the aforementioned machines would be approximately $16.50, $14.46, $11.57, $17.29 and $61.49 (PLS approach), respectively. For this reason, running the designed approach on a single-board computer, such as Nvidia Jetson Nano, provides better performance per cost than executing in more robust machines.

Recommended Scenarios. After conducting a series of thorough evaluations, the MLP-based approach has obtained the most efficient results. It proved to be more robust on most literature benchmarks as it attained the lowest presentation error rates. On the other hand, PLS and SVM approaches accomplish the best performance in terms of computational cost, capable of processing more frames per second. The three proposed approaches hand over a tradeoff between higher frame rates and accuracy. The method consisting of an ensemble of MLP classifiers is recommended to those scenarios in which spoofing detection demands higher precision and recall whereas PLS/SVM ensembles are indicated when it is possible to hand over a little bit of accuracy in exchange for a reduced computational cost per frame.

5 Conclusions

This study details a simple low-memory detection algorithm and demonstrates how it performs to emulate real-world scenarios in an experimental setup. The proposed method proved to be fast, working well on single-board computers and handling high-resolution video recordings. Favorably, it was able to achieve state-of-the-art performance on widely explored databases.

An objective investigation showed how promising spatial and frequency-based descriptors can be when combined with an array of learning algorithms. We work out three approaches (i.e., ensembles comprised of Multi-Layer Perceptrons, Partial Least Squares or Support Vector Machines) to conclude that the combination of long-established feature descriptors accomplishes impressive performance in same-database settings. Experiments trained and tested on different datasets show that the accuracy tends to degrade significantly, mainly due to their inherent bias.

Regardless of the expressive advances in numerous areas of biometric science, current presentation attack detection approaches have shown lack of generalization in cross-dataset settings, which best represents real-world scenarios. In the next steps, we plan to add extra feature descriptors, include other relevant spoofing datasets and learn spatial-temporal representations.

[1] Prices taken from BestBuy Retail store, and official Raspberry Pi and Nvidia resellers in July 2020.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development – CNPq (Grants 438629/2018-3 and 309953/2019-7), the Minas Gerais Research Foundation – FAPEMIG (Grants APQ-00567-14 and PPM-00540-17), and the Coordination for the Improvement of Higher Education Personnel – CAPES (DeepEyes Project).

References

1. Information technology – biometric presentation attack detection – part 1: Framework. international organization for standardization. Tech. rep., ISO/IEC JTC 1/SC 37 Biometrics (2016)
2. Bhattacharjee, S., Mohammadi, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. In: Conference on Biometrics: Theory Applications and Systems. IEEE (2018)
3. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S., Ouafi, A., et al.: A competition on generalized software-based face presentation attack detection in mobile scenarios. In: International Joint Conference on Biometrics, pp. 688–696. IEEE (2017)
4. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: International Conference on Image Processing, pp. 2636–2640. IEEE (2015)
5. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. Transactions on Information Forensics and Security **11**(8), 1818–1830 (2016)
6. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: Oulu-npu: A mobile face presentation attack database with real-world variations. In: International Conference on Automatic Face & Gesture Recognition. IEEE (2017)
7. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
8. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: International Conference of Biometrics Special Interest Group, EPFL-CONF-192369 (2012)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886–893. IEEE (2005)
10. Efitorov, A., Dolenko, S., Dolenko, T., Laptinskiy, K., Burikov, S.: Use of wavelet neural networks to solve inverse problems in spectroscopy of multi-component solutions. In: International Conference on Neuroinformatics, pp. 285–294. Springer (2019)
11. Feng, L., Po, L.M., Li, Y., Xu, X., Yuan, F., Cheung, T.C.H., Cheung, K.W.: Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation **38**, 451–460 (2016)
12. Garcia, D.C., de Queiroz, R.L.: Face-spoofing 2d-detection based on moiré-pattern analysis. Transactions on Information Forensics and Security **10**(4), 778–786 (2015)
13. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment **32**(14-15), 2627–2636 (1998)
14. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: European conference on information retrieval, pp. 345–359. Springer (2005)
15. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. Transactions on Systems, Man, and Cybernetics (6), 610–621 (1973)
16. Ilin, R., Kozma, R., Werbos, P.J.: Beyond feedforward models trained by backpropagation: A practical training tool for a more efficient universal approximator. IEEE Transactions on Neural Networks **19**(6), 929–937 (2008)
17. Kumar, S., Singh, S., Kumar, J.: A comparative study on face spoofing attacks. In: International Conference on Computing, Communication and Automation, pp. 1104–1108. IEEE (2017)
18. Li, L., Feng, X., Boulkenafet, Z., Xia, Z., Li, M., Hadid, A.: An original face anti-spoofing approach using partial convolutional neural network. In: International Conference on Image Processing Theory, Tools and Applications, pp. 1–6. IEEE (2016)
19. Liu, S., Yang, B., Yuen, P.C., Zhao, G.: A 3d mask face anti-spoofing database with real world variations. In: Conference on Computer Vision and Pattern Recognition Workshop (2016)
20. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Conference on Computer Vision and Pattern Recognition, pp. 389–398 (2018)
21. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: International Joint Conference on Biometrics, pp. 1–7. IEEE (2011)
22. Manjani, I., Tariyal, S., Vatsa, M., Singh, R., Majumdar, A.: Detecting silicone mask-based presentation attack via deep dictionary learning. Transactions on Information Forensics and Security **12**(7) (2017)
23. Menotti, D., Chiachia, G., Pinto, A., Schwartz, W.R., Pedrini, H., Falcao, A.X., Rocha, A.: Deep representations for iris, face, and fingerprint spoofing detection. Transactions on Information Forensics and Security **10**(4), 864–879 (2015)
24. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
25. Pereira, T., Anjos, A., Martino, J.M., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: International Conference on Biometrics, pp. 1–8. IEEE (2013)
26. Pinto, A., Pedrini, H., Schwartz, W.R., Rocha, A.: Face spoofing detection through visual codebooks of spectral temporal cubes. Transactions on Image Processing **24**(12), 4726–4740 (2015)
27. Pinto, A., Schwartz, W.R., Pedrini, H., de Rezende Rocha, A.: Using visual rhythms for detecting video-based facial spoof attacks. Transactions on Information Forensics and Security **10**(5), 1025–1038 (2015)
28. Plataniotis, K.N., Venetsanopoulos, A.N.: Color image processing and applications. Springer Science & Business Media (2013)
29. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. In: International Statistical and Optimization Perspectives Workshop, pp. 34–51. Springer (2005)

30. Schwartz, W.R., Rocha, A., Pedrini, H.: Face spoofing detection through partial least squares and low-level descriptors. In: International Joint Conference on Biometrics, pp. 1–8. IEEE (2011)
31. Siddiqui, T.A., Bharadwaj, S., Dhamecha, T.I., Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Face anti-spoofing with multifeature videolet aggregation. In: International Conference on Pattern Recognition, pp. 1035–1040. IEEE (2016)
32. Steinwart, I., Christmann, A.: Support vector machines. Springer Science & Business Media (2008)
33. Valle, E., Lotufo, R.: Transfer learning using convolutional neural networks for face anti-spoofing. In: International Conference Image Analysis and Recognition, vol. 10317, p. 27. Springer (2017)
34. Vareto, R.H., Diniz, M.A., Schwartz, W.R.: Face spoofing detection on low-power devices using embeddings with spatial and frequency-based descriptors. In: Iberoamerican Congress on Pattern Recognition, pp. 187–197. Springer (2019)
35. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
36. Xiong, Q., Liang, Y.C., Li, K.H., Gong, Y.: An energy-ratio-based approach for detecting pilot spoofing attack in multiple-antenna systems. Transactions on Information Forensics and Security **10**(5) (2015)
37. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: International Conference on Biometrics, pp. 26–31. IEEE (2012)