Video-Based Face Spoofing Detection through Visual Rhythm Analysis

Allan da Silva Pinto[1], Helio Pedrini[1], William Robson Schwartz[2], Anderson Rocha[1]

[1]Institute of Computing
University of Campinas
Campinas-SP, Brazil, 13083-852

[2]Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte-MG, Brazil, 31270-901

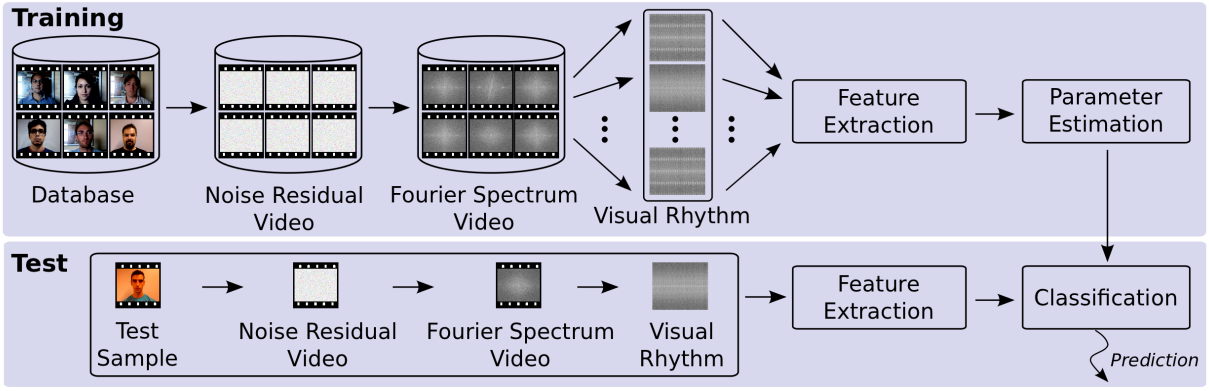


Fig. 1. Given a training set consisting of videos of valid accesses, video-based spoofs, and a video for testing, first we extract a noise signature of every video (training and testing) and calculate the Fourier Spectrum on logarithmic scale for each video frame. Thereafter, we create visual rhythms for each video and train a machine learning classifier using either the pixel intensities directly as features or a summarized version of the visual rhythms using gray level co-occurrence matrices. With a trained classifier, we are able to test a visual rhythm for a given video under investigation and point out whether it is a valid access or a video-based spoof.

***Abstract*—Recent advances on biometrics, information forensics, and security have improved the accuracy of biometric systems, mainly those based on facial information. However, an ever-growing challenge is the vulnerability of such systems to impostor attacks, in which users without access privileges try to authenticate themselves as valid users. In this work, we present a solution to video-based face spoofing to biometric systems. Such type of attack is characterized by presenting a video of a real user to the biometric system. To the best of our knowledge, this is the first attempt of dealing with video-based face spoofing based in the analysis of global information that is invariant to video content. Our approach takes advantage of noise signatures generated by the recaptured video to distinguish between fake and valid access. To capture the noise and obtain a compact representation, we use the Fourier spectrum followed by the computation of the visual rhythm and extraction of the gray-level co-occurrence matrices, used as feature descriptors. Results show the effectiveness of the proposed approach to distinguish between valid and fake users for video-based spoofing with near-perfect classification results.**

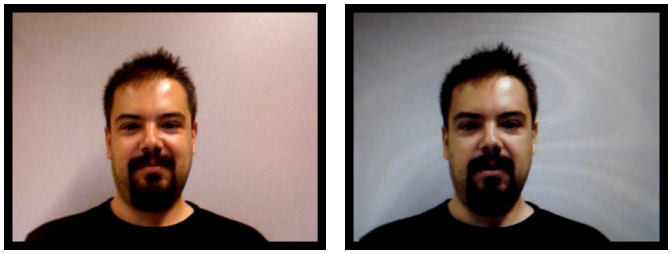
***Keywords*-Face Biometrics, Video-based Face Spoofing, Visual Rhythm, Gray-Level Co-occurrence Matrix.**

I. Introduction

Biometrics provide tools and techniques based on behavior, physical and chemical traits to recognize humans in an automatic and a unique manner. The most common cues are fingerprint, face, iris, hand geometry, hand vein, signature, voice and DNA [1]. Due to recent pattern recognition advances applied to face recognition, biometric systems based on facial characteristics have been largely applied to problems, including access control, surveillance and criminal identification [1], [2], [3].

At the same time that significant advances have been achieved in biometrics, several spoofing techniques have been developed to deceive the biometric systems, and the security of such systems against attacks is still an open problem. Spoofing attacks occur when a person tries to masquerade as someone else falsifying the biometrics data that are captured by the acquisition sensor in an attempt to circumvent a biometric system [4], [5]. Therefore, there is an increasing need to detect such attempts of attacks to biometric systems.

In addition to spoofing attacks, there are other ways to



(a) Frame of a valid video. (b) Frame of an attack video.

Fig. 2. Frames illustrating the video-based face spoofing attack to a biometric system using an LCD technology monitor. (a) example of a real access; (b) attempt of attack.

attack a system [6], [7]. If an impostor (user that does not have permission to access the system) has access to scores of the recognition system, the user can easily circumvent the system [8]. However, this type of attack is more difficult to be performed. Since the acquisition sensor is the most vulnerable part (any user has easy access to this part of the system), spoofing attack techniques have become more attractive for impostor users. Moreover, unlike the authentication systems based on passwords and smart cards, some of our biometric data such as faces are widely available in social networks, personal web sites, and can be easily sampled directly with a digital camera.

In the context of face biometrics, an impostor tries to access the system as a valid user with three approaches [9]: (1) showing a photography of a valid user; (2) showing a video of a valid user, or (3) showing a 3D facial model of a valid user. If any of these approaches succeeds, the uniqueness characteristic of the biometric system will be violated and the system will become fragile [1].

In this context, this work presents a solution to detect attempts of video-based face spoofing. Such attacks can be accomplished presenting a video of a valid user through any display device. Fig. 2 shows a frame of an attack video and a valid (non-attack) video. To the best of our knowledge, this is the *first attempt of dealing with video-based face spoofing* in the literature based on the analysis of global information that is invariant to the video content. We explore the artifacts added to the biometric samples during the viewing process of the videos in the display devices and noise signatures added during the recapture process performed by the acquisition sensor of the biometric system. Our hypothesis is that both noise and artifacts are sufficient to detect the face liveness.

The difficulty in spoofing detection based on video lies on the fact that with a video it is much easier to circumvent the authentication system compared to an image, once the dynamics of the video (e.g., small head movements or blink of eyes) makes the fake biometrics data more real. Also, the ease of incorporating other biometrics data on a single media (e.g., face and voice) allows attacks in multi-modal systems.

We organize this paper into six sections. Section II discusses the current state-of-the-art in anti-spoofing for face biometrics. Section III presents the proposed method. Section IV shows the experimental setup. Section V presents the results and discussion. Finally, Section VI concludes the paper with final remarks and future work.

II. Related Work

There are four major categories of anti-spoofing methods: data-driven characterization, user behavior modeling, user interaction need, and the presence of additional devices [4]. In this section, we review the literature on non-intrusive methods without extra devices and human interaction, since such methods are preferable in practice because they are easily integrated to existing face recognition systems.

Regarding the data-driven characterization, some methods are based on the analysis of skin properties such as texture and reflectance. Li et al. [10] proposed an anti-spoofing solution to photo attacks under the assumption that the size of photos is smaller than a live face and the expressions and poses of the face contained in photos are invariant. These characteristics are detected by analyzing the 2D Fourier spectrum, because photos certainly contain fewer high frequency components, and a threshold is used to detect a spoofing attack. Although the reported results have been satisfactory, in practice these assumptions do not hold. Movements are easy to simulate rotating or bending the photos. Furthermore, the method will probably fail for photos with high quality.

In [11], Tan et al. proposed a solution based on the Lambertian reflectance properties to distinguish between valid and fake users under the assumption that the surface roughness of both classes is different. The authors use two methods for extracting latent reflectance features: variational retinex-based and difference-of-Gaussian (DoG). The authors reported promising results on a publicly available database *(NUAA Database)* composed of true accesses and attacks of 15 subjects using both photo-quality and laser-quality prints.

In a recent work, Peixoto et al. [12] extended the technique proposed in [11] to an image-based spoofing detection based on the fact that the brightness of the LCD screen affects the recaptured image, which makes the image edges more susceptible to a blurring effect. To capture this information, the authors propose an intermediate step before extracting latent reflectance features that consists in applying an adaptive histogram equalization to the images. The reported results on the publicly available *NUAA Database* and *Yale Face Database* show that the proposed extension reduced the classification error in more than 50% for high-quality printing spoofs in the *NUAA* database and 65% for images recaptured from an LCD monitor for the *Yale Face Database*.

Inspired by image quality assessment, characterization of printing artifacts and by differences in light reflection, Määttä et al. [5] proposed an anti-spoofing solution based on micro-texture analysis. The authors use the Local Binary Patterns (LBP) texture analysis operator for describing the micro-textures and use the feature vectors in a Support Vector Machine classifier which determines whether an extracted micro-texture pattern belongs to a fake person (non-live) or a live person.

Tronci et al. [13] proposed a method that is based on multiple cue analysis to detect photo-based spoofing. For that, two techniques are considered: static- and video-based analysis. Static analysis is based on the fundamental idea that during the manufacturing process of a photo attack a certain loss of information occurs and also peculiar noise is introduced. Some used descriptors are color and edge directivity, fuzzy color and texture histogram, MPEG-7 descriptors, RGB and HSV histograms. The video analysis is performed as a combination of simple measures of movements such as eye blinks, mouth alterations and changes in facial expression. With these descriptors in hand, a classifier is trained to each feature descriptor and then a fusion is made to decide whether the biometric data is fake or real.

In order to incorporate temporal information from videos captured from an image, Schwartz et al. [14] presented a holistic method for describing faces combining several feature descriptors. Considering only the facial region, the authors extract several features using descriptors that capture different characteristics of the images, such as shape, color and texture. They reported improved results, but the combination of these descriptors generated high dimension feature spaces that may not be suitable for standard classification methods.

Optical flow analysis has also been considered in the literature. Bao et al. [15] obtained a reference field from the actual flow field data on fake and valid images to estimate their differences. Kollreider et al. [16] presented a method based on the optical flow algorithm for capturing and tracking subtle movements of different facial parts, assuming that facial parts on real faces move differently than on photos.

Considering behavior modeling, some works have focused on eye blinking [17], [18] and small movements of parts of the head and face [9] to detect specifically photo-based spoofing. Considering that a person blinks approximately once every two to four seconds, Pan et al. [17] proposed the use of an undirected conditional random field framework to represent eye blinking from hidden Markov models that relax the independence assumption of generative modeling, with the advantage that the method allows to relax the assumption of conditional independence of the observed data.

In [19], Pan et al. extended upon the work in [17] by adding counter-measures to include a scene context matching in stationary face-recognition systems. For this, the authors analyze inside face cues such as eye blinking and outside face cues, since the background scene is known by the recognition system. The authors reported that their method works well to photo-based, video-based and 3D-based spoofing techniques, because inside-face clues of spontaneous eye blinks can be employed to detect photo-based spoofing and 3D models and outside-face clues of scene context to detect video-based spoofing. However, since a video or image background can be easily changed, their method may fail and a 3D model can incorporate the action of blinking similar to a real eye blinking. A private dataset, in which they obtained almost perfect results, was created but not released to the public.

Even though the described approaches present interesting results, they do have some weaknesses. Small facial movements may be easily simulated by tilting a photography used in the attack, therefore, methods assuming static images will fail. In addition, such approaches are not suitable for attacks performed using videos of a valid user due to movements present in both valid and fake accesses. The method proposed in this work aims at overcoming such difficulties by capturing noise information features generated by the recapturing of videos in such a way that it is independent of their content.

III. Proposed Method

The first task performed on any facial biometric system is the data acquisition to authenticate the user. This acquisition is performed by a camera that has an imaging sensor with thousands of photosensitive transducers capable of converting light energy into electrical charges. The camera lenses allow light reflected by the objects in the scene to focus on the imaging sensor, transforming light energy into electrical charges, which are converted into digital signals by an A/D converter [20].

During this process of transforming an analog signal into a digital signal, the appearance of noise in the resulting image is inevitable. The analysis of noise in images has been widely explored in the digital document forensic analysis area, more specifically, the problem of identifying the specific camera that acquired a document. In this case, the main goal is to estimate the type and manufacturer of the cameras with just one image. Lukas et al. [21] discuss two types of noise present in images: the fixed pattern noise (FPN) and the noise resulting from the photo-responsiveness of non-uniform light-sensitive cells (PRNU).

FPN noise is caused by the presence of *dark currents* that can be defined by accumulated electrons in the inverse joints of the light-sensitive cell pins of the imaging sensor. This accumulation of electrons occurs mainly due to a thermal action and is independent of the amount of light incident on the sensor [20]. On the other hand, PRNU noise is defined by the difference in sensitivity of the light sensitive cells caused by the non-homogeneity of the silicon wafer and other imperfections inserted during the manufacturing process of the sensor [20].

Another noticeable fact is the appearance of artifacts generated by means of videos captured from other videos, which do not exist in videos generated from the capture of real scenes. These artifacts, addressed in this paper as *noise*, are generated mainly during the process of creation and exhibition of the frames on monitor screens, producing undesirable effects such as distortion, flickering, moiring, among others [22]. Thus, the biometric samples extracted from videos submitted to the biometric system (referred to as *attack videos*) will likely have more noise than the biometric samples captured directly from live people (referred to as *valid videos*). Fig. 3 shows the effect of moiring due to the screen capture of three different monitors with a digital camera.

The first step of the proposed method for capturing such differences and solving the problem is to isolate the noise

Fig. 3. Moiring effects in videos shown on three different monitors and captured with a digital camera.

information contained in the videos generated during the data acquisition.

A video V in the domain $2D + t$ can be defined as a sequence of frames t, each frame as a function $f(x, y) \in \mathbb{N}^2$ of the luminous intensity of each pixel at position (x, y) of the scene. To isolate the noise of the t-th frame in a video V, a copy of this frame is submitted to a filtering process in order to eliminate noise. Then, a subtraction is performed between the original and the filtered frame, generating a new frame containing only the noise, as formalized in Equation 1. The collection of the new frames is called noise residual video

$$V_{noise}^{(t)} = V^{(t)} - f(V_{copy}^{(t)}) \qquad \forall\, t \in T = \{1, 2, \ldots, t\}, \quad (1)$$

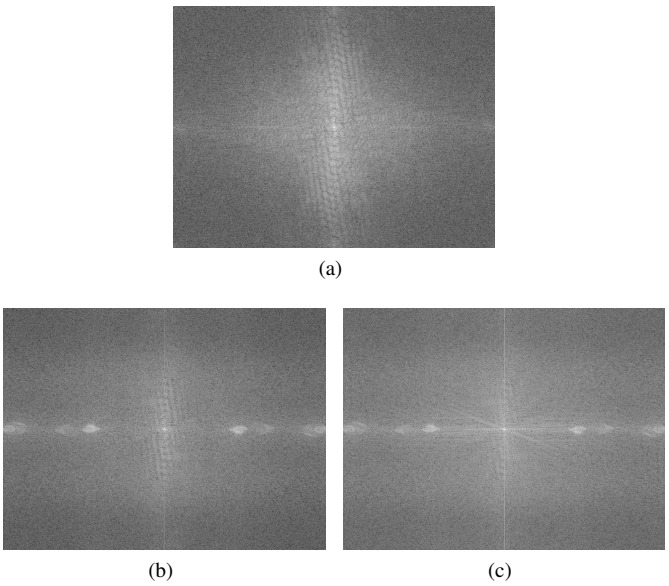
where $V^{(t)} \in \mathbb{N}^2$ is the t-th frame of V and f a filtering operation.

We can analyze the noise pattern and possible artifacts contained in the video by applying a 2D discrete Fourier transform on each frame of the noise residual video $V_{noise}^{(t)}$ and calculating its spectrum in the logarithmic range and with origin at the center of the frame using Equations 2 and 3, respectively. As a result of this process, we end up with a video of the spectra. Fig. 4(a) shows the logarithm of the Fourier spectrum for a video frame obtained from a valid video, and Fig. 4(b-c) show the logarithm of the Fourier spectrum for an attack video using the Gaussian and Median filter, respectively.

$$\mathcal{F}(\upsilon, \nu) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} V_{(noise)}(x, y) e^{-j2\pi[(\upsilon x/M)+(\nu y/N)]} \quad (2)$$

$$|\mathcal{F}(\upsilon, \nu)| = \sqrt{\mathcal{R}(\upsilon, \nu)^2 + \mathcal{I}(\upsilon, \nu)^2}$$
$$\mathcal{S}(\upsilon, \nu) = \log(1 + |\mathcal{F}(\upsilon, \nu)|) \quad (3)$$

Note that the logarithm of the Fourier spectrum, shown in Fig. 4(b-c), contains the highest responses concentrated in the abscissa and ordinate axes, whose origin is at the center of the frame, unlike the logarithm of Fourier spectrum shown in Fig. 4(a). This situation occurs in practically every frame. This is an important fact, since the occurrence or not of these components in the axes allows us to decide whether the video is fake or valid. To have this decision made automatically, we design a space-time descriptor that captures this information used in a classification algorithm for associating input patterns



(a) (b) (c)

Fig. 4. Example of video frame of the spectra generated from (a) a valid video and (b)-(c) an attack video considering a Gaussian and Median filter, respectively.

in certain classes or categories. In this work, we consider the Support Vector Machine [23] (SVM) and the Partial Least Squares regression [24] (PLS) to classify the patterns that are extracted from the visual rhythm of the video. We shall define visual rhythms shortly.

A. Video Characterization

The construction of the descriptors used in the classification process is done using the concept of visual rhythms [25], which can efficiently capture temporal information and summarize the video contents in a single image. The application of the visual rhythm can be found on the works including the one by Chun et al. [26] for fast text caption localization on videos and by Guimarães et al. [27] for identifying cut and gradual transitions on videos.

Considering a video V in the domain $2D + t$ with t frames of dimension $M \times N$ pixels, the visual rhythm $\Re$ is a simplification of the video V, in which lines or columns of each frame t are sampled and concatenated to form a new image, called visual rhythm. Fig. 5 illustrates a visual rhythm and its sampled 2D image acquired from a video.

Given that the highest responses for our problem are concentrated on the abscissa and ordinate axes of the logarithm of the Fourier spectrum, we consider two regions of interest of the frames that form the spectrum video in the construction of two types of visual rhythms: (i) the horizontal visual rhythm formed by central horizontal lines; and (ii) the vertical visual rhythm formed by central vertical lines. In both cases, we can summarize relevant content of the spectrum video in a single image. Fig. 6 depicts the visual rhythms generated by two regions of interest considering a valid (Fig. 6(a) and Fig. 6(c)) and an attack video (Fig. 6(b) and Fig. 6(d)).

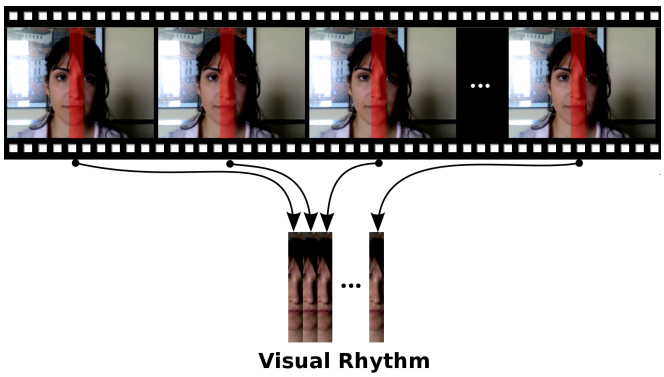
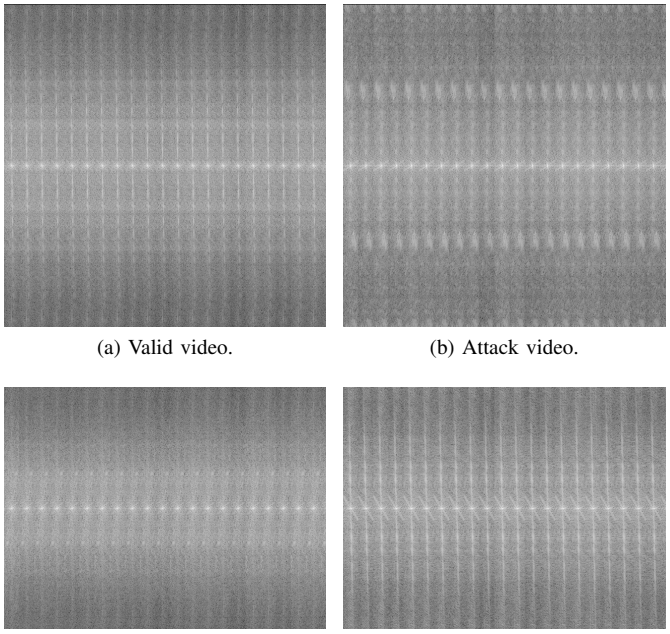


Fig. 5. Example of a simplification of a video by means of the visual rhythm using vertical blocks.



(a) Valid video. (b) Attack video.

(c) Valid video. (d) Attack video.

Fig. 6. Examples of visual rhythms constructed from (a)-(b) central horizontal lines and from (c)-(d) central vertical lines. Note that the visual rhythm obtained from horizontal lines has been rotated 90 degrees.

Once the visual rhythms are computed, we can employ a machine learning classifier to automate the process. However, if the intensity of the pixels composing the visual rhythms are directly considered, the dimensionality of the feature space will be extremely high and most of the classification methods to date will not work properly. Therefore, we need to extract a compact set of feature descriptors that best discriminate the visual rhythms generated from the fake and valid videos.

In our work, we consider the visual rhythms as a texture map, such that we can apply the well-known gray-level co-occurrence matrices (GLCM) [28] to extract textural information from them, since this descriptor provides spatial distribution and brightness variation of the image regions [14]. A GLCM is a structure that describes the frequency of occurrence of gray levels between pairs of pixels. When normalized, the co-occurrence matrix becomes the estimation of joint probabilities between pairs of pixels at a distance d in a given orientation θ. After calculating the co-occurrence matrices for four orientations, we extract 12 measures summarizing textural information from each matrix: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and directionality [28].

Finally, we use either SVM or PLS regression method to classify the patterns that are extracted from the visual rhythms and GLCM. The SVM algorithm [29] uses a linear or non-linear mapping, depending on the type of space used to transform the original data onto a higher dimensional. Within this new space, the SVM finds an optimal hyperplane that separates the input data into classes. The algorithm finds this hyperplane of separation through support vectors and a margin. The support vectors are essentially training tuples that are close to the decision boundary of the classes and the margin is defined as the perpendicular distance between the decision boundary and the closest tuple for each class.

PLS regression method [30], [31] is based on the linear transformation of a large number of descriptors to a new space based on a small number of orthogonal projection vectors. In other words, the projection vectors are mutually independent linear combinations of the original descriptors. These vectors are chosen to provide maximum correlation with the dependent variables, that are the labels of the training classes. Fig. 1 summarizes the main steps of the proposed method.

IV. Experimental Setup

In this section, we describe the details of the experiments performed to validate the proposed method for detecting video-based spoof attacks. We created a dataset comprising valid access and video-based spoof videos given that the benchmarks publicly available, such as NUAA Database [11] and the Print-Attack Database [32], are meant only for photo-based spoofing attacks.

All experiments were conducted on an Intel Xeon 5160, 3GHz dual core processor with 8GB of RAM running Windows 7 operating system.

A. Dataset Creation

The dataset created in this paper expands upon the Print-Attack Database [32], which consists of 200 videos of valid accesses of 50 different users and 200 videos of spoof attacks using printed photographs in 320×240 pixel resolution, which were constructed by presenting printed photographs without movement to the acquisition sensor. Therefore, such videos are indicated to evaluate photo-based anti-spoofing solutions. Given that our goal is to deal with video-based spoofs, we only consider the 200 valid access videos and expand upon them to create a video-based dataset.

First, we upsample the 200 valid access videos to 640×480 pixels in resolution. Thereafter, we play 100 valid access videos in six monitors in a controlled environment to minimize

the illumination changes between a valid and attack video. Finally, we recapture them using a Sony *CyberShot* camera also in 640×480 pixel resolution that represents the acquisition sensor of a biometric system. The 100 selected videos were used only to create the attack videos and were therefore discarded. Table I summarizes some characteristics of the monitors considered in the experiments. We encode all the videos with YV12 codec at 30 frames per second. The final dataset, which we will publicly upon acceptance [1], comprises 700 videos (100 valid accesses and 600 spoofs).

TABLE I
CHARACTERISTICS OF THE MONITORS CONSIDERED IN THE EXPERIMENTS FOR CREATING THE VIDEO SPOOFING SAMPLES.

ID	Manufacturer	Technology used in an image formation	Screen type
Monitor 1	Itautec	LCD	Glossy
Monitor 2	LG	LCD	Matte
Monitor 3	Samsung	LCD	Matte
Monitor 4	LG	LCD	Matte
Monitor 5	AOC	LED	Matte
Monitor 6	LG	LCD	Matte

B. Analysis of the Filtering Process and Visual Rhythm

To extract noise of the videos as shown in Equation 1, we consider a linear and non-linear spatial filter: a Gaussian with $\mu = 0$, $\sigma = 2$, and size 7×7 and a Median filter of size 7×7. All this parameters were obtained empirically.

After computing noise signatures using Equations 2 and 3, we extract the visual rhythm of each video (horizontal and vertical) using the first 50 frames and a block of either 30 rows (for horizontal) or 30 columns (for vertical) of pixels. The horizontal visual rhythms are in a $640 \times 1,500$-d while the vertical ones are in $480 \times 1,500$-d space.

From the calculated visual rhythms, we can work directly on this high dimensional space, which we call pixel intensity analysis, or we can consider visual rhythms as texture maps and calculate the textural patterns from them using the GLCM.

Given that the horizontal and vertical visual rhythms extracted from each video form different texture maps, we can assess the two types of visual rhythms as well their combination.

C. Discriminating Power and Classification Techniques

With this experiment, our objective is to assess the discriminability power of the visual rhythms for classifying spoof attempts in videos. We evaluate two different sets of features: the direct pixel intensities in the visual rhythms and a compact representation of the visual rhythms using the gray level co-occurrence matrix, with orientations $\theta \in \{0°, 45°, 90°, 135°\}$, distance $d = 1$ and 16 bins. In this work, 12 measures of texture were extracted from the generated four co-occurrence matrices. Table II shows the dimensionality of each type of features (individually and combined).

In order to evaluate the extracted features, we can use them to train a machine learning classifier and generate a model capable of distinguishing valid and attack videos, and test the effectiveness of the model. In this paper, we use two classification techniques: SVM and PLS. For SVM, we analyze two different kernels: linear and radial basis function kernels using the LibSVM [33] implementation.

TABLE II
NUMBER OF FEATURES (DIMENSIONS) USING EITHER THE DIRECT PIXEL INTENSITIES AS FEATURES OR THE GLCM-BASED TEXTURE INFORMATION FEATURES.

Nome	Descriptor Dimensionality		
	Horizontal	Vertical	Horizontal + Vertical
Pixel Intensity	960,000	720,000	1,680,000
GLCM	48	48	96

D. Data Set Partitioning

The dataset was divided into four sets: (1) *Valid #1*, comprising 50 valid access videos; (2) *Valid #2*, comprising 50 valid access videos; (3) *Attack #1*, with 300 attack videos created by using the monitors 1, 2 and 3; (4) *Attack #2*, comprising 300 attack videos created by using the monitors 4, 5 and 6. The partitioning considering different monitors for both Attack sets has been chosen to avoid the classifier to take conclusions over images coming from monitors it already had seen during training. In the protocol we devised, a classifier is trained with images from a set of monitors and tested with images of monitors it never had access to.

Therefore, we design two configurations for the experiments. In the first configuration, we use *Valid #1* and *Attack #1* groups to train the classifiers and *Valid #2* and *Attack #2* groups to evaluate the model found by the classifiers. In the second configuration, we use *Valid #2* and *Attack #2* groups to train the classifiers and *Valid #1* and *Attack #1* to test. The results reported in the Tables III, IV, V and VI are in terms of average and standard deviation of the two configurations. We do not show the ROC curve because the obtained results are near-perfect.

V. RESULTS AND DISCUSSION

Tables III and IV show the experimental results considering the SVM classification technique for Gaussian and Median filters. Similarly, Tables V and VI show the experimental results for the PLS classification technique.

The results show that visual rhythms calculated on a logarithmic scale Fourier Spectrum represent an effective alternative to summarize videos and an important forensic signature for detecting video-based spoofs.

The results allow us to assert that the filtering process does not have influence on our method, since that the obtained results using the Median and Gaussian filters, using either the SVM or PLS classifiers, are statistically comparable. Although

[1] http://www.ic.unicamp.br/∼rocha/pub/communications.html

the standard deviations showed in the Tables III are 1.60% and 0.50% using the vertical and horizontal visual rhythms, respectively, with the combination these features, we obtained an AUC with 100.0% and standard deviation of 0.0%.

In addition, we notice that visual rhythms can be interpreted as texture maps that can be summarized using simple texture descriptors, such as the gray level co-occurrence matrices. This result is important since many classification techniques (e.g., SVM) have memory allocation problems when dealing with high-dimensional feature spaces. Finally, it is important to mention that combining the horizontal and vertical visual rhythms indeed represent a boost in the classification, providing better results than each individual set of features.

TABLE III
OBTAINED RESULTS IN TERMS OF AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) CONSIDERING THE SVM CLASSIFICATION TECHNIQUE AND GAUSSIAN FILTER. SVM WAS NOT ABLE TO CALCULATE A CLASSIFICATION HYPERPLANE WHEN USING DIRECT PIXEL INTENSITIES AS FEATURES.

Visual Rhythms	SVM Linear		SVM RBF	
	Intensity	GLCM	Intensity	GLCM
Vertical	–	$\overline{x} = 98.4\%$	–	$\overline{x} = 99.9\%$
	–	$\sigma = 1.60\%$	–	$\sigma = 0.10\%$
Horizontal	–	$\overline{x} = 99.6\%$	–	$\overline{x} = 99.7\%$
	–	$\sigma = 0.50\%$	–	$\sigma = 0.10\%$
Horiz.+Vert.	–	$\overline{x}$ = 100.0%	–	$\overline{x} = 100.0\%$
	–	$\sigma = 0.0\%$	–	$\sigma = 0.0\%$

TABLE IV
OBTAINED RESULTS IN TERMS OF AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) CONSIDERING THE SVM CLASSIFICATION TECHNIQUE AND MEDIAN FILTER. SVM WAS NOT ABLE TO CALCULATE A CLASSIFICATION HYPERPLANE WHEN USING DIRECT PIXEL INTENSITIES AS FEATURES.

Visual Rhythms	SVM Linear		SVM RBF	
	Intensity	GLCM	Intensity	GLCM
Vertical	–	$\overline{x} = 99.7\%$	–	$\overline{x} = 99.6\%$
	–	$\sigma = 0.20\%$	–	$\sigma = 0.10\%$
Horizontal	–	$\overline{x} = 99.9\%$	–	$\overline{x} = 100.0\%$
	–	$\sigma = 0.10\%$	–	$\sigma = 0.0\%$
Horiz.+Vert.	–	$\overline{x} = 100.0\%$	–	$\overline{x} = 100.0\%$
	–	$\sigma = 0.0\%$	–	$\sigma = 0.0\%$

VI. CONCLUSIONS AND FUTURE WORK

Due to the importance of providing secure biometric systems based on facial traits, in this paper we investigated the characteristics of the problem and presented a method for detecting video-based face spoofing by analyzing noise signatures generated by the video recapturing process through analysis of visual rhythms and textural information.

We believe that video-based spoofing attacks refer to a problem more realistic than photo-based spoofing attacks, since the algorithms based on movements and eye blinking might not be suitable for detecting such attacks.

TABLE V
OBTAINED RESULTS IN TERMS OF AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) CONSIDERING THE PLS CLASSIFICATION TECHNIQUE AND GAUSSIAN FILTER.

Visual Rhythm	PLS	
	Intensity	GLCM
Vertical	$\overline{x} = 99.9\%$	$\overline{x} = 98.2\%$
	$\sigma = 0.20\%$	$\sigma = 0.40\%$
Horizontal	$\overline{x} = 100.0\%$	$\overline{x} = 98.9\%$
	$\sigma = 0.0\%$	$\sigma = 1.50\%$
Horiz. + Vert.	$\overline{x} = 100.0\%$	$\overline{x} = 99.9\%$
	$\sigma = 0.0\%$	$\sigma = 0.10\%$

TABLE VI
OBTAINED RESULTS IN TERMS OF AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (AUC) CONSIDERING THE PLS CLASSIFICATION TECHNIQUE AND MEDIAN FILTER.

Visual Rhythm	PLS	
	Intensity	GLCM
Vertical	$\overline{x} = 100.0\%$	$\overline{x} = 99.5\%$
	$\sigma = 0.0\%$	$\sigma = 0.70\%$
Horizontal	$\overline{x} = 100.0\%$	$\overline{x} = 99.9\%$
	$\sigma = 0.0\%$	$\sigma = 0.10\%$
Horiz. + Vert.	$\overline{x} = 100.0\%$	$\overline{x} = 100.0\%$
	$\sigma = 0.0\%$	$\sigma = 0.0\%$

The experiments we carried out demonstrate that the Fourier spectrum of video noise signatures and the use of visual rhythms are able to properly capture discriminative information to distinguish between valid and fake users for video-based spoofing. In addition, the extraction of feature descriptors with GLCM provided a compact representation while keeping the method discriminability.

The compact representation achieved by using visual rhythms as texture maps has positive impacts on the implementation of the method for a real biometric system that needs to respond to inputs fast, allowing the use of our method in small and medium computational systems. Furthermore, the reduced dimensionality of the method allows the use of our method in large video databases.

Finally, directions for future work include the exploration of new video summarization approaches as well the use of more monitors and real videos. As we employed monitors with different image formation technologies (LCD and LED), additional tests could be performed considering tablets and smart phones as well as the investigation of illumination influences on the proposed method.

ACKNOWLEDGMENT

We would like to thank Microsoft Research, São Paulo Research Foundation (FAPESP), CNPq, and CAPES for the financial support.

REFERENCES

[1] A. K. Jain and A. Ross, *Handbook of Biometrics*. Springer, 2008, ch. Introduction to Biometrics, pp. 1–22.

[2] A. Jain and B. Klare, "Matching Forensic Sketches and Mug Shots to Apprehend Criminals," *Computer*, vol. 44, no. 5, pp. 94–96, 2011.

[3] N. Zamani, M. Darus, S. Abdullah, and M. Nordin, "Multiple-frames Super-resolution for Closed Circuit Television Forensics," in *Intl. Conference on Pattern Analysis and Intelligent Robotics*, vol. 1, 2011, pp. 36–40.

[4] K. Nixon and V. A. R. Rowe, *Handbook of Biometrics*. Springer, 2008, ch. Spoof Detection Schemes.

[5] J. Määttä, A. Hadid, and M. Pietikainen, "Face Spoofing Detection from Single Images using Micro-texture Analysis," in *Intl. Joint Conference on Biometrics*, Oct. 2011, pp. 1–7.

[6] C. Rathgeb and A. Uhl, "Attacking Iris Recognition: An Efficient Hill-Climbing Technique," in *International Conference on Pattern Recognition*, Aug. 2010, pp. 1217–1220.

[7] ——, "Statistical Attack Against Iris-Biometric Fuzzy Commitment Schemes," in *IEEE Computer Vision and Pattern Recognition Workshops*, Jun. 2011, pp. 23–30.

[8] A. Adler, *Handbook of Biometrics*. Springer, 2008, ch. Biometric System Security.

[9] G. Pan, Z. Wu, and L. Sun, *Recent Advances in Face Recognition*. InTech, 2008, ch. Liveness Detection for Face Recognition, pp. 235–252.

[10] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live Face Detection Based on the Analysis of Fourier Spectra," in *Biometric Technology for Human Identification*, 2004, pp. 296–303.

[11] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model," in *European Conference on Computer Vision*, 2010, pp. 504–517.

[12] Peixoto, C. Michelassi, and A. Rocha, "Face Liveness Detection under Bad Illumination Conditions," in *IEEE Intl. Conference on Image Processing*, Sep. 2011, pp. 3557–3560.

[13] R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, and F. Roli, "Fusion of Multiple Clues for Photo-Attack Detection in Face Recognition Systems," in *Intl. Joint Conference on Biometrics*, Oct. 2011, pp. 1–6.

[14] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face Spoofing Detection through Partial Least Squares and Low-Level Descriptors," in *Intl. Joint Conference on Biometrics*, Oct. 2011, pp. 1–8.

[15] W. Bao, H. Li, N. Li, and W. Jiang, "A Liveness Detection Method for Face Recognition Based on Optical Flow Field," in *IEEE Intl. Conference on Image Analysis and Signal Processing*, 2009, pp. 233–236.

[16] K. Kollreider, H. Fronthaler, and J. Bigun, "Non-Intrusive Liveness Detection by Face Images," *Elsevier Image and Vision Computing*, pp. 233–244, Feb. 2009.

[17] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera," in *IEEE Intl. Conference on Computer Vision*, 2007, pp. 1–8.

[18] J.-W. Li, "Eye Blink Detection Based on Multiple Gabor Response Waves," in *IEEE Intl. Conference on Machine Learning and Cybernetics*, 2008, pp. 2852–2856.

[19] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular Camera-based Face Liveness Detection by Combining Eyeblink and Scene Context," *Telecommunication Systems*, vol. 47, pp. 215–225, 2011.

[20] A. Rocha, W. Scheirer, T. Boult, and S. K. Goldenstein, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics," *ACM Computing Surveys*, vol. 26, no. 1, pp. 26–42, 2011.

[21] J. Lukas, J. Fridrich, and M. Goljan, "Digital Camera Identification from Sensor Pattern Noise," *IEEE Trans. on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, 2006.

[22] A. Beach, *Real World Video Compression*. Peachpit Press, 2008.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[24] H. Wold, "Partial Least Squares," in *Encyclopedia of Statistical Sciences*, S. Kotz and N. Johnson, Eds. New York: Wiley, 1985, vol. 6, pp. 581–591.

[25] M.-G. Chung, J. Lee, H. Kim, S. M.-H. Song, and W.-M. Kim, "Automatic Video Segmentation based on Spatio-Temporal Features," *Korea Telecom*, vol. 1, no. 4, pp. 4–14, 1999.

[26] S. S. Chun, H. Kim, K. Jung-Rim, S. Oh, and S. Sull, "Fast Text Caption Localization on Video Using Visual Rhythm," in *Recent Advances in Visual Information Systems*, ser. Lecture Notes in Computer Science, S.-K. Chang, Z. Chen, and S.-Y. Lee, Eds., 2002, vol. 2314, pp. 43–58.

[27] S. J. F. Guimaraes, M. Couprie, N. J. Leite, and A. A. Araujo, "A Method for Cut Detection Based on Visual Rhythm," in *Brazilian Symposium on Computer Graphics and Image Processing*, 2001, pp. 297–304.

[28] R. Haralick, K. Shanmugam, and I. Dinstein, "Texture Features for Image Classification," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 3, no. 6, 1973.

[29] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[30] A. Hoskuldsson, "PLS Regression Methods," *Journal of Chemometrics*, vol. 2, no. 3, pp. 211–228, 1988.

[31] H. Abdi, "Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, 2010.

[32] A. Anjos and S. Marcel, "Counter-Measures to Photo Attacks in Face Recognition: A Public Database and A Baseline," in *Intl. Joint Conference on Biometrics*, 2011, pp. 1–7.

[33] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 2011, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.