Journal of *Imaging*

MDPI

Article

A Robust Document Identification Framework through f-BP Fingerprint

Francesco Guarnera [1,*], Oliver Giudice [1], Dario Allegra [1], Filippo Stanco [1], Sebastiano Battiato [1], Salvatore Livatino [2], Vito Matranga [3] and Angelo Salici [3]

[1] Department of Mathematics and Computer Science, University of Catania, 95125 Catania, Italy; giudice@dmi.unict.it (O.G.); allegra@dmi.unict.it (D.A.); filippo.stanco@unict.it (F.S.); battiato@dmi.unict.it (S.B.)

[2] School of Physics, Engineering and Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK; s.livatino@herts.ac.uk

[3] Raggruppamento Carabinieri Investigazioni Scientifiche, RIS di Messina, 98122 Messina, Italy; vito.matranga@carabinieri.it (V.M.); angelo.salici@carabinieri.it (A.S.)

* Correspondence: francesco.guarnera@unict.it

Abstract: The identification of printed materials is a critical and challenging issue for security purposes, especially when it comes to documents such as banknotes, tickets, or rare collectable cards: eligible targets for ad hoc forgery. State-of-the-art methods require expensive and specific industrial equipment, while a low-cost, fast, and reliable solution for document identification is increasingly needed in many contexts. This paper presents a method to generate a robust fingerprint, by the extraction of translucent patterns from paper sheets, and exploiting the peculiarities of binary pattern descriptors. A final descriptor is generated by employing a block-based solution followed by principal component analysis (PCA), to reduce the overall data to be processed. To validate the robustness of the proposed method, a novel dataset was created and recognition tests were performed under both ideal and noisy conditions.

Keywords: document identification; binary pattern; texture fingerprint

Citation: Guarnera, F.; Giudice, O.; Allegra, D.; Stanco, F.; Battiato, S.; Livatino, S.; Matranga, V.; Salici, A. A Robust Document Identification Framework through f-BP Fingerprint. *J. Imaging* **2021**, *7*, 126. https://doi.org/10.3390/jimaging7080126

Academic Editor: Raimondo Schettini

Received: 25 May 2021
Accepted: 26 July 2021
Published: 29 July 2021

1. Introduction and Related Works

The manufacturing process needed to produce common paper sheets involves the use of wood particles with subsequent application of other compounds. The intrinsic random imperfections generated make the sheet almost unique, and under certain conditions it is possible to extract a proper fingerprint. The massive demand of robust identification methods in many contexts [1–6], makes fingerprint extraction from a sheet of paper an attractive and challenging research topic. Investigative scenarios in the forensic field [7,8], could gain several advantages from the availability of such a fingerprint.

Although several techniques have been proposed, most of them require expensive industrial devices [9,10], which are not commonly affordable. Taking inspiration from the use of wood fiber patterns for fingerprint extraction [11], the main objective of this work is the design of a cheaper solution able to extract a robust fingerprint. Please note that in contrast with a biometric system [12] the main interest is in finding a robust strategy to recognize a specific paper sheet by matching it with a previously acquired version of itself. Local Binary Pattern (LBP) [13] and its variants [14] have been employed under different conditions and global experimental settings, clearly outperforming the results obtained by Guarnera et al. [15] in terms of efficiency and effectiveness. It is worth noting that since in ideal conditions the paper texture is unique, any sufficiently descriptive image-processing approach should perform with good results in terms of accuracy. This led us to further investigate the problem by performing tests on papers where some degradation was synthetically applied (e.g., stain, crop, etc.) simulating real case scenarios where part of the original information is totally or partially missing.

However, the identification of a document for legal purposes, employed to detect counterfeiting and piracy, is usually done through the use of different techniques [7]. The interest in anti-counterfeiting measures, based on the fingerprints left on the surface of the paper without any specific embedding requirement, is the core of the present study. Other techniques are based on security patterns or on properly generated features that are hidden in the substrate material or masked by special ink properties. Such identification strategies are widely used, but typically require an additional pattern to be added, and are also expensive and hard to generalize for all cases (e.g., legal documents, banknotes, etc.). By contrast, strategies that directly analyze the physical properties of the material by not adding any signal, as in the case of common active methods (e.g., watermarking) are highlighted in this paper. The underlying hypothesis for the development of a fingerprint extraction technique is the existence of low-cost physically unclonable functions (PUFs) to generate an intrinsic random physical feature for paper identification with the following two properties:

- fast and deterministic processing to obtain a response;
- the return of a unique response for the same request.

The response must be unpredictable, even for an attacker with physical access to the object, by operating as a sort of random function. The paper surface presents an inherently unique structure, as it consists of overlapping and inter-twisted wood fibers. Hence, the imperfections of a paper sheet caused by the manufacturing process can be exploited to uniquely identify such sheet. The use of a fingerprinting technique for document identification was proposed for the first time by Buchanan et al. [16]. It has been proven that is extremely unlikely that two document surfaces created with the same raw materials will be identical, although they will present some similarities. This fingerprint makes forgery unfeasible, given that it is unique and virtually impossible to modify. To extract a fingerprint from the paper structure, the authors in [16] employed laser irradiation from four different angles and acquired the reflected energy. Inspired by Buchanan et al., the authors in [17] proposed an improvement based on correlation metrics between the acquired energy signals. Cowburn introduced the use of laser speckle for product identification ([18,19]). Clarkson et al. [20] proposed the extraction of 3D paper structure by scanning different orientations and employing a Voronoi distribution to build the fingerprint. Samsul et al. [9] proposed a fingerprint extraction method, which exploits CCD sensors and laser speckle, to employ the visible pattern of bright and dark spots generated by interference of two or more light beams with different phases. A similar approach has been proposed by Sharma et al. [10]. In contrast to [9], they employed a microscope to acquire the speckle pattern. In recent years, CNN-based methods have achieved great performance in image recognition and classification, but have high complexity and require GPUs to perform training.

The aforementioned approaches work well for paper fingerprint extraction, but they require industrial and specific equipment. Recently, this limitation was overcome by the works of Toreini et al. [11] and Wong et al. [21]. In [21] the authors proposed a strategy to extract paper surface imperfections by exploiting multiple shots taken by a mobile camera under semi-controlled light conditions; subsequently, they investigated selected candidates through ad hoc mathematical models for each camera-captured image [22]. Unlike previous works, Toreini et al. [11] did not detect surface imperfections, but captured the random arrangement of the wood fibers within the paper sheet. To extract the paper pattern, they exploited a consumer camera and a backlit surface. However, they printed a bounding box on the analyzed paper to simplify the automatic texture registration. Since in real scenarios this registration strategy is not applicable, a different acquisition framework is needed. Based on the same filter of [11], Chen et al. in [23] exploited the microscopic features of wood fibers to obtain similar patterns, using expensive equipment based on double cameras. As already demonstrated in [24], the random disposition of wood fibers on paper sheets makes possible the construction of a fingerprint virtually impossible to tamper with; hence, given the limits of the previous works in terms of costs, acquisition constraints, and robustness, in [15] the authors presented a novel fingerprint extraction strategy using

specific low-cost image-acquisition equipment and a simpler and faster method based on local binary patterns. In this paper, further experiments on LBP variants are carried out, such as Local Ternary Pattern (LTP) [25], Statistical Binary Pattern (SBP) [26] and Complete modeling of Local Binary Pattern (CLBP) [27] to find the most descriptive binary pattern for fingerprint identification and tampering to achieve a more robust solution. In contrast to [9,10,16–20,23], the present work proposes a fingerprint extraction method that does not require expensive industrial equipment (e.g., laser, microscopes), but solely cheaper devices such as an RGB camera, as it is based on wood fiber translucent patterns. Additionally, the proposed approach overcomes a method based on translucent patterns such as the one by [11]; in fact, LBP descriptors used in [15] have already been proved to outperform Gabor filters employed in [11] in terms of effectiveness and efficiency. Since the intrinsic advantages of [15] over [11] were confirmed in previous research, we compare the proposed approach only with [15] and not with [11].

The main contributions of this paper can be summarized as follows:

1. a new fingerprint extraction method, based on LBP variants, which outperforms existing approaches in the field;
2. an optimization of a BP-based fingerprint that employs block subdivision and Principal Component Analysis (PCA);
3. a new public dataset that includes images acquired with both low-cost and high-end devices, showing wood fiber patterns, which is the one available to the best of our knowledge;

The remainder of this paper is organized as follows. The next section is devoted to briefly summarizing the state of the art in the field. Section 2 details the proposed fingerprinting extraction strategy; Section 3 describes the acquisition procedure of image data and the overall organization of the employed dataset; in Section 4, experimental results are reported together with a deep analysis on the results obtained by the proposed approach. Conclusions are reported in Section 6.

2. Fingerprint Extraction Process

Illuminating the surface to highlight the wood fibers is mandatory to properly extract the pseudo-random pattern which is unique for each sheet of paper. However, such patterns must be digitalized and properly modeled mathematically to implement a robust document identification system, which is the goal of this paper. Given a certain physical paper document d_i, the aim of this work is to obtain a digital fingerprint F_i, namely a sequence of K ordered values $\{f_i^{(1)}, f_i^{(2)}, ..., f_i^{(K)}\}$, which is solely determined by correctly processing the digital image s_i, which is the acquisition of the document d_i. The overall proposed pipeline is summarized in Figure 1.

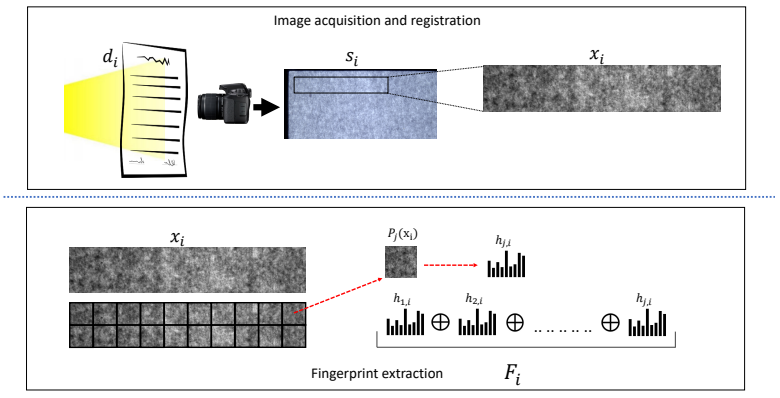


Figure 1. Overall pipeline of the proposed framework. First row describes the process to acquire documents; second row shows the fingerprint extraction process.

2.1. Document Digitization and Image Registration Considerations

The physical set of N documents $\mathfrak{D} = \{d_1, d_2, ...d_N\}$ was acquired using devices that are able to capture the wood fiber pattern by exploiting the translucent properties of the paper. In this work, two different acquisition environments were employed to compare the performance of low-end and high-end equipment. Details about devices and related settings are provided in Section 3. The acquisition of a physical document d_i was carried out in a semi-constrained environment; specifically, the documents must be roughly aligned regarding the capturing device to guarantee an effective consequent registration. For the sake of readability, the set of the digitized versions of the documents $\mathfrak{D} = \{d_1, d_2, ...d_N\}$ can be defined as $\mathfrak{S} = \{s_1, s_2, ...s_N\}$.

To successfully analyze the wood fiber pattern of a document d_i, the related digital image s_i must be registered. This step is critical as the paper fingerprint strongly depends on spatial information; hence, one must ensure that if a given document is acquired multiple times under the same setup, the system will process exactly the same region of the paper surface. To this aim, reference points were exploited (e.g., black bands in the acquired image) to rotate and properly crop s_i (see Section 2.2 for more details). After registration, a $W \times H$ sample from each document s_i was obtained, defined ad x_i, and the related set $X = \{x_1, x_2, ..., x_N\}$ was employed to build the fingerprint.

2.2. Extracting a Unique Fingerprint

The extraction of a unique fingerprint from a sample x_i is the process that encodes the texture information in such a way as to satisfy the following properties: (i) low complexity; (ii) encoding capabilities; (iii) robustness with respect to the missing parts. To this aim, the LBP descriptor and its variants [14] are employed, which are demonstrated to satisfy all the aforementioned requirements. These descriptors guarantee high capabilities in terms of discriminative power by maintaining low computational complexity and working almost perfectly even in the presence of slight variations on textures. In particular, LBP is a local descriptor that compares a pixel, called a pivot, to its n neighbors along the circle defined by a certain radius r [13]. In recent years the use of LBP for texture classification has grown, and a wide set of LBP variants has been proposed [14]. Hence, the so-called f-BP variant has the goal to improve the accuracy and the robustness for a specific task. The well-known local property makes the f-BP a flexible descriptor even in the presence of small perturbations, which is the fundamental requirement of the fingerprint we are looking for. Regardless of the f-BP, after pattern extractions the final descriptor is obtained by counting the times each pattern occurs, namely by computing a histogram.

Histograms are compact and effective descriptors for a various number of tasks; nevertheless, they heavily discard spatial information. To face this issue, x_i is first divided in M non-overlapping $p \times p$ patches and the histogram is separately calculated for each patch $P_j(x_i)$ with $j = \{1, 2, ..., M\}$; hence the histogram $h_{i,j}$ represents the histogram of the j patch of the sample x_i. The importance of spatiality is easily guessed: if the document presents some types of fault (e.g., missing parts, tears, holes, noise) it is important they do not affect the whole fingerprint, but just a portion. For this reason, the choice of the patch size p and the hyperparameters θ_f of the employed f-BP variant (e.g., the number of neighbors n and the radius r) have consequences on the performance. The size T of the histogram depends on the number of possible patterns the f-BP variant led. For example, if one employs classical LBP with $n = 8$ and $r = 1$ the number of possible patterns, and the histogram size T, is 256. As far as the patch size p is concerned, large patches decrease spatial information while small patches make the BP excessively local and increase the complexity of the obtained fingerprint.

The final fingerprint F_i for document d_i can be obtained by concatenating all the histograms $h_{j,i}$ for $j = 1, 2, ..., M$:

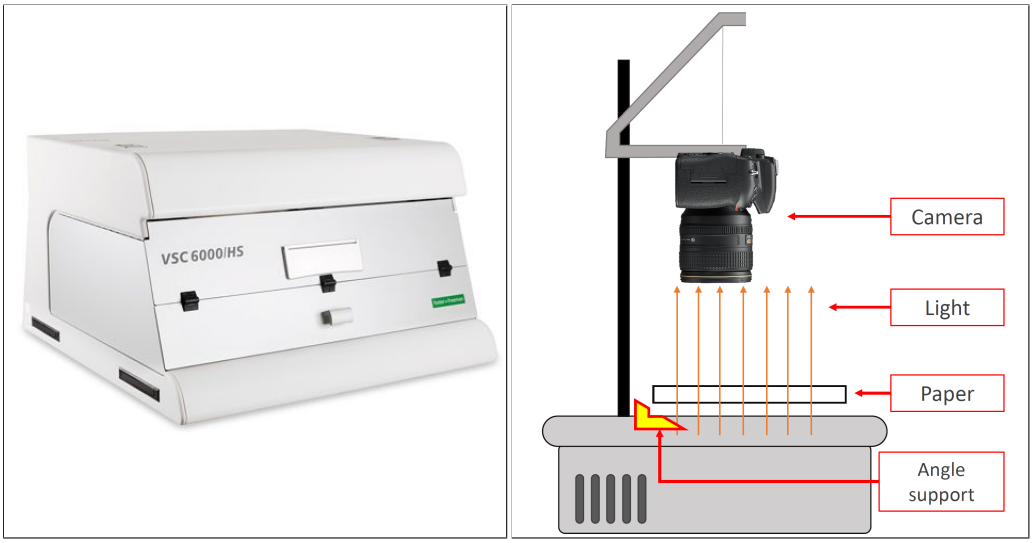
$$F_i = \bigoplus_{j=1}^{M} h_{j,i} \tag{1}$$

The size K of F_i is $K = M \times T$, as M patches are obtained from M histograms of size T. The goal of this study is to test different f-BP variants and look for the parameters $\{W, H, p, \theta_f\}$ which led to the most robust fingerprint.

3. Datasets for Document Identification and Fingerprint Testing

To evaluate the proposed approach and provide a great contribution to this research field, a new dataset is introduced, which is composed by 200 A4 paper sheets arranged in groups of 40 and divided into 5 non-overlapping classes. Each class is defined by two attributes: the manufacturer of the paper $b \in \{b_1, b_2, b_3, b_4\}$ and the weight or grammage (measured in g/m^2) $g \in \{80, 160, 200\}$. Thus, the obtained classes are the following: $(b_1, 80)$, $(b_2, 80)$, $(b_3, 80)$, $(b4, 160)$, $(b4, 200)$.

All the 200 documents in $\mathfrak{D}$ were then acquired multiple times using two different devices as described in Figure 2 and detailed in the next subsections. The dataset will be made available online after this paper is accepted and a download link will be placed in this section.



(**a**) Video Spectral Comparator 6000 (**b**) Backlight Imaging Tool

Figure 2. Devices employed for acquisitions.

3.1. Devices

To compare the performances obtainable with high-end and low-end equipment, each document is digitized using two different devices. For the high-end case the Video Spectral Comparator 6000 (VSC) was employed while for the low-end one we used the Backlight Imaging Tool (BIT): a cheap overhead projector combined with a digital camera that we accurately designed.

The VSC consists of a main unit (Figure 2b) connected to a standard workstation. It provides several functionalities and a set of different light sources to highlight paper details normally not visible in standard conditions. Table 1 shows VSC acquisition settings.

The BIT consists of an overhead projector which serves as source light and a consumer RGB camera hung on the projector arm. The employed camera is a Nikon D3300 equipped with a Nikon DX VR 15 mm–55 mm 1: 3.5–5.6 GII lens. Settings details are listed in Table 2.

Table 1. VSC Settings.

Light	Longpass	Mag	Exposure	Brightness
Transmitted	VIS	2.18	Auto	60

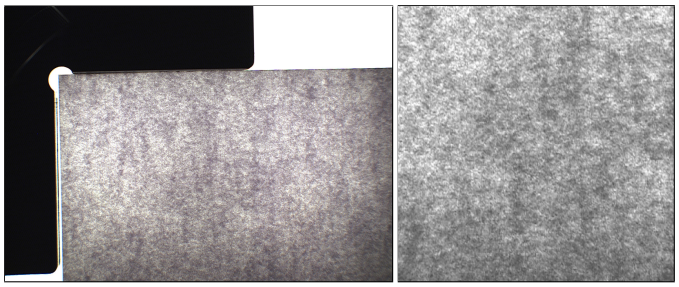
Table 2. BIT Settings.

Acquisition	Exposure Time	Opening/ISO/VR	Exposure Compensation	White Balance
RAW + JPEG	1/25	F29/100/ON	−5.0	Incandescence

3.2. Dataset Acquisition

For the sake of clarity, the terms $\mathfrak{S}_{VSC}$ and $\mathfrak{S}_{BIT}$ will be employed for referring to the digital acquisitions made by the VSC and the BIT respectively. The overall dataset acquisition pipeline is depicted in Figure 1. As expected, $\mathfrak{S}_{VSC}$ and $\mathfrak{S}_{BIT}$ show different contrast and sharpness.

$\mathfrak{S}_{VSC}$ consists of 200 documents acquired twice, for a total of 400 acquisitions (Table 3). The result of a single acquisition is a bitmap image of 1292 × 978 pixels and 300 dot per inch (dpi), as reported in Figure 3a. $\mathfrak{S}_{BIT}$ consists of 200 documents acquired 8 times. However, the insufficient power of light in the BIT does not allow the extraction of the translucent pattern from paper with grammage 160 or 200. Thus, only the 120 documents with grammage 80 were considered for a total of 960 acquisitions with a resolution of 6000 × 4000 pixels and 300 dpi (Table 3). Figure 4a shows a raw acquisition, where the black bands, used for image registration, are visible.



(**a**) Before registration (**b**) After registration

Figure 3. Document acquisition with VSC before registration (**a**) and after registration (**b**).

(**a**) Before registration (**b**) After registration

Figure 4. Document acquisition with BIT before (**a**) and after registration (**b**).

3.3. Image Registration

The acquisition of the black bands outside the paper area surface was voluntarily performed to distinguish selectively the pixels from the external area and easily obtain a registered set of images. All the raw images in $\mathfrak{S}_{VSC}$ and $\mathfrak{S}_{BIT}$ were converted into grayscale. First, a luminance threshold is used to find the top-left corner (y_0, y_1) of the sheet of paper. Secondly, the image anchored in position $(y_0 + u, y_1 + u)$ is cropped, where u is the minimum offset to perform a cropping by excluding the external area. The value of u is variable: the larger the external area acquired is, the greater will be its value. Images acquired by means of the VSC are cropped into patches of 400 × 400, while the ones acquired with the BIT are cropped into patches of 5000 × 1000 pixels. Finally,

one obtained X_{VSC}, the set of 400 registered samples from VSC and X_{BIT}, the set of 960 registered samples from BIT. Source examples are shown in Figures 3b and 4b.

Table 3. Dataset Table.

	VSC	BIT
CLASS	**ACQUISITIONS**	**ACQUISITIONS**
{e,80}	2	8
{f,80}	2	8
{u,80}	2	8
{m,160}	2	-
{m,200}	2	-
DEVICE IMG	$10 \times 40 = 400$	$24 \times 40 = 960$

4. Experiments and Discussion

To evaluate the proposed fingerprint extraction approach in depth, analysis of the datasets described in Section 3 were performed in terms of recognition tests. Since each document was acquired multiple times (i.e., twice for the VSC and 8 times for the BIT), a fingerprint reference dataset was built to face the recognition task; such reference datasets consist of only one sample per document while the rest of the samples were used for querying it. A certain document d with extracted fingerprint F_a will have a correct match with the closest element in the reference dataset F_b, if both F_a and F_b "belong" to the document d; in other words, a correct match occurs if s_a and s_b are different acquisitions of the same document. The recognition test performances are measured using the well-known accuracy metric defined as the rate of queries, which obtain a correct match. The adopted similarity measure for fingerprints was the Bhattacharyya distance [28], which is typically and effectively employed for problems where probability distribution must be compared. However, to better assess the effectiveness of the proposed fingerprint, four different recognition experiments are performed as detailed in the following. First, the original LBP was employed to compare the recognition accuracy on both datasets (VSC and BIT) obtaining the demonstration of device invariance. Given this result, a comparison was performed only on the BIT dataset employing LBP fingerprints computed as in [15] vs. the three other LBP variants, i.e., LTP [25], SBP [26] and CLBP [27]. Moreover, also the fingerprint robustness was investigated. To this aim, a challenging scenario was created where the query samples were intentionally altered by removing some pixels from the digital image to simulate physical damage of the paper (e.g., tears, holes). Finally, an optimization in terms of fingerprint dimensions was carried out and tested as well by exploiting principal component analysis (PCA) [29].

4.1. Dataset Comparison

To demonstrate the goodness of the LBP-based fingerprint extraction method, we started from the work of Guarnera et al. [15], our previous work, which represents the state of the art. Table 4 shows the overall accuracy obtained in the recognition tests performed on both datasets: 96.5% and 99.2% for VSC and BIT, respectively. Although samples from different datasets have different patch sizes, the best for both datasets was 100×100. This is a reasonable trade-off to preserve local spatial information. The accuracy on the BIT dataset is slightly higher than the accuracy obtained on the VSC. This demonstrates that the robustness of the fingerprint does not depend on the acquisition settings nor device.

Table 4. Best configuration parameters and accuracy of recognition test in VSC and BIT datasets.

Dataset	Patch Size	Number of Neighbors	Radius	LBP Type	Accuracy
VSC	100	32	42	uniform	96.5%
BIT	100	24	12	default	99.2%

4.2. Comparisons among LBP Variants

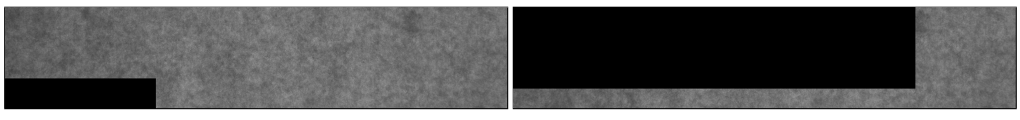
As introduced in Section 2, many LBP variants were proposed for texture analysis. Among them, LTP [25], SBP [26] and CLBP [27] were selected for the experiments described in this section. In the previous section, the independence of the proposed fingerprint from the acquisition device was demonstrated. Starting from this evidence, in the next experiments, only the BIT dataset will be employed given the higher number of available samples. The results in terms of accuracy are reported in Table 5 where CLBP and SBP show an improvement in terms of performance vs. LBP, by achieving an accuracy of 99.7% and 99.4%, respectively. It is worth noting that LBP is the employed method of [15] to extract the fingerprint, so the aforementioned results represent the overperformance with respect to the state of the art. As described in the literature, LTP tends to work better than LBP when the texture presents regions that are uniform (i.e., low variance). It is worth noting that the wood fiber patterns show a high variance, thus explaining the worse results of such descriptor. SBP, which is a generalization of the common binary pattern, as expected, obtains accuracy results of (99.4%) that are slightly better than LBP. Finally, the best performance was obtained by CLBP (99.7%) even if it delivers the largest fingerprint in terms of histogram dimensions (number of bins).

Table 5. Recognition test accuracy of the test carried out in BIT dataset employing LTP, SBP and CLBP.

	LTP	CLBP	SBP
Accuracy	90.83%	99.7%	99.4%

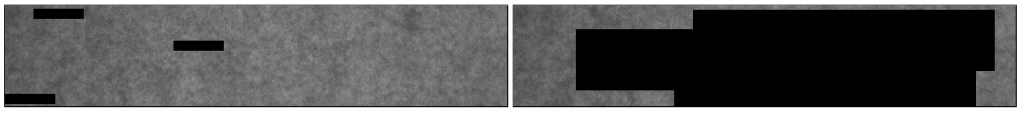
4.3. Tests on Noisy Environment: Synthetically Altered Documents Are Introduced

The proposed method for fingerprint extraction was tested under controlled conditions to properly assess what was expected to happen in real cases, namely when a document experienced some alteration between the first fingerprint extraction and the successive ones. Hence, the original fingerprint of the document may be very dissimilar from the latter one. To this aim, two types of damages on paper were simulated: tears and stain. The "tear" simulates a loss of information which starts from one angle of a sheet sample x_i by replacing such loss with black pixels, while the "stain" introduces random black blocks on the sample to simulate holes or stains. For both, the so-called degree represents the size of black area: the maximum degree corresponds to about 75% of the full sample to be removed (see Figures 5 and 6). Given the aforementioned alterations, a new recognition test on the BIT dataset was carried out, which includes 120 samples without any alterations on the fingerprint database and other 960 samples with alterations that were used to query the database. The results are reported in Figures 7 and 8 further proving the robustness of the proposed fingerprint, specifically the CLBP-based one which achieves best performance once more.



(**a**) Degree 1. (**b**) Degree 11.

Figure 5. Examples of altered documents with simulations of tear damage; in particular (**a**) represents the first degree of damage while (**b**) the last (e.g., 11).



(**a**) Degree 1. (**b**) Degree 11.

Figure 6. Examples of altered documents with simulations of stain damage; in particular (**a**) represents the first degree of damage while (**b**) the last (e.g., 11).

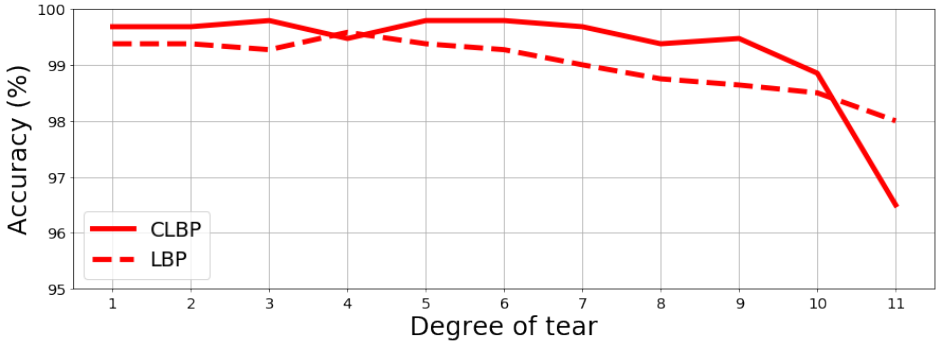


Figure 7. Accuracy employing CLB and LBP VS degrees of tear alteration.

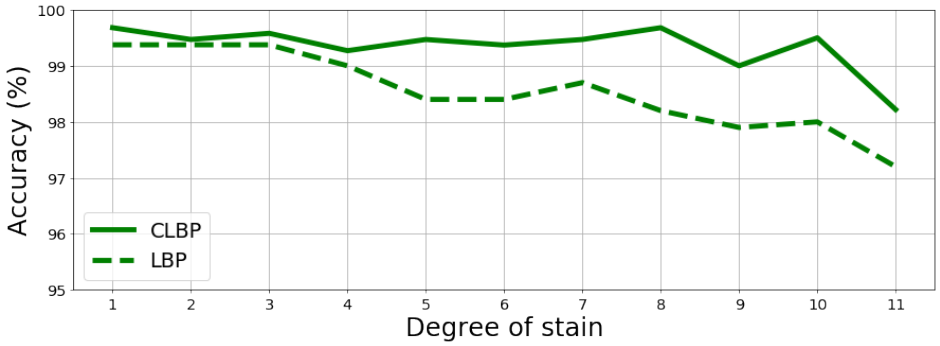


Figure 8. Accuracy of CLB and LBP VS degrees of stain alteration.

4.4. Fingerprint Dimensions Optimization

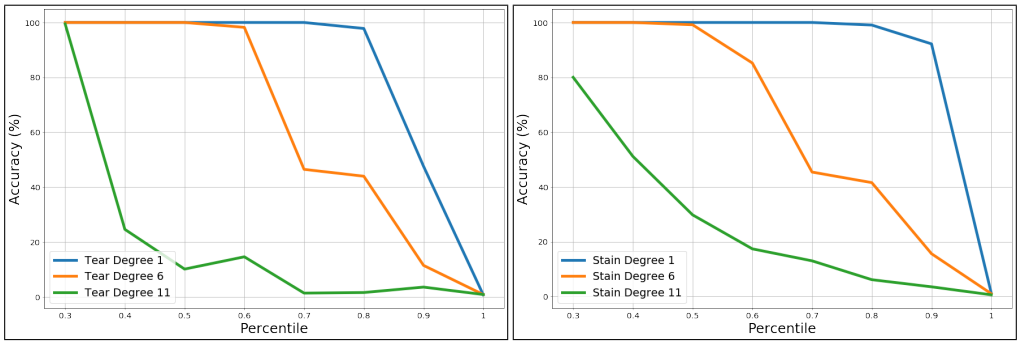
All the tests described in the previous sections were performed employing the pipeline described in Figure 1 with the following settings: images were cropped into patches of 100×100 pixels; number of neighbors for CLBP were $n = 12$ and radius was $r = 6$. These settings brought to 500 patches from the BIT dataset; thus, a histogram of 8194 elements was computed for each patch. This results in a fingerprint with dimension of $500 \times 8194 = 4{,}097{,}000$ elements whose storage occupancy is about 8.3 MB. Since the fingerprint with the proposed method could be even larger depending on parameters and since large fingerprints decrease efficiency, some optimization strategies to reduce it were explored.

The simplest strategy to reduce the fingerprint size was the increment of the patch size p; however, this could not guarantee the same accuracy performance. Table 6 shows the results obtained using larger values of p while monitoring the fingerprint size. The analysis of the results showed that the setting with $p = 200$ presents a performance similar to $p = 100$ (i.e., only a drop of 0.4% of accuracy) reducing the size by 25%, from 4,097,000 to 1,024,250 elements, that can be stored in 2.2 MB. However, as stated in the previous sections, employing bigger patches does not preserve spatial information and actually shows a tremendous accuracy drop (e.g., 67.6% for $p = 500$).

Table 6. Accuracy and size of CLBP fingerprints to vary of patch size.

Patch Size	Number of Bin	Accuracy	Storage Occupancy (MB)
100	4,097,000	99.7%	8
200	1,024,250	99.3%	2.2
250	655,520	97.9%	1.5
500	163,880	67.6%	0.4

To optimize the size of the fingerprint preventing a large loss in terms of accuracy, we employed the Principal Component Analysis (PCA) [29]. As we know, PCA reduces the dimensions by projecting each data point onto only some of the principal components to obtain lower-dimensional data while preserving most of the data variance; in a nutshell, it reduces the dimensions by preserving most of the information, which better describes a certain phenomenon. PCA is applied to each histogram $h_{j,i}$ previously obtained using CLPB. Hence, such histograms are drastically reduced in terms of dimensions. First, for testing purposes, all the 120 samples included in the fingerprint database are used to fit the PCA model. By employing the well-known explained variance analysis, we found that 95% of the information can be preserved using the first 32 principal components (also known as features), despite the original 8194. However, PCA moves histograms in a geometric space where the Bhattacharyya distance becomes less efficient; to face this problem, the recognition test was performed by means of the Euclidean distance. To verify the quality of reduction, the same recognition tests, as described in the previous sections, were carried out with the now-reduced fingerprints, delivering an accuracy of 97.97% with only 16,000 elements while maintaining the excellent performance of the not-reduced fingerprints case. It is worth noting that the PCA model was built using all the samples of each of the 120 documents in BIT. This could generate a PCA model overfitted on the data. Thus, a further test was performed using only the 50% of the dataset (60 documents) to fit the PCA model, while and we queried the reference dataset with the samples which come from the remaining 50%. In this case, it was found that 95% of information can be preserved using the first 40 principal components for each patch. recognition tests confirmed the results obtained with the PCA model built on all the 120 documents (97.97% of accuracy). It is important to note that although in the fingerprint comparison we also consider the missing parts when an alteration occurs, this does not heavily affect the Bhattacharyya distance between two fingerprints. On the contrary, the Euclidean distance is affected by this. In fact, the Euclidean distance calculated between an unaltered fingerprint of a document and an altered fingerprint of the same document exhibits extremely higher values, which impacts on the accuracy performance. To overcome this latter problem, a custom Euclidean distance was employed, where only a part of the fingerprint elements is considered in distance computation. Specifically, the differences between each element of the two fingerprints is computed and, subsequently, we sorted those differences by considering only a certain percentage of the lower ones. This percentage depends on the dimensions of the missing part, but this information is known by the operator during the identification phase, because in a real document the altered parts are visible. Figure 9a,b report the accuracy (vertical axis) when varying the percentage of elements included in distance computation (horizontal axis). The obtained results also suggest how to maintain a high accuracy according to the alteration degree. For example, in an average scenario of damage (orange lines) the 50% of distance is needed to maintain the accuracy over the 99%.



(**a**) Tear damage. (**b**) Stain damage

Figure 9. Accuracy variability of different percentiles on tear (**a**) and stain (**b**) damages.

5. Fingerprint Robustness Analysis

The carried-out recognition tests started from the hypothesis that every query fingerprint F_q could find a correspondent fingerprint F_x into fingerprints database previously extracted from the same document and stored. A real case scenario could present some differences: the query fingerprint F_q could not find a correspondent F_x and then the nearest one has no meaning (it is the most similar but it is a fingerprint extracted from another document). Hence, additional information is needed: given the distance $\bar{e}$ between two samples. To solve this problem, $\bar{e}$ was analyzed in all the previously presented experiments; in particular, starting from the fingerprints extracted by the images acquired with BIT device (e.g., 960), the distances obtained in the tests without simulated damages employing CLBP and LBP were analyzed, considering three kinds of distances:

- $\bar{e}_0$: distance obtained between F_q and F_x, both extracted from the same document, when F_x is the closest fingerprint in the recognition test.
- $\bar{e}_1$: distance obtained between F_q and F_x, both extracted from the same document, when F_x is not the closest fingerprint in the recognition test.
- $\bar{e}_{null}$: distance obtained between F_q and F_x, extracted from different documents.

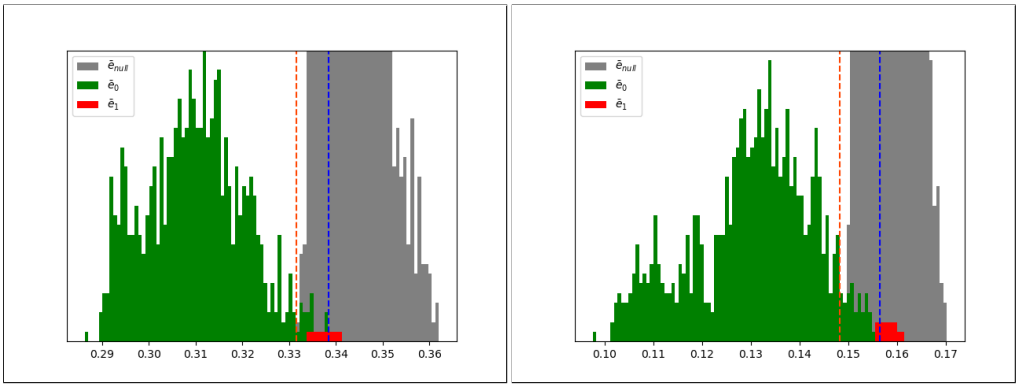
Given 840 different F_q, 120 distances have been computed for each of them. For every F_q analyzed, two results were obtained:

- the closest fingerprint F_x is extracted from the same document of F_q, and then the distance between them is classified as $\bar{e}_0$ and the others 119 distances are classified as $\bar{e}_{null}$.
- the closest fingerprint F_x is not extracted from the same document of F_q, and then the distance between them is classified as $\bar{e}_1$ and the others 119 distances are classified as $\bar{e}_{null}$.

It is easy to figure out that the population of $\bar{e}_{null}$ is much bigger than $\bar{e}_0$ and $\bar{e}_1$, whose sum is exactly 840.

Figures 10a,b represent the plot of distances $\bar{e}_0$, $\bar{e}_1$, $\bar{e}_{null}$ in both tests (LBP and CLBP). The plots have been cut because the populations are unbalanced and because the focus of the analysis is on the intersections of the two curves. In those plots it is possible to detect two Gaussians almost fully separated. The intersection between them (the tail of green Gaussian, delimited by orange and blue lines) represents an uncertainty zone. It is worth noting that the position of $\bar{e}_1$ in both cases (LBP and CLBP) is within this zone that confirms the meaning of distance: lower will be the distance with the nearest fingerprint and greater will be the possibility that the fingerprints are extracted from the same document. Naturally, the concept of low/great depends on the descriptor employed; in the forensics domain it is important the measure of the degree of uncertainty whenever it is available. The percentage of uncertainty zone z and the percentage r of $\bar{e}_0$ inside it gives a further degree of confidence and it is variable for each descriptor. Given a descriptor the couple (z,r) can be employed to describe the robustness of it. CLBP has the $\bar{e}_0$ range between 0.286 and 0.338 and uncertainty zone between 0.331 and 0.338 and then $z = 13.46\%$, while $r = 2.62\%$

due to 22 $\bar{e}_0$ inside uncertainty zone on 837 total; LBP has $z = 13.56\%$ and $r = 4.92\%$. Table 7 shows the analysis for every binary pattern tested.



(**a**) Tear damage. (**b**) Stain damage

Figure 10. Accuracy variability of different percentiles on tear (**a**) and stain (**b**) damages. For both the plots x-axis represents the values of the distances obtained and y-axis the number of occurrences. $\bar{e}_{null}$, $\bar{e}_0$ and $\bar{e}_1$ are represented by gray, green and red respectively.

Table 7. Percentage of uncertainty zone (z) and percentage of $\bar{e}_0$ inside it (r) for each analyzed descriptor.

Descriptor	z	r
LBP	13.93 %	4.91 %
CLBP	13.32 %	2.62 %
SBP	15.72 %	3.13 %
LTP	91.38 %	91.05 %

Moreover, a cross-dataset analysis has been conducted to understand if there is a correlation between input and descriptor efficiency. The textures with the distance within z have been analyzed: CLBP has 22 distance on 837 while LBP 41 on 835. 13 are shared while others are close to z meaning that bad texture (in terms of acquisition) will have a bad distances (close or within z), independently from the descriptor.

6. Conclusions

In this paper, a novel approach for document identification was proposed. The method employs variants of binary pattern descriptors (e.g., LBP, LTP, SBP, CLBP) to obtain a proper fingerprint to uniquely recognize the input document, but at the same time, be easily manageable. For this reason, an additional analysis was conducted to optimize the fingerprint in terms of dimensions; it was based on PCA which has confirmed almost the same degree of confidence, reducing the fingerprint size to less than 1/100 of the original. To demonstrate the robustness of the method, the dataset was expanded by including more noisy samples, demonstrating the value of the proposed technique in real case scenarios and better results with respect to the state of the art. Finally, a further analysis on the meaning of distances was conducted, to generalize the recognition test.

Author Contributions: Conceptualization, F.G. and O.G.; Data curation, F.G. and D.A.; Investigation, F.G., O.G. and D.A.; Methodology, F.G. and O.G.; Resources, O.G., V.M. and A.S.; Software, F.G.; Supervision, O.G. and S.B.; Validation, F.G., O.G. and S.B.; Writing original draft, F.G.; Writing review and editing, O.G., D.A., F.S., S.B., S.L., V.M. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank Raggruppamento Carabinieri Investigazioni Scientifiche, RIS di Messina for providing the VSC®6000 instrumentation and support and iCTLab s.r.l. (Spinoff of University of Catania) for help during the dataset creation. Both were fundamental also for their insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cheddad, A.; Condell, J.; Curran, K.; Mc Kevitt, P. Combating digital document forgery using new secure information hiding algorithm. In Proceedings of the International Conference on Digital Information Management, London, UK, 13–16 November 2008; pp. 922–924. [CrossRef]
2. Ahmed, A.G.H.; Shafait, F. Forgery Detection Based on Intrinsic Document Contents. In Proceedings of the International Workshop on Document Analysis Systems, Tours, France, 7–10 April 2014; pp. 252–256. [CrossRef]
3. Berenguel, A.C.; Terrades, O.R.; Lladós, J.C.; Cañero, C.M. Banknote Counterfeit Detection through Background Texture Printing Analysis. In Proceedings of the IAPR Workshop on Document Analysis Systems, Santorini, Greece, 11–14 April 2016; pp. 66–71. [CrossRef]
4. Bruna, A.R.; Farinella, G.M.; Guarnera, G.C.; Battiato, S. Forgery Detection and Value Identification of Euro Banknotes. *Sensors* **2013**, *13*, 2515–2529. [CrossRef] [PubMed]
5. Gill, N.K.; Garg, R.; Doegar, E.A. A review paper on digital image forgery detection techniques. In Proceedings of the International Conference on Computing, Communication and Networking Technologies, Delhi, India, 3–5 July 2017; pp. 1–7. [CrossRef]
6. Kumar, M.; Gupta, S.; Mohan, N. A computational approach for printed document forensics using SURF and ORB features. *Soft Comput.* **2020**, *24*, 13197–13208. [CrossRef]
7. Berenguel, A.C.; Terrades, O.R.; Lladós, J.C.; Cañero, C.M. Identity Document and banknote security forensics: A survey. *arXiv* **2019**, arXiv:1910.08993.
8. Battiato, S.; Giudice, O.; Paratore, A. Multimedia forensics: Discovering the history of multimedia contents. In Proceedings of the 17th International Conference on Computer Systems and Technologies 2016, Palermo, Italy, 23–24 June 2016; pp. 5–16.
9. Samsul, W.; Uranus, H.P.; Birowosuto, M.D. Recognizing Document's Originality by laser Surface Authentication. In Proceedings of the International Conference on Advances in Computing, Control and Telecommunication Technologies, Jakarta, Indonesia, 2–3 December 2010; pp. 37–40. [CrossRef]
10. Sharma, A.; Subramanian, L.; Brewer, E.A. PaperSpeckle: Microscopic fingerprinting of paper. In Proceedings of the ACM Conference on Computer and Communications Security, Chicago, IL, USA, 17–21 October 2011; pp. 99–110. [CrossRef]
11. Toreini, E.; Shahandashti, S.F.; Hao, F. Texture to the Rescue: Practical Paper Fingerprinting based on Texture Patterns. *ACM Trans. Priv. Secur.* **2017**, *20*, 1–29. [CrossRef]
12. Petrovska-Delacrétaz, D.; Jain, A.K.; Chollet, G.; Dorizzi, B. *Guide to Biometric Reference Systems and Performance Evaluation*; Springer: London, UK, 2009.
13. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [CrossRef]
14. Brahnam, S.; Jain, L.C.; Nanni, L.; Lumini, A. (Eds.) *Local Binary Patterns: New Variants and Applications*; Springer: Berlin/Heidelberg, Germany, 2014. [CrossRef]
15. Guarnera, F.; Allegra, D.; Giudice, O.; Stanco, F.; Battiato, S. A New Study On Wood Fibers Textures: Documents Authentication Through LBP Fingerprint. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 4594–4598. [CrossRef]
16. Buchanan, J.D.; Cowburn, R.P.; Jausovec, A.V.; Petit, D.; Seem, P.; Xiong, G.; Atkinson, D.; Fenton, K.; Allwood, D.A.; Bryan, M.T. Forgery: Fingerprinting documents and packaging. *Nature* **2005**, *43*, 475. [CrossRef] [PubMed]
17. Van Beijnum, F.; Van Putten, E.G.; Van der Molen, K.L.; Mosk, A.P. Recognition of paper samples by correlation of their speckle patterns. *arXiv* **2006**, arXiv:physics/0610089.
18. Cowburn, R. Laser Surface Authentication-natural randomness as a fingerprint for document and product authentication. In Proceedings of the Optical Document Security Conference, San Francisco, CA, USA, 23–25 January 2008.
19. Cowburn, R. Laser surface authentication—Reading Nature's own security code. *Contemp. Phys.* **2008**, *49*, 331–342. [CrossRef]
20. Clarkson, W.; Weyrich, T.; Finkelstein, A.; Heninger, N.; Halderman, J.A.; Felten, E.W. Fingerprinting blank paper using commodity scanners. In Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA, USA, 17–20 May 2009; pp. 301–314. [CrossRef]
21. Wong, C.W.; Wu, M. Counterfeit Detection Based on Unclonable Feature of Paper Using Mobile Camera. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1885–1899. [CrossRef]
22. Liu, R.; Wong, C.W.; Wu, M. Enhanced Geometric Reflection Models for Paper Surface Based Authentication. In Proceedings of the IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 11–13 December 2019. [CrossRef]
23. Chen, D.; Hu, Q.; Zeng, S. An Anti-Counterfeiting Method of High Security and Reliability Based on Unique Internal Fiber Pattern of Paper. In Proceedings of the 2020 IEEE 14th International Conference on Anti-Counterfeiting, Security, and Identification (ASID), Xiamen, China, 30 October–1 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 174–178.

24. Haist, T.; Tiziani, H.J. Optical detection of random features for high security applications. *Opt. Commun.* **1998**, *147*, 173–179. [CrossRef]
25. Tan, X.; Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. In Proceedings of the Analysis and Modeling of Faces and Gestures, Rio de Janeiro, Brazil, 20 October 2007; pp. 168–182. [CrossRef]
26. Nguyen, T.P.; Vu, N.S.; Manzanera, A. Statistical binary patterns for rotational invariant texture classification. *Neurocomputing* **2016**, *173*, 1565–1577. [CrossRef]
27. Guo, Z.; Zhang, L.; Zhang, D. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **2010**, *19*, 1657–1663. [CrossRef] [PubMed]
28. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.
29. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *2*, 559–572. [CrossRef]