Local Binary Patterns for Printer Identification based on Texture Analysis

Weina Jiang,
Anthony TS Ho,
Helen Treharne,
Yun-Qing Shi

September 2011

University of Surrey

Department of Computing

Computing Sciences Report

CS-11-05

Local Binary Patterns for Printer Identification based on Texture Analysis

Weina Jiang, *Student Member, IEEE,* Anthony TS Ho, *Senior Member, IEEE,*
Helen Treharne,Yun Qing Shi, *Fellow, IEEE*

***Abstract*—This paper proposes a texture analysis of the printed document based on Local Binary Pattern (LBP) descriptor for the application of printer identification. The LBP provides a statistical description of the pixels' gray level differences within their neighborhoods. The occurrence histogram of local binary patterns is able to capture the document's texture modifications by the distortion during the printing-and-scanning process, such as halftoning, geometric distortion, and mechanical defects. The most frequently appeared local binary patterns represent bright or dark flat regions. Furthermore, Gou *et al.* proposed an approach based on the combination of three different types of statistical features for scanner identification. We deconstruct their approach in order to evaluate the effectiveness of each type of features for printer identification.**

Our proposed LBP descriptor based model provides an excellent identification rate at approximately 99.4%, with a low variance. These results were achieved by Support Vector Machine (SVM) classification via n-fold cross validation and *leave one out*. They exceed any of the results obtained using the features, employed by the Gou *et al.* approach either singularly or in combination. Our experiments were conducted on 350 printed images, as well as 350 printed text documents, by a set of similar printers, two of which were exactly identical. The proposed model remains robust against common image processing, including averaging filtering, median filtering, sharpening, rotation, resizing, and JPEG compression.

***Index Terms*—Authentication, Printer Identification, Digital Forensic, Local Binary Patterns, SVM**

I. Introduction

DEVICE identification focuses on the problem of which device produced a given media. To identify such a device for a given media, i.e., the type, brand, model and other characteristics of the device, has a significant contribution to legislation, insurance claims and fraud detection [1]. Forensic tools that unveil the origin and discover the authenticity are essential to forensic examiners. Forensic applications in law enforcement rely on a higher level of reliable and accurate source information. Thus, reliability, accuracy, and computational simplicity are highly desired. Device identification can be categorized into three classes: 1) Senor pattern noise feature based schemes [2], [3]; 2) Mechanical defect based schemes [4], [5]; 3) Statistical feature based schemes [6], [7].

In the literature, these three classes of features have been applied to various devices. The senor pattern noise feature based schemes were initially proposed for camera identification [2], [3]. Subsequently, they were applied to scanner identification [8]. Lukas *et al.* initially proposed a source camera identification scheme based on the pattern noise of an imaging sensor [2]. This sensor pattern noise provides the unique stochastic characteristic of the imaging sensor. The factors that generate the pattern noise can be pixel nonuniformity, optical interference, on-chip and off-chip noise [9]. Pattern noise can be extracted by subtracting a denoised image from its original version. A reference pattern noise that serves as an intrinsic signature is established from averaging pattern noise of the same scene images taken from the digital camera. The camera is detected based on a correlation approach by matching a maximum correlation between the image noise pattern and the camera reference pattern. An improved method for source camera identification is proposed on joint estimation and detection of photo-response nonuniformity (PRNU) [3] for the forgery detection in digital camera images.

This sensor noise pattern has also been adapted for scanner forensics [8]. For scanner forensics, sensor pattern noise of an image [2] is substituted by calculating locally, i.e., along row direction and along column direction. The average of the noise pattern in rows (row reference pattern), the average of the noise pattern in columns (column reference pattern), the correlation between the noise pattern in row and row reference pattern, the correlation between the noise pattern in column and column reference pattern, are defined as the noise patterns. The statistics, which include mean, standard deviation, skewness and kurtosis, are applied to these noise patterns to obtain their forensic features for scanner identification. In this paper we will analyze the key factors that preclude senor pattern noise feature based schemes being applied to printer identification.

The mechanical defect based scheme proposed for printer identification in [4] [10]–[12] investigated printing artifacts due to EP printers mechanisms. These artifacts are as a result of variations between mechanical devices involved in the printing process, such as optical photoconductor (OPC) drum, polygon mirror, gear eccentricity and gear backlash. Mikkilineni *et al.* [4], [10] and Khanna *et al.* [11], [12] reported the banding defect as a consequence of space variation between raster lines. They believed texture descriptions of printed document would capture the defect during EP printing process. A texture analysis based on Gray Level Co-occurrence Matrix (GLCM) together with pixel based measurements i.e., variance and entropy are employed as intrinsic features for SVM classification.

Bulan *et al.* also observed that the mechanical defect

Weina Jiang is with the Department of Computing, University of Surrey, Guildford, , GU2 7XH, UK e-mail: w.jiang@surrey.ac.uk.

Anthony TS Ho and Helen Treharne are with University of Surrey, Guildford, UK.

Yun-Qing Shi is with New Jersey Institute of Technology, Newark, NJ, US.

changes the spacing between printed halftone dots [5] which is originally generated by a halftoning algorithm. The cluster dithered halftoning algorithm is commonly employed by an EP printer by periodically varying the size of halftone dots to generate the illusion of different gray levels. The impact of the *halftoning algorithm* [13] is to generate a texture pattern from a dithering matrix or an error diffusion kernel. However, the texture pattern will be modified by a number of distortions during printing-and-scanning as discussed in this paper.

For the third class, *the statistical feature based scheme* is proposed for scanner identification [6]. In [6], three different types of statistical features were extracted on scanned images to perform scanner identification and forensic analysis. The features do not consider the source of scanning noise, which includes image denoising filtering, wavelet analysis and neighborhood prediction. Gou *et al.* believed that the features are able to capture the variations of pixel values in scanned images which could be induced by different types of noise. Similarly, Jiang *et al.* [7] proposed a forensic technique based on the multi-sized Benford's law (MBL) to identify the brands and models of printers from the printed-and-scanned images, at which the first digit probability distribution of multi-sized block DCT coefficients were extracted to constitute a feature vector as the input of SVM classifier. MBL can achieve almost the same classification accuracy as the mean and standard deviation of the denoised image [6] does for printer identification. However, it is less accurate at identifying printers of similar models [14].

In this paper, we analyze the statistical feature based scheme proposed by Gou *et al.* [6] as a generic algorithm to determine its applicability to printer identification. It is based on the assumption that even though the noises induced by a number of distortions in printing-and-scanning may corrupt the pixel values at a random amount, the statistical characteristics should be constant for documents printed by a single printer. The validation of the assumption is confirmed by our successful experimental results where in particular, we identify different combinations of features proposed by this scheme and analyze the performance of these combinations via an SVM classifer. From this we can confirm that the statistical features proposed in [6] is a generic algorithm and we identify the combination that achieves the highest accuracy for printer identification.

In this paper, we observe that the image pixel values vary among different printers and the pixels spread into different spatial structures. This observation complements what has been found in [5] and [4], [10], [12] that the effects of pixel size, the spacing between inter pixels and the spacing between raster lines contribute to the texture in the printed documents. The texture of a printed document established a unique pattern, which belongs to a specific printer. Therefore, we investigate the texture of printed documents based on Local Binary Pattern (LBP) descriptor for printer identification. LBP provides a statistical description of pixels gray level differences in a small neighborhood provided by an angular space (P, R), at radius R and angle of $360^\circ/P$. Certain local binary patterns, which represent bright and dark flat regions, are the most frequent patterns of printed documents. The LBP descriptor is found to be able to capture texture modifications due to different distortions during the printing-and-scanning. The occurrence histogram of these local binary patterns provides a powerful discrimination capability as an intrinsic feature for printer identification.

The printer identification is conducted based on identifying either images or text documents which are printed by different printers. In order to analyze these physically printed images or text document that may contain forensic characteristics of the printers, both the printed images and the printed text documents are scanned into digital format by one scanner. This is assumed that the interference of the scanner can be minimized. Therefore, we use printed-and-scanned images and printed-and-scanned text document to perform an empirical investigation and forensic analysis throughout the paper.

A novelty of the paper is to conduct a texture analysis based on LBP descriptor for printer identification. The proposed printer identification scheme is sucessfully demonstrated on both printed images and printed text documents. The overall workflow of our proposed approach is shown in Figure 1. Moreover, the identification is successfully verified between similar printers in the same model group providing the ability to perform fine-grained identification of printers. The identification accuracy of our LBP based scheme is presented and compared to the statistical features based scheme [6] using two SVM techniques: n-fold cross validation and *leave one out*. As a comparison of performance to [6], our proposed LBP descriptor based scheme can achieve an excellent classification accuracy with low variance, also with the benefit of computational efficiency.

The rest of this paper is organized as follows: In Section II, related work is analyzed. We address the infeasibility of sensor pattern noise feature to printing-and-scanning channel for printer identification; the statistical feature based forensic analysis for scanner and scanned images [6] is presented in Section II-B. In Section III, a texture analysis of the printed document based on LBP descriptor is proposed. In Section IV, we evaluate the significance of the features in [6] and apply feature selection for its application in printer identification. The classification results are analyzed and comprehensively compared with our proposed LBP descriptor based scheme. The clarification of the SVM classification between n-fold cross validation and *leave one out* is presented in Section IV-A, where we address the issue on how to construct a fair classification measure of a reliable and stable prediction accuracy.

II. Related Work

In this section, we analyze techniques for device identification summarized in the previous section. Sensor pattern noise [2] was proposed as an intrinsic feature for imaging sensor based device identification. This model has been applied in camera identification [2], [3] and scanner identification [8]. In section II-A, we examine two key factors of this model which highlight the infeasibility in adapting the model to printer identification.

In section II-B, we show that the use of statistical features can be adapted for use in printer identification. We do not

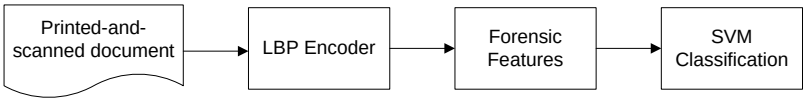


Fig. 1. Flowchart of Forensic Analysis for Printer Identification

believe that the features identified in [6] can only be considered applicable to scanner identification. Hence, we discuss the statistical features in more detail and the experimental results in Section IV-A2 will show that the use of these features can achieve a 93.14% accuracy of printer identification. The applicability of these features to printer identification extends the results of [6] and is one of the contributions of this paper.

A. Sensor Pattern Noise Feature

The sensor pattern noise is considered to be an intrinsic characteristic generated by imaging sensors. This pattern noise is independent of image content. A reference sensor pattern noise N_{ref} [2] of individual camera set is defined and calculated by averaging the sum of noise residuals of multiple images taken by the same camera as expressed in Equation (1). The noise residual $n(i,j),(i,j) \in [m,n]$ is obtained as being the subtraction between an image $I(i,j),(i,j) \in [m,n]$ and the denoised version $f(I(i,j))$ as defined in Equation (2). Two factors that are crucial to construct an accurate identification system, are the denoising filter and the averaging solutions to remove the random noise.

- *the denoising filter* The denoising filter f is based on a spatially adaptive statistical model where image wavelet coefficients are modeled to be independent Gaussian random variables with zero-mean and variance [15]. Assuming the image signal is transmitted over an Additive Gaussian Noise (AWGN) channel, the ratio between the noise clear image signal and noisy output is the filter transfer function. The variance of input image signal is estimated by the data points in square windows of local neighborhood. The noisy imaging sensor output consists of the noise pattern of the camera that captures the image, and the image content effects. Designing/applying an appropriate denoising filter for the underlying channel of camera imaging sensor, is crucial to removing the image content effects, which are dominant in the noisy output. In [2], Lukas *et al.* chose this denoising filter as it outperformed other filters, such as Wiener filter or median filter. As for printer identification, the image is present as a printed-and-scanned digital image. The assumption made previously is not valid, as the print-and-scan channel is not simply AWGN.

$$N_{ref} = \frac{1}{N} \sum_{i \leq M, j \leq N}^{k=1,..,N} n^{(k)}(i,j) \tag{1}$$

$$n^{(k)}(i,j) = I^{(k)}(i,j) - f(I^{(k)}(i,j)) \tag{2}$$

- *the averaging solution* Applying the averaging solution to denoised images aims to remove the random noise components from the fixed pattern noise [2]. This operation requires pixel-wise alignment. In the camera, the color filter array (CFA) is used to measure primary colors, i.e., red, green and blue. For the CFA generated camera pattern noise, it performs auto-alignment of image-to-image every time it is captured by a camera. However, the printer does not have a CFA, and the Equation (1) will not be satisfied. As images $I^k(i,j)$ are printed every time by the same printer, the averaging of the sum of noise residuals $n^{(k)}(i,j)$ of the k-th image, where $k \in [1,N]$, will require it to be the pixel-wise, i.e., (i,j), where $(i,j) \in [m,n]$, aligned correspondingly.
The sensor noise pattern feature based scheme in [2] assumes an AWGN channel. Peticolas *et al.* in [16] describes a resampling process that introduces errors found during a high resolution printing-and-scanning process. These errors are non-uniform geometric distortion and they do not fit an AWGN channel. Therefore, since the research in this paper is concerned with printer identification, it is not appropriate to use a senor noise pattern feature scheme as a basis for printer identification.

B. Statistical Feature

In [6], three different types of statistical features were proposed to capture the characteristics of scanner noise, without distinction between fixed pattern noise and random noise. The features includes the mean and standard deviation of the noise obtained by a set of denoising filters, the goodness of Gaussian fitting to the actual wavelet coefficients distribution in high frequency bands, and pixel value neighborhood prediction in scanned image smooth regions.

1) *the denoising filtering* The idea of using a denoising filter is explained in Equation 2 in Section II-A. Without knowing the explicit channel model, Gou *et al.* proposed using the denoising filters comprising linear filtering, median filtering and Wiener filtering, as shown in Table 1. The logarithm function, i.e., $\log_2$ is applied to the mean and the standard deviation of the noise which is captured by Equation (2). As for printing-and-scanning, we apply the filters as indicated in Table 1 to printed-and-scanned images. The SVM performance will be discussed in Section IV.

TABLE I
STATISTICS BASED ON THE DENOISING FILTERS

Denoising Filter	Statistics
Liner averaging filter Liner Gaussian filter Median filtering Wienar 3×3 Wienar 5×5	$\log_2(Mean)$ $\log_2(Standard\ deviation)$

2) *wavelet analysis* Once applying one-step Digital Wavelet Transform (DWT), an image is converted into four

subbands: LL subband, LH subband, HL subband and HH subband, respectively. It is noted that the HH, LH and HL subbands of DWT does not obey Gaussian distribution [17] [18] [19]. Gou *et al.* observed that the scanned digital photograph in the high frequency subbands of the DWT domain follows a Gaussian distribution as a consequence of scanning noise introduced by scanner [6]. As such, we applied the Gaussian fitting to the high frequency subbands of DWT coefficients to the printed-and-scanned images. The goodness of Gaussian fitting is varied as it depends on the brands and models of the printers.

3) *neighborhood prediction error* In [6], based on the assumption made by Gou *et al.*, an image pixel values in a smooth region, can be constantly predicted from its eight neighboring pixels with high precision. The noise may corrupt an image pixel value that may lead to some neighborhood prediction error. The mean and standard deviation of the predicted error was used as their last categorized features.
Gou *et al.* noted that these statistical features remained constant for a particular scanner. In this paper we assume that a similar observation can be made of printers. Thus, for any given printer, the statistical features representing the printer is the same. Therefore, in this paper we apply the approach of Gou *et al.* for printer identification. The results in Section IV-A2, which shows successful printer classification confirms that our assumption about the uniqueness of statistical features for a giving printer is a valid one. In the paper, we evaluate the significance of the three types of statistical features on the classification accuracy of printer identification. In particular, in Section IV-A2 we present classification accuracy results of different combinations of the three types of features. Additionally, a comparison of its performance with our proposed LBP algorithm is given in Section IV.

III. Local Binary Pattern for Printer Identification

Resolution can be identified by the width and height of the image together with the total number of pixels in an image. According to Nyquist sampling theorem [20], two times resolution scanning can capture the details of any pixel modified by the print defects. However, these modifications may present locally. Therefore, local texture analysis would be ideal to analyze the distorted documents during printing-and-scanning.

Texture analysis has been applied to image classification and segmentation [21]. The first order texture measures include mean and variance. The second order texture measures, such as GLCM [22] consider the relationship between groups of two pixels. A texture analysis based on GLCM [4] [10]–[12] is proposed to analyze gray level occurrences in a vertical direction covering pixels between one row to ten rows distance of text document for EP printer identification. Even though the direction can be extended into horizontal and diagonal directions, the description capability of gray levels of the number of pixels involved is much lower compared with LBP descriptor, at which the joint gray level differences distribution of pixels in a local neighborhood is defined in an angular space (P, R). The angular space is P equally spaced pixels on a circle of radius R, which forms a circularly symmetric neighborhood.

A. LBP Overview

The texture T in a local neighbor of a gray-scale image is defined [23] as the joint distribution of gray levels of P image pixels as shown in Equation (3), where g_c is the gray scale value of the central pixel of the P (from 0 to $P-1$) neighborhoods. P pixels are equally distributed on a circle of radius R which forms a circularly symmetric neighbor set, defined as $LBP_{P,R}$. Assume that g_c is located at (0,0), $g_p, p \in [0, P-1]$ will be placed at $\{R\cos(2\pi p/P), R\sin(2\pi p/P)\}$. As an example, for $(P, R) = (8, 1)$, the circular distribution of P pixels is shown in Figure 2.

$$T = t(g_c, g_0, ..., g_{P-1}) \tag{3}$$

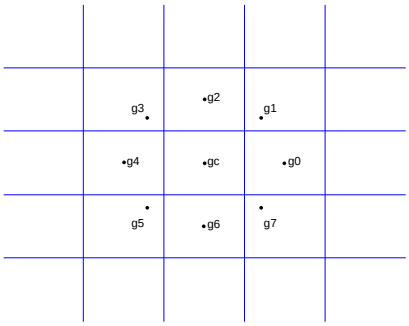


Fig. 2. Circularly symmetric neighbor pixels for P=8, R=1

- Properties of Grayscale Invariance
  By subtracting the gray value of the center pixel g_c from the gray values of the circularly symmetric neighborhood g_P (p=0,..,P-1), the local image texture becomes $t(g_c, g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c)$. Assuming g_c is independent of $g_p - g_c$, T can be approximated as

$$T \approx t(g_c)t(g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{4}$$

  In Equation (4), $t(g_c)$ describes the overall luminance of the image, unrelated to the local image texture. Therefore, the texture information in a local neighborhood is interpreted as a joint difference distribution as shown in Equation (5),

$$T \approx t(g_0 - g_c, g_1 - g_c, ..., g_{p-1} - g_c) \tag{5}$$

  Since taking the signed function $S(x)$ over the differences $g_p - g_c$,where $p \in [0, P-1]$, $S(g_p - g_c)$ would remain constant regardless of any changes of the main luminance. The invariance is achieved by scaling the gray scale where the signed differences substitute the differences of real pixel values. Therefore, the joint difference distribution is invariant towards any shifts in gray scale, as expressed in Equation (6).

$$T \approx t(s(g_0 - g_c), s(g_1 - g_c), ..., s(g_{p-1} - g_c)) \tag{6}$$

where

$$s(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0. \end{cases} \quad (7)$$

Finally, the joint difference distribution of P pixels becomes binary values by multiplying a factor of 2^P, as expressed in Equation (8).

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p. \quad (8)$$

The Equation (8) defines the local binary pattern derived from the image texture characteristics, by thresholding the pixels of a local neighborhood at its central pixel value. The pixel, which does not located at the centre of the pixels, is calculated by an interpolation algorithm as defined in [23]. The signed differences guarantee LBP invariance against any monotonic transformation of the gray scale.

B. Design of Uniform Sampling in LBP for Printer Identification

1) An example of printer distortions: The technologies used in printers vary among manufacturers, and on the specific application. Laser printers usually adapt a dithering halftoning algorithm due to its computational efficiency and its widely use for document processing. In contrast, inkject printers adapt an error diffusion halftoning algorithm which is based on a neighborhood operation. It is found to achieve a better quality as compared with dithering. Hence its application is commonly found in image processing. Phaser printers use Xerox Solid Ink technology. The print quality is considered to be excellent with early applications in the graphic design industry. Phaser printers are now mostly found in industrial markets.

A printed image carries information about the printer from where it originated. Besides the content it conveys, the information may include printing technologies, geometric distortion, print quality defects, toner usage and paper types. All of these factors will result in the ultimate texture reflected in the printed document.

During a printing-and-scanning process, a number of distortions may cause the degradation of an image from its original profile, in the form of gray level shifts and spatial location displacement. In order to discover the differences of a digital image before and after printing-and-scanning, an image of size 4000×4000 pixels with gray scale value of 128, was printed in 300 dpi by printers HP4250, HP4500, Xerox8500, respectively. This was followed by Infotec ISC 3535 scanning at 600dpi in JPEG format. For display purpose, three printed-and-scanned images of size 256×256 pixels are presented in Figures 3(a) 3(b) and 3(c). Starting at the top left of each image at pixel coordinates (10,11), the pixel values of each 8x8 image block of the corresponding Figures 3(a), 3(b) and 3(c), are displayed in Figure 3(d), 3(e) and 3(f) respectively. As observed, the pixel values in each block of the images printed by three distinct printers are significantly different. The original gray scale values of the pixels have changed from 128 and varied randomly between 0 to 255.

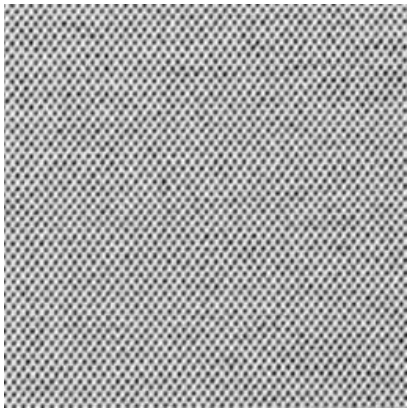
For the image in Figure 3(c) printed by the Xerox Phaser8500 printer, the visual dot structure is rather different compared with the other two images shown in Figure 3(a) and Figure 3(b), which are printed by HP laser printers. This may be due to the solid ink technology used by the Phaser printer, which is different from the laser or inkjet printer. In solid ink technology, ink or toner powder is softened, melted from a heated drum before it is transferred to the paper. The composition of the ink is uniquely made from food-grade processed vegetable oils, which is covered by a wax with a glossy surface [24].

The printed patterns in Figure 3(a) and Figure 3(b) share some similarities as they appear as a structure printed in dithered dots. However, the size of printed dots, the spacing between the dots and the spacing between the raster lines appear differently as the dots are more condensed in the image shown in Figure 3(b) than in image shown in Figure 3(a). This is due to the fact that before printing, image pixels are first converted into halftone patterns according to a fixed spatial structure determined by a specific halftoning algorithm embedded in the types of printers. For a laser printer, an ordered dither algorithm generates a binary halftone image by comparing pixels with a threshold value determined from a dithering matrix. The visual pattern would depend upon the size of the dithering matrix, the values of dithering matrix and also the spatial structure if it appears dispersed or clustered as named dispersed dithering or cluster dithering [13].

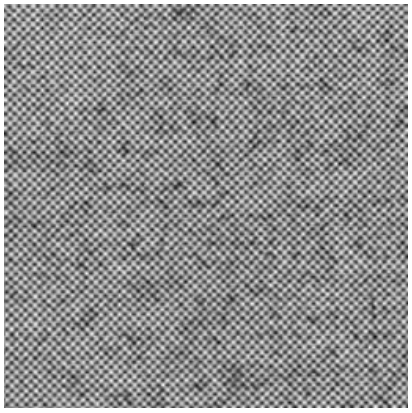
Another interesting observation in the images shown in Figure 3(a) and in Figure 3(b) is that the spacing between raster lines varied at some random locations, which looks like "scratches" in white lines. In [5] non-uniform spacing between raster lines is explained as the geometric distortion caused by the variations in laser scanning speed over a scan line, and variations in the velocity of the OPC drum. It is these differences between halftone dot positions before and after printing that form the basis of the approach to printer identification. In [4] these variations are referred to as banding defects. These variations are reflected in the texture of a printed document and a GLCM matrix is proposed to measure the texture.

The GLCM in [4] calculates the probability of gray level (grayscale intensity) value n and gray level value m occurred between two spatially placed pixels where one pixel is located vertically from one row to ten rows distance to another pixel. Besides of some basic statistics applied in GLCM, together with variance and entropy of the pixel values in printed area of character, a 22 dimensional features are proposed for texture analysis in printed text documents.

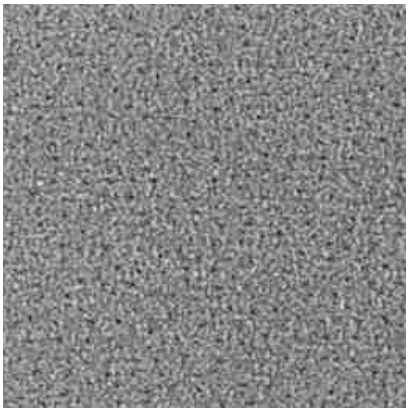
In comparison to GLCM, LBP provides graylevel spatial distribution in multidimension, since it counts the gray levels in a circularly symmetric neighborhood which is determined by an angular space (P, R), with P equally distributed pixels around a circle of radius R. As it considers the signed differences of gray levels rather than the gray levels of pixels in a local neighborhood, the LBP descriptor is also invariant against any monotonic transformation of the grayscale. Therefore, the LBP based descriptor suits perfectly in printing-and-scanning scenario, where the pixel values of an image



(a) Printed-and-scanned Midtone Grayscale Patch Displayed at Size 256x256 by HP LaserJet4250



(b) Printed-and-scanned Midtone Grayscale Patch Displayed at Size 256x256 by HP LaserJet4500



(c) Printed-and-scanned Midtone Grayscale Patch Displayed at Size 256x256 by Xerox Phaser8500

156	231	228	151	80	86	127	205
142	232	232	159	94	69	126	213
192	231	216	184	171	152	184	217
223	187	154	167	204	222	209	171
194	113	86	141	212	234	180	105
163	72	66	153	241	236	161	79
176	144	150	188	222	193	194	165
172	196	203	189	171	142	171	194

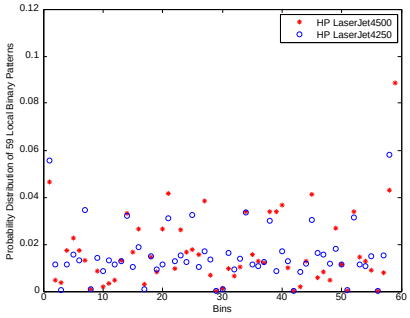
(d) Printed-and-scanned Midtone Grayscale Values at Size 8x8 by HP LaserJet4250

176	145	127	121	148	186	198	172
202	141	94	66	105	178	208	167
165	172	124	86	103	159	195	182
96	155	185	155	113	110	106	131
72	105	162	181	139	84	61	106
139	126	145	178	178	142	136	162
194	143	76	74	131	181	222	198
196	151	71	59	115	173	220	195

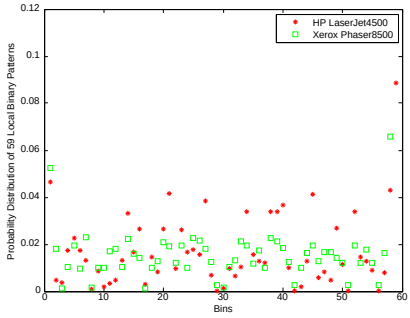
(e) Printed-and-scanned Midtone Grayscale Values at Size 8x8 by HP LaserJet4500

150	134	123	124	138	155	156	143
147	127	144	138	143	155	151	136
149	139	133	140	153	162	154	144
151	147	140	138	144	149	144	137
153	155	152	143	140	144	143	138
151	155	161	149	144	151	153	145
144	143	154	145	146	156	157	147
143	138	141	136	140	149	148	138

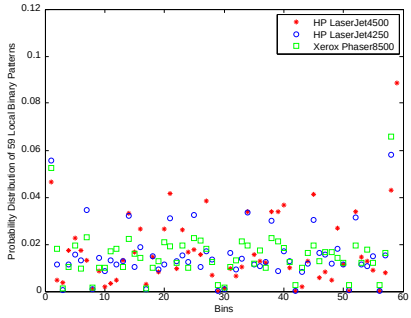
(f) Printed-and-scanned Midtone Grayscale Values at Size 8x8 by Xerox Phaser8500



(g) Probability Distribution of LBPs at (P,R)=(8,1) of Printed-and-scanned Midtone Grayscale Patch by HP LaserJet4250 and HP LaserJet4500 versus 59 Bins



(h) Probability Distribution of LBPs at (P,R)=(8,1) of Printed-and-scanned Midtone Grayscale Patch by HP LaserJet4500 and Xerox Phaser8500 versus 59 Bins



(i) Probability Distribution of LBPs at (P,R)=(8,1) of Printed-and-scanned Midtone Grayscale Patch by HP LaserJet4250 and HP LaserJet4500 and Xerox Phaser8500 versus 59 Bins

Fig. 3. Demonstration of Printed-and-Scanned Midtone Patches and Corresponding LBP Descriptions at (P,R)=(8,1)

will change and vary arbitrarily. Even though the printed-and-scanned image itself corrupts due to a number of causes, such as local geometric distortion, noise and mechanical defect, the device from which the image is generated, i.e., the printer, will leave a unique mark or signature that would be captured more clearly by a multidimensional texture analysis, such as LBP, than by GLCM based analysis. The construction of LBPs and parameter selection are presented in the following section.

2) Uniform Sampling and Parameter Selections: As shown in Equation (8), 2^P binary patterns will be generated in a P pixel neighborhood. Ojala *et al.* observed the pattern in the P pixel neighborhood has a uniformity value U [23]. This value is calculated based on the number of spatial transitions from "0" and "1" and vice versa. Ojala *et al.* refered to the pattern that have a value of at most 2, i.e., $U \leq 2$, as a uniform pattern. Therefore, instead of Equation (8), they

propose an LBP descriptor for texture classification as shown in Equation (9).

$$LBP_{P,R}^{descriptor} = \begin{cases} \sum_{p=0}^{P-1} s(g_p - g_c)2^p & U(LBP_{(P,R)}) \leq 2, \\ P+1 & \text{others.} \end{cases} \tag{9}$$

where

$$U(LBP_{(P,R)}) = |s(g_{p-1} - g_c) - s(g_0 - g_c)| \tag{10}$$

$$+ \sum_{p=1}^{P-1} |s(g_p - g_c) - s(g_{p-1} - g_c)| \tag{11}$$

In this paper, we propose to use LBP descriptor for printer identification by classifying the texture of printed document. An LBP encoding process is applied to every pixel of a printed document. Therefore, every grayscale pixel of a document is converted into a binary codeword. It is illustrated by a single block image with eight neighborhood pixels surrounded, i.e., $(P, R) = (8, 1)$, as shown in Figure 4. An LBP codebook is constructed based on LBP descriptor presented in Equation (9). As shown in Table II, an LBP codebook is demonstrated with $(P, R) = (8, 1)$ which consists of 58 uniform unique patterns plus one single pattern. In total, a 59-dimensinal histogram is constructed as an intrisic feature of a printed document.

The patterns of a printed document have U values which vary from 0, 2, 4 and 6. For example, the patterns "0000 0000" and "1111 1111" have U value of 0; the pattern "0000 0010" has U value of 2; the pattern "0000 1010" has U value of 4 and the pattern "0010 1010" has U value of 6. The patterns of which $U \leq 2$ are sequentially encoded from 1 to 58. These 58 patterns are the uniform pattern which stands for the smooth region in printed-and-scanned document. The remaining patterns with $U = 4$ and $U = 6$ are regarded as one single pattern. In Table II, the pattern "0000 1010" is enumerated as one of the patterns of which U value is equal to 4 and the pattern "0010 1010" is enumerated as one of the patterns of which U value is equal to 6. The feature employed in texture analysis is a histogram of these 59 patterns. There are 59 bins in the histogram. Each of the 58 uniform patterns falls into one bin. 58 uniform patterns are sequentially assigned from bin 1 to bin 58. The 59th bin of the histogram allocates the rest of patterns with $U = 4$ and $U = 6$. Hence, the LBP codebook is constructed with 58 individual uniform local binary patterns and a single local binary pattern which counted by a number of local binary patterns with the uniformity value $U > 2$ as presented in Table II.

The selection of (P, R) determines the number of pixels and their spatial structure enclosed in a small neighborhood. As the value of (P, R) becomes large, the number of the corresponding LBPs increases, and so does the value of uniformity U. As an example, for $(P, R) = (10, 1)$, the uniformity value U varies from 0, 2, 4, 6, 8, 10. The uniform patterns ($U \leq 2$) of LBPs are chosen and counted from 1 to 92. The rest of LBPs are treated as one single pattern, with the bin index 93. As demonstrated in Section III-C, we found a selection of (P, R) at which P varies from 8 to 16 and R varies from 1 to 2, and slightly vary the discrimination capability of LBP descriptor. Therefore, we choose a selection of $(P, R) = (8, 1)$ throughout the paper because of its computational simplicity. This is validated by our hypothesis study in Section III-C, where LBP descriptor is demonstrated on 350 printed images as well as 350 printed text documents.

For the printing-and-scanning process, we argue that the uniform pattern which represents the smooth region would capture the texture information. The smooth region, where the gray level differences in a local neighborhood are small, by LBP descriptor, is defined as a local binary pattern in which the number of spatial transitions is small, i.e., the pattern with $U \leq 2$. This pattern dominates the texture of a printed document which is confirmed by our hypothesis study shown in Section III-C. The non-smooth region, where the gray level differences in a local neighborhood are large, given by the LBP descriptor, is defined as a local binary pattern in which the number of spatial transitions is large, i.e., the pattern with $U > 2$. This pattern is more likely to change into a different pattern caused by distortions during a printing-and-scanning process. Therefore, a histogram of LBP descriptor provides an excellent discrimination capability for printer identification as demonstrated in Section IV.

C. Hypothesis Study

Considering the pixels in a 3x3 neighborhood, i.e.,$(P, R) = (8, 1)$, the uniform patterns allocated from the 1st bin to the 58th bin and the rest of the patterns falls into the 59th bin. Therefore, the 256 gray levels ranging from 0 to 255 is scaled into 59 gray levels ranging from 0 to 58. The histogram of 59 binary patterns with the bin index from 1 to 59, constitutes an intrinsic feature to identify the texture patterns in a printed-and-scanned document.

We present LBP descriptor at $(P, R) = (8, 1)$ of printed-and-scanned midtone patches in Figures 3(g) 3(h) and 3(i). The x-axis indicates the 59 bins which are constructed from 59 patterns of LBP descriptor at $(P, R) = (8, 1)$. The y-axis is the normalized histogram which counts the number of patterns that fall into each bin. As shown in Figure 3(g), from bin 1 to bin 59, there is only one case where the values generated by printers of HP LaserJet4250 and HP LaserJet4500 at y-axis that are identical, i.e., 0.0113 at $Bin = 50$. In Figure 3(h), from bin 1 to bin 59, there are no caes where the values generated by printers of HP LaserJet4500 and Xerox Phaser8500 at y-axis that are identical. In Figure 3(i), there are no cases where the values generated by the three printers are identical. Therefore, the LBP descriptor provides good discrimination capability for identifying the textures of images printed by three different printers. This is further verified by testing 350 images on seven printers, i.e., 50 images on each printer, as discussed below.

The occurrence probability of the 58 uniform patterns of total patterns is presented in Figure 6. The result is based on 350 images randomly selected in Uncompressed Image Database (UCID) [25], and printed by seven printers as shown in Table III. The selected uniform patterns with $U \leq 2$, contribute at an average rate of 88.54% of 59 total patterns for seven printers. For each printer, the percentage accuracy rate of a uniform pattern for 50 images is as follows:

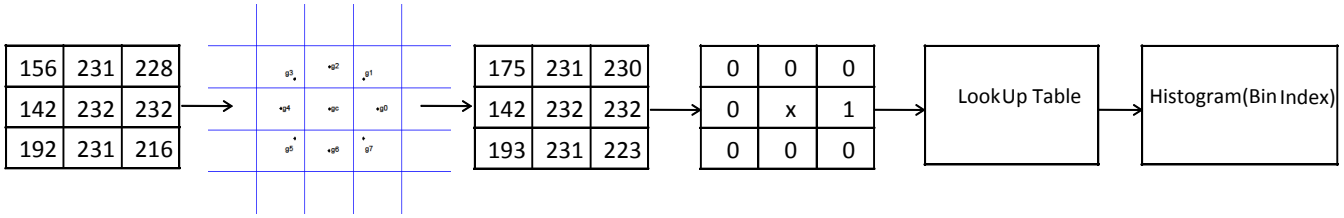


Fig. 4. an Illustration of LBP Encoder at $(P, R) = (8, 1)$ Converting an Image Block of 3×3 Pixels into a Binary Codeword

TABLE II
LBP CODEBOOK

Patterns	U Value	Bin Index	Patterns	U Value	Bin Index
00000000	0	1	00000011	2	31
00000010	2	2	00000111	2	32
00000100	2	3	00001111	2	33
00000110	2	4	00011111	2	34
00001000	2	5	00111111	2	35
00001100	2	6	01111111	2	36
00001110	2	7	10000001	2	37
00010000	2	8	10000011	2	38
00011000	2	9	10000111	2	39
00011100	2	10	10001111	2	40
00011110	2	11	10011111	2	41
00100000	2	12	10111111	2	42
00110000	2	13	11000001	2	43
00111000	2	14	11000011	2	44
00111100	2	15	11000111	2	45
00111110	2	16	11001111	2	46
01000000	2	17	11011111	2	47
01100000	2	18	11100001	2	48
01110000	2	19	11100011	2	49
01111000	2	20	11100111	2	50
01111100	2	21	11101111	2	51
01111110	2	22	11110001	2	52
10000000	2	23	11110011	2	53
11000000	2	24	11110111	2	54
11100000	2	25	11111001	2	55
11110000	2	26	11111011	2	56
11111000	2	27	11111101	2	57
11111100	2	28	11111111	0	58
11111110	2	29	00001010	4	59
00000001	2	30	00101010	6	

$87.66\% \pm 0.0098$ for HP 4200, $89.13\% \pm 0.013$ for HP4100(1), $88.74\% \pm 0.0205$ for HP4250, $88.56\% \pm 0.018$ for HP4100(2), $89.15\% \pm 0.0232$ for HP4015, $87.91\% \pm 0.0236$ for HP4500 and $88.65\% \pm 0.0107$ for Xerox8500 as shown in Figure 5.

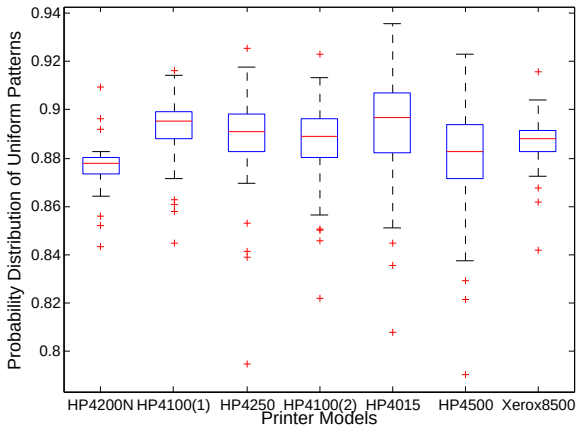
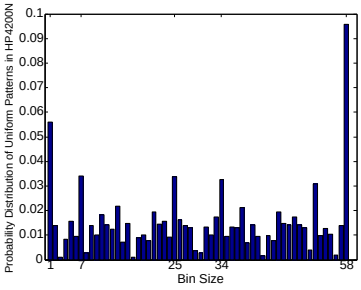


Fig. 5. Probability Distribution of Uniform Patterns

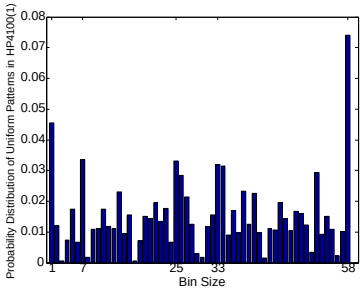
For each printer, the occurrence histogram of 59 LBPs is presented with the bin index given at the x-axis, which indicates the most frequently appeared uniform patterns out of 58 uniform LBPs as shown in Figure 6(a) to Figure 6(g). For seven printers, the most frequently appeared uniform pattern is "0000 0000" with an average rate at 4.45% and "1111 1111" with an average rate at 8.32% of total 58 uniform LBPs.

Furthermore, as discussed in Section III, the selection of (P, R) will cover a different number of neighborhood pixels located in a space determined by radius R and angle $360°/P$. Therefore, the LBP will result in different local texture binary patterns. As presented in Figure 7, we demonstrated the SVM classification rate at angular space of (P, R), where R=1 with P = 8,10,12,14,16, and where R=2 with P=8,10,12,14,16. As the number of P and R increase, the computation complexity rises accordingly. This is also true for purely text documents which are downloaded from copyright-free ebook Gutenberg project [26] as presented in Figure 8.

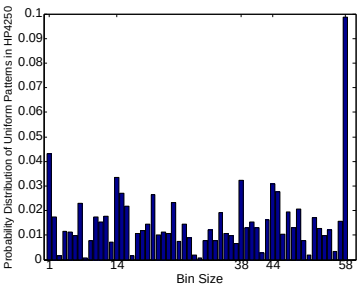
The contribution of our work focused on the LBP decriptor based texture analysis is able to capture the texture information



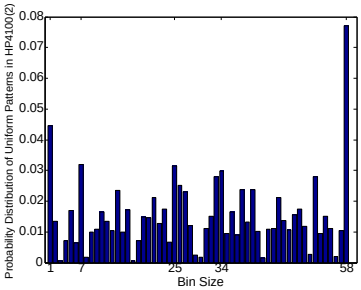
(a) image datasets at (P,R)=(8,1) by HP laserJet4200N



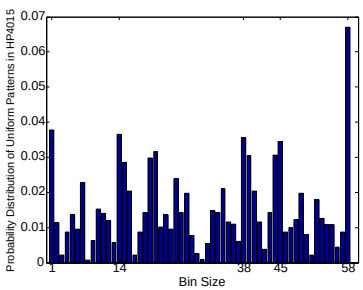
(b) image datasets at (P,R)=(8,1) by HP laserJet4100(1)



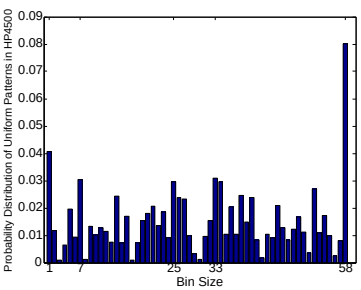
(c) image datasets at (P,R)=(8,1) by HP laserJet4250



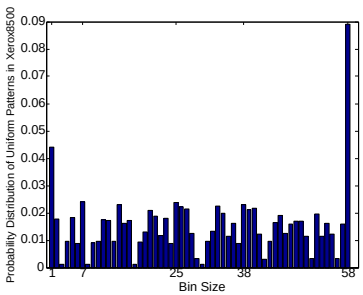
(d) image datasets at (P,R)=(8,1) by HP laserJet4100(2)



(e) image datasets at (P,R)=(8,1) by HP laserJet4015



(f) image datasets at (P,R)=(8,1) by HP laserJet4500



(g) image datasets at (P,R)=(8,1) by Xerox Phase 8500

Fig. 6. Occurrences Histogram of Uniform LBPs at $(P, R) = (8, 1)$ for Seven Printers

of printed documents. The occurrence histogram of the LBP decirptor is sufficiently suited as a feature detector for printer identification. For $(P, R) = (8, 1)$, The LBP descriptor scales gray levels from 0 to 255 to 1 to 59. The LBP descriptor divides the local texture patterns into 58 uniform patterns and one single pattern. The uniform patterns dominate at average 88.54% of total texture patterns. The most frequently appeared local binary patterns of uniform patterns represents as dark flat spots at an average rate of 8.32%, corresponding to "1111 1111" and bright flat spots at an average rate of 4.45%, corresponding to "0000 0000".

IV. Performance Evaluation and Comparisons

In this section, we demonstrate the performance of proposed LBP descriptor at $(P, R) = (8, 1)$ for printer identification with a comparison study of the statistical features based forensic analysis. The statistical features based forensic analysis was investigated and a set of features were selected for identifying printed-and-scanned images based on the criteria that the combination of selected features achieved a maximum SVM classification rate. Furthermore, the robustness of the proposed scheme was also tested against common post processing as presented in Section IV-D. Finally in Section IV-E, we present the computation complexity analysis with a comparison between our LBP descriptor and selected statistical features

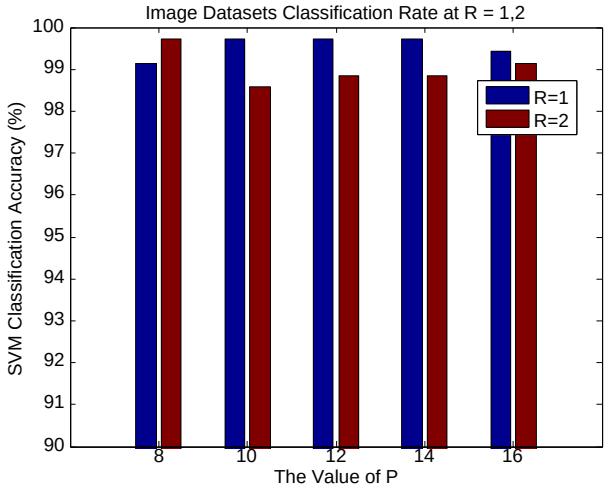


Fig. 7. SVM Performance of LBP Descriptor based on (P,R) for Printed Image Datasets

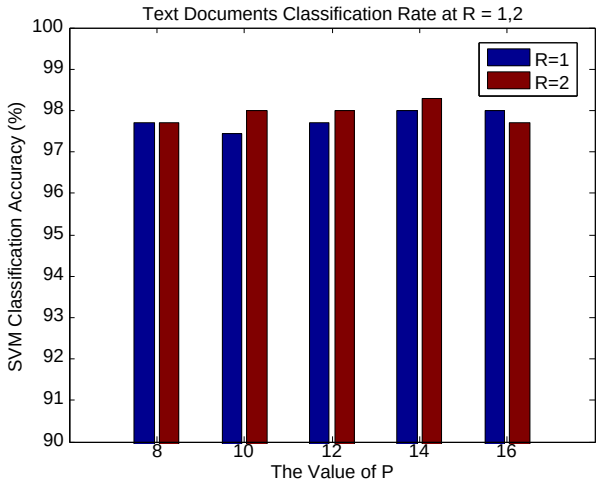


Fig. 8. SVM Performance of LBP Descriptor based on (P,R) for Printed Text Documents

based scheme.

TABLE III
FINE PRINTERS IN EXPERIMENT

Brand	Model Parameters	DPI
HP LaserJet	4200N	300
HP LaserJet	4100	300
HP LaserJet	4250	300
HP LaserJet	4100	300
HP LaserJet	4015	300
HP LaserJet	4500	300
Xerox Phaser	8500	300

A. The Choice of Classifier

SVM is commonly used as a classifier which maps an unlabeled instance to a label by learning a selected kernel function. It constructs a hyperplane which achieves a good separation by finding the largest distance to the labeled instances of any classes. We consider an n-fold cross validation and *leave one out* for learning the parameters of the selected kernel function on a training set so as to predict the labels on a test set [27]. The accuracy of a classifier is the probability of correctly classifying a randomly selected instance out of the overall number of instances in the dataset. The accuracy of the classifier is ideally to be an estimation with low bias and low variance, i.e., the classification accuracy is stable with confidence. The *leave one out* is almost unbiased especially in the case where the number of sample is small, but at an extra computational cost [28]. In the following section, the estimation of the classifier accuracy is demonstrated both by 5 fold cross validation and *leave one out*.

1) SVM:Five-Fold Cross Validation vs Leave One Out: As defined by cross validation [27], the training set is randomly divided into two sets which includes a set of data samples to be trained with, and together with a validation set. The validation set that is used as part of training is not the same as the test set. The test set is used to evaluate how well the SVM classifer performs. It is not correct to use the test set as part of training or validation. The classifer that is trained on a training set and validated on a validation set, is expected to be able to predict the dataset that is unseen, i.e., to predict the accuracy on the test set. Therefore, the test set cannot be used for training or validation [27].

In the v-fold Cross Validation, the training set D is randomly split into v exclusive folds $D_v, v \in [1, v]$. The classifier is trained and validated v times. At each time, it is trained on v-1 folds and validated on 1 fold. The parameters used for SVM training on v-1 fold, achieves the highest prediction accuracy on validation 1 fold, is used for training the whole training set D to generate a model. The model is used to analyze how good the classifier by predicting the accuracy on the test set. Unlike the *leave one out*, which is the complete v fold cross validation, splits datasets into a single fold, i.e.,the classifier is trained on $m - 1/m$ instances and tested on a single one instance, assume overall m instances available in the dataset. The *leave one out* is expensive compared with cross validation, while it is feasible and reliable for a small samples datasets.

In our experiments, fifty images are chosen from the UCID image database for each printer. In total there are 350 printed-and-scanned images in our image dataset. Fifty pages of pure text document from copyright-free ebook Gutenberg project [26] are selected for each printer, totally 350 printed-and-scanned text documents in our text dataset.

In order to estimate a classification accuracy with reliability, for five fold cross validation, we randomly permute 350 instances dataset 10 times and split them into 210 instances for training and validation, and 140 instances for testing. The 210 instances are further randomly permuted 10 times into 5 folds, where 168 instances are used for training, and 42 instances are used for validation. LIBSVM [29] provides a MATLAB interface for SVM-train function and SVM-predict function. The RBF kernel is selected with hyperparameter (C,γ) that needs be tuned to find the best classification performance over the 210 instances to generate the classifier. The selection range of c is (-5:2:15) and the selection range for γ is (-15:2:3). The classifier is further to predict over the 140 instances. Therefore, the accuracy of the classifier is the average of 100 predictions, i.e., the number of correctly predicted randomly

selected instances out of overall instances in the dataset.

For *leave one out*, 349 instances are used for training, with one instance for testing. After 350 times of training and testing, the labels of 350 instances are predicted to calculate the number of correctly predicted instance labels out of the overall number of instances in the dataset of 350. By 5 fold cross validation and *leave one out*, the performance of our proposed LBP descriptor $(P, R) = (8, 1)$ is evaluated with confidence under different attacks. A comparison study to statistical features based forensic analysis, also revealed that the proposed LBP scheme is able to identify printers at a high accuracy of printed-and-scanned image or text documents with a low variance. In Section IV-A2, the statistical features based forensic analysis is evaluated. A subset of relevant features is selected for printing and scanning scenario based on the best classification accuracy of the combined feature set.

2) Feature Selection and Statistical Significance Analysis: In [6], Gou *et al.* proposed a statistical feature based forensic analysis on scanners and scanned images by three different types of features. These features are denosing filtering, wavelet analysis and neighborhood pixel prediction error, giving a total of 60 dimensional features. The dimensionality of the features is reduced to ten by Linear Discriminant Analysis (LDA) [8]. However, our goal is to acquire a better understanding about the effectiveness of the individual features and how they relate with each other. The results presented in [6] are for the combination of all three features only. Given three different types of features: A, B, and C, we identify the accuracy of:

1) the features independently, i.e., $acc(A)$, $acc(B)$ and $acc(C)$.
2) the features in pair-wise combinations, i.e., $acc(A + B), acc(A + C)$ and $acc(B + C)$.
3) the combination of all three features, i.e., $acc(A + B + C)$.

We tabulate the results of these accuracies and choose the highest value to be the most optimum choice of features. In table IV, for each of the independent features, wavelet analysis feature in row two is the most significant feature that achieves the highest accuracy of 86.60% by cross validation and 90.00% by *leave one out*. The denoising filtering feature in row one is served as a complementary feature to wavelet analysis feature, increasing approximately 4% accuracy by adding another 10 dimensional features. Therefore, for features in pair-wise combinations, performance accuracy of the combination of wavelet analysis feature and denoising filtering feature in row four, achieves at 91.09% by cross validation and 94.86% by *leave one out*. This is even higher than the performance accuracy of the combination of all three features in row seven, which is 89.45% by cross validation and 93.14% by *leave one out*. It means that adding another 4 dimensional features of neighborhood prediction into the pair-wise combination of wavelet analysis feature and denoising feature, degrades the performance of the classifier. We will also show in Section IV-E that this 4 dimensional neighborhood prediction feature costs a computational complexity of $O(N^2)$. Therefore, the features selected for printer identification is the pair-wise combination of wavelet analysis feature and denoising filtering feature, giving a total of 16 dimensions. These features will be compared with our proposed LBP descriptor. The performance accuracy of LBP descriptor in row eight, is 97.92% by cross validation and 99.43% by *leave one out*. The confusion matrix of selected features, and confusion matrix of our proposed LBP descriptor is presented in Section IV-B2.

B. Printed-and-Scanned Image Identification Apart from Attacks

1) SVM performance: The performance is evaluated by SVM classifier of five-fold cross validation and *leave one out*, over 350 printed-and-scanned images and 350 printed-and-scanned pure text documents. Image were downloaded from UCID database. The text datasets were downloaded from copyright-free ebook Gutenberg project [26]. The first fifty pages of the ebook titled $\ll Emma \gg$ which are purely text document is used. They were printed at 300 dpi by seven printers which include two identical models from HP LaserJet 4100 as shown in Table III. The printouts were scanned into A4 size at 600 dpi grayscale JPEG format by Infotec ISC 3535. A bounding box is applied to remove the white margin of the scanned images and documents.

For $(P, R) = (8, 1)$ LBP descriptor, 59 dimensional features were extracted for each printed-and-scanned image. In total 350 instances with 59-dimensional feature space are used for SVM classification. As indicated in Section IV-A1, 350 instances feature space is randomly divided into 210 instances for 5 fold cross validation and 140 instances for testing. During 5 fold cross validation, 210 instances randomly split into 5 fold for training and validation. In total 100 random permutations were performed to estimate the classification accuracy, which is the average of the number of correctly predicted labels out of 140 instances. The classification accuracy achieved was 97.92%, with the standard deviation of 1.08. The confusion matrix is presented in Table V. In addition, we also experimented with *leave one out* to the 350 instances with 59 dimensional feature space with the proposed LBP descriptor at $(P, R) = (8, 1)$. An accuracy rate of 99.43% was achieved. As given by the confusion matrix in Table VI, the most confusion was between the two HP 4100 printers at a 2% mis-classification rate, which was the same as the confusion result obtained by 5 fold cross validation presented in Table V.

2) Selected Features for Printer Identification: We analyze the performance of utilizing the features investigated in Section IV-A2 for identifying the printers as shown in Table III. As shown in Table IV, the wavelet analysis and denoising filtering have been selected for identifying printed-and-scanned image by seven distinct printers. For each image, a 16-dimensional feature vector was extracted. In total, a 350x16-dimension feature space was constructed for SVM classification. The five fold cross validation was the same as presented in Section IV-A1. An accuracy rate of 91.09% was achieved by 100 time random permutations of dataset, with a standard deviation of 1.85. As presented by the confusion matrix in Table VII, the HP 4100(1) has a 23.3% confusion with HP4200(2). HP4200(2) only achieved 68.65% accuracy rate to identify itself. As compared with the proposed LBP

TABLE IV
FEATURES SELECTION AND COMPARISON OF SELECTED FEATURES TO LBP

	Features	Feature Dimension	5 Fold Cross Validation(Avg%)	Standard Deviation	*Leave One Out*(%)
Independent [6]	Denoise Filtering	10	72.21	3.48	80.00
	Wavelet Analysis	6	86.60	2.51	90.00
	Neighborhood Prediction	4	63.86	2.74	69.43
Pair-wise [6]	Denoise Filtering + Wavelet Analysis	16	91.09	1.85	94.86
	Denoise Filtering + Neighborhood Prediction	14	75.49	3.09	83.14
	Wavelet Analysis + Neighborhood Prediction	10	88.48	2.84	93.71
All [6]	Denoise Filtering + Wavelet Analysis + Neighborhood Prediction	20	89.45	2.72	93.14
Our Feature	LBP Descriptor	59	97.92	1.08	99.43

TABLE V
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON LBP DESCRIPTOR BY CROSS VALIDATION

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	100%						
	HP Laserjet 4100(1)		91.8%		5.85%			
	HP Laserjet 4250			100%	0.5%			
	HP Laserjet 4100(2)		7.8%		93.65%			
	HP Laserjet 4015					100%		
	HP Laserjet 4500		0.4%				100%	
	Xerox Phaser 8500							100%

TABLE VI
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON LBP DESCRIPTOR BY *Leave One Out*

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	100%						
	HP Laserjet 4100(1)		98%		2%			
	HP Laserjet 4250			100%				
	HP Laserjet 4100(2)		2%		98%			
	HP Laserjet 4015					100%		
	HP Laserjet 4500						100%	
	Xerox Phaser 8500							100%

descriptor, the confusion was more widely dispersed between the HP models. For the *leave one out*, the identification rate achieved was 94.86%, and the confusion matrix is presented in Table VIII.

C. Printed-and-scanned Text Documents Classification

The performance of our proposed LBP descriptor is also validated over pure text document dataset. An accuracy rate of 98.06% was achieved with a standard deviation of 1.24. As presented in Table IX, the most confusion was still between the two HP 4100 printers. A 98% accuracy rate was achieved by *leave one out* with the confusion matrix results given in Table X.

D. The Robustness Against Arbitrary Image Processing Operations

The proposed scheme was tested robustness against a series of image processing operations, including averaging filtering, median filtering, sharpening, rotation, JPEG compression and resizing.

- Average Filtering Operations
  Average filtering filters the image with the predefined averaging filter of size 3x3, 5x5 and 7x7. In the experiment, we perform averaging filtering on 350 printed-and-scanned images, resulting as three different filtered dataset. The accuracies achieved by *leave one out* of the filtered datasets are presented in Table XI. The perfor-

TABLE VII
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON SELECTED FEATURES IN [6] BY CROSS VALIDATION

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	99.9%		0.85%		0.05%		
	HP Laserjet 4100(1)		72.6%	0.05%	26.6%		0.05%	
	HP Laserjet 4250		0.05%	98.05%	0.5%	0.05%	0.45%	
	HP Laserjet 4100(2)		23.3%	0.35%	68.65%		0.4%	
	HP Laserjet 4015		1.5%	0.6%	2.05%	99.85%	0.55%	
	HP Laserjet 4500	0.1%	2.55%	0.1%	2.2%	0.05%	98.55%	
	Xerox Phaser 8500							100%

TABLE VIII
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON SELECTED FEATURES IN [6] BY *Leave One Out*

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	100%						
	HP Laserjet 4100(1)		82%		18%			
	HP Laserjet 4250			100%				
	HP Laserjet 4100(2)		16%		82%			
	HP Laserjet 4015					100%		
	HP Laserjet 4500		2%				100%	
	Xerox Phaser 8500							100%

TABLE IX
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON TEXT DOCUMENTS OF LBP DESCRIPTOR BY CROSS VALIDATION

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	97.15%	0.25%	1.05%	0.95%			
	HP Laserjet 4100(1)	0.85%	97.65%	0.05%	2.8%	0.2%		
	HP Laserjet 4250	0.25%	0.3%	98.7%	0.2%			
	HP Laserjet 4100(2)	0.55%	1.8%	0.05%	95.4%	0.2%		
	HP Laserjet 4015	0.1%		0.1%		98.35%	0.8%	
	HP Laserjet 4500	1.1%		0.05%	0.65%	0.25%	99.2%	
	Xerox Phaser 8500					1.00%		100%

mance of our proposed LBP descriptor is consistent for smaller filter sizes while it only dropped 4% classification rate for filter size 7x7. This is because average filtering is a linear operation. Therefore, LBP takes the signed differences to the neighborhood pixel values, which remains invariant towards any linear operation.

- Median Filtering Operations
  In the experiment, median filtering of order 3 , 5 and 7 [30] was applied to 350 printed-and-scanned images. The performance of *leave one out* reveals that our proposed LBP descriptor maintains a good robustness against median filtering as presented in Table XI.
- Sharpening Operation
  The sharpening operation, filters an image with a 3-by-3 unsharpened contrast enhancement filter, with shaping factor 0, 0.2, 0.4, 0.6. In the experiment, 350 printed-and-scanned images are sharpened by a shaping factor 0, 0.2, 0.4 and 0.6 respectively to form four sharpened image datasets for classification by *leave one out*. The accuracies of the classifier decreases as the images sharpened from 0 to 0.4, while it improves by 0.3% accuracy with shaping factor 0.6 as shown in Table XI.

TABLE X
CONFUSION MATRIX OF SEVEN PRINTERS CLASSIFICATION BASED ON TEXT DOCUMENTS OF LBP DESCRIPTOR BY *Leave One Out*

		Predict						
	Brand	HP Laserjet 4200N	HP Laserjet 4100(1)	HP Laserjet 4250	HP Laserjet 4100(2)	HP Laserjet 4015	HP Laserjet 4500	Xerox Phaser 8500
Train	HP Laserjet 4200N	96%		2%				
	HP Laserjet 4100(1)		98%		2%			
	HP Laserjet 4250	2%		98%				
	HP Laserjet 4100(2)		2%		98%			
	HP Laserjet 4015					98%	2%	
	HP Laserjet 4500	2%					98%	
	Xerox Phaser 8500					2%		100%

- JPEG Compression Operation
  For JPEG compression, 350 printed-and-scanned images in TIFF format are compressed with Quality factor of 70, 80, 90, and 100 [31]. The *leave one out* accuracy is tested over each of these four compressed image datasets. As presented in Table XI, our proposed LBP descriptor achieves superior robustness against JPEG compression, since the random binary texture pattern is generated in the way constant to any monotonic gray scale transformation III-B. It is worth mentioning that the classifier accuracy is maintained even at the JPEG compression Quality factor of 75, at which most of other forensic schemes began to degrade [6].
- Rotation Operation
  Rotation operation, rotates an image by an angle in degrees in a counterclockwise direction around its center point, using the bilinear interpolation method. In our experiment, we tested our proposed LBP descriptor for rotation angles of 1 degree, 5 degrees, 10 degrees and 20 degrees. As presented in Table XI, the accuracies of our proposed LBP descriptor remained constant at an average of 99.1425% with ± 0.2327 variation.
- Resizing Operation
  Resizing operation, scales an image at a scaling factor of 0.5, 0.75, 1.25 and 1.5 by bicubic interpolation method. In the experiment, 350 printed-and-scanned images were resized by four different scaling factor. *Leave one out* was applied to predict the classification accuracy of seven printers. As demonstrated in Table XI, the classification accuracy remained the lowest at the scaling factor of 0.5 and achieved the maximum at scaling factor of 1.5.
  In summary, these operations have a minimum impact on the classification performance of our proposed scheme, which further confirms that the LBP descriptor is a unique and stable forensic feature for printing-and-scanning.

E. Computation Complexity Analysis

Low-complexity of the feature extraction is always an important consideration. We have analyzed the computation complexity for all features mentioned in this paper as presented in Table XII. Without loss of generality, we assume the images are square size. For denoise filtering, which uses a filtering operation, typically has a complexity of $O(N \log N)$. Then mean and standard deviation are applied. Each of them costs a complexity of $O(N^2)$. Therefore, the maximum complexity is $O(N^2)$. Similarly, for wavelet feature extraction, the complexity is similar to the Discrete Fourier Transform (DFT) process. The fast operation of DFT will cost a complexity of $O(N \log N)$. Neighborhood Prediction needs to solve nonnegative least-square problem (NNLS) for each pixel of the image. The NNLS complexity is $O(N * B^3)$, where B is the 3x3 block size coefficients estimation. Then mean and standard deviation are applied to the estimated image with a complexity of $O(N^2)$. Therefore, Neighborhood Prediction has a complexity of $O(N^2)$. The computational complexity of the features proposed in [6], in total is counted as $O(N^2)$. The feature selected for printer identification which is the combination of denoising filtering and wavelet analysis, have a computational complexity of $O(N^2)$. Compared with their features, LBP descriptor only calculates (P, R) neighboring patterns for each pixel of the image. Then a histogram costs a constant computation. Therefore, the complexity of LBP is a polynomial $O(N)$.

TABLE XII
COMPUTATIONAL COMPLEXITY ANALYSIS COMPARISON TO [6]

Features	Complexity
Denoising Filtering	$O(N^2)$
Wavelet Analysis	$O(N \log N)$
Neighborhood Prediction	$O(N^2)$
Denoising Filtering+ Wavelet Analysis+ Neighborhood Prediction	$O(N^2)$
Denoising Filtering+ Wavelet Analysis	$O(N^2)$
LBP descriptor	$O(N)$

V. CONCLUSION AND FEATURE WORK

In this paper, we investigated the texture of printed documents based on an LBP descriptor. The local binary patterns were determined to be the representatives of local texture description of printed documents. The occurrence histogram of local binary patterns provided a powerful discrimination capability as an intrinsic feature for printer identification. The effectiveness of the proposed LBP based features was

TABLE XI
ROBUSTNESS DEMONSTRATION OF LBP DESCRIPTOR BY *Leave One Out* CLASSIFICATION ACCURACIES

Attacks	Parameter	Accuracy(%)	Attacks	Parameter	Accuracy(%)
Average Filtering	3	98.00	Median Filtering	3	99.43
	5	98.86		5	98.86
	7	94.86		7	98.57
Sharpening	0	99.71	Rotation	1	98.86
	0.2	99.43		5	99.14
	0.4	98.57		10	99.43
	0.6	98.86		20	99.14
JPEG compression	100	99.71	Resizing	0.5	96.57
	90	99.14		0.75	98.86
	80	98.86		1.25	97.43
	70	99.14		1.5	99.43

achieved by evenly sampling the neighborhood pixels by an angular space, determined at radius R and angles of $360^\circ/P$. The efficiency of the scheme was also achieved by characterizing the probability distribution of gray levels from 256 ($0 \sim 255$) to a limited number of gray levels, i.e., 59, when $(P, R) = (8, 1)$. The local binary pattern representing the bright or dark flat region was the most frequently appeared pattern that contributed significantly to the classification in the printed images and printed text documents. Since by definition, LBP takes the signed differences to pixels in a neighborhood, the local texture patterns remained constant for any monotonic gray scale transformation. This was verified by the classification performance which showed LBP based features were superior in terms of invariance under a series of image operations, especially at the higher JPEG compression rates. By either n-fold cross validation or *leave one out*, our proposed LBP based scheme was able to improve the classification accuracy up to 97.9214% and 99.4%, which was a 6% improvement in accuracy to the statistical feature based scheme [6]. More importantly, our computational complexity was determined to be O(N). It could be inferred that a feature selection scheme which reduces the dependencies of the current multi-dimensional features into lower dimension, would further improve the classification accuracy.

ACKNOWLEDGMENT

The authors would like to thank Dr Dan He from the Center of Communication System and Research for the insightful discussions, also for supporting the printers to conduct the experiments, Dr Chris Culnane from the Department of Computing, University of Surrey, UK for his valuable comments. The authors wish to thank Dr Fei Yan from the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, for sharing the knowledge of SVM and feature selection, and in particularly the precise instruction of implementing cross validation and leave one out.

REFERENCES

[1] H. Pearson, "Image manipulation: Csi: cell biology," *Nature*, vol. 434, pp. 952–953, 2005.

[2] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, June 2006.

[3] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, March 2008.

[4] A. K. Mikkilineni, P.-J. Chiang, G. N. Ali, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, "Printer identification based on graylevel co-occurrence features for security and forensic applications." in *Security, Steganography, and Watermarking of Multimedia Contents*, 2005, pp. 430–440.

[5] O. Bulan, J. Mao, and G. Sharma, "Geometric distortion signatures for printer identification," in *Proc. IEEE Intl. Conf. Acoustics Speech and Sig. Proc.*, Taipei, Taiwan, 2009, pp. 1401–1404.

[6] H. Gou, A. Swaminathan, and M. Wu, "Intrinsic sensor noise features for forensic analysis on scanners and scanned images," *Information Forensics and Security, IEEE Transactions on*, vol. 4, no. 3, pp. 476 –491, Sep 2009.

[7] W. Jiang, A. T. S. Ho, H. Treharne, and Y. Q. Shi, "A novel multi-size block benford's law scheme for printer identification," in *The Pacific-Rim Conference on Multimedia (1)*, 2010, pp. 643–652.

[8] N. Khanna, A. K. Mikkilineni, and E. J. Delp, "Scanner identification using feature-based processing and analysis," *Trans. Info. For. Sec.*, vol. 4, no. 1, pp. 123–139, 2009.

[9] J. R. Janesick, "Scientific charge-coupled devices," 2001.

[10] A. K. Mikkilineni, O. Arslan, P.-J. Chiang, R. M. Kumontoy, J. P. Allebach, and G. T.-c, "Printer forensics using svm techniques," in *Proceedings of the IS&T's NIP21: International Conference on Digital Printing Technologies*, vol. 21, Baltimore, MD, October 2005, pp. 223–226.

[11] N. Khanna, A. K. Mikkilineni, P.-J. Chiang, M. V. Ortiz, V. Shah, S. Suh, G. T.-C. Chiu, J. P. Allebach, and E. J. Delp, "Printer and sensor forensics," in *IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, Washington D.C., USA,, April 11-13 2007.

[12] N. Khanna, A. K. Mikkilineni, G. T. Chiu, J. P. Allebach, and E. J. Delp, "Survey of scanner and printer forensics at purdue university," in *IWCF '08: Proceedings of the 2nd international workshop on Computational Forensics*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 22–34.

[13] D. L. Lau and G. Arce, *Modern Digital Halftoning*. New York, NY, USA: Marcel Dekker, Inc., 2001.

[14] W. Jiang, "Forensic analysis in printed-and-scanned image and its application to printer identification," in *PhD Thesis*, 2011, p. 150.

[15] M. Kivanc Mihcak, I. Kozintsev, and K. Ramchandran, "Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising," in *Proceedings of the Acoustics, Speech, and Signal Processing. on 1999 IEEE International Conference*, ser. ICASSP '99, vol. 06. Washington, DC, USA: IEEE Computer Society, 1999, pp. 3253–3256.

[16] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Attacks on copyright marking systems," in *Proceedings of the Second International Workshop on Information Hiding*. London, UK: Springer-Verlag, 1998, pp. 218–238.

[17] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," *IEEE Trans. Image Processing*, vol. 9, pp. 1522–1531, 2000.

[18] ——, "Adaptive wavelet thresholding for image denoising and compression," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 9, no. 9, pp. 1532–1546, 2000.

[19] E. Simoncelli and E. Adelson, "Noise removal via bayesian wavelet coring," in *Image Processing, 1996. Proceedings., International Conference on*, vol. 1, Sep 1996, pp. 379 –382 vol.1.

[20] C. E. Shannon, "Communication in the presence of noise," *Proceedings of Institute of Radio Engineers*, vol. 37, no. 1, pp. 10–21, Jan 1949.

[21] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of

texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.
[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610–621, November 1973.
[23] O. T. and P. M. . M. T., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." vol. 24, no. 7, pp. 971–987, 2002.
[24] "Xerox launches revolutionary colour printer," *Wall Street*, 2009-05-06.
[25] G. Schaefer and M. Stich, "Ucid - an uncompressed colour image database," in *In Storage and Retrieval Methods and Applications for Multimedia 2004, volume 5307 of Proceedings of SPIE*, 2004, pp. 472–480.
[26] *Free eBooks by Project Gutenberg*. [Online]. Available: http://www.gutenberg.org/wiki/Main_Page
[27] D. Poole and A. K. Mackworth, *Artificial Intelligence - Foundations of Computational Agents*. Cambridge University Press, 2010.
[28] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 1995, pp. 1137–1143.
[29] P.-H. Chen and C.-J. Lin., *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[30] W. Pratt, *Digital Image Processing*. John Wiley and Sons, 1978.
[31] G. K. Wallace, "The jpeg still picture compression standard," *Commun. ACM*, vol. 34, pp. 30–44, April 1991.

PLACE PHOTO HERE

Weina Jiang Biography text here.

Anthony TS Ho Biography text here.

Helen Treharne Biography text here.

Yun Qing Shi Biography text here.