International Journal of Computer Applications (0975 8887)
Volume 105 - No. 15, November 2014

# Statistical Measures for Differentiation of Photocopy from Print technology Forensic perspective

M. Uma Devi
Student
Department of Computer Science and Engineering
Universal College of Engineering and Technology, Dokiparru

C. Raghvendra Rao
Professor
Department of Computer and Information Sciences
University of Hyderabad

M. Jayaram
Associate Professor
Department of Computer Science and Engineering
Universal College of Engineering and Technology

## ABSTRACT

Forensic document examination plays an important role in providing the evidence to the court related to disputed documents. Emerging print technologies are posing challenges to document examiner in identification of source of document. Recent trends suggest the need for good preprocessors and post analysing tools which characterize printed text for identification of print technology. Each printing technology differs in their process of placing marking material on the target. Image analysis methods along with statistical tools are applied to study class characteristics of document for identifying the source of the document. This paper focuses on frequently used word like 'the' as test sample for characterizing printed text. The proposed algorithm is based on analysis of histogram of printed text image. Statistical measures skewness and kurtosis of histogram are used as features for distinguishing inkjet print from its photocopy.

### General Terms:

Forensic document examination, source of document

### Keywords:

Histogram, Skew , Kurtosis

## 1. INTRODUCTION

Printed material like documents generally contains information[18] about transactions like agreement related to authority or ownership of properties, identity cards. Documents have text content which is often forged by antisocial elements for performing criminal activities. Forgery is done by altering any of the contents of the document or reproducing the whole document with evolving digital imaging techniques. one can easily add text to the margin of document or with in the document. As technology tends to evolve, the methods to forge documents are ever sophisticated challenging the skills of forensic examiners. Document examiner needs an excellent eye sight for examining such fine details in the document. High quality forged documents are produced using scanner printers which makes identification of the document more complex. There fore determining the genuineness of document is critical and needs to established. Often the document examiner needs to answer questions about the genuineness and consistency of the document like,

(1) Whether the given document is printed or photocopied or what is the source of the document.
(2) whether the content in the document is printed using one source printer or more than one printer.

There is need for an extensive knowledge of emerging print technologies to identify the class characteristics of the document. Printed text characterization assists forensic examiner in identifying printing process or techniques by identifying spatial pattern of text produced by that printing mechanism.
As the document is combination of text and images, identification of source of document depends on both the image and text of that document. Identification of print technology using Gaussian Variogram Model[27] is based on homogeneous colour regions of printed document. The methodology proposed in [27] Gaussian Variogram Model identifies the print technology of document based on uniform colour region of image. In case of text documents, it is difficult to find enough region in the text for Gaussian variogram analysis to identify print features contained in the document. Print Index [28] generic measurement proposed based on which printed text in the document is classified. Characteristics of printing mechanism are the features that distinguishes spatial pattern of one printing mechanism from another mechanism. Identification of such characteristics are derived from spatial statistics of printed text and it is referred to as printed text characterization.
As printed text is a combination of ink or toner on the printed material (which is generally white paper), the word and characters of the text in the questioned document has influence of the printing technology used. This text is a collection of various dots and style of printing. Hence, the variations are observed in the image format of the printed text. The printed image can be viewed as combinations of basic features like background (paper or printed material), foreground on the printing style of a printer as well as some distortions or noise. Selected segment of the text as a combination

of standard normal distribution with mean, mixed proportion and variation are used finding the print pattern. Thus, the pixels of image of the text region can be modeled as a mixture of distributions with three classes. The statistical analysis of word count reveals that 'the' is the most frequently used word in English [4]. Hence 'the' word region has been considered for the printed text characterization which induces document segmentation and then recognizing region of 'the' are sub challenges. The exploratory analysis of the parameters of mixing model for printed text 'the' from various printer like inkjet printer and laser printer motivated us to propose a novel index method. This index measure is based on Expectation Maximization mixed distribution characteristics of print and it is referred to as Print Index. This Print Index itself can be grouped, which in turn formulates as rules for classification of print technology as inkjet or laserjet.
While making the copies of printed document the distribution of text, background and noise is maintained similar to the original document. Hence, proposed Print Index methodology could not distinguish between print and photocopy. Print index of sample 'the' and its photocopy are shown in Figure 1 and it is observed that Print Indices of print and its photocopy are almost similar. Hence, the methodology is developed to classify print from photocopy and is discussed in this section.

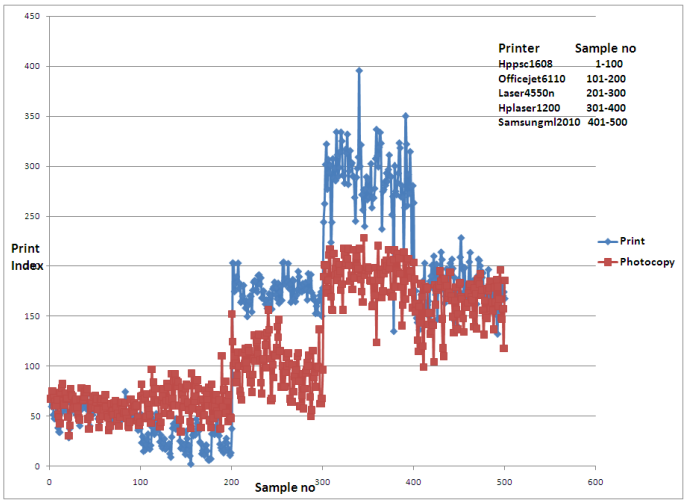



Fig. 1. Comparision of Printindex of print and its photocopy

This paper propose the methodology for differentiating photocopy from its print and it is extension of printed text characterization. Section 2 discusses about related work in the field of forensic analysis of printed document for identification of source of document. Section 3 addressing the problem and statistical measures used for differentiation of print technology. Section 4 explains procedure for differentiating photocopy from its inkjet. Section 5 presents experimental results and section 6 concludes with the recommended methodology for printed document source identification.

## 2. LITERATURE SURVEY

Recent research publications demonstrate various approaches suggested for discriminating printing techniques like ink jet or laser jet, photocopy. Research activities at Purdue University are on characterization of Electro photographic printers[19], the gray level co-occurrence feature[21] of the most frequently occurring letter 'e' and Gaussian Mixture Model(GMM) [6] are used for printer identification.
In GMM, Principal component analysis is used as dimension reduction technique by 1-D projections of the extracted text character. These researches are exclusively for the identification of Electro photographic printers.
Machine Identification code project by Electronic Frontier Foundation identifies presence of pattern of yellow dots [2] in colour laser printouts which represent printer serial number. This is not applicable to all Electro-photographic printers as some printers do not show the presence of these yellow dots.
Ink and toner analysis for identification of printing process using HSV color space by Haritha[16][9], is based on hue histogram for identification of printing process and photocopy. Hue contrast, periodicity and ink over spray are the features selected for classifying ink jet, laser jet and photocopies. This is color image processing technique using HSV color space.
The work discussed in[15], [14] for identification of fraud documents by comparison of characters in the given documents and needs high resolution microscope to capture these characters.
Gray level features proposed by Lambert[20] for discriminating ink jet from laser jet print is dependent on high resolution scanned images, 3200 dot per inch. Recent research is concentrated on evaluation of gray level features like perimeter based edge roughness of the text[25] for print technique classification, based on low resolution image for high through put document management system.
In Beusekom [8] proposed a method to detect misalignment of text lines that are additionally inserted in a document. If the forged text line alignment is same as original text line, it is difficult to find out the forged document.
The works discussed in[6],[7],[24],[11],[10],[22],[5],[12] are concentrated on identification of Electro-photographic printers. The embedding techniques [23],[5], [17], [26] is useful for only few group of printers for identification. Significant amount of work is done in identification of electro photographic printers, but still there is need for techniques to identify various printing techniques like inkjet, laserjet, photocopy from forensic perspective.

## 3. CLASSIFICATION OF PHOTOCOPY FROM INKJET PRINT

### 3.1 Problem definition

Photocopier [1] is a machine, which produce documents similar to the original documents. These are electrostatic machine like laser printer in all aspects, except laser printer produces grid pattern. Most photocopiers use dry process called Xerography. Several copies of original document is produced in low cost using photocopiers compared to printouts. Hence, the use of photocopies is more compared to the printed documents. Major portion of the printed material used by people are photocopies, which appears similar to the printed one. Using these copying machines several copies of identity proofs, passes and tickets can be created for performing illegal activities. Hence the identification of photocopy or differentiating photocopy from printed document are significant problems in document examination field.
Some of the forged documents are produced combining more than one printing technique. Identifying the technique by which the document is created is a challenge to the forensic examiner. Document investigation starts with identifying whether the document is produced by a printer or it is a photocopy of a printout.
Forensic discrimination of photocopy from print is often done by chemical analysis techniques like infra-red spectroscopy. However,

it is possible to differentiate photocopy from its printout using digital image processing techniques along with statistical analysis without resorting to expensive instruments. These techniques are non destructive. Determining the general class characteristics of document helps in classifying print out from its photocopy. While making a copy of original document, some disturbances occur which results in photocopied document. These disturbances in photocopies are exploited and characterized using statistical analysis techniques.

### 3.2 Samples

Two type of text documents are used for sample collection. One is having most frequently occurring word 'the' and the other is 'The'. In the proposed method, the text documents contain text samples of font type Times New Roman and size 12pt. These text documents are printed at 600 dpi on printers listed in Table 1. The printed documents are photocopied using photocopiers listed in Table 2. These prints and its photocopy documents are scanned at high resolution 2400 dots per inch. Text samples are collected from the high resolution images of printed text document and its photocopy.

### 3.3 Statistical measures used to differentiate photocopy from inkjet print

Document produced using inkjet printer and the photocopy of the same inkjet document are similar in normal view. High resolution images of text printed on inkjet and photocopy of the text has shown that there are significant differences between these two images. One can observe the difference between these scanned images of print and its photocopy as shown in Figure 2. It is observed from inkjet print and its photocopy that the roughness in printed text is smoothed while producing the photocopy and noise is clearly visible as dark spots in the background area of the photocopy. These differences can be explained with aid of histogram of print and its photocopy. This is clearly visible from Figure 3, which compares histograms of inkjet print and its photocopy. The sharp features of histogram of photocopy are peakedness and tilt of peak, which resembles the disturbances that occur while copying. These disturbances can be modelled as features of histogram.


Fig. 2. Print out and its photocopy

The purpose of the histogram[13] is to graphically summarize the distribution of univariate data set. The histogram of an image shows

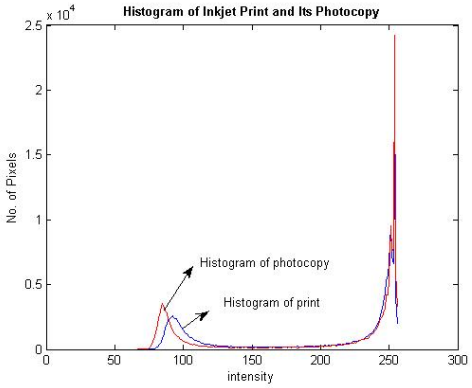



Fig. 3. Histogram of inkjet(Hppsc1608)print and its photocopy

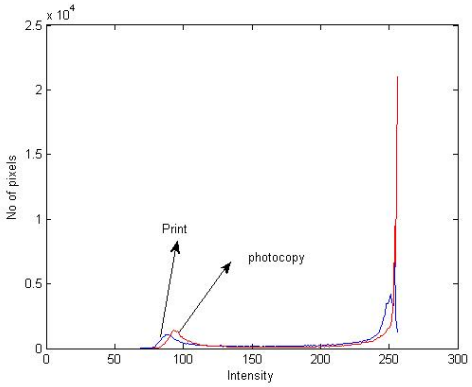



Fig. 4. Histogram of Laser(Hplaser1200)print and its photocopy

the distribution of gray levels in the image. Histogram analyses the following features of the data: the centre of the data, spread of the data, skewness of the data, presence of the outliers and presence of multiple modes in the data. These features provide the distribution model for the data.

The statistical analysis techniques that characterize symmetry and peakedness of dataset are known as skewness and kurtosis[3]. Skewness is a measure of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point. For symmetric distribution, the body of the distribution refers to center of the distribution. The tail of the distribution refers to the extreme regions of the distribution, both left and right. The tail length of the distribution indicates how the distribution of data reaches zero. The short tailed distribution is where probability is constant in a certain range. Moderate tailed distribution is Gaussian distribution where the tail declines to zero moderately. For skewed distribution, one tail of distribution is always longer than the other tail. The occurrence of skewness is due to the lower or upper bounds of data. Hence, data having lower bounds will result in skew right distribution and the data having upper bounds will result in skew left distribution.

$$Skew = \beta_1 = \mu_3^2/\mu_2^3 \quad (1)$$

Kurtosis is a measure of peakedness of data relative to a normal distribution. Data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.

Table 1. Printers used.

| P. id | No. of words 'the' Collected | No of general 3 letter words | printer | Print technology |
|---|---|---|---|---|
| 1 | 100 | 34 | Hppsc1608 | Inkjet |
| 2 | 100 | 34 | Officejet6110 | Inkjet |
| 3 | 100 | 34 | Hplaser4550N | Colorlaserjet |
| 4 | 100 | 34 | Hplaser1200 | Laserjet |
| 5 | 100 | 34 | SamsungML2010 | Laserjet |

Table 2. List of photocopiers used.

| S.no | Photocopier |
|---|---|
| 1 | Konika Minolta bizhub210 |
| 2 | XeroxWCM118 |

Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

$$Peak = \beta_2 = \mu_4/\mu_2^2 \tag{2}$$

Hence, these features of histogram are used to identify disturbance that occur in a sample while photocopying. Skew and kurtosis of images shown in Figure 2 are presented in the below Table 3.
The following section explains preparation of data set of print and its photocopy and analysis of skewness and kurtosis of those dataset to differentiate photocopy from print.

## 4. PROCEDURE FOR DIFFERENTIATING PHOTOCOPY FROM INKJET PRINT

(1) Selected sample is a text printed at 600 dpi for both photocopy and printed text.
(2) Scan the sample at 2400 dpi
(3) From the scanned image, fix text using MBR. Pixels greater than 150 are considered as back ground and are filtered out.
(4) Generate histogram for the foreground (printed text).
(5) Calculate skew(beta1) and peakedness(beta2) of histogram as shown below.

$$Skew = \beta_1 = \mu_3^2/\mu_2^3 \tag{3}$$

$$Peak = \beta_2 = \mu_4/\mu_2^2 \tag{4}$$

where $\mu_k = 1/N * \sum_{i=1}^{N}(x_i - m)^k$ and $m = 1/N * \sum_{i=1}^{N} x_i$

## 5. EXPERIMENTAL RESULTS

Standard photocopy machines listed in Table 2 are used to produce photocopy of text samples collected from the printers. Text documents having the sample 'the' are printed on printer listed in Table 1 and are photocopied using XeroxWCM118 photocopier. Text documents having 'The' are printed on printers listed in Table 1 and photocopied on Konika Minolta photocopier. The skewness and peakedness of samples are calculated as mentioned in Equation 3 and Equation 4.
The skewness and peakedness of the printed text sample 'the' and its photocopy are shown in Figure 5 and Figure 6. First, 200 samples skewness(symmetry) and kurtosis(peak) of print is clearly distinguishable from its photocopy. The samples from 1 to 200 are produced using inkjet printing mechanism, samples from 201 to 300 are from colour laser printer. Samples

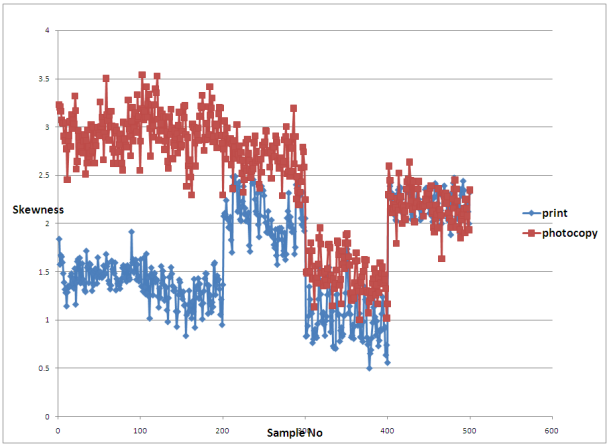



Fig. 5. Skewness of print and its photocopy for text sample 'the'

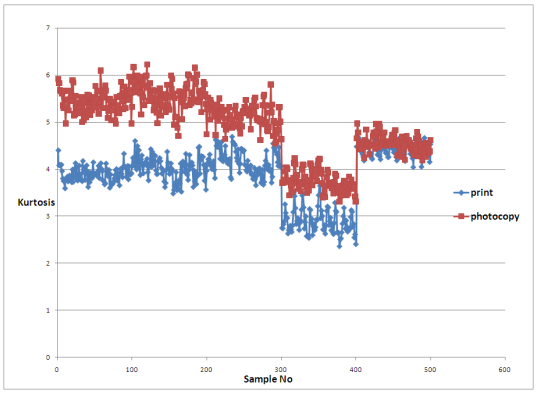



Fig. 6. Kurtosis of print and its photocopy for text sample 'the'

from 301 to 500 are from two black and white laser printers. Skewness and kurtosis for the black and white laser are difficult to differentiate as both black and white laser printing mechanism and photocopy printing mechanism is based on electrostatic printing technique.
For text sample 'The' and its photocopy, first 500 samples peak and skewness of print is clearly distinguishable from its photocopy. Samples from 1 to 400 are produced using inkjet printing mechanism and samples from 401 to 500 are from colour laser printer. The samples from 501 to 700 are from two black and

Table 3. Skew and Kurtosis of Print and its Photocopy.

| P.id | Print | | Photocopy | |
|---|---|---|---|---|
| | Skew | Kurtosis | Skew | Kurtosis |
| 1 | 1.809792809 | 4.257259987 | 2.918359538 | 5.40452616 |
| 2 | 1.91484902 | 4.680384175 | 2.94265574 | 5.366635791 |
| 3 | 2.007055235 | 4.232307006 | 2.982989282 | 5.512413205 |
| 4 | 2.561714282 | 4.803079783 | 2.51990513 | 4.904508484 |
| 5 | 1.092584645 | 3.120035086 | 1.391041767 | 3.441752727 |

white laser printers. Skewness and kurtosis for the black and white laser are difficult to differentiate as both black and white laser printing mechanism and photocopy printing mechanism is based on electrostatic printing technique. General three letter word sample in Figure 9 are photocopied using Konika Minolta photocopier. Skew and kurtosis of print and photocopy of these sample are shown in Figure 10. The skew and kurtosis of sample 'the' are consistent with skew and kurtosis of the general three letter word.

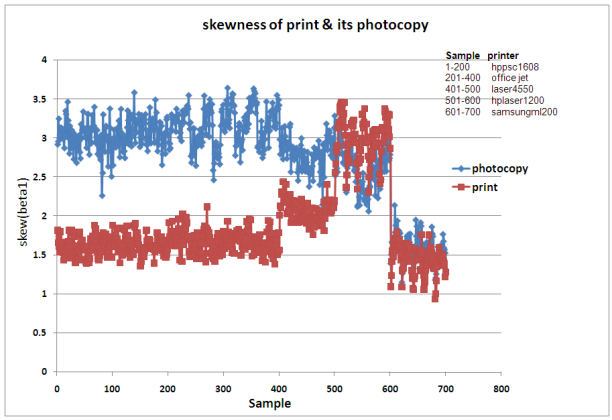



Fig. 7. Skewness of print and its photocopy for text sample 'The'

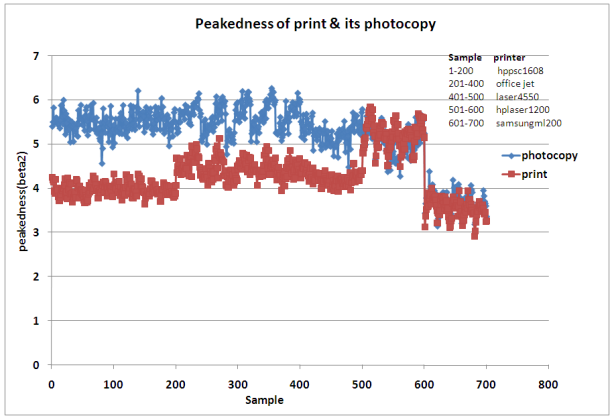
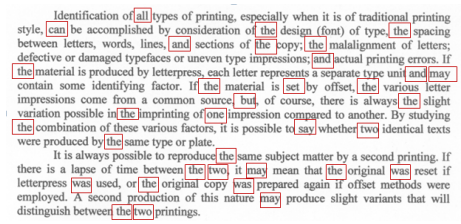



Fig. 8. Kurtosis of print and its photocopy for text sample 'The'


Identification of all types of printing, especially when it is of traditional printing style, can be accomplished by consideration of the design (font) of type, the spacing between letters, words, lines, and sections of the copy; the malalignment of letters; defective or damaged typefaces or uneven type impressions; and actual printing errors. If the material is produced by letterpress, each letter represents a separate type unit and may contain some identifying factor. If the material is set by offset, the various letter impressions come from a common source, but, of course, there is always the slight variation possible in the imprinting of one impression compared to another. By studying the combination of these various factors, it is possible to say whether two identical texts were produced by the same type or plate.

It is always possible to reproduce the same subject matter by a second printing. If there is a lapse of time between the two, it may mean that the original was reset if letterpress was used, or the original copy was prepared again if offset methods were employed. A second production of this nature may produce slight variants that will distinguish between the two printings.

Fig. 9. Three letter words contained in general text documents

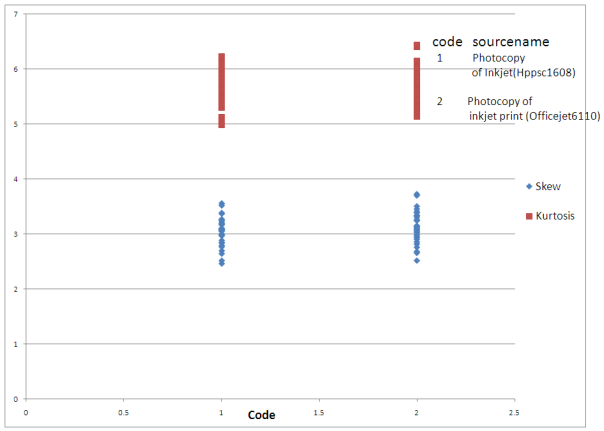



Fig. 10. Skew and Kurtosis of general three letter word photocopy samples

## 6. RECOMMENDED METHODOLOGY FOR PRINTED DOCUMENT SOURCE IDENTIFICATION

For the given questioned document, select the text word like 'the' or 'The' or any three letter word. Fix the sample to minimum bounded rectangle and resize the sample for data generation. The features to be calculated are

1. Print Index
2. Skew
3. Kurtosis

The flow chart shown in Figure 11 demonstrates identification of source of the printed text based on these three features. Skew measure for inkjet print and its photocopy forms two non-overlapping sets. Hence, if any instance leads to inconsistency with skew and kurtosis measures such as skew is 1.5 and kurtosis is 5.2, then it is recommended to identify print technology based on skew measure.

## 7. REFERENCES

[1] http://en.wikipedia.org/wiki/Photocopier.

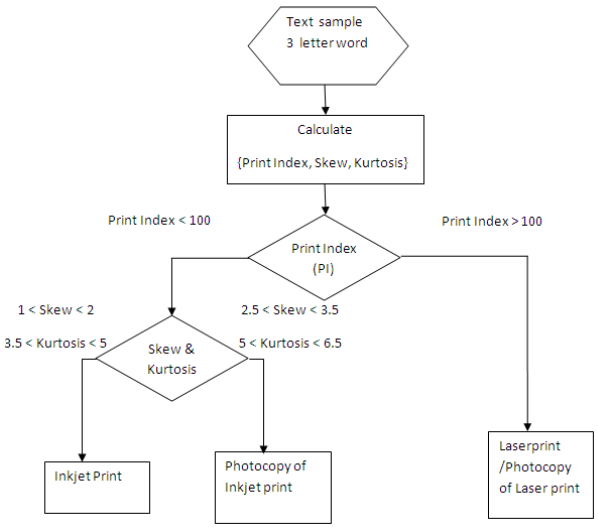



Fig. 11. Printed document source identification

[2] http://www.eff.org/issues/printers.

[3] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm.

[4] http://www.world-english.org.

[5] G. N. Ali, P. J. Chiang, A. K. Mikkilineni, J. P. Allebach, G. T. C. Chiu, and E. J. Delp. Intrinsic and Extrinsic Signatures for Information hiding and Secure printing with Electrophotographic Devices. *Proc. IST's NIP19: International Conference on Digital Printing Technologies,Vol.19*, pages 511–515, 2003.

[6] G. N. Ali, P. J. Chiang, A. K. Mikkilineni, G. T. Chiu, E. J. Delp, and J. P. Allebach. Application of Principal Components Analysis and Gaussian Mixture Models to Printer Identification. *Proceedings of the IS & T's NIP20: International Conference on Digital Printing Technologies, Volume 20*, pages 301–305, Nov 2004.

[7] G. N. Ali, P. J. Chiang, A. K. Mikkilineni, G. T. Chiu, E. J. Delp, and J. P. Allebach. Application of Principal Components Analysis and Gaussian Mixture Models to Printer Identification. *Proceedings of the IS & T's NIP20: International Conference on Digital Printing Technologies, Volume 20*, pages 301–305, Nov 2004.

[8] J. V. Beusekom, F. Shafait, and T. M. Breuel. Document Inspection Using Text-Line Alignment. *Document Analysis Systems,*, pages 263–270, 2010.

[9] C. Bhagvati and D. Haritha. Classification of Liquid and Viscous Inks using HSV Color Space. *Proceedings of Eight International Conference on Document Analysis and Recognition*, pages 660–664, 2005.

[10] P. J. Chiang, A. K. Mikkilineni, R. M. Kumontoy O. Arslan, G. T. C. Chiu, E. J. Delp, and J. P. Allebach. Extrinsic Signature Embedding in Text Document using Exposure Modulation for Information Hiding and Secure Printing in Electrophotography. *Proc. IST's NIP21: International Conference on Digital Printing Technologies, vol. 21*, pages 231–234, 2005.

[11] J. H. Choi, D. H. Im, H. Y. Lee, J. T. Oh, J. H. Ryu, and H. K. Lee. Color Laser Printer Identification by Analyzing Statistical Features on Discrete Wavelet Transform . *ICIP*, pages 1505–1508, 2009.

[12] J. H. Choi, H. K. Lee, H. Y. Lee, and Y. H. Suh. Color Laser Printer Forensics with Noise Texture Analysis . *MMSEC'10*, pages 19–24, September 2010.

[13] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Pearson Education Inc, second edition, 2002.

[14] G. Gupta, C. Mazumdar, M. S. Rao, and R. B. Bhosale. Paradigm Shift in Document related frauds: Characteristics Identification for Development of a Non-destructive Automated System for Printed Documents . *Digital Investigation, Vol. 3*, pages 43–55, 2006.

[15] G. Gupta, S. K. Saha, S. Chakraborty, and C. Mazumdar. Document Frauds: Identification and Linking Fake Document to Scanners and Printers. *Proceeding of the International conference on Computing Theory and Applications,ICCTA07, IEEE*, pages 497–501, 2007.

[16] D. Haritha and C. Bhagvati. Identification of Printing Process using HSV Colour Space. *Asian Conference on Computer Vision*, pages 692–701, 2006.

[17] Z. He and C. A. Bouman. AM/FM halftoning: Digital Halftoning through Simultaneous Modulation of Dot size and Dot density. *Journal of Electronic Imaging*, 2004.

[18] O. Hilton. *Scientific Examination of Questioned Documents*. CRC Press, 1993.

[19] Nitin Khanna, Aravind K, Mikkilineni, Anthony F. Martone, Gazi N. Ali, George T.C. Chiu, Jan Allebach, and Edward J. Delp. A survey of forensic characterization methods for physical devices. *Digital Investigation3s*, pages s17–s28, 2006.

[20] C. H. Lampert, L. Mei, and T. M. Breuel. Printing Technique Classification for Document Counterfeit Detection. *IEEE International Conference on Computational Intelligence and Security*, pages 639–644, Nov 2006.

[21] A. K. Mikkilineni, P. J. Chiang, G. N. Ali, G. T. Chiu, J. P. Allebach, and E. J. Delp. Printer Identification based on Graylevel Co-ocuurence Features for Security and Forensic Applications. *Proceedings of the SPIE International Conference on Security, Volume 5681*, pages 430–440, Mar 2005.

[22] A. K. Mikkilineni, P. J. Chiang, S. Suh, G. T. C. Chiu, J. P. Allebach, and E. J. Delp. Information Embedding and Extraction for Electrophotographic Printing Processes. *Proc. SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VIII, Vol. 6072*, pages 385–396, 2006.

[23] A. K. Mikkilineni, P.J. Chiang, G. T. C. Chiu, J. P. Allebach, and E. J. Delp. Channel Model and Operational Capacity Analysis of Printed Text Documents. *Proceedings of SPIE International Conference on Security , Stegnography and Watermarking of multimedia contents IX, Vol 6505*, pages 65051U.1–65051U.11, January 2007.

[24] S. J. Ryu, H. Y. Lee, D. H. Im, J. H. Choi, and H. K. Lee. Electrophotographic Printer Identification by Halftone Texture Analysis . *ICASSP*, pages 1846–1849, 2010.

[25] C. Schulze, M. Schreyer, A. Stahl, and T. Breuel. Evaluation of Graylevel-Features for Printing Technique Classification in High-Throughput Document Management Systems. *International Work shop on Computational Forensics*, pages 35–46, Aug 2008.

[26] Y. S. Subramaniam, B. Narayanan, K. Viswanathan, and K. Anjaneyulu. Detecting Modifications in Paper Documents: A Coding Approach. *Document Recognition and*

*Retrieval XVII, Proc. of SPIE-IST Electronic Imaging, SPIE Vol. 7534*, pages 75340A1 – 75340A12, 2010.

[27] M. Umadevi, A. Agarwal, and C. R. Rao. Gaussian Variogram Model for Printing Technology Identification . *International Conference on Asian Modelling Symposium*, pages 320–325, 2009.

[28] M. Umadevi, A. Agarwal, and C. R. Rao. Printed Text Characterization for Identifying Print Technology Using Expectation Maximization Algorithm. *Multidisciplinary trends in Artificial indtelligence*, pages 201–212, 2011.