A Survey of Image Processing Techniques for Identification of Printing Technology in Document Forensic Perspective

M. Uma Devi
Research Scholar
DCIS, University of Hyderabad,
Hyderabad-500046, INDIA

C. Raghavendra Rao
Professor
DCIS, University of Hyderabad,
Hyderabad-500046, INDIA

Arun Agarwal
Professor
DCIS, University of Hyderabad,
Hyderabad-500046, INDIA

ABSTRACT
This paper discusses about various image processing techniques and tools which are available for identification of printing technologies. Printing technology identification and associated problems in document forensics have been projected as challenges in image processing application. Various image processing approaches based on textures, spatial variation, HSV color space, spatial correlation, and feature based on histogram and some of the pattern recognition methods, like gray level co-occurrence matrix, roughness of the text, perimeter of edge are highlighted. This paper devotes more on one of the recent contribution, namely, Gaussian Variogram Model (GVM) for printer classification.

General Terms
Image processing, Document forensics-printing technique classification, spatial statics

Keywords
Questioned Documents, Document Forensics, Variogram, Gaussian Variogram, Reduct, Reduct based decision tree (RDT).
.

1. INTRODUCTION
Document is any material that contains printed information conveying some meaning or a message [1]. Document generally contains information about transactions like agreements, wills, ownership of properties etc. Being the legal evidence of transaction, it's necessary to state the genuineness of a document. Questioned documents are those suspected of being fraudulent or whose source is unknown? For the questioned document whose authenticity is doubtful, it's necessary to identify the source of the document. For assisting interpretation of evidence in courts, new field of document forensics has emerged.

Document forensics deals with getting evidence from the questioned documents. Creation of fraud documents is increasing with the growth of new technology. These days' personal computers, scanners and printers are good enough to generate fraud documents, like certificates, agreements, identity cards and lottery tickets etc. Printed document is spatial distribution of marking material. A printer produces a document to the extent it can match the pattern of the document. Scanner produces images by capturing document information according to the calibration and specifications of the scanner. This image will be subjected to tampering in order to produce fraudulent document, which may be printed by the same or another printer. Hence, a forged document contains composite features of the above processes. Thus, printer identification of the questioned document is a highly involved and complex process. A printed questioned document examination, by forensic scientist starts with the identifying task and the printer or source from which document has been created.

The scope of document forensics is listed below.

- Identification of handwriting and signatures.
- Identification of a document as forgery
- Identification of typewriters check, writers, photocopies
- Detection of alterations, additions, deletions.
- Deciphering obliterations, alterations, erasures.
- Identification and deciphering of indented writing.
- Comparison of inks and identification of type of writing instrument.
- Printer Identification of the document
- Two documents are similar, i.e. printed using same technology

In the context of printed questioned documents examination, forensic analyst has to answer questions like

a. Is the document consistent? Which means whether the printed content in the document is prepared from a single source?

b. Identification of source printer or printing techniques like ink jet, electro-photography printing etc.

Printed documents contain features of a printer depending on the specified procedure used by manufactures for placing the marking material on the paper. Printed documents exhibit differences in the print pattern: number of drops per dot and technology used to print, like drop on demand thermal printing, Hp photo Ret[2] technology, Laser technology, etc. Instruments used by Document Examiner to distinguish genuine document from forged are high resolution microscopes and Video Spectral Comparator (VSC000)[3]. VSC is multispectral imaging system which works on concept of separation of wavelengths of light spectrum ranging from ultraviolet to infrared. The principal functions of VSC are manipulation of visual contrast for revealing evidence of document tampering, measurement; and

comparison for detecting small differences within or between documents. It has an extensive range of facilities for detecting irregularities on altered documents. High resolution microscopes like LEICA MZ 8, LEICA MZ 12.5 are used to observe the pattern in the document. These instruments are useful in identifying the characteristics of a document but they have no mechanism for classification. These instruments are expensive.

As the current methods and instruments used are expensive in capturing the data as well as in analyzing (time and space), there is a great need to develop alternative solutions for forensic characterization of print in terms of cost, space and time.

2. OVERVIEW OF RECENT RESEATCH IN DOCUMENT FORENSIC

Recent research publications demonstrate various approaches suggested for discriminating printing techniques. Research activities on characterization of electro photographic printers in [4], gray level co-occurrence feature in [5] and most frequently occurring letter 'e' and Gaussian mixture model(GMM) in [6] are the techniques used for printer identification.

In GMM, principal component analysis is used as dimension reduction technique to obtain 1-D projections of the extracted text character. These researches are exclusively for the identification of electro photographic printers.

Features of the printed document, which are unique to printer model or manufacturer's product, is referred to as intrinsic signature. Characterization of electro photographic printers [6] is for finding intrinsic signature of printers based on banding signals to discriminate various electro-photographic printer models.

Machine identification code project [7] identifies presence of pattern of yellow dots in color 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. Some printers like Samsung clp-510 series, Hp laser 8550 series are not showing any yellow dots. Still there is some forensic information to keep track of the printer model. Identification of printing process using HSV color space by Haritha [8][9], is based on hue histogram for identification of printing process and photocopy. Hue contrast, periodicity and ink overspray are the features selected for classifying ink jet, laser jet and photocopies. This is based on color image processing technique using HSV color space.

Identification and linking fake documents to scanner by Gaurav Gupta[10] proposed new method for identifying fraudulent documents and linking it to color laserjet priter or color inkjet printer. In this proposed work they captured images text using high resolution cameras LEICA MZ 8, LECIA MZ 12.5 to capture magnified image of single character and directly transfer the images to computer. Unique color count and texture feature uniformity and intensity variation are used as parameters for distinguishing fraud documents.

Gray level features, proposed by Lampert [11] for discriminating ink jet from laser jet print, are based on high resolution scanned images, e.g. 3200 dot per inch. Recent research is concentrated on evaluation of gray level features like perimeter based edge roughness of the text [12] for print technique classification, based on low resolution image for high throughput document management system.

From the literature review, one finds that, research in identification of print technology is a very challenging area and there is a great need to develop forensic examination techniques to characterize the documents based on the printing processes.

3. PRINT TECHNOLOGIES

With emergence of print technology various types of printers are available. Commercially available printers are categorized as Electrostatic, Inkjet, Thermal and Photography [13]. Electrostatic print technology uses static charge pattern or selective charge pattern to attract toner particles. Thermal printers use heat to transfer marking material. Ink Jet printers use continuous stream of droplets selectively towards paper or use drop-on-demand process.

The way the marking material is placed on the paper changes with technology. This spatial distribution of marking material on the paper can be used for characterization of a printed document. Print technology classification involves identification of associated print patterns as features or characteristic of printed document. Hence they vary in print pattern. The inkjet printers include continuous, drop-on demand, thermal print types. Drop-on-demand thermal ink jet print technology uses heat to generate vapor bubble to eject a single drop of ink through print head nozzles only when activated. Conventional color inkjet printer places up to 8 different colors for dot. But HP Photo REt technology uses tiny drops to produce photorealistic images. It can place maximum up to 29-32 drop per dot by decreasing size of drop to 4-5 Pico liters. Study of spatial statistics of homogeneous color regions of images printed by various printers is used for its identification. Documents like certificates, identity cards, and letter heads containing uniform color region, reveal spatial features of print pattern of source printer.

4. CHALLENGES IN DOCUMENT FORENSICS

Most of the image processing research techniques is concentrated on classifying the different print documents. Different categories of document are shown in Figure 1a and 1b.

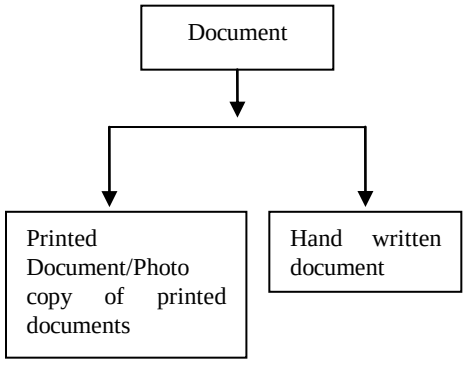

Figure 1a. Categories of documents

Identification of printed document and photocopies is based on the techniques of Color image processing. Color image

processing techniques used for identification of printing process, employed HSV color space. In this work, hue histogram is used to identify between the printed or photocopied document. Generally, the hue histograms are bi-modal and wider for photocopied documents, whereas it is uni-modal and narrower for printed document.

Identification of inkjet print and laser print are done by using features of hue contrast and edge detected hue and saturation images. Inkjet print has large number of isolated dots near the strokes and it has no variation in contrast on opposite sides of the strokes, in an edge detected hue images. While laser print has less number of isolated dots, alternating low and high contrast in opposite sides of the stroke in edge detected hue images. Another distinguishing feature defined by Haritha [8] & [9] for identification of laser jet print is periodic variation in column wise intensity profile of edge detected saturation images.

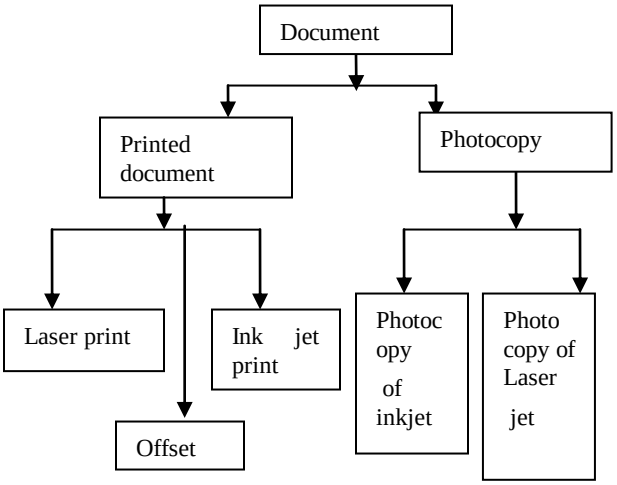

Figure 1b. Detailed classification of Documents

Generally, the printed document contains information both in form of pictures and printed text. In a picture, uniform color region of an image is taken for analysis of spatial pattern. The techniques used for classifying printed document categories are based on uniform color region of the image or based on text of the image.

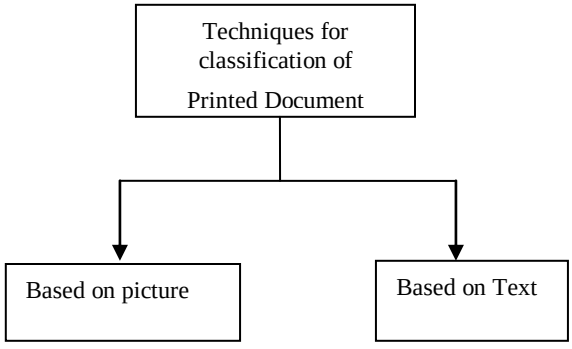

Figure 2. Techniques for classification of Documents

4.1. Method based on uniform color region of an image

Identification of electro-photographic printers is based on frequency analysis of banding signal in large mid tone area. This method has proved that different printers have different banding frequencies based on the brand and model of the printer. These results are reliable for 12.5-50% filled gray level patches.

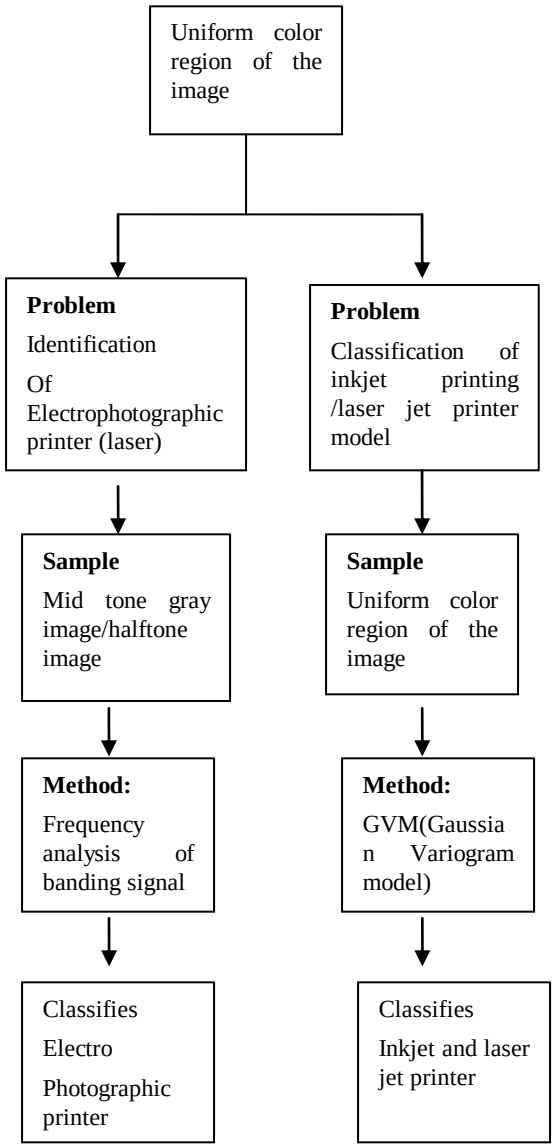

Figure 3. Classification based on uniform color region

The technique that identifies image based on uniform color region of the image is Gaussian variogram model [14], which characterizes the spatial characteristics of the print pattern in the form of variance. This technique is used for classification of inkjet print versus laser jet print and is shown in Figure 3.

4.1.1. Algorithm for Classification of Print technology using GVM

1. Select uniform color region of the image as sample
2. Convert sample to gray level image
3. Generate Variogram for the specified direction
   a. Plot Variogram as lag or distance versus variance
   b. Identify range , sill, nugget
4. Model Variogram using Gaussian curves of 5th order
5. Select parameter of Gaussian Variogram as feature set
6. Normalize feature set data using Z-score normalization method
7. Use normalized dataset for different printers as training data
8. Reduct based decision tree classification involves
   a. Calculate reduct
   b. Generate rules based on Reduct to identify the print technology of test data

4.1.1.1 Samples

Same image is printed at 600 dpi on different printers mentioned in Table 1 and scanned at 2400dpi using Hp Scanjet scanner. A uniform color region is cropped as sample of size of 127x127 pixels. Selection of a sample of a uniform color region of an image is shown in Figure 4a.

4.1.1.2 Variogram

Variogram [15] is a statistical tool that characterizes spatial continuity or roughness of data set. Variogram function gives an average dissimilarity between points separated by distance in specific direction in the form of parameter "h". Variogram represents both structural and random aspects of data under consideration. The separation distance is usually referred to as lag. Variogram has been widely used for remote sensing applications [16] and classification of geo-statistical textures [17]. The range of variogram represents structural part of Variogram model.

Representation of Variogram is in the form of graphical representation, x-axis representing the lag and y-axis representing the semi variance. Hence the variogram is plotted as distance/or lag versus variance.

$$V(h) = 1/n \sum_{i=0}^{i=n} (f(x_i + ha, y_i + hb) - f(x_i, y_i))^2 \quad \text{--(1)}$$

Where V(h) is the variance at location (x_i, y_i) with lag h in the direction (a, b), when a=0, b=1 the direction is OX (x-axis).

Parameters related to variogram are range, sill and nugget. Each parameter has its own significance. Range of the variogram represents structural model of the data and size of texture [18] that is contained in the data. Sill is the maximum variance of that variogram from where the variogram level falls off. Variogram analysis deals with experimental variogram calculated from the data. Homogenous color region of image is selected as sample for GVM analysis, as shown in Figure 4a. Samples of the same image printed on different printer and their corresponding variograms are shown in Figure 4b and 4c, respectively.

Table 1. List of printers used for study of spatial patterns

Pid	Manu-facturer	Model	Dpi	Print technology
1	HP	Deskjet930c	600	HP Photo REt III
2	HP	Deskjet840c	600	Drop-on-demand thermal inkjet
3	HP	Hppsc1608	600	Drop-on-demand thermal inkjet
4	HP	Hppsc1315	600	Drop-on-demand thermal inkjet
5	HP	Officejet6110	600	HP PhotoREt III
6	HP	Photosmart3188	600	Drop-on demand thermalinkjet
7	HP	Laser 4650	600	Laser
8	HP	Colorlaserjet4550N	600	Laser
9	Samsung	CLP-510	600	Laser

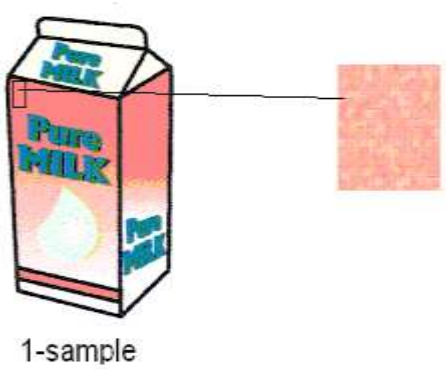

Figure 4a. Selection of homogeneous color region as sample

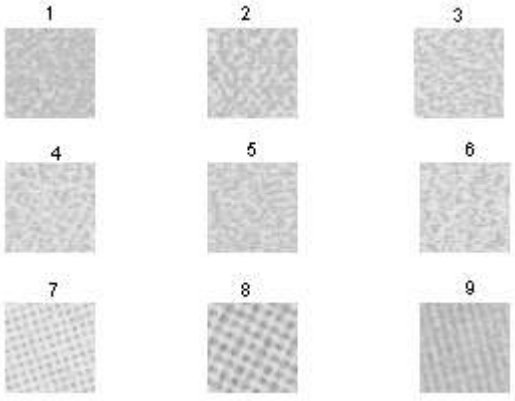

Figure 4b. Gray level converted sample which is printed on different printers listed in Table 1

Figure 4c shows different variogram pattern for different printer models. Variogram of laser print with id 7, 8, 9 shows the periodicity in variogram.

Variogram of an image corresponding to a laser printer show peaks at regular intervals which depend on the size of texture in

print patterns. We can observe here local as well as global periodicity in laser technology. Range of variogram is directly related to the size of texture in print pattern. Sill is a maximum variance in variogram and it is directly proportional to the global object variance**.**

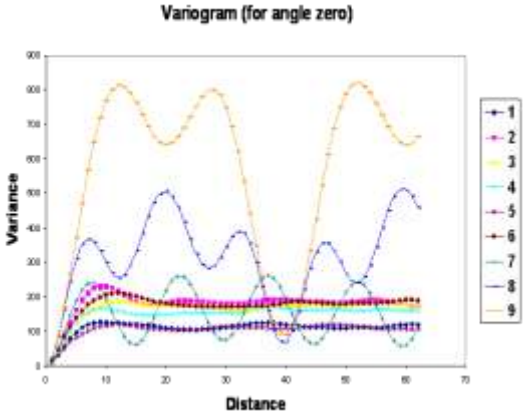

Figure 4c. Variogram of image in figure 4b

4.1.1.3 Gaussian Variogram Model

Variogram analysis is related to concept of modeled variogram. The maximum lag selected for the calculation is half of the field size. Variogram cannot be calculated at every lag distance due to variations in the estimation and it is not ensured to be valid. For ensuring validity, variograms should be modeled by mathematical function. This empirical variogram is fitted to mathematical model function is called Model Variogram. Spatial statistics study begins with construction of model to characterize spatial patterns of the sample under study. There is need to interpolate variogram function [19] for specified distance. Covariance part of the variogram must have positive definiteness of mathematical functions. Then only it is possible to use variogram for krigging and stochastic simulation applications. These are few reasons why the variogram is to be modeled. The most popular models used to fit variogram are exponential model, spherical model and Gaussian models. The variogram fitted to mathematical model of Gaussian function are called Gaussian model. Modeled Gaussian function is shown in equation 2. Gaussian function is one of the mathematical function used to model the empirical variogram. Such modeled variogram is called as Gaussian Variogram Model.

$$f(x) = a_i \exp - ((x - b_i)/c_i)^2 \quad \text{------ (2)}$$

4.1.1.4 Selection of feature set

Gaussian Variogram Model (GVM) characterizes print technology based on spatial variability of homogeneous color regions. The parameters from Gaussian Variogram Model GVM are $\{a_1, b_1, c_1, a_2, b_2, c_2, a_3, b_3, c_3, a_4, b_4, c_4, a_5, b_5, c_5\}$. These parameters along with Sill and Nugget of variogram are taken as feature set. Each sample gives feature set consisting of 17 parameters. Samples collected from each printer are modeled as GVM, which form feature set of corresponding print technology. GVM data set obtained from each sample is taken as training data set. This data set has to be normalized before it can be used as training data.

4.1.1.5 Normalization of data

Normalization is a procedure which transforms distribution of data into standard normal form. Z-score normalization method, transforms selected GVM data set into normalized data, which are used as training set for classification. This transformation makes data more comparable. Z-score normalization discretizes the data based on scaling factor which is referred to as discretization factor. For example scaling factor k=0.5, increases the interval of the data by 2 times. Features of Gaussian variogram model are given as input to reduct based decision tree (RDT) [20] to generate rules for fixing the printing technology.

4.1.1.6 Reduct based decision tree classification

Reduct based decision tree consists of two steps, first step is reduct computation and second step is decision tree construction. Reduct based decision tree construction combines merit of rough set and decision tree construction algorithm. Datasets can be discrete or continuous, but here we have used discrete data set (after discretization). The predominant attributes of the data grouped together is called as reduct. This reduct of the training set is taken to generate decision rules, which are used for classification of print technology. This method classifies inkjet printing technologies as well as laser technologies.

4.1.1.7 Experiment results

In our experiment, a total 159 samples are collected. Samples are collected from each of the printer listed in Table 2. 116 samples from 4 printer, (that is 29 samples from each printer) is taken as training set and 43 samples from the same 4 printers (11 samples from printers with pids 1,4 and 5 and 10 samples from printer with pid 6) are taken as test data set.

Table 2: Printers Used for Identification

Printers used for identification			
Pid	Manu-facturer	Model	Print technology
1	HP	Photosmart3188	Drop-on-demand thermalinkjet
4	HP	Officejet6110	HP PhotoREt III
5	HP	Colorlaserjet4550N	Laser
6	Samsung	CLP-510	Laser

Table 3: RDT results for GVM data

RDT RESULT FOR GRAY GVM ALONG X-AXIS			
ANGLE	K	REDUCT1	ACCURACY
0	0.1	'b2' 'c3' 'nugget'	37.2
0	0.2	'b2' 'c4' 'c5'	41.8
0	0.3	'b2' 'b4' 'c5'	51.16
0	0.4	'b1' 'nugget' 'c2' 'b5'	67.4
0	0.5	'b2' 'b5' 'a1' 'c3'	62.7
0	0.6	'b2' 'c4' 'nugget' 'b5' 'c5'	62.79
0	0.7	'b1' 'sill' 'c2' 'b3' 'b5'	60
0	0.8	'b1' 'sill' 'c1' 'b2' 'b5'	79
0	0.9	'b4' 'b2' 'b1' 'sill' 'c5' 'c4'	65
0	1	'b4' 'b2' 'sill' 'b1' 'c1' 'c4'	72

Identification of print technology using reduct based decision tree (RDT) for different discretization factors, the reduct selected and identification accuracy obtained are shown in Table 3. Number of samples identified correctly out of total number of samples is represented as percentage of accuracy.

4.2. Methods for identifying printing technology based on text

Forensic characterization of printed text in [5] is on particular character 'e'. The letter 'e' is extracted from the document scanned at 2400 dpi. The texture of printed region of character is modeled using gray level co-occurrence texture features and pixel based features. These features are classified using SVM classifier.

Application of principal component analysis and Gaussian mixture models (GMM) is another method for identification of the electro-photographic printers [4]. In this identification method each printer is represented as Gaussian distribution. Gaussian mixture model is combination of several different Gaussian distributions. Principal component analysis is used as dimension reduction technique before classification by GMM.

Identification and linking of fake documents to scanner and printers by Gaurav Gupta [10] identified specific characteristics of printed color text. This work is based on the Exchange principle of forensic science, which is known as Locard principle.

Essentially Locard's principle is applied to crime scenes in which the perpetrator(s) of a crime comes into contact with the scene, so the perpetrator(s) will both bring something into the scene and leave with something from the scene, ie. every contact leaves a trace. To quote

"Wherever he steps, whatever he touches, whatever he leaves, even unconsciously, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibers from his clothes, the glass he breaks, the tool mark he leaves, the paint he scratches, the blood or semen he deposits or collects. All of these and more, bear mute witness against him. This is evidence that does not forget. It is not confused by the excitement of the moment. It is not absent because human witnesses are. It is factual evidence. Physical evidence cannot be wrong, it cannot perjure itself, it cannot be wholly absent. Only human failure to find it, study and understand it, can diminish its value".

Therefore, the process of creation of fraud documents introduces impurity leading to variation in the intensity. Hence total number of unique color count varies. Based on this fundamental principle, identified characteristics of printed character is unique to specific printer/scanner.

In this study [10] two color printers used are referred to as p1, p2 and two scanners are referred to as s1 and s2 to produce fraudulent document of original document. Images of single character 'a' and 'e' were captured and magnified using high resolution cameras like LEICA MZ8, LEICA MZ12.5. Variance of Intensity, texture based parameters like gray level co-occurrence matrix and distribution of pixels in color cube are the parameters used to distinguish original document from fraudulent document. Variance of Intensity and total number of unique color count increases for the fraud document as compared to genuine one. Uniformity measure of gray level co-occurrence matrix decreases for fraud document as compared to genuine one. In this work they also proposed energy of difference histogram as an important parameter which remains constant with print technology. This parameter is useful for identification of printing technology like inkjet and laser jets.

Printing technique classification for detection of counterfeit [11] classifies letters printed by inkjet and laser jet. At low resolution, the roughness of contours of inkjet and laser jet is equal. At high resolution, contours of LaserJet print have low roughness and inkjet print has a rougher contour. So the ratio of contours length between high resolution and low resolution is taken as feature for classification. Another important feature to consider is area difference. As we slightly increase the size of the binary image, the increase in the number of additional pixels is larger for inkjet print.

Recent research on extraction of gray level features from low resolution image for identification of printing technique in document management system has also been used for print technique classification [12].

5. CONCLUSION

The techniques and methods discussed in above sections are based on the high resolution images. Most of the tools developed are for color images and few categories (types) of printers. One should also focus and develop tools for identification of offset printers, gray scale images in place of color images, enhancement in GVM approach in particular to low resolution environmental setup.

6. REFERENCES

[1] Ordway Hilton. Scientific Examination of Questioned Documents. CRC Press, 1993.

[2] http://www.hp.com.

[3] http://www.fosterfreeman.com

[4] 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. In Digital Investigation3s, pages s17–s28, 2006

[5] Mikkilineni A. K., Pei-Ju. Chiang, Ali G. N., Chiu G.T., Allebach J. P., and Delp E. J. 'Printer Identification based on Graylevel Co-ocuurence Features for Security and Forensic Applications'. In Proceedings of the SPIE International Conference on Security, Volume 5681,, pages 430–440, Mar 2005.

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

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

[8] Harith D and Chakravarthy B. 'Identification of Printing Process using HSV Colour Space'. In Asian Conference on Computer Vision, pp 629-701,2006.

[9]Chakravarthy B and Haritha D. 'Classification of Liquid and Viscous Inks using HSV Color Space'. In Proceedings of Eight International Conference on Document Analysis and Recognition, 2005. pp 660-664.

[10]Gaurav Gupta, Sanjoy Kumar Saha, Shayok Chakraborty, Chandan Mazumdar, Document Frauds: Identification and Linking Fake Document to Scanners and Printers, Proceeding of the International conference on Computing Theory and Applications,ICCTA'07, IEEE, pp 497-501, 2007.

[11] Christoph H. Lampert, Lin Mei, and Thomas M. Breuel. 'Printing Technique Classification for Document Counterfeit Detection'. In IEEE International Conference on Computational Intelligence and Security, pages 639–644, Nov 2006.

[12] Christian Schulze, Marco Schreyer, Armin Stahl, and Thomas Breuel. 'Evaluation of Graylevel-Features for Printing Technique Classification in High- Throughput Document Management Systems'. In International Work shop on Computational Forensics, pages 35–46, Aug 2008.

[13] Gary K.Starkweather. Electronic color printing technology. In IEEE Proceedings of COMPCON'96-41st IEEE International Computer Conference, pages 435–439, 1996.

[14] M Uma Devi, Arun Agarwal and C.Raghavendra Rao 'Gaussian Variogram Model for Printing Technology Identification' International Conference on Asian Modeling Symposium , pp 320-325,2009.

[15] http://www.goldensoftware.com

[16] A. Wijaya, P.R.Marpu, and R.Gloaguen. Geostatistical Texture Classification of Trophical Rainforest in Indonesia. In 5th ISPRS International Symposium on Spatial Data Quality, 2007.

[17] C.A.Coburn and A. C. B. Roberts. A multiscale texture analysis procedure for improved forest stand classification. International Journal of Remote Sensing, Vol.25:4287–4308, 2004.

[18] A. Jkomulska and K.C. Clarke. Variogram derived measures of textural image classification-application to large scale vegetation mapping. In Proceedings of the Third European Conference on Geostatistics for Environmental Applications, pages 345–355, 2000.

[19] E. Gringarten and C. V. Deutsch Teachers Aide: Variogram Interpretation and Modeling, Mathematical Geology, Vol.33 (4) pages 507-534, 2001.

[20] Y. Ramadevi, C. R. Rao, and Vivekchan Reddy. Decision tree induction using roughset theory comparative study. In Journal of Theoretical and Applied Information Technology, pages 110–114, 2007.