Identification of Printing Techniques: A Document Forensics Approach

A thesis submitted during 2011 to the University of Hyderabad in partial fulfillment of a Ph. D degree in Department of Computer and Information Sciences, School of Mathematics and Computer & Information Sciences

by

Maramreddi Umadevi

Department of Computer and Information Sciences
School of Mathematics and Computer & Information Sciences

University of Hyderabad
(P.O.) Central University, Gachibowli
Hyderabad - 500 046
Andhra Pradesh
India

C E R T I F I C A T E

This is to certify that the thesis work entitled **"Identification of Printing Techniques: A Document Forensics Approach"** submitted by **Mrs. M. Uma Devi** bearing (Regd. No. **05MCPC02**) in partial fulfillment of the requirements for the award of **Doctor of Philosophy in Computer Science** is bonafide work carried out by her under our supervision and guidance.

The thesis has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Prof. Arun Agarwal
(Supervisor)
Department of Computer and
Information Sciences,
University of Hyderabad,
Hyderabad.

Prof. C. R. Rao
(Supervisor)
Department of Computer and
Information Sciences,
University of Hyderabad,
Hyderabad.

Head of the Department

Dean of the School

DECLARATION

I, M. Uma Devi hereby declare that this thesis entitled **"Identification of Printing Techniques: A Document Forensics Approach "** submitted by me under the guidance and supervision of **Prof. Arun Agarwal** and **Prof. C. Raghavendra Rao** is a bonafide research work. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma.

M. Uma Devi

Date:

Signature of the Student:

Regd. No.05MCPC02

ABSTRACT

Forensic document examination plays an important role in providing the evidence to the court related to disputed documents. Emerging printing technologies are posing challenges to document examiner in identification of source of document. A printed questioned document examination by forensic scientist starts with identifying the printer or source from which document has been created. As the current methods and instruments used are expensive in capturing the data as well as in analysing, there is a great need to develop alternative solutions for forensic identification of print technology that is most effective in cost, space and time.

This thesis focussed on implementing methodologies for providing sophisticated tools for assistance in document examination. Printed documents contain features of a printer depending on the specified procedure used by them for placing the marking material on the paper. In this Image analysis methods along with statistical tools applied to study the class characteristics of document for identifying the source of the document are discussed. To address some of the issues in forensic document examination, we designed and developed methodologies to differentiate various print techniques such as inkjet, laserjet and photocopy. We determine the important features of the document that distinguish between various print technologies. The techniques proposed are both region and text based and it helps in checking the consistency of the document.

The present work achieved 75 % and 96% accuracies for homogeneous region based and text based identification of print technology. The developed approach is based on the benchmark data set prepared which is set of homogeneous regions and most frequently used words like 'the'. The demonstration of the proposed methodology and its robustness for general three letter words is presented. Robustness of Print Index measure for varying font type and size to be exercised.

Acknowledgements

I owe my deepest gratitude to my supervisors Prof. Arun Agarwal and Prof. C. Raghavendra Rao whose valuable guidance and support from preliminary to concluding level enabled me to develop an understanding of the subject. I am thankful to them for their patience and constant encouragement in motivating me with the words of hope throughout this work.

It is pleasure to thank DRC committee, Dr. Chakravarthy Bhagvati and Dr. Bapi Raju for their valuable suggestions in completing this work. I sincerely thank to Dr. Rajeev Wankar for spending his valuable time and giving advice concerning this thesis work.

I wish to thank Head, Department of Computer and Information Sciences, Prof. C. Raghavendra Rao for providing me with the facilities. I would like to express my sincere thanks to Dean, School of Mathematics and Computer & Information Sciences for his cooperation.

I would like to thank Directorate of Forensic Science and GEQD office for providing the fellowship Ref. No. 87(1)/GEH/JRF/SRF for the period of Dt: 28-02-2005 to 27-02-2009 to pursue this research in the field of forensic science. I am thankful to officers of GEQD, Hyderabad, Shri. Mohinder Singh, Government Examiner of Questioned Documents, Shri. Y. S. Prasad, Shri. P. Krishna Sastry and Shri. M. Krishna for their valuable suggestions and for allowing me to use required equipment, whenever I needed.

I would like to avail this opportunity to thank my parents, Shri M. Srirama Reddy and Shrimati Sambrajyam whose good wishes enabled me to purse and achieve my goal. I gratefully thank my uncle and aunt, Shri R. Nagashankar Reddy and Shrimati M. Malleswari for their concern and moral support through out my academics.

I would also like to thank my elder brother, sister and sister-in-law for encouraging me with their best wishes. I am also thankful to my husband Shri Y. Mallikarjuna Reddy for his effort and cooperation throughout this journey.

The warm support of all my friends in University namely, Sumagna, Sapna, Sandhya, Mini, Nagalakshmi and Vani enabled me to complete this thesis and have a wonderful time along the way.

Uma Devi. M

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Document forensics

Documents play a vital role in financial, legal, social and personal life of any human being. From the time of birth to death of a human being everything is in the form of documents like birth certificate, identity of a person, ownership of properties, business transactions and death certificate. Modification in any part of a document alters the meaning of the document. It causes loss to one of the party involved in the transaction. Documents suspected of being fraudulent or whose source is unknown are called Questioned documents. For those Questioned documents whose authenticity is doubtful, it is necessary to identify the source of the document. Being the legal evidence of transaction it is necessary to state the genuineness of the document. For assisting interpretation of evidence in courts new field of Document forensics has emerged and it deals with getting evidence from the Questioned documents.

1.1.1 Documents

Document is any material that contains written or printed information conveying some meaning or a message to someone [1]. It contains symbols, marks or signs. Depending on the material and instruments used in preparation of the document

they are categorised. This section describes the various categories of document and materials used in preparation of the same.

A. Categories of documents

Documents are generally paper based and classified as handwritten, typed, printed or photocopied dependent on the technology used to produce them. These categories of documents are shown in Figure 1.1. Hand written documents are pro-

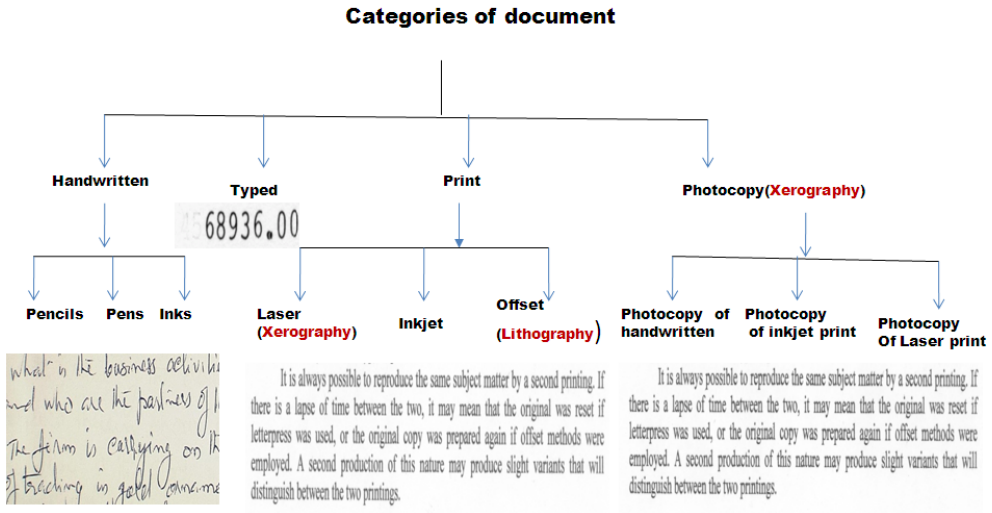


Figure 1.1: Categories of documents

duced by individuals using writing instruments and materials. Many styles are used in handwritten documents over centuries. Writing instruments include pens, pencils, inks and marker.

Typed documents are produced using typing machine. Typed documents vary depending on the typefaces of the machine. Each type machine has type letters particular to that machine. They will have specific physical feature depending on the make of the machine.

Printed documents are produced using various printing mechanisms. Printed documents contain features of a printer depending on 1) the specified procedure used by them for placing the marking material on the paper and 2) inks or toners used in that process. They differ in the print pattern, number of drops per

dot and technology used to print, like Drop on demand thermal printing e.g., HP Photo REt [2] technology, Laser technology etc. Printing instruments commercially available in market today are inkjet printers, laserjet printers and offset printers. Toners used in the printer can be classified as liquid or powdered toners.

Photocopied documents are produced by copying original printed or handwritten documents by electro photography procedure. Photocopied documents look similar to printed document for the untrained eye. Photocopiers are machines, which produce documents similar to the original documents. Photocopiers, like laser printers are electrostatic machines in all respects, only difference being that laser printer produce grid pattern in document. Most of forged documents like counterfeit currency are produced using colour photocopiers or colour printers.

B. Material used to prepare documents

i. Pens: Pens are available in various models like ball point, gel, porous tip pen, roller or fountain pens. The marking tip of ball point pens has a small freely rotating ball bearing that rolls the ink onto the paper. Many of these pens use highly viscous inks. Porous tip pen has porous writing point, which spreads fluid ink on the paper. The ink tends towards intense colours and these pens deliver heavy and quick drying stroke [1]. These strokes are distinctive and can be distinguished from strokes of ball point and fountain pens. Roller pens are ball point pens which use fluid ink. Differences in inks distinguishes it from ball point pens. Fountain pens are mostly known as nib pointed pens and their writing characteristics vary. The width of the nib point and its degree of flexibility are factors in this variation.

ii. Paper Type: Various kinds of papers are used for writing or printing purposes. All papers are composed of closely matted fibers. Specially treated wood fibers are used. Certain papers undergo special finishing to smooth the surface by coating with clay or starch. In type writing paper such special coating is used to facilitate easy erasing. Watermarks on writing papers are the identifying

characteristics that are designed according to manufacturer.

iii. Ink type: Writing ink is liquid used to produce writing on the surface of paper and its colour pigment gives colour to the writing. Until ball point pens became popular the common writing instrument was ink. Earliest material to produces permanent ink was carbon; later various natural dyes were used for preparing ink. Inks can be classified as viscous or non-viscous inks.

1.1.2 Questioned documents

Questioned documents are documents whose authenticity is in doubt, that is documents which are suspected of being fraudulent or whose source is unknown. All the questioned documents may not be fraud documents. A questioned document may be genuine. They may be partially or completely faked by obliterating, erasing or altering the content in the document. A partially fake document is produced by adding or erasing the content of a document by which meaning of the document is changed. Examples of completely faked documents are counterfeit currency, bus tickets and lottery tickets. Documents of less value like bus tickets, bus pass can be reproduced using simple and cheaper production techniques. Modification in the original form of an artifact by which one person gains or intends to gain an advantage over another person is termed as a "forgery".

1.1.3 Forensic science

Forensic science is the application of a broad spectrum of sciences to answer questions of interest to a legal system [3]. The discipline divides neatly into halves, like the term that describes it. The word "forensic" comes from the Latin adjective forensis meaning " of or before forum" and the word "science" is the collection of systematic methodologies used to increasingly understand the physical world [4]. Forensic science applies various aspects of scientific and technological methods in collection of evidence, reconstruction of crime scene and provides scientific

explanation of evidence such that it convinces courts, both the parties (accused of crime and accuser) and the general public.

Evidence collection is the starting step of forensic investigation. Forensic evidence is sometimes known as hard evidence [5] as it never gets confused, it never forgets and it never lies. Tracing evidence is based on the fact that when two objects come into contact with each other they exchange trace evidence. In its simplest form it states that "Every contact leaves a trace". This is the basic principle of forensic science known as Locard's Exchange Principle [6]. It is formulated by Dr. Edmond Locard and other works explains this principle as "Wherever he steps, wherever he touches, whatever he leaves, even without consciousness, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibres 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." [7]. Evidence collection and its analysis uncover the truth to provide justice.

Within forensic science there are number of individual disciplines. Criminalistics, Forensic Anthropology, Computational Forensics, Forensic Chemistry, Forensic Botany, Forensic DNA analysis, Forensic Serology, Forensic Toxicology, Digital Forensics, Forensic Device Forensics, Forensic Document Examinations are some areas, which come under Forensic science. Each discipline has its own technological methods for tracing evidence. Forensic document examination is the application of allied sciences and analytical techniques to questions concerning documents. Examination is the act of making a close and critical study of any material and with questioned documents it is a process necessary to discover the facts about them. Various types of examinations are undertaken including microscopic, visual, photographic, chemical, ultraviolet, and infra red examinations.

1.1.4 Problems related to questioned documents

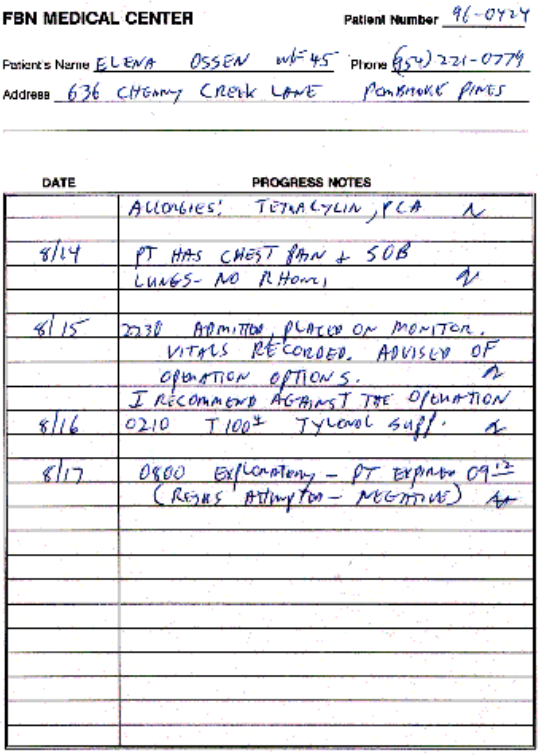
The scope of document forensics in different document problems is listed below.

a. Identification of handwritten documents: Handwritten identification [8] is based on the scientific principle "no two people write exactly alike". Writing is affected by many influences which results in a unique style of writing ability in each person and these influences continues through out the life of an individual. In extended writing samples of single person, no two writings will be exactly the same, this being the second basic principle of hand writing identification. So a document examiner must have the skills to distinguish between natural variation and a different writer. Document examiner compares the questioned signatures with the admitted signature to give his opinion that these questioned signatures are genuine or forged. Figure 1.2 shows an original questioned medical record suspected of being changed at a specific area "I recommend against the operation". A document examiner is equipped with special training to recognize and evaluate characteristics of document.

b. Identification of forged documents: Document examiner answers questions like: is the document genuine or forged?. If it is a forged document, then there is need to identify whether the document is partially or completely forged.

c. Identification of typewriter: This involves identification of the make and model of type writer on which typed question document was produced. Typed document examination involves comparing slight variations in alignment of the letters and uneven wear of type faces. Most type writers have their own signature that is it's own type faces which allows typed question document to be linked to the type writer on which it was produced.

d. Deciphering obliterations, alterations and erasures: Obliteration is the act of hiding information from or removing all signs of authentication of a



FBN MEDICAL CENTER Patient Number 96-0424

Patient's Name ELENA OSSEN WF 45 Phone (954) 221-0774

Address 636 CHERRY CREEK LANE POMPANO PINES

DATE	PROGRESS NOTES
	ALLERGIES: TETRACYLIN, PCN
8/14	PT HAS CHEST PAIN + SOB
	LUNGS- NO RHONCI
8/15	2230 ADMITTED PLACED ON MONITOR.
	VITALS RECORDED. ADVISED OF
	OPERATION OPTIONS.
	I RECOMMEND AGAINST THE OPERATION
8/16	0210 T 100.2 TYLENOL GIVEN.
8/17	0800 EXPLORATORY - PT EXPIRED 09:12
	(RESULTS ATTACHED - NEGATIVE)

Figure 1.2: Questioned document

document by adding strikeouts with pen. Figure 1.3 shows an obliterated document in which 303 is overwritten hiding the original information contained in the document.

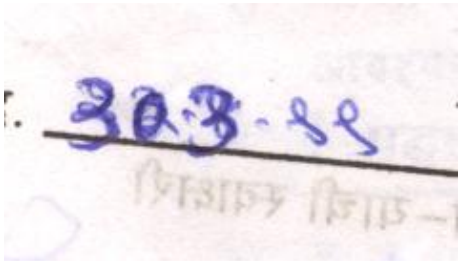


Figure 1.3: Obliterated document

Changing or modifying text so that the overall meaning of the text is changed and interpreted wrongly by the user is known as alteration in a document. For example making '0' as '8' changes the meaning of the content in the document. Obliterations [9] and alterations in the document is identified by Band pass/reject,

Channel analysis and Stroke analysis algorithms. The removal of writing, typewriting, or printing from a document is known as an Erasure. It can be accomplished by either of two means, one is chemical eradication in which the writing is removed or bleached by chemical agents, e.g., liquid ink erasure and two, physical erasures in which removal of information by using a blade/knife by abrasion, dye or white fluid. Chemical erasures are detected by ultraviolet examination of the questioned documents and an example of chemical erasure is shown in Figure 1.4. Physical erasures creates disturbances in paper fibers and are detected by microscopic examination of the document. Physical erasure is shown in Figure 1.5.

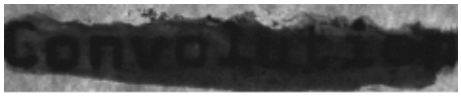


Figure 1.4: Chemical erasure document

Figure 1.5: Physical erasure document

e. Identification of inks and writing instruments: Differentiation of inks based on the fluidity of ink plays a major role in identification of writing instruments like ball pens, gel, roller and fountain pens which produced the document. Sample writings of ball, gel and roller pens collected from [10] is shown in Figure 1.6. For printed document, toner analysis identifies whether it uses liquid toner or powdered toner.

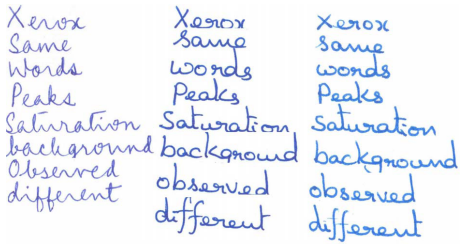


Figure 1.6: Sample writing of ball,gel and roller pens

f. Printer identification of the document: In the context of printed questioned documents examination, a forensic analyst has to answer questions about the source of the printed document. The content in the printed document shown in Figure 1.7 is not from single source, i.e. first paragraph is printed on inkjet printer and the second paragraph is printed later using laser printer.

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 two, it may mean that the original was reset if letter press 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.

Figure 1.7: Printed document

g. Photograph tampering: It involves techniques to trace manipulations in photograph [11], to trace presence of hidden messages embedded in the digital document, authentication of digital images [12] and to differentiate between photographic and photo-realistic images. Photo-realistic images are computer generated images. Picture shown in Figure 1.8 is an example of photograph manipulation. It was created by splicing the head of Oprah Winfrey onto the body of actress Ann-Margret.

Figure 1.8: Tampered photograph

1.2 Motivation of the work

The technology used to produce documents continues to evolve; the methods used to produce forgeries are ever more sophisticated; the expectations of lawyers and courts are yet more demanding. When the document's legitimacy is in question, methods are needed to non intrusively analyse distinguishing features of a document in order to learn more about its origin. The knowledge of spatial pattern produced by various print technologies are helpful in determining the source device that produced these patterns.

Printed document is spatial distribution of marking material. A printer produces a document to the extent of matching the pattern of the document. A scanner produces images by capturing document information according to the calibration and specifications of the scanner. This image will be subjected to tampering to produce a fraudulent document, which may be printed by the same or another printer. Hence, a forged document will posses composite features of the above processes. Thus printer identification of the questioned document is a highly involved and complex process. A printed questioned document exam-

ination, by forensic scientist starts with identifying the printer or source from which the document has been created. Printed documents contain features of a printer depending on the specified procedure used by them for placing the marking material on paper.

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

1. Is the document consistent, implying whether the content printed in the document is prepared from a single source?

2. Identification of source printer or printing techniques like ink jet, electrophotography printing etc.

3. Are two documents similar, i.e., printed using same technology or printer?

4. Differentiate between photographic and photo-realistic images

Questioned document examiner must be able to distinguish between genuine and forged documents [13]. Classification and identification are based on comparison of class characteristics of documents. Hence a document examiner should know the class characteristics of the document based on the materials and methodology used in preparation of the document. Instruments used by document examiner to distinguish genuine document from forged are high resolution microscopes, Electro Static Detection Apparatus(ESDA) and Video Spectral Comparator(VSC) [14]. ESDA is used to reveal indented impression on paper and it is a non destructive technique [15]. VSC is a multi spectral imaging system which works on the 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.

As current methods and instruments used are expensive for capturing the data as well as in analysing, there is a great need to develop alternative solutions for forensic identification of print technology that is more effective in cost, space and time. Further, these documents are material evidence in courts and therefore are to be subjected to non-invasive techniques, like image processing.

1.3 Contributions

Printed document forensics is mostly off line activity but sometimes can be soft real time activity. To give an apt solution, the technology has to be built by considering the limitations of equipment at investigation location and time. Some times the data available for analysis will be imprecise as well as partial due to limited resolution. To address these problems, the thesis develops various methods based on computing tools by considering the scanned image of moderate resolution. Our contributions can be categorized as discussed below:

1. **Identifying print technology of source document based on uniform colour region or homogeneous colour region:** Developed a Gaussian Variogram Model (GVM) model, for identifying the print technology which produced the given document. This method characterizes print technology based on spatial variability. Homogeneous colour region of images are taken as samples for the GVM data generation. The generated GVM data is taken as input to form the Reduct based Decision Tree(RDT) [16], which gives rules to identify the source printer for the given test data. Performance analysis of the model is also presented. Developed method assists the document examiner in finding basic print pattern of printers and it is also helpful in classifying different print technologies.

2. **Identifying print technology of source document based on printed**

text: We have formulated a Print Index for classification of inkjet printed text versus laser jet printed text. This work focuses on a frequently used word like 'the' as test sample for characterizing printed text. The novelty of the proposed approach is that the selected printed text is modelled as mixture of three Gaussian models namely text, noise and background. The associated patterns and features of the models are derived using Expectation Maximization(EM) algorithm [17] and few indices are proposed based on these parameters. One of the indices called Print Index(PI) for text is used for basic print technology discrimination. EM algorithm is also used as dimension reduction technique to characterize printed text.

3. **Differentiation of inkjet print from its photocopy:** Statistical measures are discussed as features for distinguishing print technology like inkjet print from its photocopy.

4. **Identification of tampered document:** Questioned document in general may be printed by various printers. Thus identification of tampered document involves the following

   a. Identification of tampered region Sliding window protocols are proposed and demonstrated for the purpose of identifying tampered document regions and also the nature of tamper. This moving window generates variograms to reveal various textures or spatial pattern underlying the sample, that are used for detecting tampered ROI(Region of Interest).

   b. Identification of combination of print technology Questioned document may have mixed print technology i.e., a document may be reproduced by scanning original document and then reprinting on another printer. In this way the reproduced document contains the features of both the previous print as well as the present printing characteristics. Window-wise analysis of variogram with varying window size captures mixed/combination of

print pattern existing in the documents. Results for varying window are demonstrated.

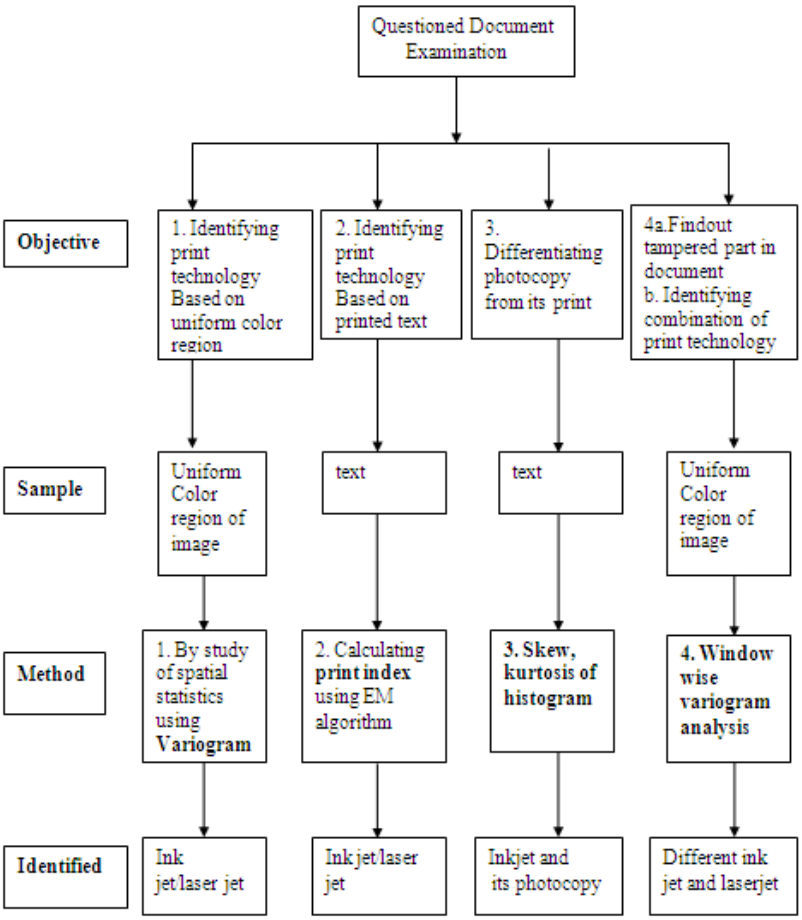


Figure 1.9: Problems identified and proposed methodology

The above contributions are depicted in Figure 1.9. Once the ROI is identified then any one of the above suggested techniques can be applied. Figure 1.10 depicts system overview of the hybrid system developed in the present study for addressing problems mentioned above associated with document forensics. Apt experiments are designed, for developing datasets as well as validating the proposed methodologies in this thesis, for demonstrating the proof of concept. The following section presents organization of thesis.

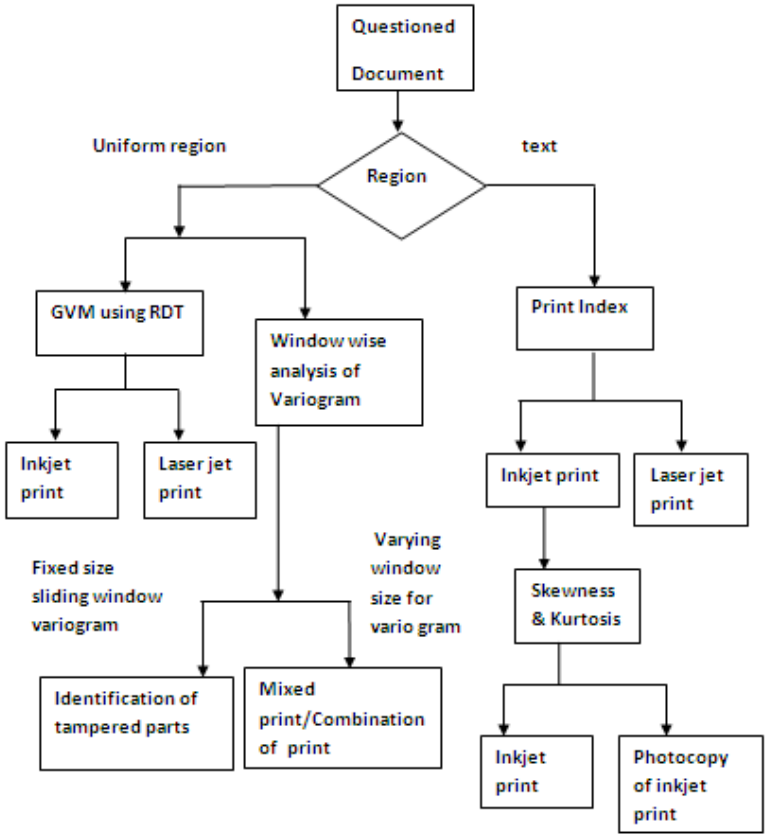


Figure 1.10: System overview

1.4 Organization of thesis

The chapters of the thesis are organized as follows: Chapter 1 introduces areas of forensic science with emphasis on Document Examination and its scope. It describes instruments used to prepare document, tools used by forensic examiner for examination of documents. Chapter 2 presents survey of various print technologies used to produce documents and brings out current state of tools or methodologies existing in identifying printed document source mechanism. Review of the techniques are also organised as ink and toner analysis, texture, region based, text based techniques and embedding techniques. Our contributions in the present study are spread over Chapter 3, 4 and 5. Chapter 3 develops techniques for identification of printer technology based on uniform colour region by hybrid methodologies integrating the merits of GVM as well as Roughset classification

techniques. Gaussian Mixture Models through Expectation Maximisation have been aptly moulded for identification of the printer technology based on printed text is the subject matter of Chapter 4. A novel print index is proposed and its applications and limitations are demonstrated. Histogram features like skewness and kurtosis are augmented to enhance the ability of these techniques for discriminating source of questioned document as photocopier from the printer. In Chapter 5 the challenging problem of identifying tampered regions of tampered document has been addressed by sliding window directional variogram based procedures through characteristics of empirical directional variogram. Chapter 6 concludes the thesis with future directions.

Appendix A is devoted for standard datasets for validating methodologies developed for this environment. Appendix B contains various results for the methodologies developed in Chapter 3 for the datasets, by using the homogeneous dataset given in Appendix A.

Chapter 2

Literature Survey

2.1 Types of printing technologies

In perpetrating a crime, it is often only the output from a computer in the form letter, set of accounts, counterfeit document or pornographic picture, which is recovered by the investigator. Hence document examination skills can be used to link these documents with a particular printing technology [18]. In the market there are many types of printers available, from daisy wheels to laser printers. All these printers leave traces in the hard copy they produce to link the document to a particular print technology. In [19] four basic printing technologies that are used for printing are 1. Electrostatic 2. Inkjet 3. Thermal 4. Photography. Printer types within a technology category are also discussed. Electrostatic technology category contains the printers types Xerography, Ionography and Electrophotography. Inkjet category contains Continuous, Thermal and Drop on demand inkjet printers. Thermal printing technology contains thermal transfer and dye migration. Photography category contains wet or dry processed silver and Cycolor technologies.

Electrostatic marking technology known as Xerographic, utilizes static charge pattern to attract polymeric toner particle. Ionographic printers create required charge pattern with selective charging rather than optical patterning of a uni-

formly charged surface and subsequent patterning via optical exposure. Electrophotographic printers use a coated paper stock instead of a photo conductor and selectively charge and develop the image on this coated paper. Electrophotographic(EP) printers are often known as xerographic and/or laser printers[20]. The term laser refers to the light exposure technology used in this process. Steps involved in the process of EP printing are shown in Figure 2.1 collected from [21]. The image to be printed is formed by the action of laser on light sensitive drum. Toner particles are electrostatically picked up by the drums charged areas. The drum then prints the image onto paper by direct contact and heat, which fuses the toner to the paper.

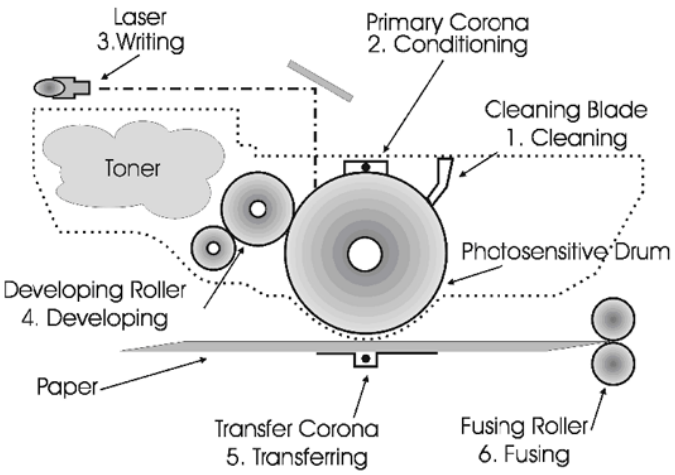


Figure 2.1: Laser print technology

Inkjet printers forms colour images by depositing droplets of coloured ink on the paper. Bubble jet uses small heater near the nozzle to produce a steam bubble that ejects a single drop of ink. Cyan, Magenta, Yellow and Black colours of ink are used to produce desired images. Drop on demand print uses solid or hot-melt technology. Solid ink printers melt a wax like material and eject the molten drop to the paper where it freezes. Drop on demand method used two types of droplet formation one is thermal, another is piezo electric effect. Figure 2.2 collected from [22] explains these two basic methods of drop formation in inkjet printers. The

thermal Drop on demand inkjet technology is used by HP, Canon and others. In this, droplets of ink are forced out of the nozzle by heating a resistor, which causes an air bubble to expand. When the bubble collapses, the droplet breaks off and the system returns to its original state. In piezo electric inkjet printer the walls in the nozzle filling chamber are made up of piezo electric material. When voltage is applied, the deflection of walls forces drop of ink out of nozzle.

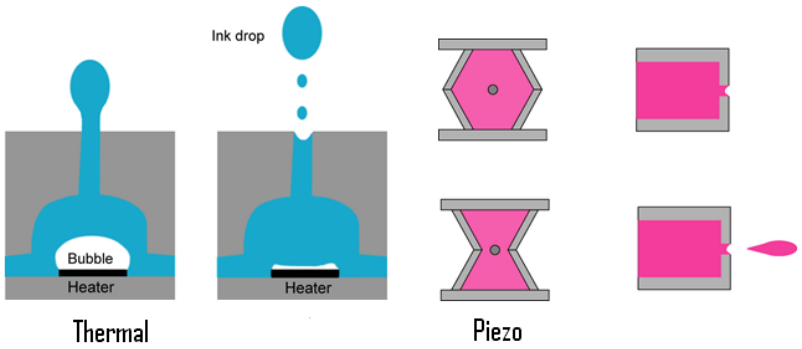


Figure 2.2: Drop formation in inkjet print technology

It can be noticed in water based ink jet printout that the dye is present in the dried ink on the uppermost part of the paper. Inkjet print should be protected from ultra violet light to ensure colour stability.

Thermal printers are simple form of printer which use heat to transfer marking material from some intermediate donor surface to the paper. The colourant donor support is typically Mylar ribbon. In case of thermal transfer printer, the ribbon is coated with a wax material of the appropriate colour. Thermal transfer printers usually require a special paper that has a very smooth printing surface to minimize heater-ribbon-paper gaps. It contains heater head compromised of 2500 miniature heaters. The second type of thermal printer, dye migration printer, uses ribbon coated with dye. Usually thermal printers are binary but recent thermal printers are producing gray scale images. The individual heaters are often driven by pulse width modulated signals that provide variable heat to produce a varying amount of dye to the receiver sheet or paper. The paper used is a special type that often has the appearance and texture of photo paper.

Photographic printing is a familiar method of printing where the image captured in photographic film is to be processed in a dark room for viewing the final image on paper. The process involves development of exposed image. It reduces the silver halide in the latent image to metallic silver and then stops development by removing the developing chemicals. The image is fixed by dissolving undeveloped silver halide from the light sensitive emulsion and washing thoroughly to remove processing chemicals[23]. The materials used are expensive and time to process film is a constraint in photographic printing. Dry processed colour photography has been a very limited alternative due to restricted image quality from dry or heat-developed technology. In Cycolor process, cyan, magenta and yellow capsules or cyliths are embedded in film like substrate. When these cycliths are exposed to light, they harden. Exposed film is put through a set of heated rollers that squeeze the capsules while unexposed capsules yield their coloured contents to a receiver material such as paper to produce a direct colour image.

Marking techniques[24] are categorised into major classes as Photographic, Lithographic, Xerographic and Inkjet. Photographic reproduction is continuous tone, whereas other using halftone techniques in printing. Inkjet uses primarily dispersed dot aperiodic halftoning whereas lithographic and xerographic reproduction use periodic rotated clustered dot screens[25].

Understanding different printing mechanisms, reproduction process and spatial characteristics for scanned image forms the basis for identifying type of printing technology used for reproduction of particular document. Hence, recent researches are concentrated on simple, non invasive and inexpensive imaging techniques which study the characteristics of various marking techniques for identification of source of the documents [26]. This chapter briefly explains various research contributions in document examination field.

Recent research publications demonstrate various approaches suggested for discriminating printing techniques. These approaches are based on the materials used in producing the documents like ink and toner. Some of the approaches like texture analysis and embedding techniques are based on arrangement of marking material

in documents. Texture analysis reveals the printing methodology employed in the document and these methods are again categorized based on the content of the document used in this analysis like homogeneous regions or text which is colour or gray.

2.2 Ink analysis and toner analysis

Ink and toner analysis in forensic document examination determines the nature of ink/toner and printing process by which ink/toner is transferred to the paper. The absorption characteristics of ink/toner in paper, form the basis for classification of inks and toners[27]. Laser and photocopier use solid ink where as inkjet use liquid ink. Differences in ink absorbency and glossy appearance of ink in case of laser printer, pixel distribution in case of inkjet printer are features used for the questioned document examination. Sobel edge detected image[28] captures ink spray characteristics and help to distinguish inkjet printers and photocopiers.

Ink and toner analysis for identification of printing process employed HSV colour space [29]. This work distinguishes whether the document is printed or photocopied. It uses hue histograms which 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, edge detected hue and saturation images [30]. For example inkjet print has large number of isolated dots near the strokes and it has no variation in contrast on opposite sides of the stokes in edge detected hue images, while laser print has less number of isolated dots, alternating low and high contrast in opposite sides of the stroke [30][10]. Another distinguishing feature for identification of laserjet print is periodic variation in column wise intensity profile of edge detected saturation images.

Various texture analysis models are employed for identification of printing techniques. These models are based on print patterns and spatial relationship between these patterns.

2.3 Texture analysis for document classification

There is no generally agreed upon definition for texture, one of the definition is addressed here :"We may regard texture as what constitutes a macroscopic region. Its structure is simply attributed to the repetitive patterns in which elements or primitives are arranged according to a placement rule" [31]. Texture refers to the properties that represents the surface or structure of the object and something consist of mutually related elements. The main objectives of texture analysis are texture recognition and texture based shape analysis. Texture features are found in tone and structure of the texture but these features are scale dependent. Texture primitive is a contiguous set of pixels with some tonal and/or regional property and can be described by its average intensity, maximum or minimum intensity, size, shape, etc. Tone is based mostly on the pixel intensity properties in the primitive. Image texture has number of perceived qualities which plays an important role in describing texture. Some of these perceived qualities are not independent. Properties identified by Laws[32] in describing texture are uniformity, density, coarseness, roughness, regularity, linearity, directionality, direction, frequency and phase.

Texture analysis [33] of documents identifies the source print technology which produced the document. Printed document is a spatial distribution of marking material to the paper using the printing mechanism. This distribution differs for each print technology based on parameters employed. The parameters affecting the print pattern are inks or toners used for printing, drop size, and number of drops per dot etc. Analysing of texture pattern in the document such as whether it contains smooth or rough text and direction of texture reveals the source printing mechanism. Several researches employed texture based analysis[34] for image classification.

The proposed models of texture are applied in the areas of texture problems like texture segmentation, texture classification, texture synthesis and shape from texture. Texture segmentation, classification and shape from texture are used

for image analysis methods. It includes recognition of documents, identifying the characteristics of printed document and processing of documents.

Identification of perceived qualities of the texture requires mathematical model of texture. Taxonomy of texture models are (a) Statistical methods (b) Geometrical methods (c) Model based methods and (d) Signal processing methods. The texture models for solving different document problems are now discussed here.

a) Statistical models are employed in printer identification based on gray level co-occurrence features for security and forensic applications [35]. This work describes analysis of image texture to identify the printer used to print a document. Statistical methods describe texture in a form suitable for statistical pattern recognition, in which each type of texture is described by feature vector properties. In electro-photography printing fluctuations in the developed toner on a printed page are directly tied to the electromechanical properties of the printer. Fluctuations in the toner can be modelled as a texture. Gray level co-occurrence features are used to describe the texture and identify the printer. Gray level co-occurrence features are determined using gray level co-occurrence matrix GLCM[36] which contains overall spatial relationships among pixels in the image. In this work, forensic characterization of printed text is based on texture analysis of particular character 'e', which is most frequently occurring character in English language. Test documents are printed on various printers and scanned at 2400 dpi. All letter 'e' s in the document are extracted and 22 features are selected from each character. These features includes texture features, histogram based features and pixel based features. Nearest Neighbour(5NN) classifier is used for classification of feature vectors.

b) Geometrical methods depends on the geometrical properties of the texture elements and are used to identify texture elements in the image. It describes the rules governing the spatial organization of texture primitives. Halftone [37] textures in printed documents are analysed for identification of Electro-photographic printer.

In [38], describes halftone texture in each colour channel is printed with identical textures having different angle. Hough transform[28] is applied for extracting angles of halftone texture in each channel of CMYK. Hough transform extracts straight lines, which are described by angle and distance. Most angles obtained by the transform are similar among images printed by the same model. Hence histogram accumulating angle values are homogeneous for the images printed by the same model printers while histograms are heterogeneous for the images printed on different model printers.

c) Model based methods are based on construction of image model. Model parameters explains the qualities of texture. Markov random fields (MRF) are popular for modelling images, and it is used to capture local information from the image. Markov random fields are used in image classification for image segmentation and texture classification[39]. Recently, MRF based neighbouring patches[40] has been used to classify different kind of text such as machine, handwritten text and noise from annotated machine printed documents.

d) Frequency analysis of the image is a methodology in signal processing models. Signal processing models like discrete wavelet analysis is employed for identification of colour laser printers [41]. Same image printed on several colour laser printers differ in colour noise. These differences can be observed in HH-subband of an image after application of 1-level 2D discrete wavelet. To characterise colour noise in HH-subband of image, statistical analysis is performed. For each colour channel, statistical features like standard deviation, skewness and kurtosis is calculated. Covariance and correlation between pair of colour channel is calculated. These statistical features are extracted using RGB and CMYK colour domain. Total of 29 features are extracted from each image and trained using SVM classifier[42] to identify the brand and model of source printer.

2.3.1 Techniques based on uniform region of image vs text:

Study of printed document in high resolution provides information of local pattern in the document. As printed document is a combination of picture and text, the techniques available for identification of printing process are based on uniform colour regions or text of the document at high resolution.

a) Region based approach: Identification of electro-photographic printers is based on frequency analysis of banding signal in large mid tone area[35]. This method has proved that different printers have different banding frequencies based on the brand and model of the printer. As process direction of printer and scanner is horizontal, the banding signal feature is present in the process direction. Hence in this method, 1-D projection of banding signal is used. Fast Fourier transform of this projection has given distinguishing number of cycles for each brand. These results are reliable for 12.5-50 percent filled gray level patches. In [43], features of colour noise and GLCM features of original image is used for identification of colour laser printers. Noise features are extracted by taking difference between Wiener-filtered CMY images and original CMY images. Statistical features of GLCM and noise features are trained using SVM classifier for identification of colour laser printer.

b) Text based approach As it is difficult to find large mid tone gray regions in the text, techniques are suggested for identification of print technology based on text. Gray level co-occurrence feature of the most frequently occurring letter 'e' and Gaussian mixture model (GMM)[35][44] are the techniques used for printer identification. In GMM, principal component analysis is used as a dimension reduction technique to obtain 1-D projections of the extracted text character. These researches are exclusively for the identification of electro-photographic printers which are usually referred to as laser printers.

Gray level features proposed in [45] for discriminating inkjet from laserjet print are based on high resolution scanned images, e.g. 3200 dots per inch. Recent

research is concentrated on evaluation of gray level features like perimeter based edge roughness of the text[46] for print technique classification, based on low resolution image for high throughput document management system.

In [47], electro-photographic printer identification based on character matching is explained. It is based on the fact that the shape of a printed character is stable compared to print quality and it presents a macro printing style. A special optical instrument having LED light source, lens and CCD camera is used to get magnified printed document images. Characters are then extracted from these images. These characters are normalized and recognized using OCR techniques. Distance transform is applied for calculating dissimilarity measure between the same characters of pending/questioned documents and known documents in the database. Minimum distance indicates the possibility that the two documents are created by the same printer. This work is based on the condition that printing style of printer is stable over a long term. The methods mentioned above, extracted features from local character area in the document where as [48] features are extracted features from the whole document image. The rows in one page of the document are strictly parallel. For printed document the slope of rows changes regularly along printing direction and it varies according to the printer. This phenomena is called geometric distortion of printed document and it is taken as an intrinsic feature of a printer. Projective transformation model is used to model geometric distortion. Least square method is used to get parameters of the models and these are considered as geometric distortion features. Work mentioned in [35], [44], [45], [46] and [47] identifies the print technology based on gray text only.

Colour is one of the important features of the document which can be exploited for the identification of printed document and various colour image models [49] are employed in analysing overlapped pattern in document examination. Colour image processing techniques exploit features of the colour spaces for identification of printing process. Identification of the printing process, using HSV colour space [30][10] is based on hue histogram to ensure whether a document is print or photocopy. Hue contrast, periodicity and ink are the features selected for classifying

ink jet, laser jet and photocopies. In [50] and [51] identification of fraudulent documents are demonstrated using printer and scanner combinations which produced the fraud documents. In this work they captured images of text using high resolution cameras LEICA MZ 8, LECIA MZ 12.5 and directly transferred the images to computer. Unique colour count, texture feature uniformity and intensity variation are used as parameters for distinguishing different print technologies.

c) Combination of region and text: Image qualities measure extracts region, text, colour based features of document and are applied to distinguish between genuine and fake document[52] . Image quality measures[53] are categorised into pixel based difference measures, correlation based difference measures, edge based measures, spectral distance based measures, content based measures and human visual system based measures according to the information they use. This is experimented on fraud gift vouchers printed using several colour printer. Forgery detection with SVM classifier uncovers the inherent device features on which forged document is produced. The main objective of this classification is based on discriminating original from fraud documents subjected to printing distortion. Visual quality measures[54] are employed to measure the difference between the two images. SVM classifier is trained with image quality measures of original and fraud documents. It is ensured that feature vector from same class have similar value even though they have geometrical distortion. Image quality measures are evaluated statistically for their sensitivity and consistency behaviour. Image quality measures selected as features for training are Mean Square Error (MSE) in the selected region of two images, Maximum value among pixel differences, MSE in Lab colour space, structural content, normalized cross correlation, image fidelity, absolute mean and variance of angle between the pixel vectors. Measure from frequency domain are MSE of magnitude, phase and combined measure. Measures related to HVS are also used. Total 17 image quality measures are selected. All these measures are calculated as in each block of the region of interest. Hence, the distortion measures are the median of the corresponding blocks. Original gift

vouchers are scanned and printed on laserjet and inkjet printer for making fraud documents. The classification results are perfect when original document and fraud documents generated on lasers or ink jets are trained. It has shown low accuracy when fraud documents generated on both laser and inkjet are grouped in training.

Gray vs colour:

Region or text based methods could be gray/colour. The techniques proposed in [35], [44], [55] and [56] are used for identification of gray level or black and white laser printers. These techniques use gray level data or binarized image data for identifying the printing techniques or printers. The techniques mentioned in [43], [49], [30], [10], [50] and [51] are colour image processing techniques.

2.4 Embedding techniques

Copyright protection is an important issue to intercept and prevent forgery in printed material such as currency notes, bank checks and to track and validate sensitive and secret printed material. Now, the method of embedding watermark[57] in the printed image is taking the advantage of printing process to track and validate sensitive and secret printed materials. In general printing of image with finite number of inks is performed distributing ink dots in various densities and patterns throughout the print. The embedded code locally varies the pattern of ink dots such that correct distribution is still maintained. This pattern is tightly constrained to allow high probability of correct decoding of watermark. This also reduces the visibility of texture artefacts and hence this method is visually not obtrusive to the printed image. Watermark is a binary sequence reordered into an array. A single dither cell is selected for each entry in the watermark. Encoding of this watermark is performed using halftoning techniques which uses a set of dither cells. Number of different dither cells are used to create threshold pattern in the halftoning image. For example, dither cell c_0 for black and c_1 for white.

Watermark is used to select dither cells from the predefined set of dither cells. Decoding of water mark involves the scanning of halftoned image and determining the sequence of dither cell used to create a halftone pattern.

Copyright protection of the document ensure the authenticity to trace the forged documents. Laser amplitude modulation is used in electro-photographic printer[58] to embed information in the text of the document. In [55], [56], [59] two strategies are proposed for printer identification. One of those strategies is passive which characterizes the printer by finding the intrinsic features in printed document. These features are characteristics of a particular printer model or brand. These features are referred to as intrinsic signatures of the printer. Knowledge of intrinsic signature of the printer helps in determining the signature in a printed document. Another strategy is active which embeds the information in the printed text by modulating the process parameters in the printing mechanism. The information embedded in this method are known as extrinsic signature, which contains identifying information like printer serial number and date of printing. As embedding density increases, it is required to have good decoding algorithms. For embedding data in printer level mechanism Amplitude/Frequency Modulated (AM/FM) halftone algorithms[60] are used.

Machine identification code project by Electronic Frontier Foundation identifies presence of pattern of yellow dots[61] in colour laser printouts and these dots reveal information like 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 Laserjet-8550 series do not show any yellow dots. Still there exist some forensic information to keep track of the printer model.

Content Integrity of Printed Documents using Error Correction (CIPDEC)[62] detects any modifications in document without requiring original document for such detection. The requirements of CIPDEC algorithm are that the document should contain bar code with Error Correcting Code(ECC) parities and that the document must be marked with corner and dot marker for pixel precision fideli-

ties. It is also robust to photocopying, stains, markers and tears. This algorithm involves two stages: generation stage and verification stage. In generation stage, original electronic document is treated as an array of pixel and ECC is computed. Two types of markers- corner marker and dot markers are added to the original document to produce marked document. This marked document is printed using a printer. Verification stage of document involves reconstruction of document with the help of dot markers and corresponding modified pixels are identified by decoding ECC parities. Error correcting code corrects the document and it reveals the tampered parts of the document.

2.5 Summary

Survey	Requirement	Identifies		
		Inkjet	Laserjet	Photocopy
[10][27][30]	For HSV colour domain	yes	yes	yes
[35][59][55] [56][44]	Only for Laser printers	no	yes	no
[38][41][43]	Colour laser printer	no	yes	no
[47]	Need text lines in whole document	no	yes	no
[45]	Gray Images at 3200 dpi	yes	yes	yes
[50], [51]	Need high resolution microscope, this work is based on printer and scanner combination	yes	yes	yes
[58],[59], [60],[62],[61]	For Laser printer with embedding technique	no	yes	no

Table 2.1: Literature survey with details of addressed problems

In [10], [30] ink/toner analysis for identification of writing and printing instrument is based on HVS colour space techniques. The works discussed in[35],[38], [41], [44], [55], [56], [59],[43] are concentrated on identification of Electro-photographic printers. In [45] identification of printing technique for gray level text is based on 3200 dpi scanned images. The work discussed in[50], [51] for identification of fraud documents by comparison of characters in the given documents and needs high resolution microscope to capture these characters. The embedding techniques

[58],[59], [60], [62] 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.

Chapter 3

Identification of Printing Technology based on Homogeneous Colour Regions of Image

3.1 Introduction

A document contains both printed text and images. Printed document is a complex convolution of the content of the document as well as printing marks of the printer. Hence printed documents contain features of printer depending on the specified procedures used by them for placing marking material on the paper. They differ in the printed pattern, number of drops per dot and technology used to print like drop on demand thermal printing, HP photo REt technology, laser technology etc. Documents like certificates, identity cards and letter heads containing uniform colour region reveal spatial characteristics of particular print technology. Segment uniform colour region from printed document to provide clear distinguishable feature for the identification of printing technique. Segmentation divides the image into its constituent regions or sub images which are similar in property intensity,

colour, brightness. Segmentation of non trivial images one of the challenging task in image process and still under research. Existing segmentation techniques [63] are divided into seven categories: Histogram based [64], Clustering [65], Region based, Edge based, Physical model based, Fuzzy approaches based and Neural network based. Histogram thresh holding is applicable for monochrome image which can produce histogram with distinguishable peaks and valley. As several techniques are available for segmenting homogeneous regions of image, contribution in this thesis starts with selected homogeneous colour region of image.

An image is a spatial object where each pixel at a given position will be attributed with its pixel characteristics. Several image processing techniques are based on these collection of vectors. A printed image inherits the image characteristics coupled with printer technology. The features extracted from a printed image will be associated or correlated to the image, printer technology and interaction of the image and printer technology. Thus, it is expected a meaningful inferences can be derived by employing statistic tools. A variogram is a statistical tool employed in geology that has been adopted and its abstraction through GVM is employed to model the printer behaviour. To identify the print technology of the document, soft computing tools are considered for classification.

The following subsections presents a problem which is addressed in this chapter and a brief account of printing technology. Section 3.2 gives brief account of variogram, GVM, Roughset based decision tree for print technique classification. GVMPT algorithm is developed in Section 3.3 for print technology identification and presents an illustrative demonstration of the same. This section also demonstrates influences on the choice of parameters used in capturing the print technology of a document based on an uniform/homogeneous image region.

Main objective of the present study is printer identification. This chapter concentrates on printer technology identification i.e type of inkjet or laserjet printer. The study reveals that the following sub problems need to be addressed for printer technology identification based on image region of a printed document.

1. Identifying/clipping homogeneous colour region of image of a questioned document

2. Extracting spatial characteristics of the above clipped image

3. Deriving model representation for the same

4. Developing a classifier

Extracting homogeneous region is a popular problem in image processing and is addressed through segmentation methods in general. Thus, the present study starts with identified homogeneous region of the questioned document.

3.2 Variogram

3.2.1 Definition

Variogram (semivariogram which is half of the variogram)[66] is a statistical tool that characterizes spatial continuity or roughness of data set. Variogram gives an average dissimilarity between points separated by distance h in specific direction $\vec{d}$. The separation distance h is usually referred to as lag. Variogram is calculated as shown in equation 3.1.

$$V_d(h) = 1/n \sum_{i=1}^{n} (f(\vec{x_i} + h\vec{d}) - f(\vec{x_i}))^2 \tag{3.1}$$

where $V_d(h)$ is the variance in the direction $\vec{d}$ with lag h.

Variogram has been widely used for remote sensing applications [67] and classification of geo-statistical textures [68][69]. Variogram represents both structural and random aspects of the data. The variogram values increases with increasing value of the distance and at certain distance it levels off which means that at that distance it has no correlation. Thus, maximum of the variogram and its corresponding ordinate are taken as *sill* and *range* respectively. Similarly, the spatial variance at h=1 is taken as *nugget*. Hence, *sill*, *range* and *nugget* components

are referred to as parameters of variogram. The plot of lag h versus $V_d(h)$ provides graphical representation of variogram. Algorithm 3.1 explains the steps for generating variogram for an image.

The image of size 41 by 41 in Figure 3.1 and its corresponding intensity values in the form of matrix shown in Figure 3.2 are considered for demonstrating the variogram. Variogram calculated along x-axis of this sample image following the steps mentioned in Algorithm 3.1 is shown in the form of graphical representation in Figure 3.3. In this graphical representation, the variance with respect to specific orientation will display the pattern of dominant regularity along x-axis.

Figure 3.1: Sample image

3.2.2 Sensitivity of variogram

Some of the elegant properties of variogram like illumination invariant [70] makes its usage apt for print technology identification. Life of a toner, that is change in toner density has no effect on dissimilarity measures of variogram. Hence, parameters of variogram have significance in printer identification and each parameter has its own significance. *Range* of the variogram represents structural model of the data and size of texture [71] that is contained in the data, whereas *sill* represents dissimilarity(contrast) in the image. We now discuss how variogram captures underlying pattern in a sample image. In the present work unless direction is spell out $\vec{d}$ is taken as direction along x-axis.

Understanding parameters of variogram

The Images of sample 1, sample 2 and sample 3 represents patterns of texture and their corresponding variograms along x-axis are shown in Figure 3.4.

Algorithm 3.1 VARIOGRAM(Image,d)

//Calculating Variogram//

Input:

Image: an image of size M × M

d :direction specified in terms of degree

Output:

V: vector of variogram //vector of float values//

S: *sill*//float value//

N: *nugget*//float value//

R: *range*//float value//

Method:

1: Let f(x,y) be the Image, $1 \leq$ x $\leq$ M, $1 \leq$ y $\leq$ M
//Consider all 'L' pairs of points of an image such as $(x_i, y_i)$and $(x_i + h * \cos d, y_i + h * \sin d)$ separated by lag 'h', at direction 'd'.//

2: **for** $h \leftarrow 1$ *to* $M/2$ **do**

3: $V_d(h) \leftarrow 1/L \sum_{i=1}^{L} (f(x_i + h * \cos d, y_i + h * \sin d) - f(x_i, y_i))^2$

4: **end for**

5: S$\leftarrow$ Max(V_d(h)), for 1$\leq$ h $\leq$M/2

6: N$\leftarrow$ $V_d(1)$

7: **for** $h \leftarrow 1$ *to* $M/2$ **do**

8: **if** $V_d(h) = sill$ **then**

9: R$\leftarrow$h

10: **end if**

11: **end for**

12: Return V, S, N, R

Computational Complexity of VARIOGRAM	
step 2- 4	$O(M^3)$
step 7-11	$O(M)$
Algorithm Complexity: $O(M^3)$	

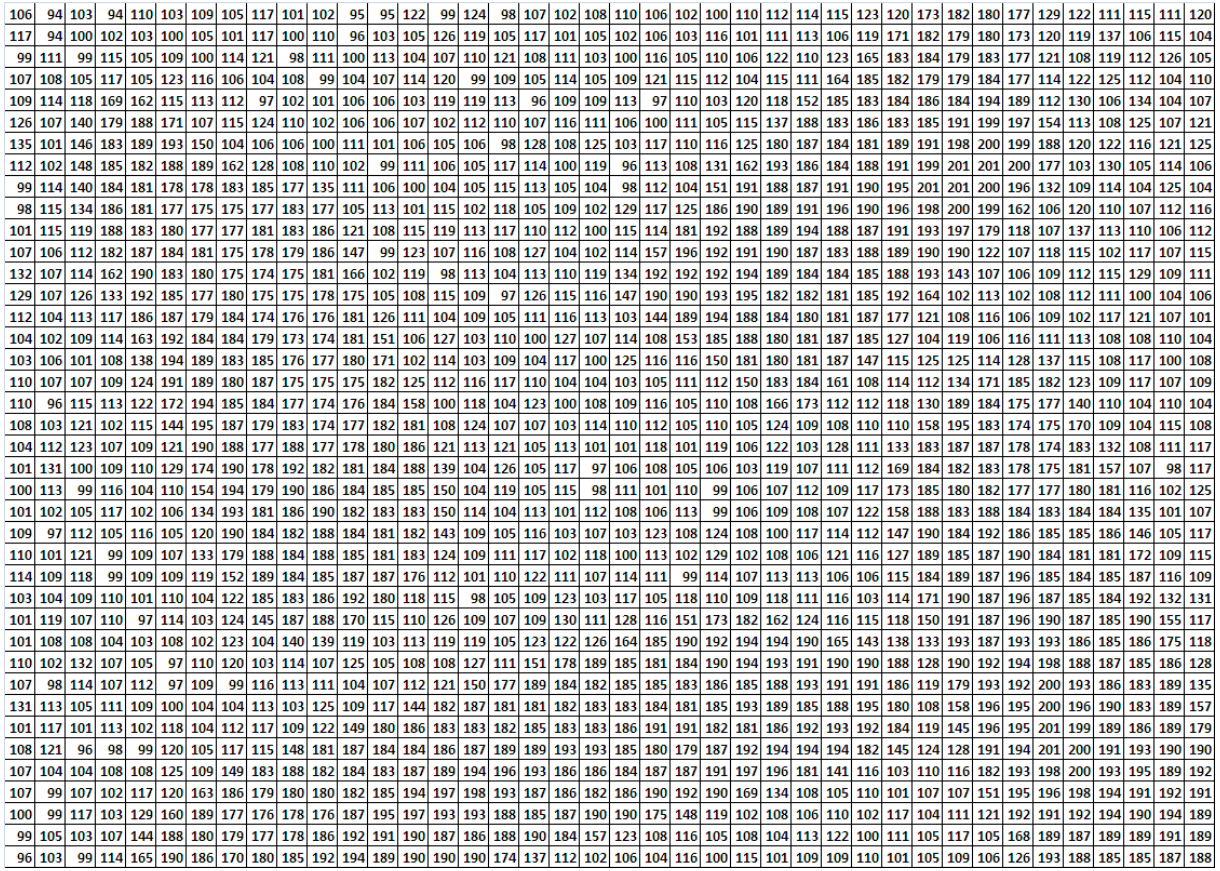


Figure 3.2: Matrix

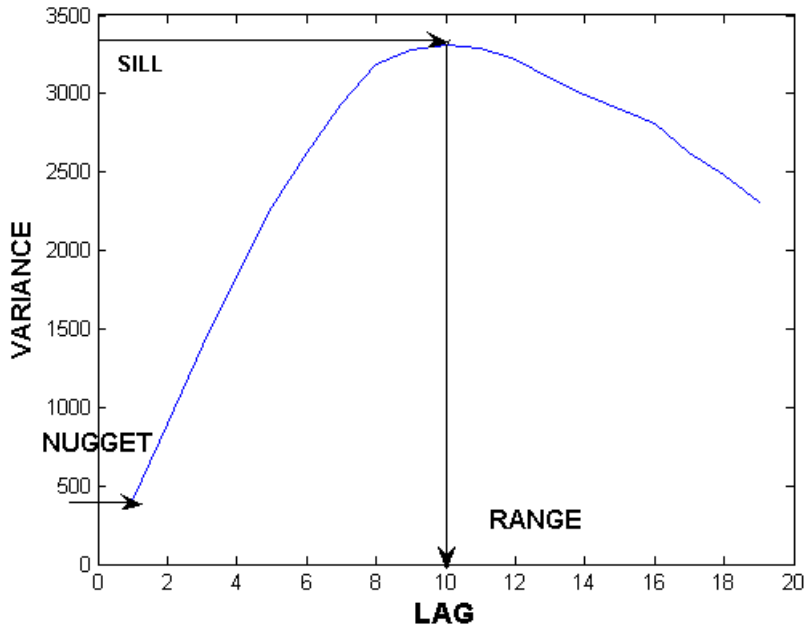


Figure 3.3: Variogram of sample image

Variogram calculated for these images quantifies the patterns contained in the images. The pattern in the sample 1 is scattered randomly and has less contrast image. These observed features can be characterized by analysing variogram of sample 1 which shows less *sill* as compared to variograms of remaining two samples. The pattern in the sample 2 is clearly distinguishable compared to the first sample and is a high contrast image which contributes to high variance. It is represented by high *sill* value as seen in the variogram of sample 2. The sample 3 shows pattern clearly arranged in a manner which resembles periodicity in its corresponding variogram. *Sill*, therefore represents contrast feature. *Range* gives the size of the texture in the image.

Let us now take three sample images of same homogeneous regions obtained by printing it using different print technology as shown in Figure 3.5(a). Here, variability in features of variogram along x-axis is linked to the underlying print technology used to produce that image. From Figure 3.5(b), Figure 3.5(c) and 3.5(d), it is observed that as sampling distance increases the variogram along x-axis is loosing its smoothness.

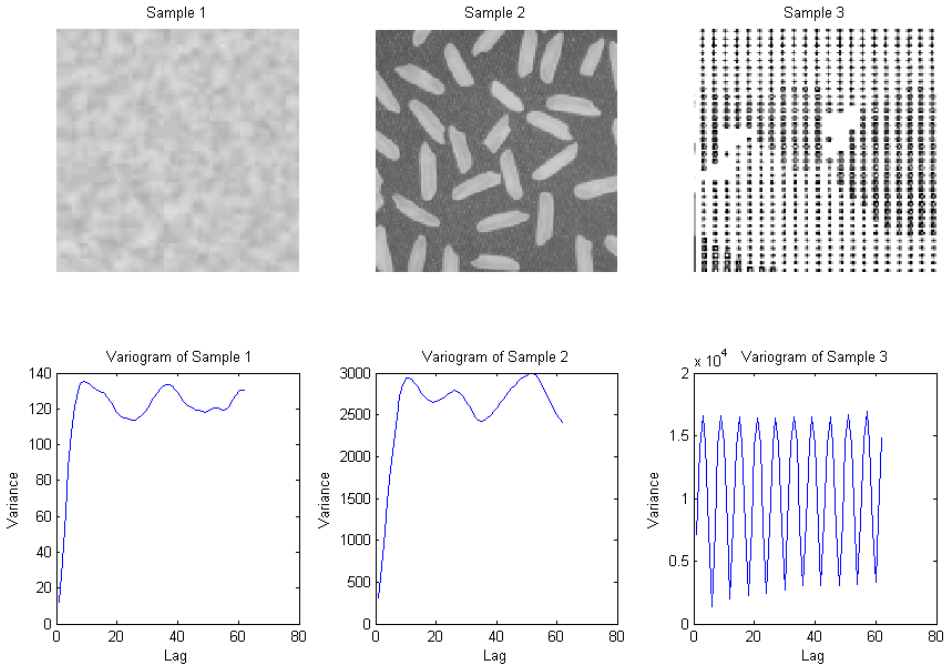


Figure 3.4: Variogram sensitivity

Study of Directional variogram

To check for directional dependence in a variogram, we have to compute variance values for data pairs falling within certain directional bands as well as falling within the prescribed lag limits. Directional variogram along specified angle for the sample images is shown in Figure 3.6. Variogram along direction 30° and 60° angle are shown in Figure 3.6(a) and 3.6(b) respectively. As the three sample images are from same source, the changes in variogram are due to the changes in the underlying print feature.

3.2.3 Modelling the variogram

Variogram has potential for developing graphical representation and visual analysis but it will not be useful for classical representation. This will need an abstraction of the variogram by producing its functional form which is close enough to it. This is called as model of the variogram [72]. Various model forms are used for

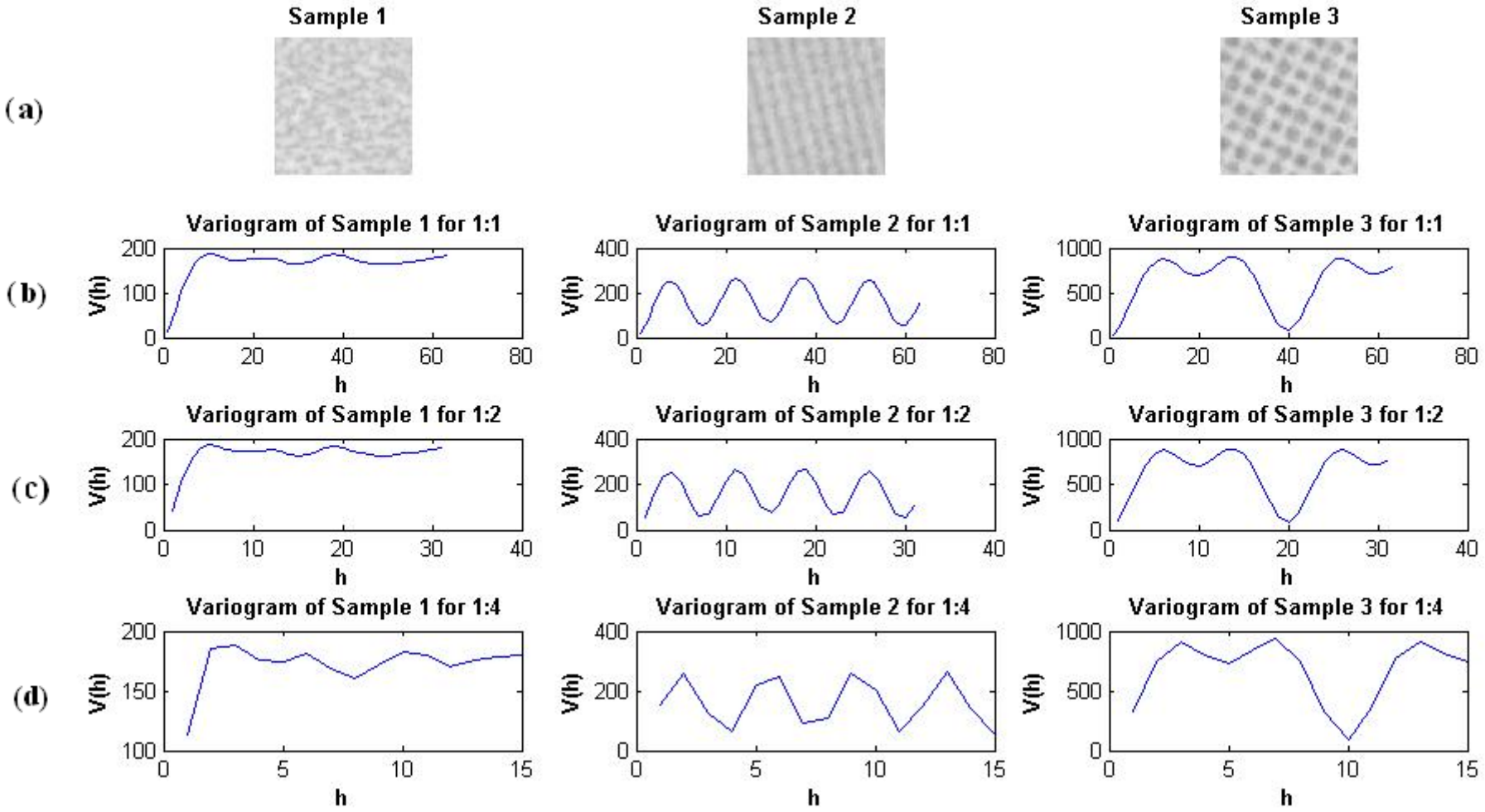


Figure 3.5: Sampling affects

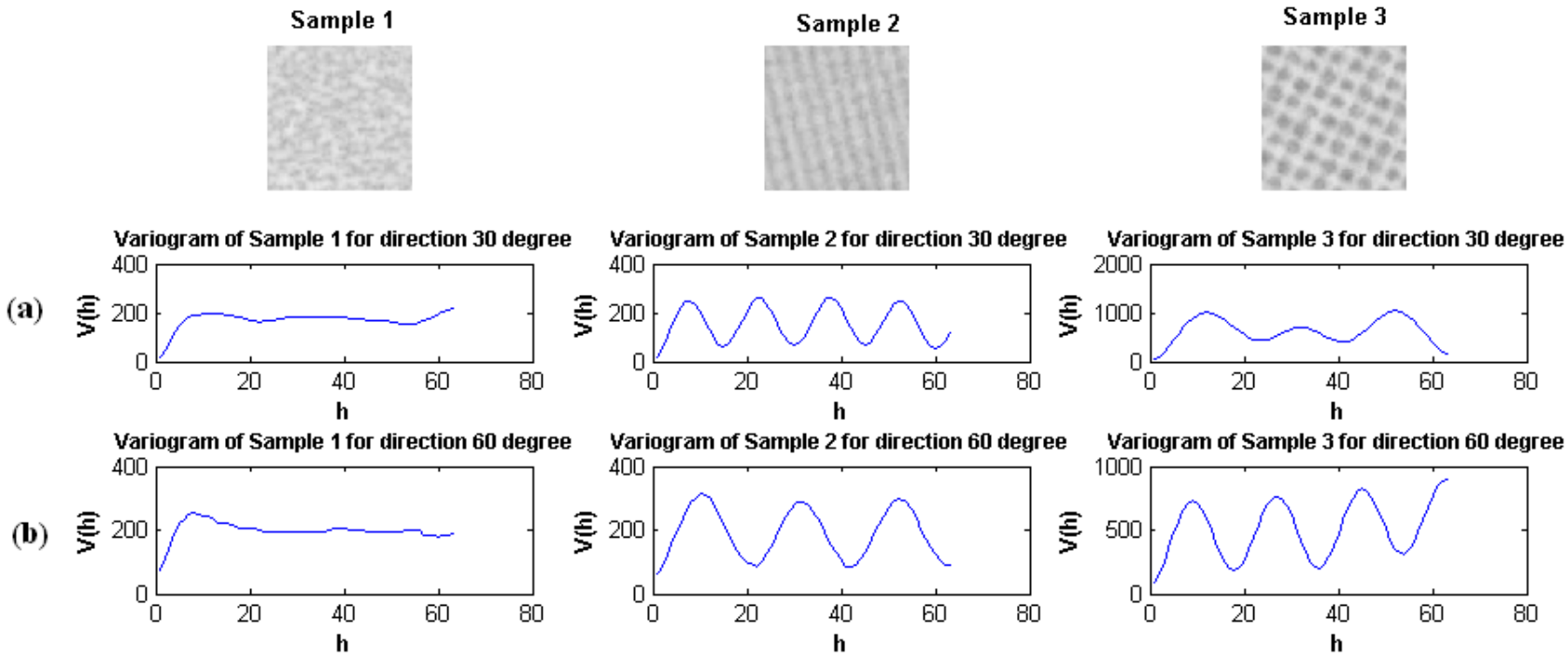


Figure 3.6: Directional variogram for sample Images

representing variogram. The most popular models used are exponential model, spherical model and gaussian models. A model form using function of gaussian is known as Gaussian variogram. A model developed based on K^{th} Gaussian functions is further referred as K^{th} order Gaussian Variogram Model(GVM). The structure of K^{th} order Gaussian model is given in Equation 3.2.

$$f(x) = \sum_{i=1}^{K} a_i exp - ((x - b_i)/c_i)^2 \tag{3.2}$$

GVM

The structure of K^{th} order Gaussian model given in Equation 3.2 involves 3K parameters. For a given variogram the problem of producing Gaussian variogram is nothing but fixing $3K$ parameters such that the model will be close enough to the empirical variogram. For a given printer for a selected direction, K^{th} order Gaussian variogram is nothing but fixing $3K$ parameter values and 3 empirical variogram parameters *sill*(S), *nugget*(N) and *range*(R). For capturing the characteristics of the object, if one needs to consider 'D' directions, then the feature vector will be $3 * (K + 1)D$.

Information System

For a given set of training objects N, one can form an information table of size $N * 3(K + 1) * D$. If printer technology is already known for the objects then appending the tag of printer technology to the above will become a decision table.

3.2.4 Classifier

Soft computing techniques are employed for building decision tree for classification. Consider learning set for Playtennis example shown in Table 3.1. The attributes and their possible values for Playtennis example are Outlook: $\{Sunny, Overcast, Rain\}$, Temperature: $\{Hot, Mild, Cool\}$, Humidity: $\{High, Normal\}$, Windy: $\{Weak, Strong\}$, Playtennis: $\{Yes, No\}$. Decision tree for Playtennis is shown in Figure 3.7. Deci-

sion tree classifies each instance by sorting them down the tree from root to leaf node. Each node in the tree specifies some attribute of instance and each branch descending from that node corresponds to one of possible values of that attribute. Decision tree [73] is a tree in which each branch node represents a choice between a number of alternatives and each leaf node represents a decision.

Learning set for Playtennis					
Day	Outlook	Temperature	Humidity	Wind	Playtennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Table 3.1: Learning set for playtennis example

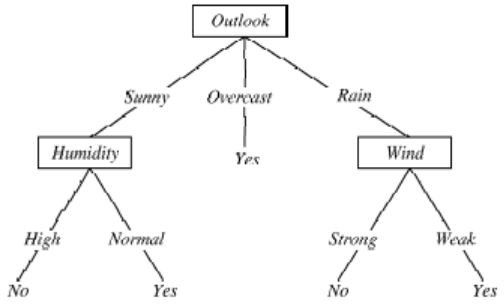


Figure 3.7: Decision tree for playtennis example

Classification can be done by considering decision table and adopting soft computing technique, hence we adopted Roughset based Decision Tree (RDT) classifier. First extract predominant features using roughset theory attaching predom-

inant attribute as a decision attribute with decision node at i^{th} level of decision tree as classifier for developing the splitting criteria. This Roughset based Decision Tree needs a set of predominant attribute which is the same as feature selection. Thus one must have feature selection method. The following subsection provides a brief introduction to roughset involving reduct computation, extraction of predominant attributes and then arriving at reduct based decision tree.

Roughsets

Rough set theory was proposed by Prof. Z. Pawlak in 1982. In Rough Set theory [74], knowledge is interpreted as an ability to classify some objects. Subset of universe of objects are called categories. Categories included in classification form a partition of the universe of objects.

In rough set theory, syntactic representation of knowledge is provided in the form of an information system. The information system is a functional representation of a classification as an attributes of objects. Let I = (U, A) be an information system, U is a non-empty finite set called universe of objects and A is a non-empty finite set of attributes such that for each attribute a belongs to A is a function $a : U \rightarrow V_a$, where V_a is the values set of the attribute a. With any P subset of A there is an associated equivalence relation IND(P), is called a P-indiscernible relation. IND(P) is defined as follows;

$$IND(P) = \{(x, y) \in \cup^2 \mid \forall a \in P, a(x) = a(y)\} \tag{3.3}$$

The partition of U is a family of all equivalence classes of IND(P) and is denoted by U/IND(P) or (U/P). If $(x, y) \in IND(P)$, then x, y are indiscernible or indistinguishable attributes of P. For example let information system shown in Table 3.2 has following equivalence relation.

The non-empty subsets of attribute set A are $\{\{a_1\}, \{a_2\}, \{a_1, a_2\}\}$. The indiscernibility relation $IND\{.\}$ defines three non trivial partitions of the universe
$1.IND\{a_1\} = \{\{x_1, x_2, x_6\}, \{x_3, x_4\}, \{x_5, x_7\}\}$
$2.IND\{a_2\} = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_6, x_7\}\}$

information system		
	a_1	a_2
x1	1	3
x2	1	0
x3	3	1
x4	3	1
x5	4	2
x6	1	2
x7	4	2

Table 3.2: Information System

3.$IND\{a_1, a_2\} = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_7\}, \{x_6\}\}$

Consider a target set X = $\{x_1, x_2, x_3\}$ represented using attribute set $P = \{a_1, a_2\}$. Here, $\{x_3, x_4\}$ are indiscernible. X cannot be expressed exactly, because the set may include or exclude objects which are indistinguishable on the basis of attributes P. Hence, the target set X can be approximated by constructing the P-lower and P-upper approximations of X. Lower approximation and positive region of X is a complete set of objects in U/P that can be positively or unambiguously classified as belonging to target set X. P upper approximation and negative region is the union of all equivalence classes. Upper approximation is the complete set of objects that are possibly belongs to a target set. P-lower and P-upper approximation are defined as follows

$$\underline{P}X = \{x \mid [x]_p \subseteq X\} \tag{3.4}$$

$$\overline{P}X = \{x \mid [x]_p \bigcap X\} \tag{3.5}$$

The set $U - \overline{P}X$ is the negative region, set of objects that can be ruled out as members of the target. The tuple $\langle \underline{P}X, \overline{P}X \rangle$ is called rough set. Rough set is composed of two crisp sets, one is lower boundary region another is upper boundary of the target X.

Reducts

In the information system the subset of attributes which fully characterizes the knowledge in the database, such as attribute set is referred to as reduct. Reduct is a subset of attributes $RED \subseteq P$, such that equivalence classes induced by reduct set is same as equivalence classes induced by full attribute set P. No attribute can be removed from reduct set without changing the equivalence classes. The attribute set which are common to all reduct is called core. These cannot be removed from an information system without disturbing equivalence relation. So these are indispensable. Discernibility matrix contains entries for a pair with different decision values is a list of attributes in which the pair differs. Reduct can be thought of as sufficient set of features to represent the category structure.

RDT Algorithm

Reduct based Decision Tree construction involves reduct computation and decision tree construction. Decision tree construction starts in recursively in top-down manner. Reduct computation starts with construction of Boolean matrix for the given input training dataset. The reduct computation is described in Algorithm 3.2 and decision tree construction in Algorithm 3.3.

3.3 GVM algorithm for print technology identification

A printed document is heterogeneous component with convolution of content as well as printer technology. As demonstrated in the previous section the spatial statistics in a variogram, in particular GVM, captures adequate information. Further we have explained the variations which are due to the content as well as printer technology. Hence, to identify printer technology, one need to filter the content associated information which itself is a challenging task. As a way out of this challenging problem we considered known components i.e., sub components

Algorithm 3.2 RCA(DT)

//Reduct computation algorithm [16]//

Input:

DT: Decision Table

Output:

Re: Reduct

Method:

1: The DT is read as input with the first column as decision column.

2: Sort the rows to get the row with the least decision attribute value on top.

3: The next step is to generate a Boolean matrix for a given decision table. For this purpose, we check if the first element of the first column in the decision table is equal to second element of the first column. If the answer is 'yes' then proceed to the next successive element along the same column until the answer is 'no'. If the answer is 'no' then assign value '1' to the corresponding entry in the Boolean matrix. Start comparing both the rows for corresponding column elements. If the column elements are same, assign '0' value otherwise assign '1' to the matrix. In a nutshell for the pair of rows in decision table having different decision attribute values, a row is generated in Boolean matrix.

4: Repeat step 3 for the second, third,.., N elements of the first column and compare with all the rest of the elements, by comparing one pair of rows at a time, and repeat this procedure till all the columns in the decision table are exhausted to construct the Boolean matrix.

5: The columns of Boolean matrix are then summed up. The column with maximum sum is picked up as a predominant attribute.

6: Take original Boolean matrix and remove those rows from it with 1 in the column corresponding to the predominant attribute selected in step 5, resulting in a reduced Boolean matrix.

7: Steps 5 and 6 are repeated until reduced boolean matrix is a null matrix, which means that the set predominant attributes picked up are sufficient enough to produce crisp rule set for discrimination.

8: These predominant attributes are grouped together and referred to as reduct, 'Re'

9: Return Re

Computational Complexity of RCA	
For data having m attributes and n instances	$O(mn^2)$
Algorithm Complexity: $O(mn^2)$	

Algorithm 3.3 RDTA(T)

//Reduct Decision tree construction algorithm[16] //

Input:

T: Training set

Output:

DecisionRule: rules from the root to each leaf node

Var:

Re: Reduct

Method:

1: Call RCA(T) //it returns reduct 'Re'. Construct a decision tree on 'T' with reduct 'Re' taking one attribute at a time//

2: Construct decision tree on training set 'T' with reduct 'Re', taking one attribute at a time and using it for splitting all nodes at the same level.

3: Generate the rules by traversing all the paths from the root to the leaf node in the decision tree.

4: Return DecisionRule

Computational Complexity of RDTA	
For m attributes and n instances	$O(mn^2)$
Algorithm Complexity: $O(mn^2)$	

of printed document in which content characteristics are more or less static.

In this chapter, we focus on the image component of the document. We further confine our work to homogeneous region of image component of the printed document. The tools discussed in the previous section are employed on this homogeneous region and a novel printer technology tool has been arrived at, which is named as GVM algorithm for print technology identification which is abbreviated as GVMPT.

3.3.1 GVMPT

For identification of print technology benchmark data is needed, which is discussed in Algorithm 3.4 and Algoirthm 3.5. Algorithm 3.4 describes steps for selection of sample from a document and Algoirthm 3.5 explains feature selection from that sample.

Preparation of training data set 'T' involves the following steps:

1. Select a document having homogeneous/uniform colour region
2. Print it on a printer at a fixed resolution 600 dots per inch
3. Scan the printed document at 2400 dots per inch
4. Identify homogeneous colour region of size 127 by 127 and convert it to gray.
5. Select features from gray converted homogeneous colour region sample along selected direction 'd' to get '$GVM\_dataset$'
6. This '$GVM\_dataset$' along with printer label forms the Training dataset 'T'. Hence, '$GVM\_dataset$' along with printer label for several known printer samples forms the training data set 'T'.
7. Normalize training data set 'T' with scaling factor 'k'.

This normalized training data set 'T_n' is used for constructing decision rules, where each leaf node contains printer label(P_id) as decision. Algorithm 3.3

Algorithm 3.4 DATAGEN(Document)

//Data generation for GVMPT//

Input:

Document: scanned image of document

Output:

g: gray converted uniform colour region of size $M \times M$ pixels

Var:

I: uniform/homogeneous colour region of size $M \times M$ pixels

Method:

1: Select homogeneous colour region say I(x, y) from Document // Use any standard technique for homogeneous region selection[63]//

2: Let I(x,y) is a uniform colour region where $1 \leq$ x $\leq$ M and $1 \leq$ y $\leq$ M

3: Convert 'I' to gray sample 'g' // Using technique proposed in [75]//

4: Return g

Computational Complexity of DATAGEN	
step 3 For I of size $M \times M$ each pixel conversion takes constant time	$O(M^2)$
Algorithm Complexity: $O(M^2)$	

Algorithm 3.5 FEATURESELECT(g, d, k)

//Feature selection from sample//

Input:

g: gray converted uniform colour region of size $M \times M$ pixels

d: direction specified in terms of angle

k: scaling factor

Output:

$GVM\_dataset$: GVM feature set having 17 parameters

Method:

1: Call VARIOGRAM(sample, d)

//Calculate Variogram along direction d of the sample//

//it returns V, S, N, R. V is vector of variogram, *sill* 'S', *nugget* 'N' and *range* 'R'//

2: Model variogram using Gaussian curves of 5th order

3: Select parameter of Gaussian Variogram Model along with 'S' and 'N' as $GVM\_dataset$.

4: Normalize $GVM\_dataset$ for scaling factor k

5: Return $GVM\_dataset$.

Computational Complexity of FEATURESELECT	
step 1, g of size $M \times M$	$O(M^3)$
step 2, number of parameters for model N_p	$O(N_p^3)$
Algorithm Complexity: $O(M^3)$ where M $\gg$N$_p$	

presents the steps for constructing decision tree on training data set 'T_n' and Algorihtm 3.6 explains classification of print technology of given test document based on decision rule. Once test document is given for identification of print technology to GVMPT, select homogeneous colour regions of the document as test sample. Extract features from test sample following the steps of Algorithm 3.5. Based on extracted features of test sample the decision rules identifies the source printer/print technology of given test document.

Algorithm 3.6 GVMPT(Testdocument, DecisionRule, k)

//GVM algorithm for print technology identification//

Input:

Testdocument: image of printed document

DecisionRule: set of rules from root to each leaf node

k: scaling factor

Output:

Label: printer id to which document belongs

Var:

g: gray converted homogeneous regions of size $M \times M$

Method:

1: Call DATAGEN(Testdocument) //it returns g, gray converted uniform colour region//

2: Call FEATURESELECT(g, d, k) //It returns testdata of sample. Here d=0°//

3: Using DecisionRule testdata will be labelled with source printer id.

4: Return Label

Computational Complexity of GVMPT	
step 1	$O(M^2)$
step 2	$O(M^3)$
step 3 m is no of attributes in reduct	$O(m)$
Algorithm Complexity: $O(M^3)$, where M $\gg$ m	

3.3.2 Illustration

An illustration of GVMPT for identification of print technology using 6 printer is explained in this section.

a) Selection of uniform colour region of image as samples

Each document is printed at 600 dots per inch and scanned at 2400 dots per inch. Uniform colour region of the scanned image is selected as sample. Selected samples are categorised as two types: Set-1 samples and Set-2 samples. Set-1 samples are uniform colour region of same source image printed on different print technology. These contain common image features and specific print features of the source print technology. Study of Set-1 sample data helps in identifying the specific features or characteristics of each print technology. The variations in samples are contributed to specific features of the print technology because the image features are common for each sample as they are from the same source or image.

Set-2 samples are uniform colour regions of different images printed using common print technology. These contain common print technology features but different image features. Set-2 samples printed on particular print technology contains common features of that print technology. Study of common characteristics among Set-2 samples reveals the common spatial statistics of print technology on which the samples are printed. These spatial statistics are considered as specific feature for identification of that print technology. Images shown in Figure 3.8 are used to produce Set-1 sample and Set-2 sample on printers listed in Table 3.3. Set-1 samples of uniform colour region Ncert-1 printed on these six printers are shown in Figure 3.9. The Set-2 samples of Deskjet 840c are shown in Figure 3.10.

Out of 66 samples due to unavailability of one print sample on Officejet6110 printer, total 65 samples are collected from 6 printer listed in Table 3.3. Uniform colour region is selected as sample. Sample of size 127 by 127 pixels was obtained by cropping and then converted to gray level image. Converted gray level image data was used as input for variogram algorithm in page 36.

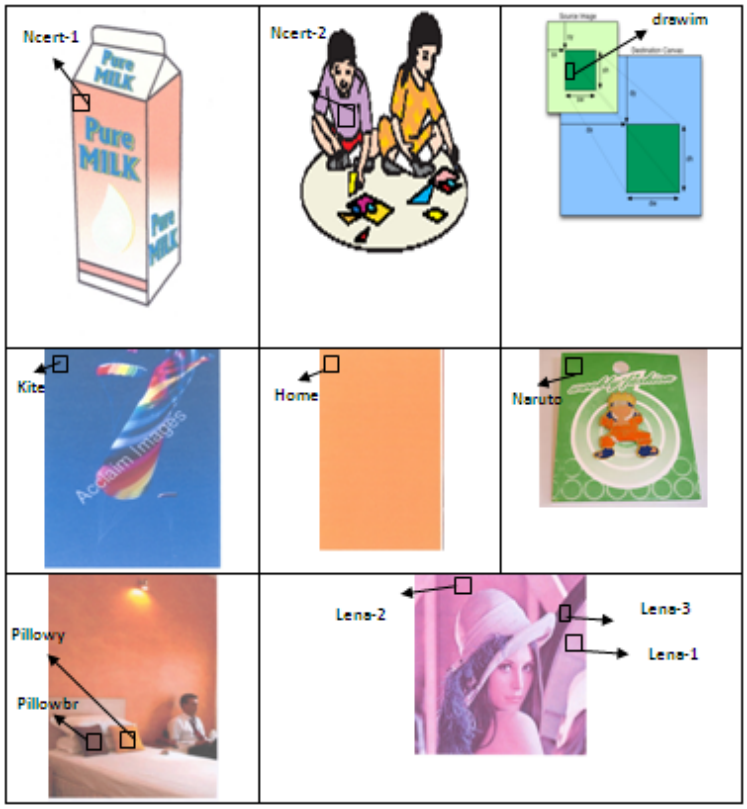


Figure 3.8: Sample images for Set-2 samples

b) Gaussian Variogram Model

Variogram generated for each sample are analysed for characterization of features which are distinguishable for identification of print technology. Here the variogram is fitted to Gaussian curve of different orders. 5th order Gaussian curve fits the variogram with low root mean square error and it has high rate of gain accuracy. Hence, the Gaussian curve of 5th order is selected to model variogram of samples. The structure of 5th order Gaussian for $K = 5$ in Equation 3.2, has 15 parameters. Non-linear Least square method is used to model variogram data [76]. The Figure 3.11 shows variogram for Ncert-1 sample of Deskjet840c and its corresponding modelled variogram using 5th order Gaussian curve. This GVM parameters are: $a_1 = 86.53$, $b_1 = 5.55$, $c_1 = 3.32$, $a_2 = 159.33$, $b_2 = 35.84$, $c_2 = 13.60$, $a_3 = 172.97$, $b_3 = 58.10$, $c_3 = 15.39$, $a_4 = 139.04$, $b_4 = 19.67$, $c_4 = 9.433$, $a_5 = 160.92$, $b_5 = 9.96$, $c_5 = 5.03$ respectively.

Figure 3.9: Set-1 samples of Ncert-1

Pid	Manu-facturer	Model	Samples collected	dpi	Print technology
1	HP	Photosmart3188	11	600	Drop-on-demand thermalinkjet
2	HP	Hppsc1608	11	600	Drop-on-demand thermalinkjet
3	HP	Deskjet840c	11	600	Drop-on-demand thermalinkjet
4	HP	Officejet6110	10	600	HP PhotoREt III
5	HP	Colorlaserjet4550N	11	600	Laser
6	Samsung	CLP-510	11	600	Laser

Table 3.3: Printers used for identification

c) Selection of feature set from Gaussian variogram model

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 15 parameters along with 'S' and 'N' of variogram is taken as feature set for quantization of the spatial features. Variability is associated with 'S' and 'N'. 'R' is associated with spatial dependency coverage which is used for the purpose of deciding window size where window wise analysis has been adopted. Thus 'S' and 'N' parameters of variogram has been adopted for analysis along with fifteen parameters of GVM. Hence, each sample gives feature set of 17 parameters. Samples collected from each printer is modelled as GVM which provides feature set of corresponding print technology. Hence, GVM data set is used as feature set for classification. GVM data along x-axis of Set-2 sample from Deskjet 840c is shown in Table 3.4.

Figure 3.10: Set-2 samples of deskjet 840c

d) Normalize GVM data for Classification

GVM data set obtained from each sample is taken as training data set. To make data more comparable, it has to be normalized before it can be used as training data set. Normalization is a procedure which transforms distribution of data into standard normal form. Z-score normalization method [77] is applied to transform GVM data set into normalized data. Z-score normalization discretizes data based on a factor which is known as scaling factor. For example for scaling factor k=0.5, attribute *range* is doubled after discretisation in contrast to k=1.

e) Construction of Reduct based Decision Tree

Reduct based Decision Tree(RDT) construction 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 algorithms. Datasets can be discrete or continuous, but here we use discrete

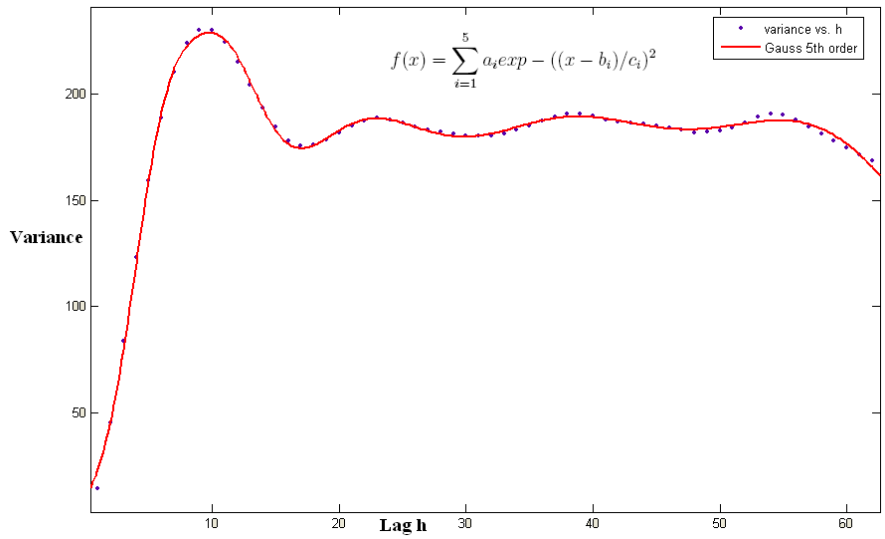


Figure 3.11: Variogram and model variogram of Ncert-1 sample of deskjet 840c

data set. The predominant attributes of the data when grouped together is called as reduct. Decision rules generated based on reduct is shown in Figure 3.12. These are then used for classification of print technology of test data set which is similarly normalized to same scaling factor as the training data.

Homogeneous colour regions of image shown in Figure 3.13 is printed on the six printers listed in Table 3.3 and taken as test sample (set-1 samples of test sample). 12 samples are taken as test data, out of which 9 are recognized accurately. Number of test samples identified accurately out of total number of test samples is taken as percentage of accuracy. For various scaling factor the percentage of sample identified correctly is shown in Table 3.5. Hence this model identifies the source printer with 75% accuracy at $k = 0.7$.

3.3.3 Directional GVMPT

In printer like laser jet technologies, they place the marking material to the paper such that it shows periodicity in particular direction like at angle 45° and it can be noticed in homogeneous colour regions of laser printer in Figure A.7. Any sample of laserjet print resembles this feature. Inkjet printers show some repeated pattern

Pid	b2	nugget	b3	b5
1	8	10	7	11
1	13	12	10	9
1	8	8	8	9
1	7	8	8	10
1	8	7	9	8
1	11	7	13	9
1	10	8	9	9
1	10	8	8	10
1	13	12	12	10
1	8	10	7	10
1	9	8	15	9
2	12	9	10	9
2	7	11	9	9
2	15	8	7	9
2	7	8	8	9
2	12	7	10	9
2	7	7	13	10
2	12	7	8	8
2	9	8	8	24
2	7	11	14	10
2	9	10	7	10
2	9	10	13	9
3	10	9	12	9
3	12	11	9	9
3	7	12	11	9
3	9	8	7	9
3	11	8	7	9
3	9	8	7	11
3	10	12	7	9
3	12	8	10	11
3	12	13	9	9
3	12	12	8	9
3	7	15	9	9
4	11	8	9	11
4	9	9	11	9
4	7	7	8	10
4	9	6	7	10
4	7	7	8	9
4	14	9	7	9
4	12	8	10	9
4	11	9	14	9
4	7	8	9	9
4	8	10	8	8
5	9	10	12	9
5	9	11	12	10
5	9	10	7	11
5	12	7	9	9
5	9	11	12	8
5	10	9	12	10
5	8	8	10	10
5	11	13	9	9
5	11	13	9	10
5	7	12	10	9
5	8	9	9	10
6	8	12	10	9
6	8	12	12	9
6	11	10	8	8
6	7	7	9	12
6	8	9	12	9
6	9	10	12	10
6	10	8	7	9
6	9	11	8	9
6	8	14	10	11
6	8	12	12	9
6	10	8	7	9

Figure 3.12: Decision Rules

GVM data with *sill* and *nugget* for deskjet 840c samples																		
Im No	P id	a1	b1	c1	a2	b2	c2	a3	b3	c3	a4	b4	c4	a5	b5	c5	*sill*	Nug -get
1	3	87	6	3	159	36	14	173	58	15	139	20	9	161	10	5	231	14
2	3	178	36	9	292	55	18	196	24	9	169	6	4	228	13	7	302	22
3	3	671	79	21	195	7	4	419	44	19	279	12	6	313	22	12	519	24
4	3	214	48	15	154	28	12	77	8	5	172	71	13	140	15	9	232	10
5	3	144	14	7	182	43	15	87	7	4	171	65	15	162	25	10	218	8
6	3	119	50	18	92	26	14	55	6	4	72	12	7	644	81	11	124	8
7	3	334	61	17	280	37	14	193	6	4	181	24	7	294	13	7	352	25
8	3	120	16	9	174	54	17	138	29	13	95	8	5	151	69	5	184	11
9	3	188	6	4	375	56	29	200	24	14	0	37	0	201	13	7	379	28
10	3	166	6	4	204	63	8	254	14	8	277	48	14	239	28	10	311	25
11	3	420	48	19	227	6	4	355	22	12	7743	137	39	251	11	6	468	37

Table 3.4: GVM data with *sill* and *nugget* for deskjet 840c samples

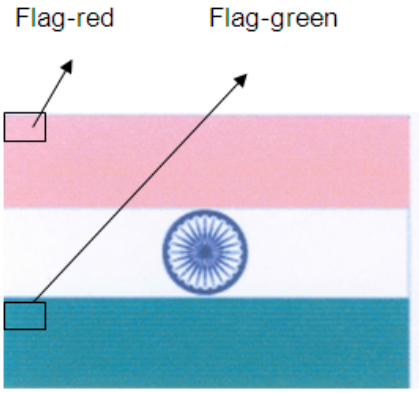


Figure 3.13: Test sample

along angle 20° in Figure A.2. Variogram generated along x-axis is not sufficient enough to quantize these features for identification. These features depend on particular direction and they can be modelled using directional variogram. To check for directional dependence in an empirical semivariogram, we have to compute variance values for data pairs falling with in certain directional bands as well as falling with in the prescribed lag limits. Algorithm 3.7 explains the steps for identification of print technology using directional variogram data.

Result for Directional GVMPT along 20° and 70° angle are shown in Table 3.6 and Table 3.7 respectively. Along 20° angle, it identifies 75% of test samples identified at k=0.5. Along 70° angle, the 83% of test samples are identified at k=0.6.

Algorithm 3.7 DGVMPT(Testdocument, d, DecisionRule, k)

// Algorithm for Directional GVMPT//

Input:

Testdocument: image of printed document

d: direction specified in terms of angle

DecisionRule: is set of rules from root to each leaf node

k: is scaling factor

Output:

Label: printer id to which Testdocument belongs

Var:

g: gray converted homogeneous/uniform region

Method:

1: Call DATAGEN(Testdocument) //it returns g, gray converted uniform colour region//

2: Call FEATURESELECT(g, d, k) //It returns Testdata of sample along direction d.//

3: Testdata is given as input to DecisionRule which returns label of printer. //Testdata selects a leaf node(label) in DecisionRule using euclidean distance measure to identify source printer//

4: Return Label

Computational Complexity of DGVMPT	
step 1	$O(M^2)$
step 2	$O(M^3)$
step 3 m is no of attributes in reduct	$O(m)$
Algorithm Complexity: $O(M^3)$, where M $\gg$ m	

RDT RESULT FOR GRAY GVMPT ALONG X-AXIS			
ANGLE	k	REDUCT	ACCURACY(%)
0	0.1	*nugget*, c2	41
0	0.2	c1, c4, *nugget*	50
0	0.3	c1, *nugget*, c3	33
0	0.4	c1, c3, c4, *sill*	58
0	0.5	b2, *nugget*, b3, b5	50
0	0.6	b1, *nugget*, c4, c3, c1	41
0	**0.7**	**b1, c4, c2, *nugget*, c1**	**75**
0	0.8	Data is not adequate	na
0	0.9	Data is not adequate	na
0	1	c1, *nugget*, b2, b3, c4, b4	66

Table 3.5: RDT results for Gray GVM data

RDT RESULT FOR GRAY DGVMPT ALONG Angle 20°			
ANGLE	k	REDUCT	ACCURACY(%)
20	0.1	b5, a5	50
20	0.2	*nugget*, c4, *sill*	58
20	0.3	*nugget* , b5, c4	50
20	0.4	b5 , b2, *nugget*	41
20	**0.5**	***nugget*, c4, b5, c5**	**75**
20	0.6	*nugget*, b5, b2, *sill*	66
20	0.7	*nugget*, b5, b2, b1, c5	16
20	0.8	*nugget*, b5, b2, a2, c5, c4	58
20	0.9	b5, b3, b2, b4, *nugget*	58
20	1	b2, b5, *nugget*, b4, c4, a5, c2	14

Table 3.6: RDT results for Gray Directional GVM data

Directional variogram along 70° angle has shown improved accuracy compared to angle 0°(along X-axis).

3.3.4 Standardised Gray Sample for building DGVMPT

To study spatial pattern that is colour independent the input sample has to be preprocessed. Preprocessing of the sample involves subtracting mean intensity of each channel from intensity of each pixel in that corresponding channel and dividing it by standard deviation of that channel. For making sample colour independent, each channel of image has to preprocessed. Such sample is called

RDT RESULT FOR GRAY DGVMPT ALONG Angle 70°			
ANGLE	k	REDUCT	ACCURACY(%)
70	0.1	b3 , *nugget*	66
70	0.2	b3, a3, c5	41
70	0.3	b3 , c5, *nugget*	33
70	0.4	*nugget*, b4, c5, b5	33
70	0.5	*nugget*, c2, b4, c5	41
70	**0.6**	**b3, c5, *sill*, c2**	**83**
70	0.7	*nugget*, b3, c5, c3, *sill*	50
70	0.8	*nugget*, b3, c5, c2, b5	41.6
70	0.9	b4, c3, c5, *sill*, *nugget*, c2	50
70	1	b4, *nugget*, c2, b3, *sill*, c5	50

Table 3.7: RDT results for Gray GVM data

standardised sample. Influence of standardised sample for GVMPT algorithm is discussed here.

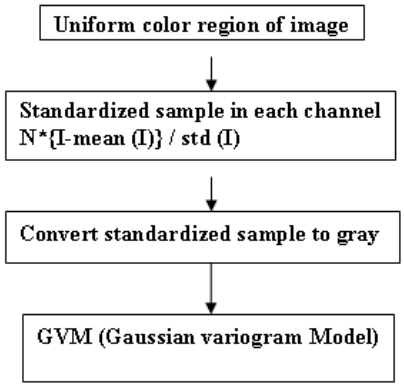


Figure 3.14: Flow chart for standardised GVMPT

The proposed Gaussian Variogram Model has given better results for preprocessed homogeneous colour regions. The preprocessing of homogeneous colour sample involves standardizing the sample in each channel. True colour images contain 3 channels namely red, green and blue. After standardising sample in the channel it has dimension of m by n by 3. This standardised sample has converted to gray level image of m by n for generating variogram.

Procedure for standardising uniform or homogeneous colour regions

1. Select uniform colour region of image as input sample
2. Standardise sample in each channel red, green and blue channel. To obtain standardised sample in each channel. It involves subtracting mean intensity of corresponding channel from intensity of each pixel and dividing it by standard deviation of the corresponding channel

$$Standardisedimage, S = N * I - mean(I)/std(I)$$

   Where N is any positive constant
3. It can be submitted to GVMPT(S, DecisionRule, k) or DGVMPT(S, d, DecisionRule, k)
4. It returns label of printer

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG X-AXIS			
ANGLE	k	REDUCT	ACCURACY(%)
0	0.1	a3, c5	41.6
0	0.2	a3, c1, c5	33.3
0	0.3	a3, c4, c1	41.6
0	0.4	c1, b3, *nugget*,c5	50
0	0.5	*nugget*, b3, c4, *sill*	41.6
0	0.6	c1, b3, c5, a3	50
0	**0.7**	**a3, *nugget*, b1, c5**	**66.6**
0	0.8	c1, a3, *nugget*, c5, *sill*	50
0	0.9	c1, b2, a3, b4, b1	41.6
0	1	b3, a2, c1, b5, c5, a3	41.6

Table 3.8: RDT results for standardised Gray GVM along X-axis

RDT results for the standardised GVMPT along x-axis is shown in Table 3.8. Comparison of result for gray GVMPT and Standardised GVMPT is shown in Table 3.9. It is observed that results for standardised sample is provide enough reduct for GVMPT while in case of gray GVMPT the at some scaling factors data is not adequate to produce reduct for identification of print technology.

ANGLE	k	Accuracy (%) for Gray	Accuracy(%) for Standardised
0	0.1	41	41.6
0	0.2	50	33.3
0	0.3	33	41.6
0	0.4	58	50
0	0.5	50	41.6
0	0.6	41	50
0	0.7	75	66.6
0	0.8	na	50
0	0.9	na	41.6
0	1	66	41.6

Table 3.9: Comparison of Gray and standardised Gray GVMPT along X-axis

3.3.5 Experimental results

In our experiment, total 159 samples were collected. These are Set-2 samples collected from each printer listed in Table 3.10. 116 samples form the training set from these 4 printers, (29 samples on each printer) and 43 samples from the same 4 printers (11 samples from each of 3 printers and 10 samples from printer with id 4) is taken as test data set. At different scaling factors, the reduct is selected and classification accuracy based Reduct based Decision Tree(RDT) is shown in Table 3.11. Percentage of number of samples identified correctly out of total number of samples is represented as percentage of accuracy. 79% of accuracy is achieved at scaling factor, k=0.8.

List of 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 3.10: List of printers used for identification

RDT RESULT FOR GRAY GVMPT ALONG X-AXIS			
ANGLE	k	REDUCT	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

Table 3.11: RDT results for Gray GVM data

3.3.6 Influence of algorithmic parameters in GVMPT and DGVMPT

The influence of algorithmic parameter in GVMPT and Directional GVMPT are summarised in this subsection. Algorithmic parameters affecting the results are input sample, direction, scaling/discretisation factor. Input sample can be gray sample or standardized gray sample. Direction is angle along which variogram is calculated, it can be along X-axis or along 10° to 85° angle with increments of 5°. Scaling factor k can be from 0.1 to 1 with increments of 0.1.

Results summary of Gray Directional GVMPT

Accuracy results of Gray GVMPT for identification of print technology along X-axis and along different directions from 10° to 85° angle for different scaling factors are shown in Table 3.12. Appendix-B presents details of reduct, identification accuracy in each direction.

Along angle 10° for different scaling factor from k=0.1 to 1, gray DGVMPT has shown 79% accuracy at k=0.8. For angle 15° the RDT results accuracy is 67% at k=0.6 and k=0.9. For angle 20°, RDT result shows that 72% of the samples are identified accurately at scaling factor k=0.2 and k=0.8. Along angle 25° at k=0.1,

RDT Result(accuracy) for Gray Directional GVMPT										
angle/k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	37.2	41.8	51.16	67.4	62.7	62.7	60	79	65	72
10	44.19	51.16	53.49	60.47	51.16	62.79	55.81	76.74	62.79	na
15	39.53	48.84	51.16	51.16	53.49	67.44	55.81	55.81	67.44	na
20	44.19	72.09	65.12	58.14	44.19	65.12	65.12	72.09	65.12	65.12
25	60.47	51.16	51.16	58.14	58.14	na	na	na	46.51	na
30	46.51	62.79	65.12	67.44	55.81	60.47	74.42	79.07	60.47	na
35	65.12	53.49	62.79	62.79	74.42	74.42	72.09	83.72	76.74	69.77
40	46.51	58.14	58.14	62.79	60.47	60.47	69.77	79.07	60.47	79.07
45	46.51	41.86	39.53	69.77	58.14	na	60.47	53.49	67.44	na
50	39.53	55.81	48.84	72.09	62.79	65.42	69.77	62.79	76.74	67.44
55	48.84	53.49	62.79	67.44	72.09	67.44	69.77	69.77	79.07	74.42
60	58.14	58.14	65.12	na	na	na	na	na	na	na
65	41.86	48.84	58.14	60.47	58.14	65.12	na	67.44	na	na
70	55.81	53.49	69.77	72.09	na	na	na	na	na	na
75	39.53	55.81	48.84	48.84	55.81	51.16	60.47	48.84	62.79	na
80	51.16	34.88	23.26	48.84	72.09	58.14	na	na	na	na
85	46.51	41.86	53.49	na	na	na	na	na	na	na

Table 3.12: RDT results for Gray DGVM data

60% of the test samples are identified correctly. GVMPT along angle 25° is not providing enough data for classification.

The Directional GVMPT along angle 30° at k=0.8 identifies 79 percent of test samples accurately. The features along angle 30° are distinguishable for identification of print technology. Along angle 35°, Directional GVMPT identifies 83% of test samples at k=0.8. Features along this direction are giving better results in identification. Along angle 40°, Directional GVMPT data using RDT classification identifies 79% of test sample accurately at k=0.8 and 1. Except k=0.1, the remaining scaling factor identifies more than 58% of test samples accurately. Angle 45° at k=0.6 and 1, directional GVMPT data is not adequate to identify print technology.

Along angle 50° Directional GVMPT identifies 76% of test samples accurately at k=0.9. Along angle 55°, directional GVMPT identifies 79% of the test sample

at k=0.9. Along angle 60°, Directional GVMPT at k=0.4 to 1, features are not enough to distinguish the samples. For angle 65° directional GVMPT data at k=0.8 identifying 67% of the test samples accurately. This direction is better than direction along angle 60°. Along angle 70°, Directional GVMPT data is not adequate for identification of the test sample. Along angle 75°, directional GVMPT identifies the test sample for scaling factor from k=0.1 to 0.9. Along angle 80°, Directional GVMPT data for k=0.7 to 1 features are indistinguishable. Along angle 85° for k=0.4 to 1, Directional Gaussian Variogram Model data is not enough for identification of the test sample.

Results summary of Standardised Gray Directional GVMPT

Along x-axis of standardised gray GVM data the RDT result are shown in Table 3.13. Comparison of RDT results for GVM (gray samples) and GVM(standardised gray sample)along x-axis is shown in Table 3.14. For various scaling factors, RDT classification accuracy does not completely coincide in both cases. The results are almost in the same *range* of accuracy. For some scaling factor, GVMPT using gray sample is giving better results like k=0.6 and 0.8. If data is not adequate for classification using gray GVM in such cases standardised gray GVMPT is gives better results. Without standardising the sample, performance of GVMPT accuracy *range*s from 37% to 79% and after standardising from 51% to 72%.

Directional GVMPT results for standardised gray sample for the training set having 116 samples and test set having 43 samples along the directions from angle 10° to angle 85° are tabulated in Table 3.15.

Along angle 10° and 15°, at k=0.9 and 1 the standardised gray Directional GVMPT data is not sufficient for identification. For the direction 20° angle, at k=0.9 it identifies 81% of test samples accurately. Along the direction of angle 25° and 30° the standardised gray Directional GVMPT identifies test samples with different accuracy for all scaling factor from k=0.1 to 1. Along angle 35°, standardised gray Directional GVMPT data at k=0.7 identifies 79% of the test

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG X-AXIS			
ANGLE	k	REDUCT	ACCURACY(%)
0	0.1	*sill*, c3 , *nugget*	51
0	0.2	*sill*, c4 , *nugget*, c5	53.4
0	0.3	c4, *nugget*, b2, *sill*	53.4
0	0.4	*sill*, b2, c4, b3	53.4
0	0.5	b3, b4, *nugget*, b2, *sill*	62.7
0	0.6	b2, *nugget*, c3, *sill*, c5	72
0	0.7	c4, *sill*, b2, b3, *nugget*, b5	62.7
0	0.8	b3, *sill*, b4, b2, b5, *nugget*	69
0	0.9	b3, c4, *sill*, b5, *nugget*, b4, c5, b2	62.7
0	**1**	**b3, b2, *nugget*, b5, b4, c3, *sill*, c5, c4**	**72.09**

Table 3.13: RDT results for standardised Gray GVM data

sample accurately. Along the direction of angle 40°, the standardised gray Directional GVMPT, at k=0.8 and identifies 79% and from k=0.5 to 1, it identifies more than 65% of the test samples.

Along angle 45° standardised gray Directional GVMPT at k=1 identifies 76% of test samples accurately. Along angle 50°, standardised gray Directional GVMPT data, at k=0.7 identifies 76 percent of test samples accurately. Except at k=0.1 for the remaining k values it identifies more than 60% of test samples accurately. Along angle 55°, standardised gray Directional GVMPT data at k=0.7 identifies 67% of test samples accurately. For values of k=0.8, 0.9 and 1, the data is inadequate for identification. Along angle 60°, at k=0.6, standardised gray Directional GVM data identifies 69% of test samples accurately. At k=0.9 and 1 the data is not enough for identification. Standarised gray GVMPT data features are indistinguishable in this direction compared to gray GVMPT.

Along angle 65°, standardised gray Directional GVMPT at k=0.6 identifies 69% of test samples correctly. Along this direction it has enough features to distinguish various print technologies for all values of k in the *range* of 0.1 to 1. Except k=0.1, 0.2 and 0.3, for the remaining k values more than 60% of samples are identified correctly. Along angle 70°, standardised gray Directional GVM data at k=0.5 identifies 79% of the test samples accurately. At k=0.9 and 1, data is

Comparision of RDT RESULT			
ANGLE	k	gray (%)	standardised gray(%)
0	0.1	37.2	51
0	0.2	41.8	53.4
0	0.3	51.16	53.4
0	0.4	67.4	53.4
0	0.5	62.7	62.7
0	0.6	62.79	72
0	0.7	60	62.7
0	0.8	79	69
0	0.9	65	62.7
0	1	72	72.09

Table 3.14: Comparision of RDT results for gray and standardised Gray GVM data

not adequate. Standarised gray Directional GVMPT data along this direction has enough features compared to gray Directional GVMPT.

Along angle 75°, standardised gray Directional GVM data is inadequate for k=0.5 to 1. Along angle 80°, standarised gray Directional GVM data at k=1, identifies 69% of the test samples correctly. Standarised gray Directional GVM data has enough features compared to gray directional GVM data along direction of angle 80°. Along angle 85° standardised gray directional GVM at k=0.7, identifies 67% of the samples accurately. Standardised gray directional GVMPT is better in this direction compared to gray directional GVMPT.

3.4 Performance analysis of DGVMPT

Performance analysis of GVMPT model depends on factors like direction of variogram and normalization factor at which GVM data is scaled. For gray level samples, the maximum accuracy of 83% is observed for the direction 35°. This is for the data normalized at factor k=0.8 that means data is scaled to 1.25 times. Here few results of gray GVMPT along different directions are listed below. Fig-

RDT Result for Directional Standardized Gray GVM										
angle/k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	51	53.4	53.4	53.4	62.7	72	62.7	69	62.7	72
10	37.21	46.51	55.81	55.81	69.77	60.47	60.47	58.14	na	na
15	37.21	41.86	51.16	44.19	51.16	53.49	53.49	67.44	na	na
20	46.51	58.14	53.49	58.14	58.14	74.42	69.77	67.44	81.3	62.79
25	37.21	48. 84	60.47	51.16	62.79	62.79	55.81	58.14	74.42	69.77
30	48. 84	51.16	48.84	69.77	72.09	69.77	69.77	69.77	67. 44	72.09
35	48.84	55.81	60.47	74.42	53.49	72.09	79.07	76.74	67.44	na
40	39.53	62.79	53.49	53.49	74.42	65.12	65.12	79.07	74.42	79.07
45	41.86	44.19	55.81	51.16	55.81	65.12	62.79	67.44	74.42	76.74
50	34.88	60.47	60.47	67.44	60.47	58.14	76.74	67.44	65.12	65.12
55	39.53	44.19	58.14	55.81	65.12	62.79	67.44	na	na	na
60	44.19	51.16	60.47	48.84	60.47	69.77	62.79	62.79	na	na
65	39.53	48.84	48.84	62.79	60.47	69.77	62.79	60.47	67.44	67.44
70	32.56	53.49	48.84	62.79	79.07	60.47	67.44	67.44	na	na
75	34.88	48.84	44.19	55.81	na	na	na	na	na	na
80	34.88	32.56	48.84	62.79	51.16	55.81	67.44	67.44	60.47	69.77
85	27.91	27.91	41.86	48.84	48.84	41.86	67.44	60.47	58.14	na

Table 3.15: RDT results(% of accuracy) for Directional Standardized Gray GVM data

ure 3.15 shows the scaling factors at which gray Directional GVMPT gives high accuracy. Figure 3.16 shows the angles at which gray Directional GVMPT gives high accuracy.

1. For along x-axis at k=0.8,accuracy is 79%
2. For angle 30° at k=0.8, accuracy is 79%
3. For angle 40° at k=0.8, accuracy is 79%
4. For angle 50° at k=0.9, accuracy is 77%
5. For angle 55° at k=0.9, accuracy is 79%
6. For angle 70° at k=0.4, accuracy is 72%
7. For angle 80° at k=0.5, accuracy is 72%

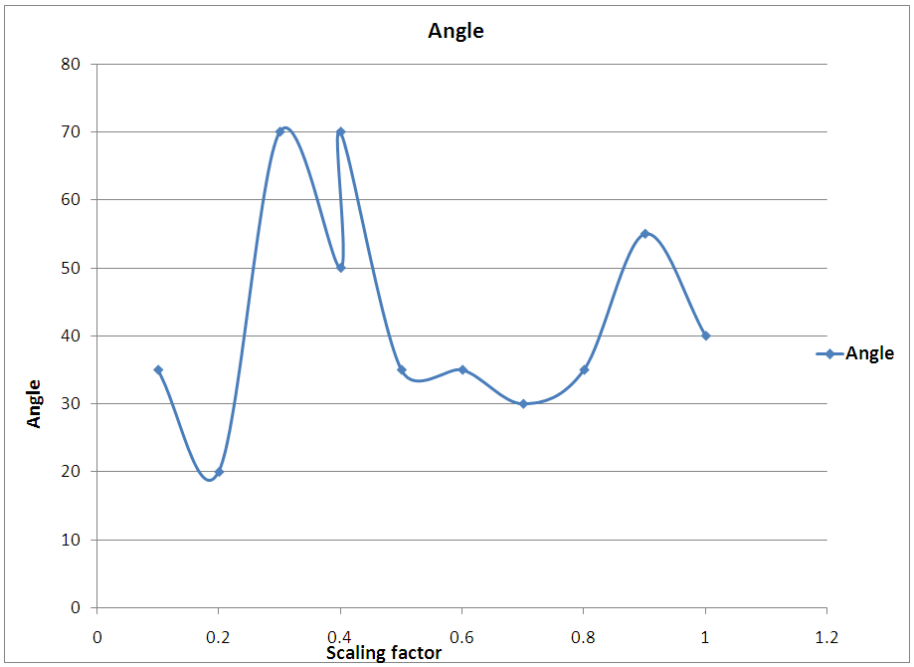


Figure 3.15: Angles at which high accuracies obtained for gray DGVMPT

For the graylevel sample along some directions at some scaling factors, the data is not sufficient to classify different print technology. This limitation is resolved by taking preprocessed sample as input to GVM. Each sample is preprocessed that means, it is standardised in each colour channel. For preprocessed sample the maximum accuracy 81% is observed for an angle 20° and for the normalized data at k=0.9. Here, few results of standardised gray GVM are listed below. Figure 3.17 shows the scaling factors at which standardised gray Directional GVMPT gives high accuracy. Figure 3.18 shows the angles at which standardised gray Directional GVMPT gives high accuracy.

1. For along x-axis at k=0.6 or k=1, accuracy is 72 %
2. For angle 35° at k=0.7, accuracy is 79%
3. For angle 40° at k=0.8 or k=1, accuracy is 79%
4. For angle 45° at k=1, accuracy is 76 %
5. For angle 50° at k=0.7, accuracy is 76 %

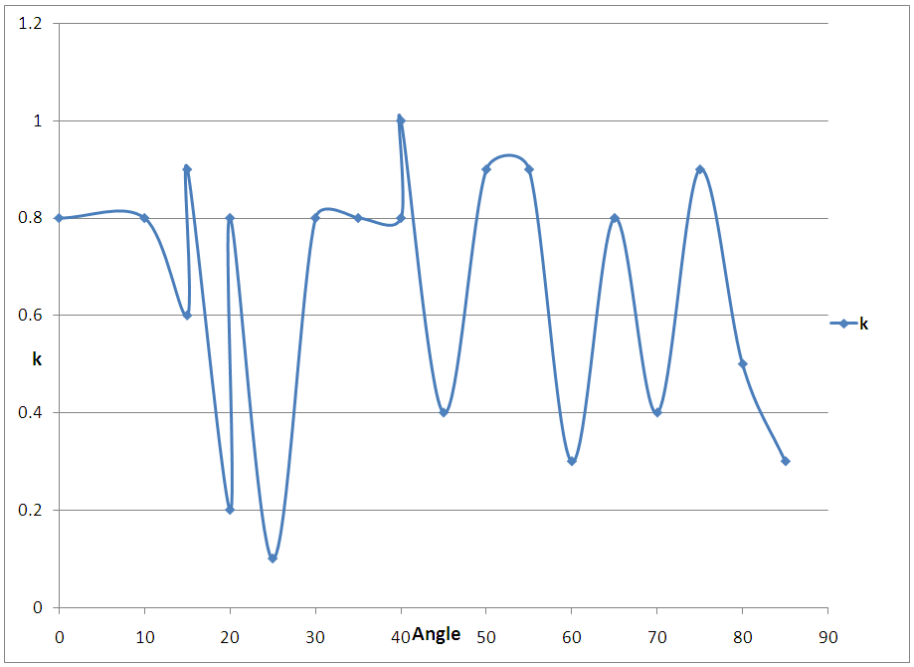


Figure 3.16: Scaling factors at which high accuracies obtained for gray DGVMPT

6. For angle 70° at k=0.5, accuracy is 79%

Selection of better direction and scaling factor provides better accuracy in identification of print technique based on homogeneous colour regions.

Recommendations

1. Select homogeneous regions of document as sample and convert to gray level image.

2. Generate directional GVM data for gray sample along direction d specified in angle.

   (a) Preferred d= 0°/20°/30°/35°/40°/50°/55° etc.

   (b) Normalize training data and test data for k=0.4/0.7/0.8/0.9.

3. Generate direction GVM data for standardised sample along direction d specified in angle.

   (a) Preferred d=20°/25°/35°/40°/45°/50°/80°

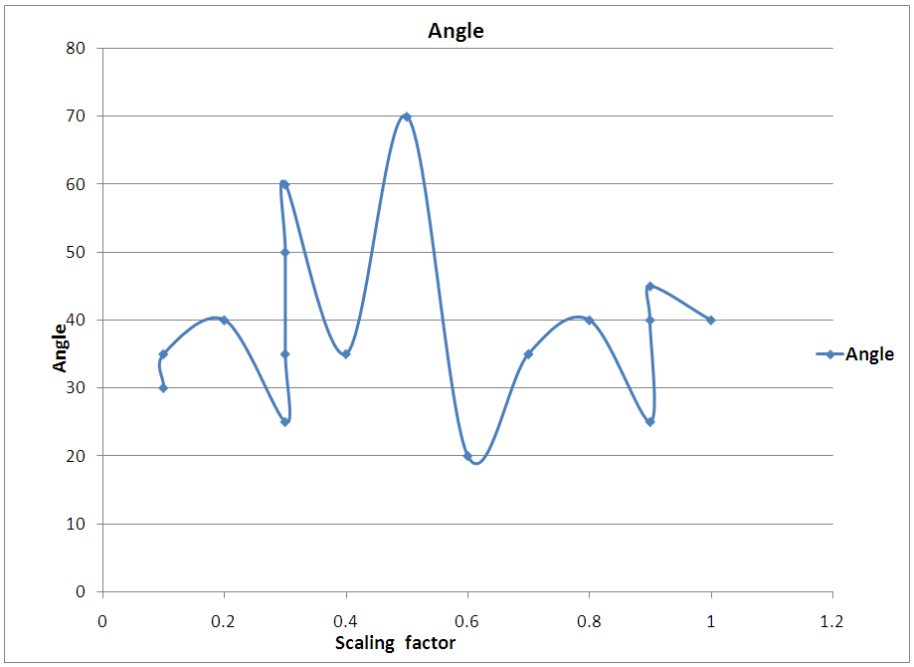


Figure 3.17: Angles at which high accuracies achieved for standardised gray DGVMPT

(b) preferred k=0.9 or 0.6 to 1.

4. Reduct generated from gray DGVMPT and/or standardized gray DGVMPT is used for identification of printing technique of test sample.

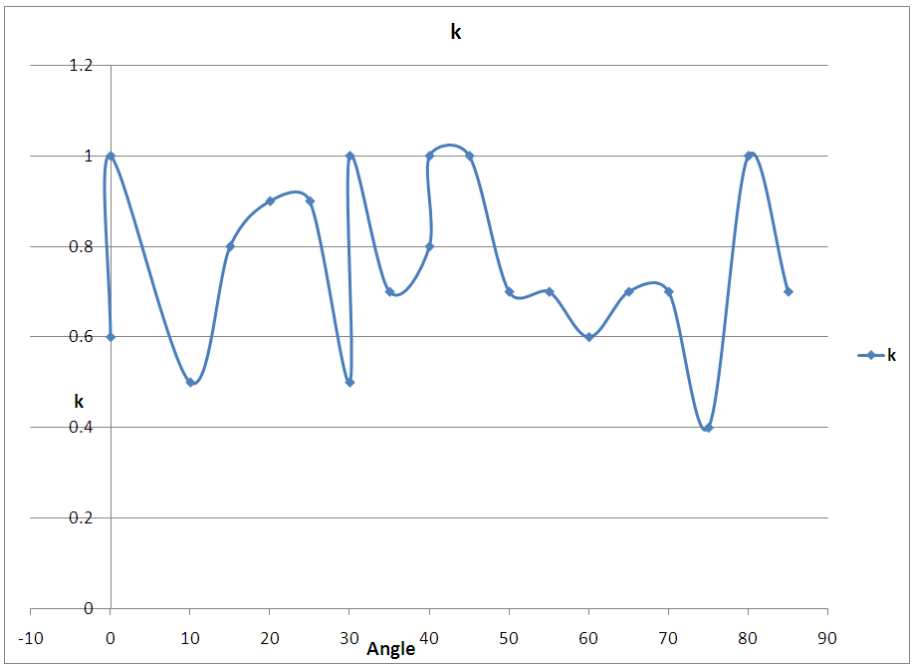


Figure 3.18: Scaling factors at which high accuracies achieved for standardised gray DGVMPT

Chapter 4

Identification of Print Technology based on Printed Text

4.1 Introduction

Printed material like documents of agreements related to authority or ownership of properties, identity cards 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 the document or with in the document. People print their names by removing genuine name in identity cards or change the names in the mark sheets for achieving certain targets.

As technology tends to evolve, the methods to forge documents are ever sophisticated challenging the skills of forensic examiners. Instinctive discrimination of document [78] is based on change in the intensity, colour, texture and structure of the document i.e., gap between lines and paragraph should be in consistent manner. In Beusekom [79] 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. Document examiner needs an excellent eye sight for examining such fine details in the document. High quality forged documents are produced using scanners and printers,

which makes identification of source of the document more complex. Therefore, determining the genuineness of document is critical and needs to be established.

This chapter is discusses the methodology for identification of print technology based on printed text. Section 4.2 discusses about the need for printed text characterization. Section 4.3 gives brief introduction to Expectation Maximization algorithm and its application to printed text for formulating a novel Print Index measure. Section 4.4 explains classification of inkjet print from laserjet print based on proposed Print Index measure and demonstrates its robustness. Section 5 describes about statistical measures used for differentiating photocopy from its print.

4.2 Need for printed text characterization

Identification of print technology is addressed in Chapter 3 and is based on homogeneous colour regions of printed document. As the document is combination of text and images, identification of source document depends on both the image and text of that document. The methodology proposed in Chapter 3, GVMPT/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. Hence, present study realises the need for generic measurement based on which printed text in the document is classified. This work focus on finding characteristics of printer from the printed text. 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.

Study of whole document in low resolution gives the whole document structure while study of local pattern in high resolution provides information of connected components or local structures in the document [78]. Characteristics of the printed

text differs based on the method of printing. These printing characteristics are connected to text content in the document. Hence the printer which produced the document is identified by characterization of the printed text content in that document. The text in the document is analysed by study of its texture which is characterized with statistical analysis of print pattern. Hence this statistical analysis is at the basis of text characterization for printing technology identification.

Document examiner needs to answer questions about the consistency of the document, i.e., 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 printed text is a combination of ink or toner over the printed material (which is generally white paper), the word and characters of the text in the questioned document has influence of the print 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. This observation made the researchers look at the segment of the selected text as a combination of standard normal distribution with mean, mixed proportion and variation.

Thus, the pixels of image of the text region can be modelled 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 [80]. 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. The contribution in this chapter can be presented in two ways: one building the classifier and second identification of print technology as inkjet or laserjet.

Forensic examination of printed document needs high resolution scanned images of the text as input for pixel level comparison. When the large number of documents are to be examined, then there is need to manage huge data for identification of print pattern. Dimension reduction of input data helps in preprocessing of the forensic analysis of printed text documents. Hence, characterization of printed text is based on feature set selection which reduces data dimension. The selected features assists the forensic examiner in identifying the basic print technology like ink jet or laser jet, which printed the text.

4.3 Expectation maximization technique

4.3.1 Introduction

Statistical partitioning of image into meaningful regions is the goal of many image analysis algorithms. The segmentation process will group the components existing in the image which are strongly related to the image objects. Expectation Maximization (EM) space partitioning algorithm is convergent and optimizes partitioning decisions based on the initial set of Gaussian Mixture Models [81]. A proper initializing condition is important. Otherwise, the algorithm will be forced to converge to numerous local minima.

Each iteration calculates the maximum likelihood estimates of the parameters of each Gaussian distribution of the data. EM [82] is an efficient iterative procedure to compute the Maximum Likelihood estimate in the presence of missing or hidden data. Maximum Likelihood estimation involves estimation of model parameters for which the observed data are most likely.

Each iteration of the EM algorithm follows two steps: E-step and M-step. E-step is an Expectation step, once observed data and current model parameters are given, expectation step involves estimation of missing data. Gaussian mixture density parameter is one of the widely used applications of EM algorithm in pattern recognition. Assume probabilistic model of mixed distribution:

$$p(x|\Theta) = \sum_{i=1}^{C} \alpha_i p_i(x|\theta_i) \tag{4.1}$$

where the parameters $\Theta = \{\alpha_1, \alpha_2, ...\alpha_C, \theta_1, \theta_2, ..., \theta_C\}$ such that

$$\sum_{i=1}^{C} \alpha_i = 1 \tag{4.2}$$

Here, C component densities are mixed together with C mixing coefficients α_i. Here i=1, 2, ...,C. where C is number of mixing components. The function p_i is the density function of Gaussian distribution and parametrized θ_i and is defined as follows:

$$p_i(x|\theta_i) = \frac{1}{\sigma_i\sqrt{(2\pi)}} exp\left[-\frac{|x - M_i|^2}{2\sigma_i^2}\right] \tag{4.3}$$

where x is the value of the observed variable, $\theta_i = \{M_i, \sigma_i^2, \alpha_i\}$, M_i, σ_i^2, and α_i are the mean, the variance and the corresponding mixing parameter for i^{th} gaussian distribution of that observed data.

The Expectation step is represented by log likelihood function as follows

$$Q(\Theta, \Theta(t)) = E[\log p(X, Y \mid \Theta) \mid X, \Theta(t)] \tag{4.4}$$

where $\Theta(t)$ are the current parameters and Θ are the new parameters that optimize the increase of Q. The M step is applied to maximize the result obtained from the E-step.

$$\Theta(t+1) = argmax Q(\Theta \mid \Theta(t)) \; and \; Q(\Theta(t+1), \Theta(t)) \;\geq\; Q(\Theta, \Theta(t)) \tag{4.5}$$

Expectation maximization steps are applied repeatedly until specified number of iterations or the log likelihood function is smaller than the specified values.

$$\alpha_i(t+1) = \frac{\sum_{j=1}^{N} p(i \mid x_j,\ \theta(t))}{N} \tag{4.6}$$

$$M_i(t+1) = \frac{\sum_{j=1}^{N} x_j p(i \mid x_j,\ \theta(t))}{\sum_{j=1}^{N} p(i \mid x_j,\ \theta(t))} \tag{4.7}$$

$$\sigma_i^2(t+1) = \frac{\sum_{j=1}^{N} p(i|\ x_j, \theta(t))|x_j - M_j(t+1)|^2}{\sum_{j=1}^{N} p(i|\ x_j,\ \theta(t))} \tag{4.8}$$

where

$$p(i|\ x_j, \phi) = \frac{\alpha_i p_i(x_j|\ \theta_i)}{\sum_{k=1}^{C} \alpha_k p_k(x_j|\theta_k)} \tag{4.9}$$

The following section discusses the application of EM for segmentation of text sample into meaningful regions.

4.3.2 Application of EM for text sample

Each test document is printed on white paper using printer where text is in the foreground, white paper in the background and pixels distributed all over the document as noise. Scanned image of printed word 'the' is taken as text sample which is shown in Figure 4.1. This scanned image 'the' is taken as input to EM algorithm for segmenting it into a strongly related component. Each word is assumed as a mixture of foreground text, back ground white colour and intermediate intensities which are known as noise. Thus, each printed text is assumed as mixture of these three Gaussian distributions. Associated patterns of these three Gaussian distributions of text is built using the EM algorithm. Gray scale image of printed text is submitted to EM algorithm considering three classes, which are fore ground text, noise and background. Each pixel in the image is assigned a posterior probability of belonging to one of the three classes. Based on posterior probability of each pixel, it is classified as belong to one of the three classes and is shown in Figure 4.2.

4.4 Classification of inkjet versus laserjet

Study of spatial pattern in homogeneous colour region employed in Chapter 3 classifies ink jet printing from laser printing[83]. In case of printed text, it is dif-

Figure 4.1: Scanned sample word 'the'

Figure 4.2: Segmented text after applying EM

ficult to get such homogeneous colour region for study of spatial pattern using Gaussian Variogram Model(GVM). Expectation Maximization method segments the printed text into strongly related components for classification of inkjet technology from laser. EM gives the parameters for segmenting the text into three clusters, which represents text, noise scattered around the edges of the printed text and background. These segments of text are useful for classification of printing techniques.

Black text with white colour as back ground is printed on white paper. Some pixels which are not part of the text are scattered around the text. These intermediate intensities which are concentrated on edges of the printed text are called as noise. Pixel intensities of noise are between the printed text and background. The amount of noise differs for various printing technology dependent on the way marking material is placed on paper and distribution of the marking material. Ink jet printers use dispersed dot dithering[24][84] in which individual pixels can be addressed whereas laser jets use clustered dot dithering[85] which is periodic

and ties several dots together as a cluster. This can be derived by characterizing printed text.

In this section, we developed a novel Print Index(PI) based on the printed text. The following systematic procedure has been developed to obtain the printer index for a given printer.

1. Print a word say 'the' on white paper at 600 dpi.
2. Scan the printed text at 2400dpi.
3. Extract the word 'the' from scanned image and fix it to minimum bounded rectangle(MBR)
4. Normalize the MBR to fixed size using nearest neighbour interpolation[86]. Each text sample "the" is resized to [300,400] and text sample 'The' is resized to [300,600].
5. Resized sample is given as input to EM Algorithm. Determine EM parameters $\{M_i, \sigma_i^2, \alpha_i\}$ for i=1,2,3 where C=3, i.e., number of classes=3.

EM algorithm segment each sample into 3 clusters namely text, noise, background (white). For each cluster we compute parameters mean intensity, mixing proportion and variance. The data generated for bench mark data set is given to EM algorithm, which returns parameters to formulate Print Index. The following sub section is devoted to methodology for classification of printer technology based on Print Index and its performance.

The steps for characterization of printed text are shown in Algorithm 4.1, which returns Print Index(PI). These PI's are for classification of print technology which produced the printed text.

4.4.1 Classification of printed text

Classification of printed text is the process which selects various text samples produced using various print technology and compute print index of these samples.

Algorithm 4.1 PRINTCHAR(Textsample)

//Algorithm for Characterization of printed text//

Input:

Textsample: resized image of printed text

Output:

PI: Print Index

Method:

1: Give Textsample as input to EM algorithm //EM parameters for text are $\{M_{text}, \alpha_{text}, \sigma^2_{text}\}$, for noise are $\{M_{noise}, \alpha_{noise}, \sigma^2_{noise}\}$ and for background component are $\{M_{background}, \alpha_{background}, \sigma^2_{background}\}$ respectively//

//Cumulative Mean, μ is defined as follows//

2: $\mu \leftarrow \mu_{text} + \mu_{noise} + \mu_{background}$

$//\mu_{text} \leftarrow M_{text} * \alpha_{text}//$

$//\mu_{noise} \leftarrow M_{noise} * \alpha_{noise}//$

$//\mu_{background} \leftarrow M_{background} * \alpha_{background}//$

3: Print Index of text is defined as $PI \leftarrow 100 * \frac{(I_{noise} - I_{text})}{(I_{text})}$

//Index of text, $I_{text} \leftarrow 100 * \frac{\mu_{text}}{\mu}$//

//Index of noise, $I_{noise} \leftarrow 100 * \frac{\mu_{noise}}{\mu}$//

//Index of background, $I_{background} \leftarrow 100 * \frac{\mu_{background}}{\mu}$//

4: Return PI

Computational Complexity of PRINTCHAR	
Complexity of EM algorithm , $p \times q$ is size of sample	$O((pq)^2)$
Algorithm Complexity: $O((pq)^2)$	

For classification, group these Print Indexes to identify the print technology of given test sample. Hence, selection of text sample is significant in building the methodology for characterization of printed text. This section discusses about preparation of text samples.

Selection of text samples

Three types of text documents are prepared for selection of text samples. Each text document has Times New Roman font with font size of 12pt. These documents are printed on white paper at 600 dpi and scanned at 2400 dpi using HP Scanjet scanner.

First type of text document contains most frequently occurring word 'the' and it is taken as Type 1 text sample. Second type of document contains word 'The' and it is taken as Type 2 text sample. Some general text document is taken as third type of text document, in which all three letter word are taken as Type 3 text sample. In all these document each sample has been viewed on the printed text as foreground, white paper as background and edge region of each character as noise. Each word in the document is selected by application of minimum bounded rectangle [87]. For sample of any small case three letter word like 'the' are resized to a fixed size of 300 by 400, where as the sample of type like 'The' is resized to 300 by 600 for characterization of text. Rapid changes in availability of printers created constraints in selection of text sample. Therefore, text sample on available printers was considered. The text samples are selected from the printers listed in Table 4.1 and Table 4.2. First and third type of text documents was printed on printers listed in Table 4.1. Second type of text documents are printed on printers listed in Table 4.2. which shows printer id, no of samples collected and printer name.

Segmentation of text samples using EM

Selection of text samples and preprocessing of samples is followed by segmentation using Expectation Maximization algorithm. Printed text is represented as

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 4.1: Printers used for printing 'the' and general three letter word

P. id	No. of words Collected	printer	Print technology
1	200	Hppsc1608	Inkjet
2	200	Officejet6110	Inkjet
3	100	Hplaser4550N	Colorlaserjet
4	100	Hplaser1200	Laserjet
5	100	SamsungML2010	Laserjet
6	100	Xeroxwcpe220	Black and white laser
7	100	Hplaser 9040	Black and white laser
8	82	CannonIR3530	Black and white laser
9	100	Cannon LBP2900	Black and white laser

Table 4.2: Printers used for printing Text sample 'The'

a mixture of components and each component is segmented using the estimated parameters of EM.

It is assumed that each word in the printed text contains three main components, namely, printed text, noise which is scattered around the edges of printed text due to the spatial distribution of the marking material and the back ground. Each component is modelled as Gaussian distribution and the printed text is represented as a mixture of these three Gaussian distributions. These models are built using EM algorithm which returns parameters with mean and variance for each of the three components.

The Print Index(PI) proposed based on these parameters is taken as a feature to characterize printed text. X is the intensity of a pixel and it is taken as data for EM model, $\theta_i = \{M_i, \alpha_i, \sigma_i^2\}$, where M_i, α_i and σ_i^2 are the mean, mixing proportion and variance for i^{th} Gaussian distribution. Here i=1, 2, 3.

EM algorithm returns parameters mean mixing proportion and variance for text $\{M_{text}, \alpha_{text}, \sigma^2_{text}\}$, for noise $\{M_{noise}, \alpha_{noise}, \sigma^2_{noise}\}$ and for background component $\{M_{background}, \alpha_{background}, \sigma^2_{background}\}$ respectively for a given printed text.

Text 'the' referred to as Type 1 samples are printed on printers listed in Table 4.1 and shown in Figure 4.3. Application of Expectation Maximization technique for segmenting each text into three components is shown in Figure 4.4. Text 'The' referred to as Type 2 samples are printed on printers listed in Table 4.2 and shown in Figure 4.5, corresponding segmented images are shown in Figure 4.6.

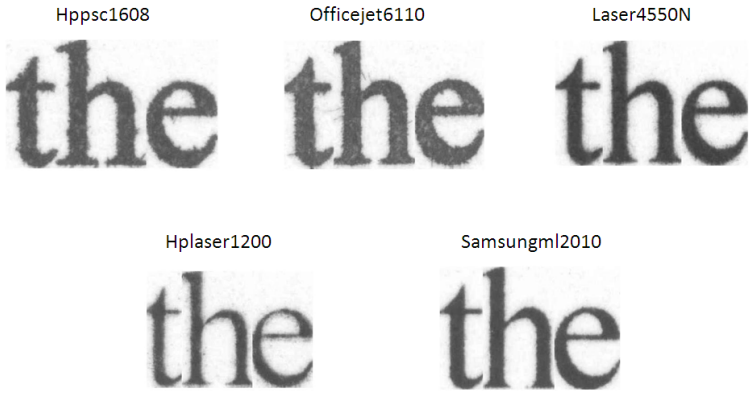


Figure 4.3: Type 1 samples printed on printers listed in Table 4.1

Figure 4.4: Segmented words of Figure 4.3

It is observed from Figure 4.4 and Figure 4.6 that the scattering noise around the text is a distinguishing feature for classification of ink jet printer from laser

Figure 4.5: Type 2 samples printed on printers listed in Table 4.2

jet printers.

Characterization of Printed Text

Each sample is characterized using Print Index measure which uses modelled Gaussian parameters of each sample. Application of EM to each printed image of text sample, it returns modelled Gaussian distribution parameters for fore ground text, background and noise. Cumulative mean for each sample and Index of text, back ground and noise of each sample is calculated as shown in Algorithm 4.1. Cumulative mean and Index of text and noise is used to formulate Print Index measure.

Total of 500 samples of 'the' are collected from 5 different printers listed in Table 4.1. Type 1 samples collected from various printers are shown in Appendix A from Figure A.14 to Figure A.18. Sample versus Index of each component in that sample is shown in Figure 4.7. From Figure 4.7 it is clearly noticed that index of text is more for inkjet samples compared to laserjet samples whereas index of noise is more in laserjet samples compared to inkjet samples.

Print Index of each sample is shown in Figure 4.8. PI for samples from each printer is plotted and it is shown in Figure 4.9. From Figure 4.8 and Figure 4.9,

Figure 4.6: Segmented words of Figure 4.5

it is observed that the PI of ink jet printers is less compared to the print index of laser jet. It is clearly visible that upper bound of print index of most inkjet printers are near value 70 while it is lower bound for the print index of laserjet sample. From the Figure 4.8 and 4.9 it is also observed that the randomness in Print Index of laserjet samples is more compared to inkjet samples. This indicates that index of noise is more and random in laser jet compared to the noise in ink jet. This is consistent with our explanation in the following section about the classification of inkjet versus laserjet. Type 1 samples, 'the' are used to build the classifier and Type 2, Type 3 samples are used for demonstrating the performance of classifier or methodology. It can be observed from the print index in Figure 4.9, that the whole Print Indexes can be divided into 2 categories of data and labelled as 1 and 2. As two classes of data is available for classification, one is for inkjet and the other is for laserjet print, first label of data represents inkjet Print Indices and the second label represents for laserjet Print Indices.

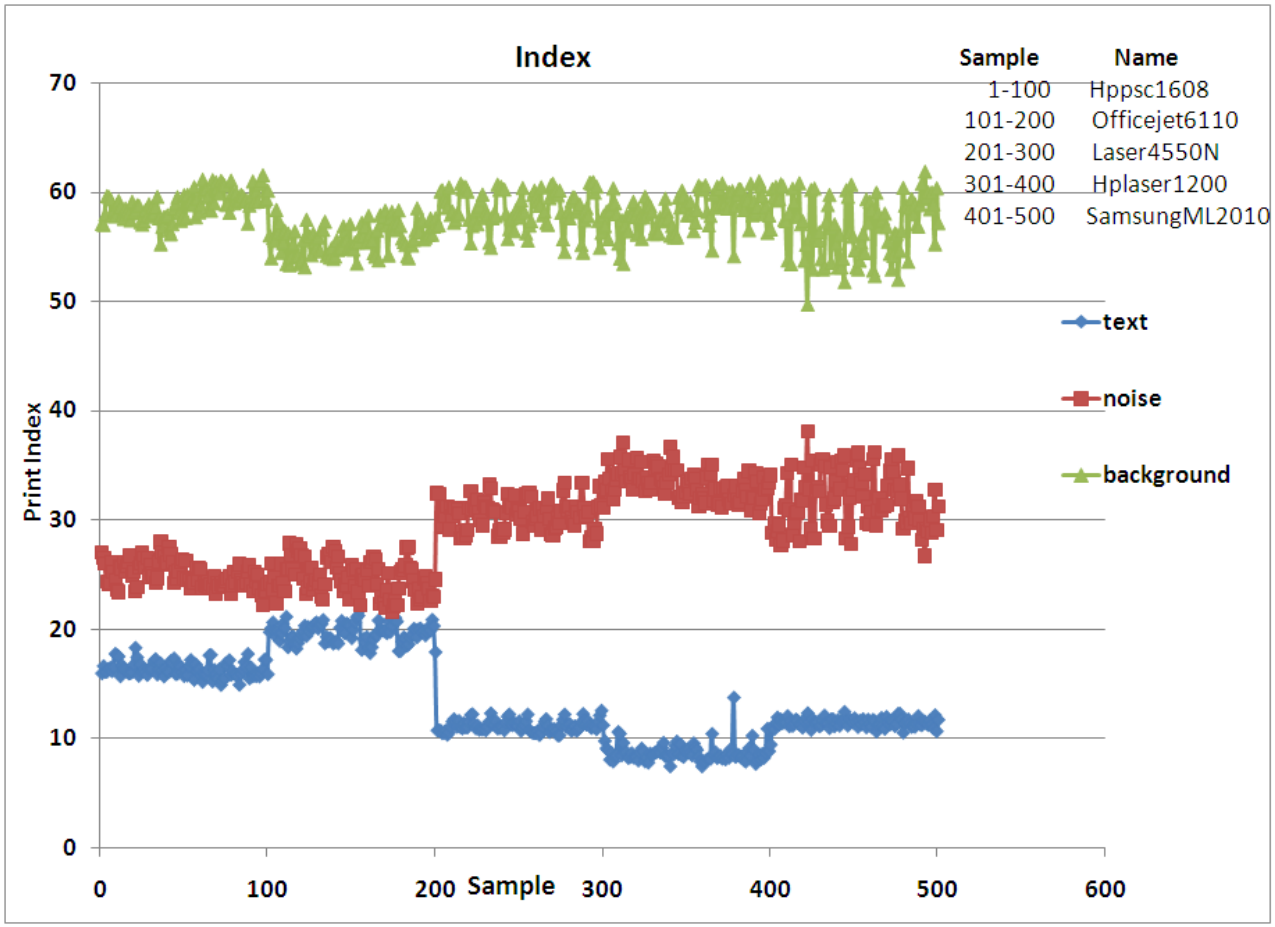


Figure 4.7: Printer sample Vs index of text, noise and back ground for Type 1 samples

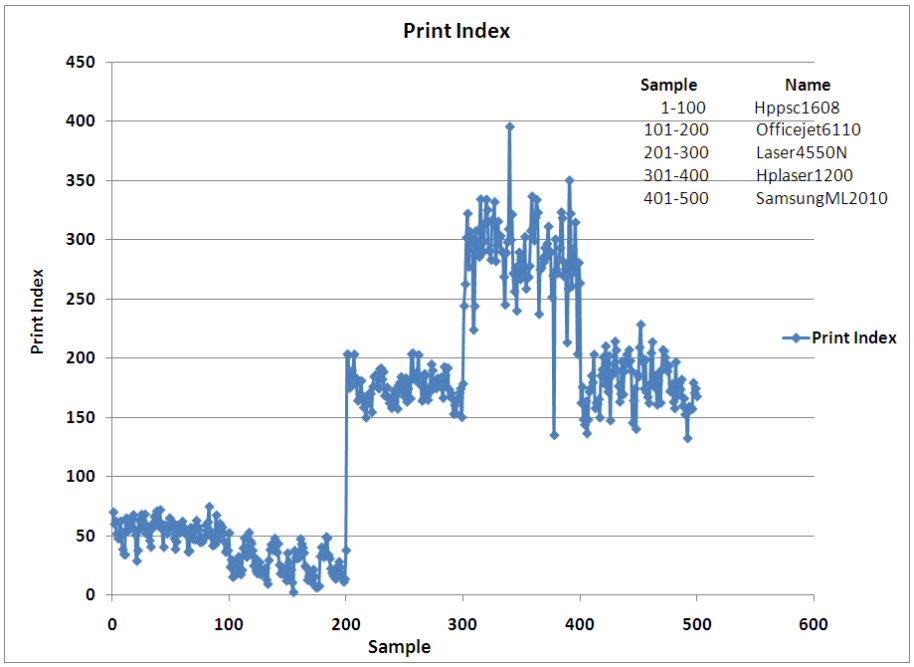


Figure 4.8: Sample Vs Print index of printers listed in Table 4.1

Experiment results

The Print Indices calculated are given as input to EM algorithm for classification of print technology. As text samples are collected from various ink and laserjet prints, the calculated print index can be grouped to form different labels based on their common class characteristics.

For Type 2 text sample 'The', classification accuracy is as follows. Out of 9 printer data, the characterized print index are labelled 1 to 2 and classified as inkjet or laser jet. For Type 2 text sample, total 1082 'The' are collected from 9 printers listed in Table 4.2 and Print Index is calculated for each resized sample of size [300,600]. Type 2 samples collected from various printers are shown in Appendix A from Figure A.19 to A.27. Print Indices of 'The' are shown in Figure 4.10. Ten fold test [88] is adopted for demonstrating performance of classification. Ten percent of the total samples are set aside as test data and remaining is taken as training data. Training data along with number of classes is given as input to the EM algorithm which returns the mixed proportion and mean of each class. Based on these parameters, the test data is classified. The test data contain

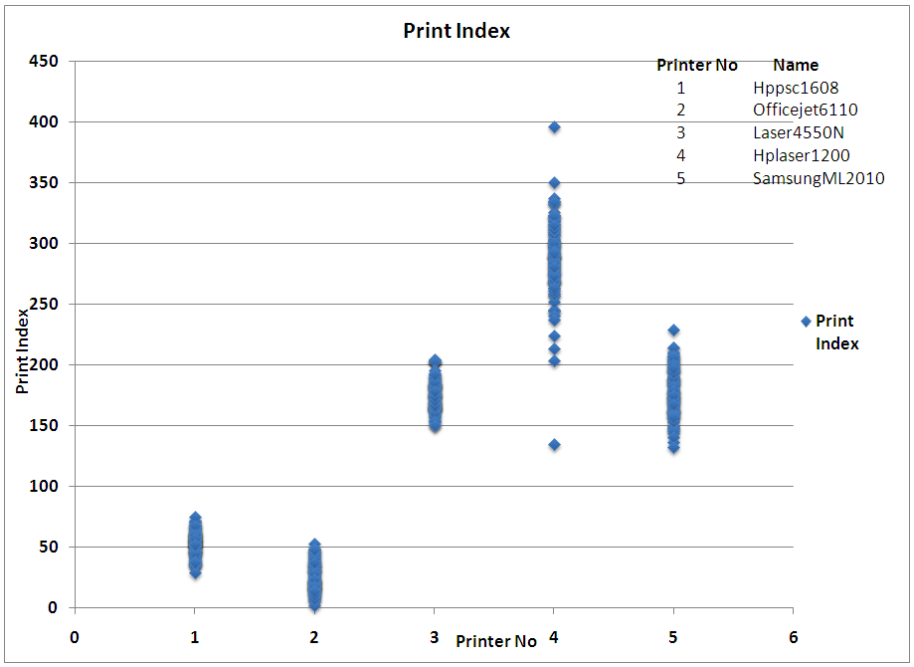


Figure 4.9: Print index for printers used for Type 1 sample

108 samples and are selected randomly out of 1082 collected samples and the experiment is repeated for each iteration. Number of classes is 2(inkjet and laser jet) and number of printers is 9. For 10 iterations, the total test data contains 1080 sample. This classification is shown in Table 4.3.

Classified as			
	Ink jet	Laser Jet	Error
Ink jet	400	0	0 in 400
Laser Jet	7	673	7 in 680
Total	407	673	1080

Table 4.3: Classifying print Technology

Out of 1080 samples, 7 samples of laserjet are misclassified as inkjet print. Performance analysis of proposed method is explained in terms of 'Kappa'[89], a statistical measure for calculating correct chance of agreement. 'Kappa' value of 98% chance of agreement is achieved and it indicates near perfect performance.

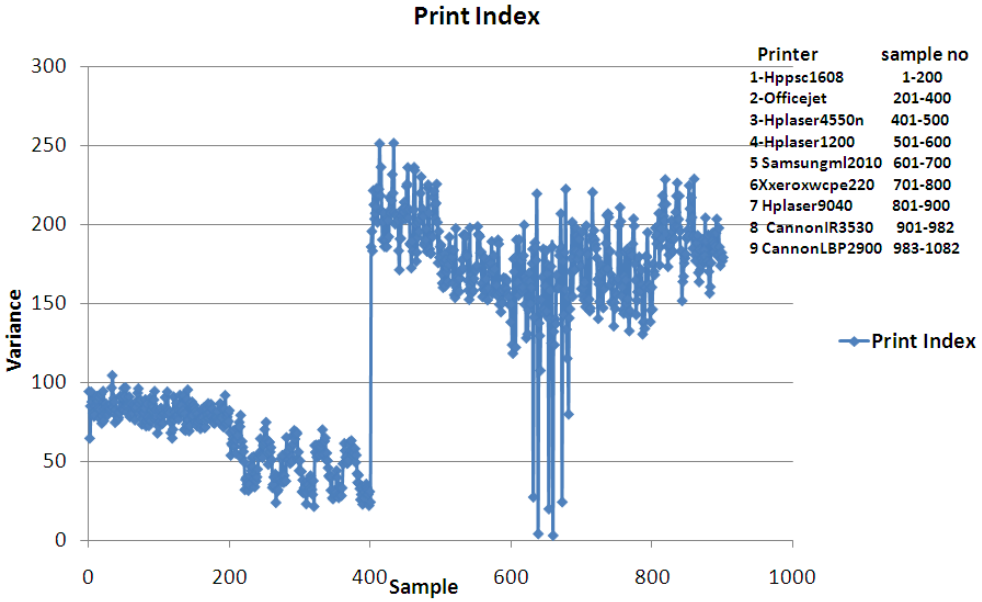


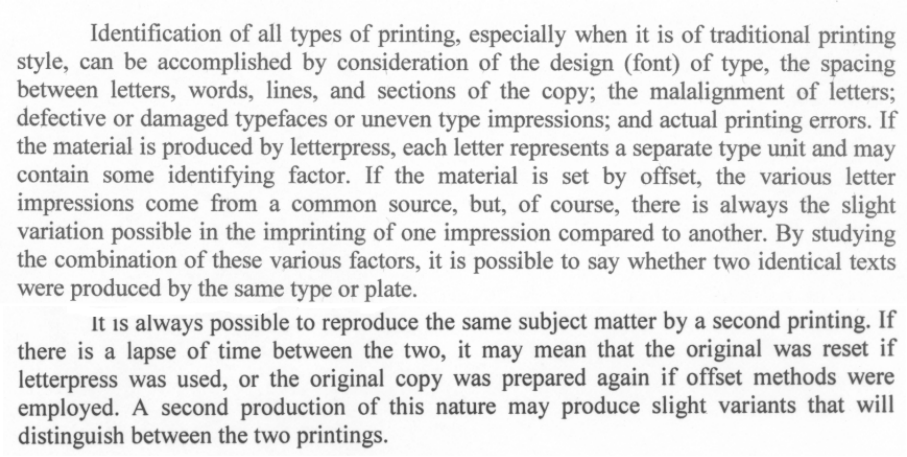
Figure 4.10: Print Indices of Type 2 sample 'The'

4.4.2 Robustness of printed text characterization

Printed text characterization is demonstrated on the general text document containing some general text which is shown in Figure 4.11. Three letter words contained in the document are shown in Figure 4.12. These 34 words are selected as text samples from this document for validating the Print Index methodology. Consider an image of printed text shown in Figure 4.11, and print indices of three letter words in that text image are computed. From Figure 4.13, one can observe that first 22 three letter words in the first paragraph are from inkjet where as 12 three letter of words in second paragraph are from laserjet.

Type 3 document was printed on five printers listed in Table 4.1. Type 3 samples are shown in Appendix A from Figure A.28 to Figure A.32. The experiment was carried out for calculating print index of each three letter word from the documents which are outputs of different source print technology. The samples numbered 1 to 68 are inkjet print samples and the samples from 69 to 170 are laserjet samples.

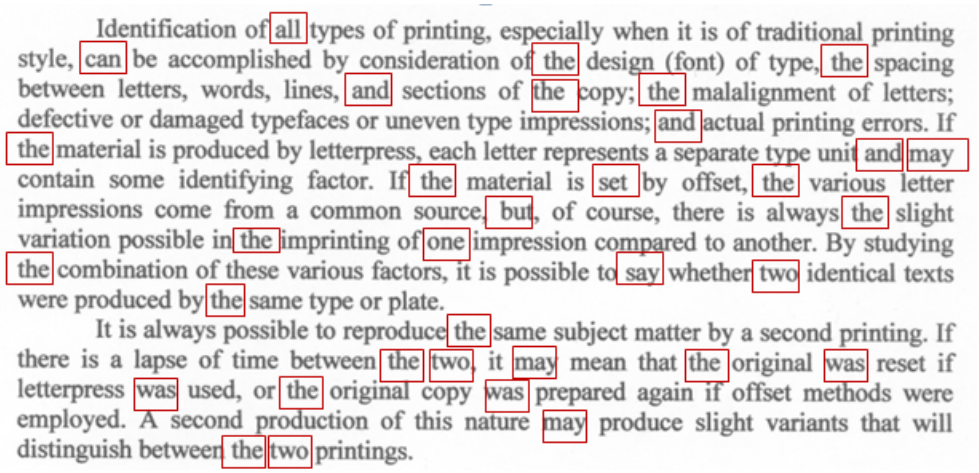
Observing the text, noise and back ground indexes of Type 3 sample, the dif-



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.

Figure 4.11: General text document



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.

Figure 4.12: Three letter words contained in general text documents

ference of text and noise index is less in inkjet print compared to the laserjet print shown in Figure 4.14. Sample 52 and 61 are from inkjet printer Officejet6110 which is not following the basic assumption that index of text is less compared to noise and these disturbance are considered as ink overflow condition. Hence, these two samples are considered as invalid samples. Print Index of Type 3 samples are plotted in Figure 4.15 which clearly forms two groups: one is from 1-68 representing inkjet print and the other is from 69 to 170 representing laserjet print. Hence, the proposed method clearly distinguishes inkjet print from laserjet for general three letter text.

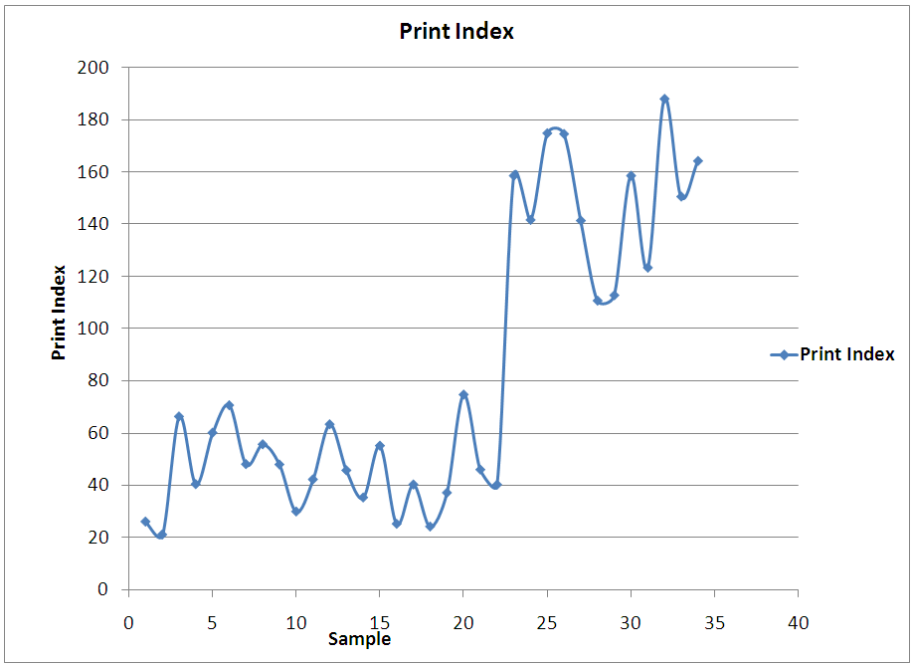


Figure 4.13: Print Index of test sample

4.5 Classification of photocopy from inkjet print

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 4.16 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.

Problem definition

Photocopier [90] 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

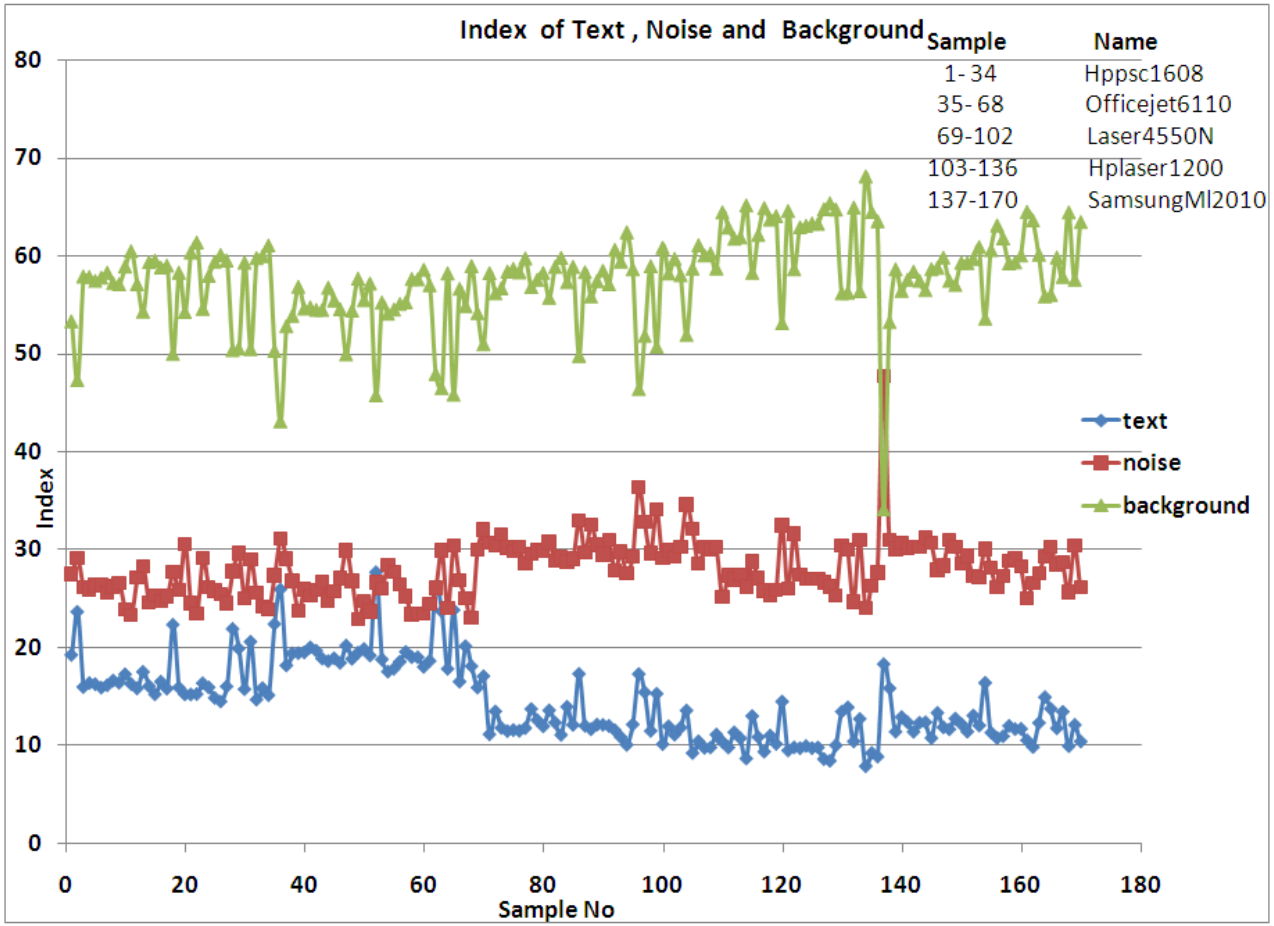


Figure 4.14: Text noise and background indexes of Type 3 sample

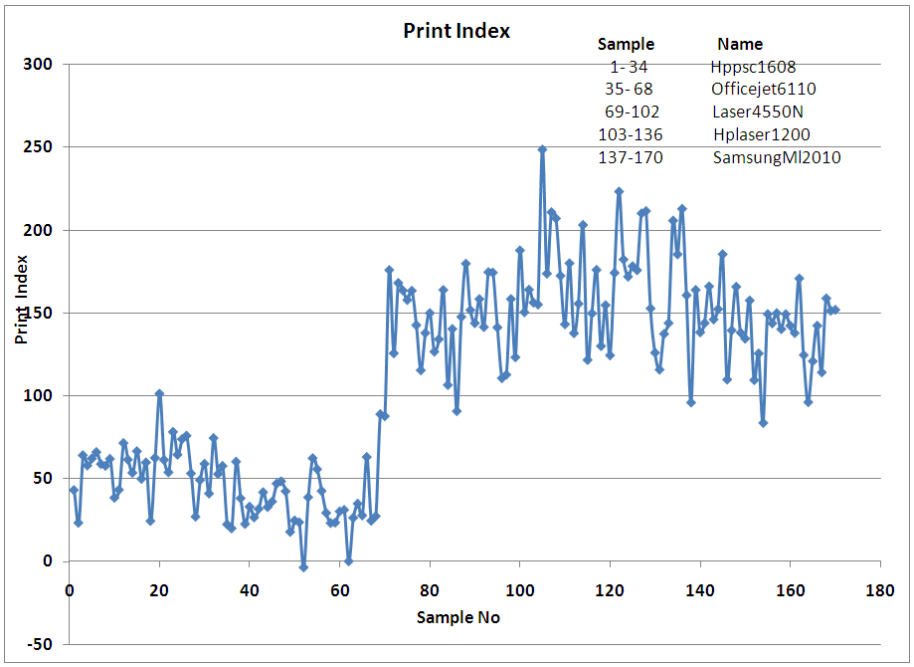


Figure 4.15: Print index of Type 3 samples

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 origi-

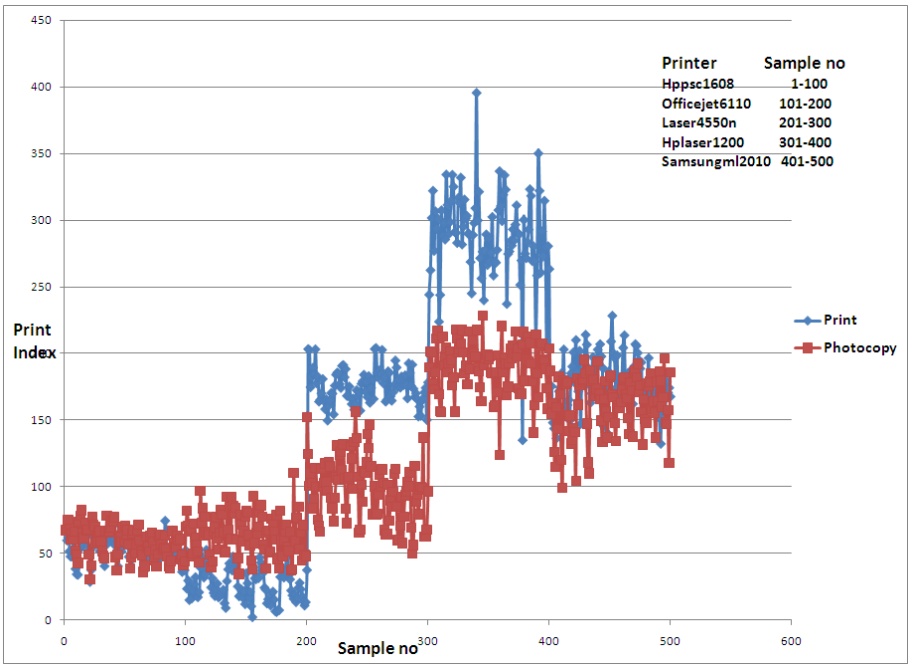


Figure 4.16: Comparision of Printindex of print and its photocopy

nal document, some disturbances occur which results in photocopied document. These disturbances in photocopies are exploited and characterized using statistical analysis techniques.

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 4.1. The printed documents are photocopied using photocopiers listed in Table 4.4. 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.

S.no	Photocopier
1	Konika Minolta bizhub210
2	XeroxWCM118

Table 4.4: List of photocopiers used

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 4.17. 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 4.18, 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.

The purpose of the histogram[28] is to graphically summarize the distribution of univariate data set. The histogram of an image shows 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[91]. 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

Figure 4.17: Print out and its photocopy

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.

$$Skew = \beta_1 = \mu_3^2/\mu_2^3 \tag{4.10}$$

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. Data sets with low kurtosis tend to have a

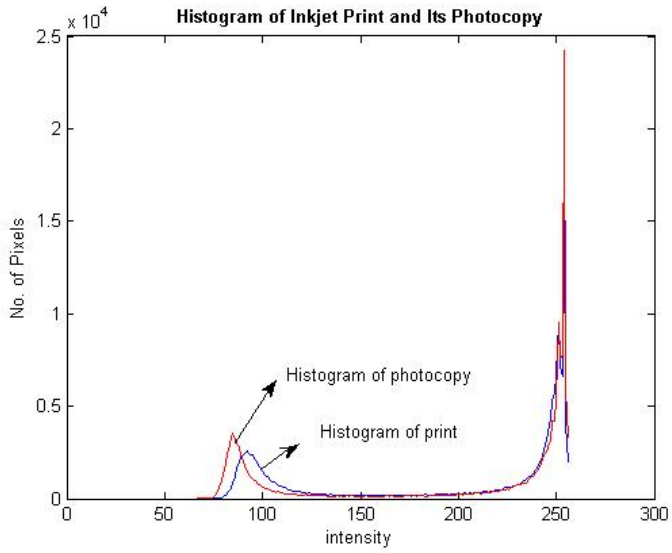


Figure 4.18: Histogram of inkjet(Hppsc1608)print and its photocopy

flat top near the mean rather than a sharp peak.

$$Peak = \beta_2 = \mu_4/\mu_2^2 \tag{4.11}$$

Hence, these features of histogram are used to identify disturbance that occur in a sample while photocopying. 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.

4.5.1 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.

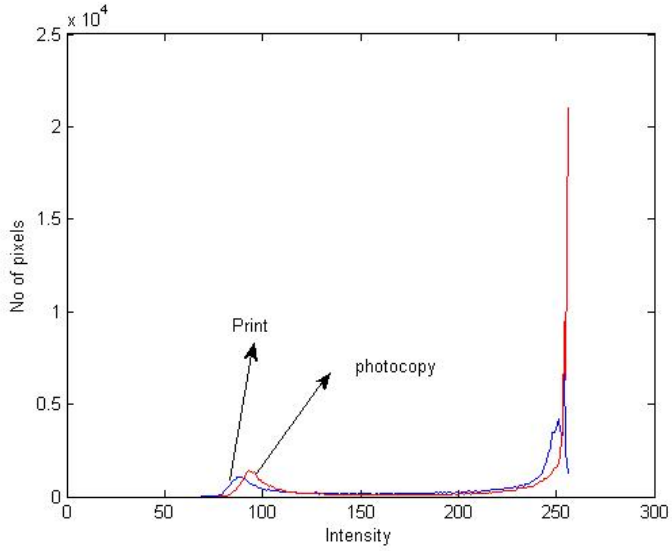


Figure 4.19: Histogram of Laser(Hplaser1200)print and its photocopy

4. Generate histogram for the foreground (printed text).

5. Calculate skew(beta1) and peakedness(beta2) of histogram as shown below.

$$Skew = \beta_1 = \mu_3^2/\mu_2^3 \tag{4.12}$$

$$Peak = \beta_2 = \mu_4/\mu_2^2 \tag{4.13}$$

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

4.5.2 Experimental results

Standard photocopy machines listed in Table 4.4 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 4.1 and are photocopied using XeroxWCM118 photocopier. Text documents having 'The' are printed on printers listed in Table 4.1 and photocopied on Konika Minolta photocopier. The skewness and peakedness of samples are calculated as mentioned in Equation 4.12 and Equation 4.13.

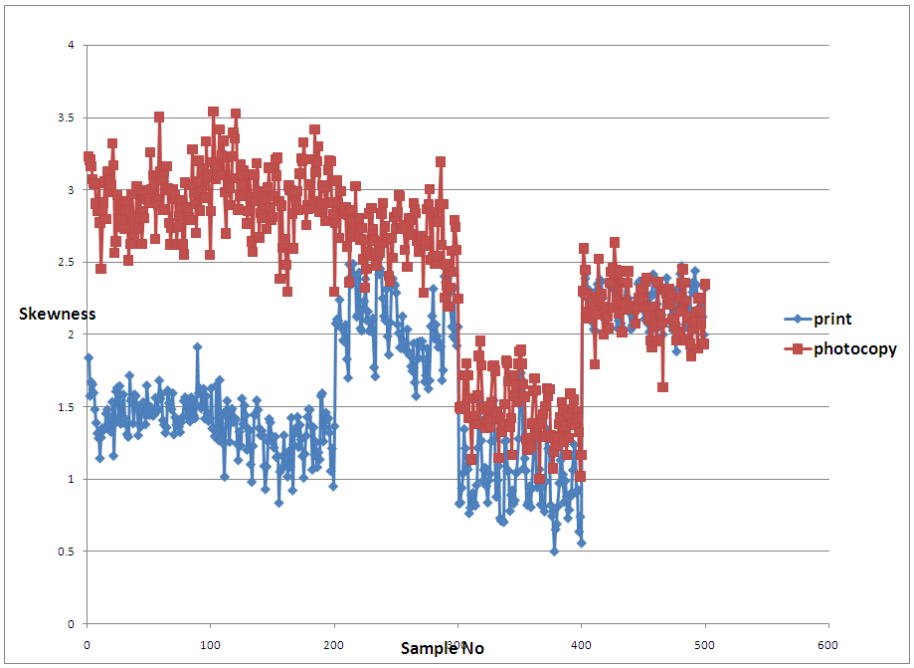


Figure 4.20: Skewness of print and its photocopy for text sample 'the'

The skewness and peakedness of the printed text sample 'the' and its photocopy are shown in Figure 4.20 and Figure 4.21. 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 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 white laser printers. Skewness and kurtosis for the black and white laser are difficult to differentiate as both black and white laser print-

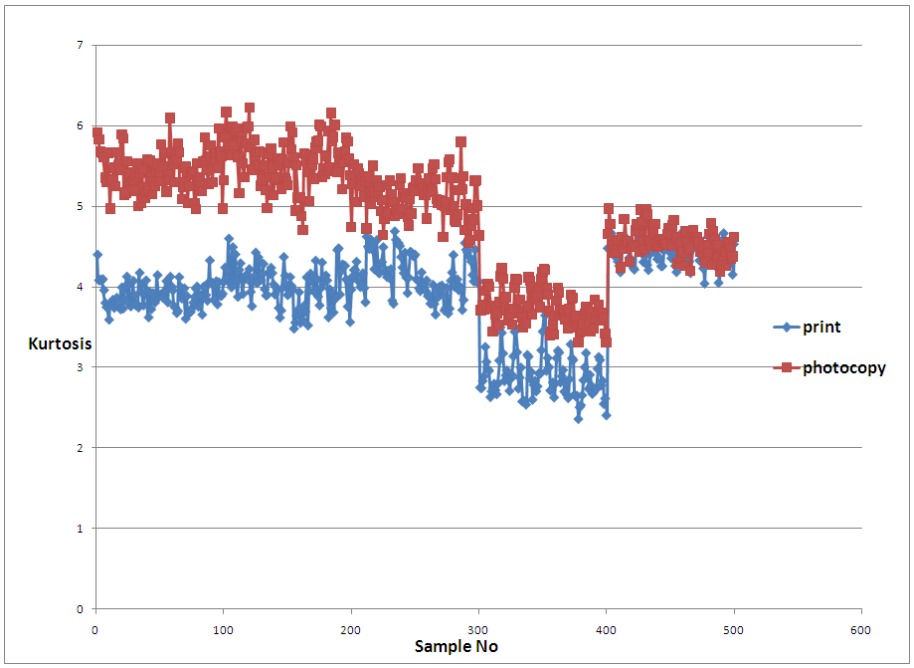


Figure 4.21: Kurtosis of print and its photocopy for text sample 'the'

ing mechanism and photocopy printing mechanism is based on electrostatic printing technique. General three letter word sample in Figure 4.12 are photocopied using Konika Minolta photocopier. Skew and kurtosis of print and photocopy of these sample are shown in Figure 4.24. The skew and kurtosis of sample 'the' are consistent with skew and kurtosis of the general three letter word.

4.6 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

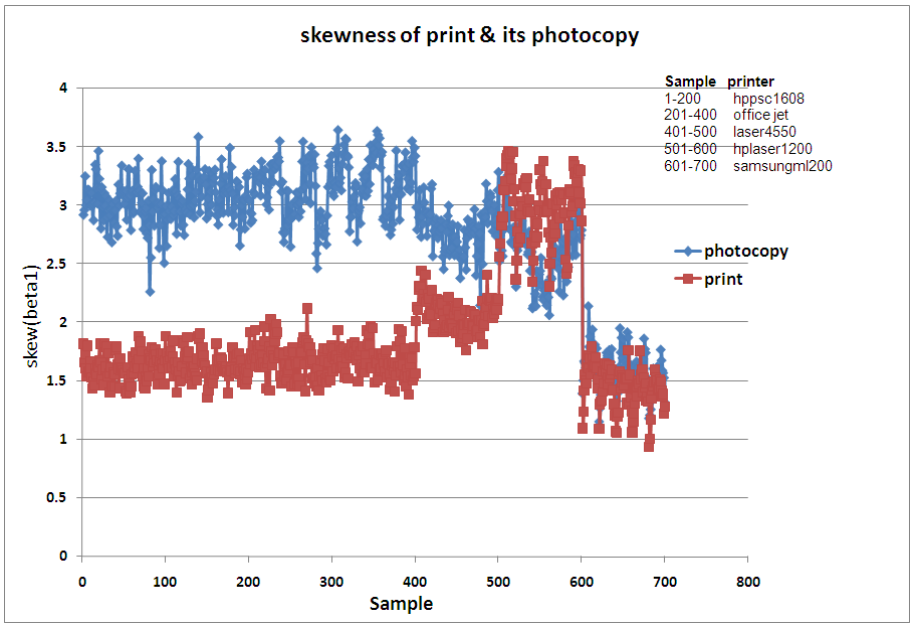


Figure 4.22: Skewness of print and its photocopy for text sample 'The'

3. Kurtosis

The flow chart shown in Figure 4.25 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.

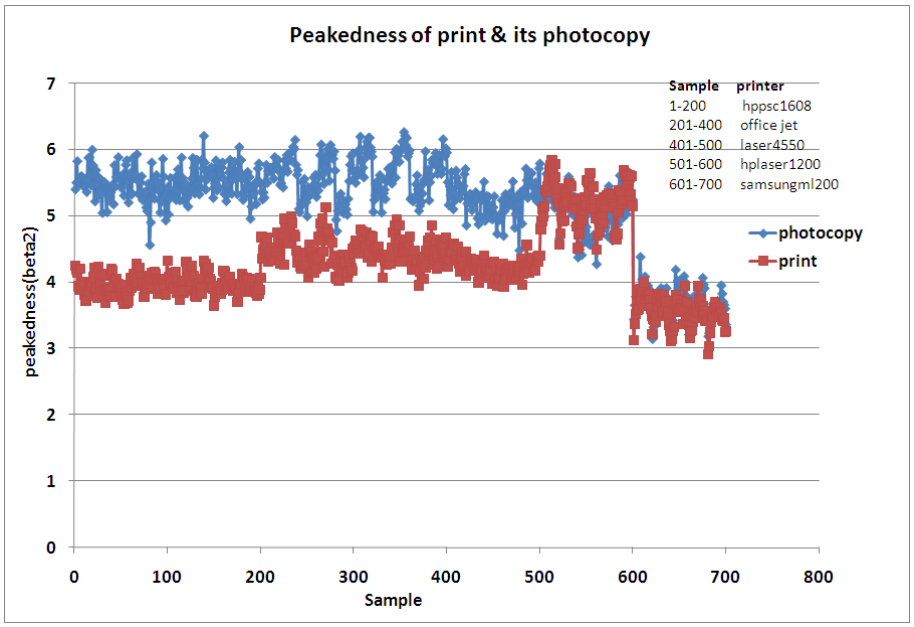


Figure 4.23: Kurtosis of print and its photocopy for text sample 'The'

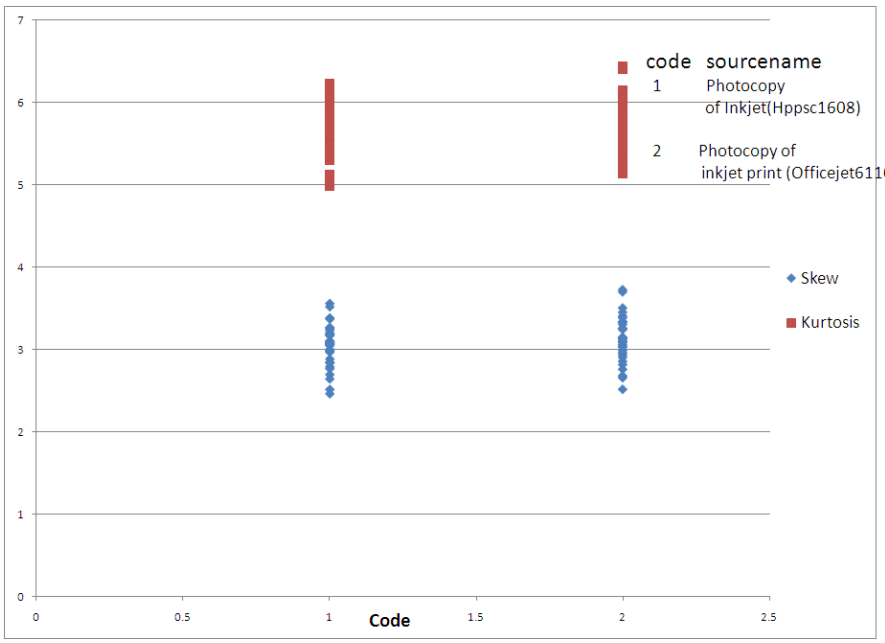


Figure 4.24: Skew and Kurtosis of general three letter word photocopy samples

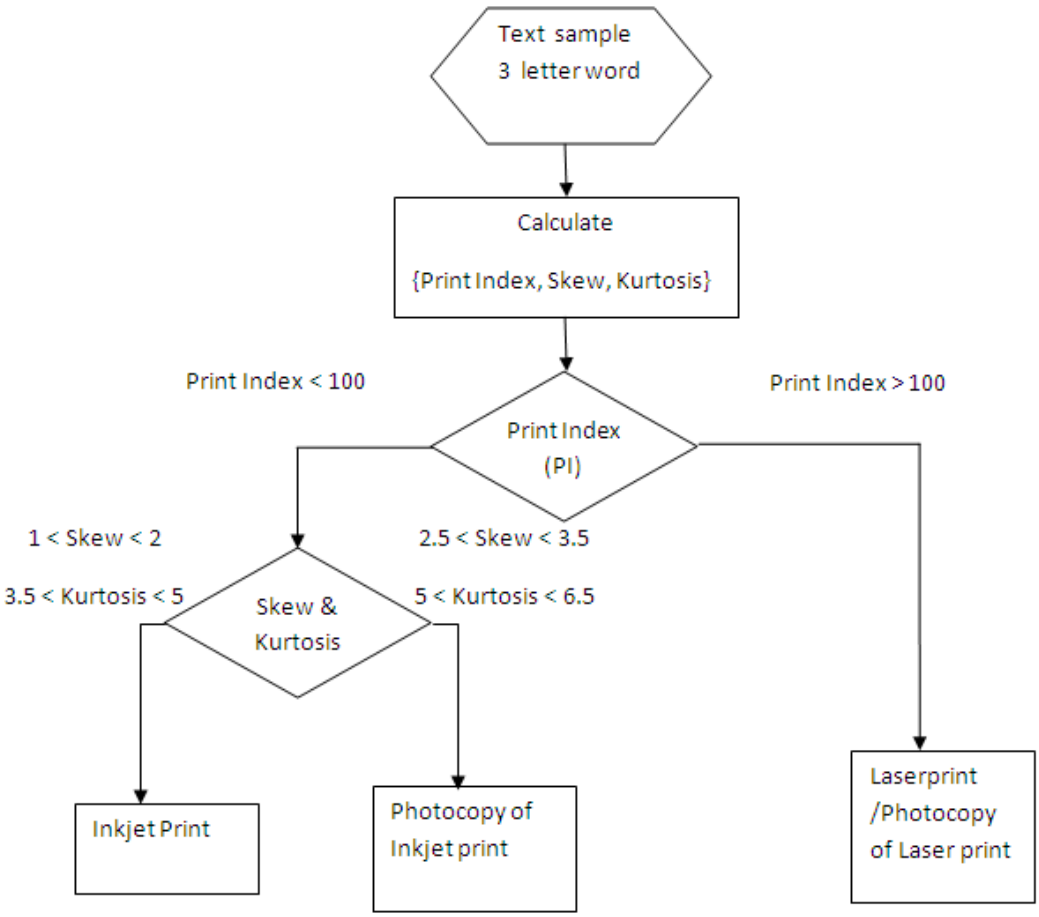


Figure 4.25: Printed document source identification

Chapter 5

Identification of Tampered Documents

5.1 Tampered documents

5.1.1 Introduction

A document which is altered by adding or deleting some parts of the document deliberately [92] is referred to as Tampered document. Most of the documents involved in malicious or illegal activities are altered form of original documents. These documents are generated by altering text content or picture content of the document. Alterations in text content of a document has been addressed in chapter 4, where we demonstrated text characterization method for checking consistency of text content in the document based on printing technology. In case of documents having both picture and text content, uniform colour regions of picture in the document are also to be studied for checking consistency. The pictures contained in the document may be generated by combining different parts, which are collected from different printed documents. It is the concern of a document examiner to analyse the document for identifying whether a document might have been

tampered with or not. The following section discusses ways the tampered documents are created from digitized print documents.

5.1.2 Types of tampered document

Tampered documents are classified into two types depending on the printing patterns or techniques contained in the document.

(a) Document having various print technology, that is some portions of the document are printed with one printing technique and the other portions of the document with some other print technique.

(b) Documents having mixing or combination of print technology that is reproducing the document by scanning original document and then reprinting on another printer. In this method the reproduced document contains the features of first print features mixed with following printing technique.

Due to lack of existing data base for experimentation, few synthetic samples are prepared and identification of tampered parts are demonstrated using these synthetic samples. This sample preparation for the experiment involves tampering a uniform colour region of an image. This involves replacing some part of an image by parts of the same image produced by another printer. Steps are:

(a) Same sample picture is printed using different printing mechanism with fixed resolution on two printers namely Deskjet930c and Hppsc1608 at 600dpi. These two samples are then scanned at 2400 dpi to get images as shown in Figure 5.1.

(b) Now uniform colour regions of these two images can be used for tampering. Figure 5.2 shows uniform colour regions of Deskjet930 image

and Hppsc1608 image. Uniform colour region of this Deskjet930c image is identified as regions 1, 2, 3 and 4. These regions are replaced with same spatial parts of uniform colour region of Hppsc1608 image. Region 1 is of size 10 by 10 and the other regions 2, 3 and 4 are 20 by 20.

(c) The resultant image is a Deskjet image containing some Hppsc parts scattered in uniform colour region, which is referred to as tampered image and is shown in Figure 5.3.

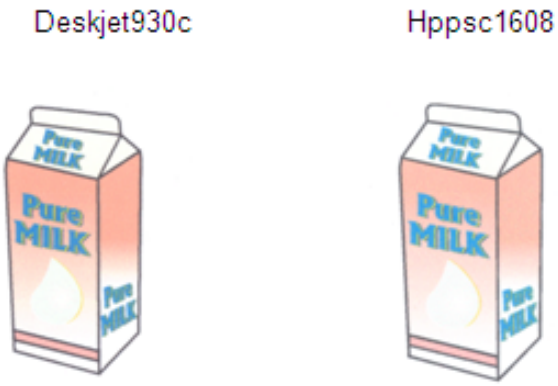


Figure 5.1: Sample images printed on Deskjet and Hppsc

5.2 Identification of tampered part of the document

Study of spatial statistics in uniform colour region of documents provides information for the identification of print technology. In our experiment, while tampering images in a document some regions of Deskjet930 image is replaced with Hppsc1608 image. These tampered documents looks close to the original and it is not possible to make out any difference with naked eye even though the spatial marking material differs. These patterns can be identified by a study of variogram generated for an image.

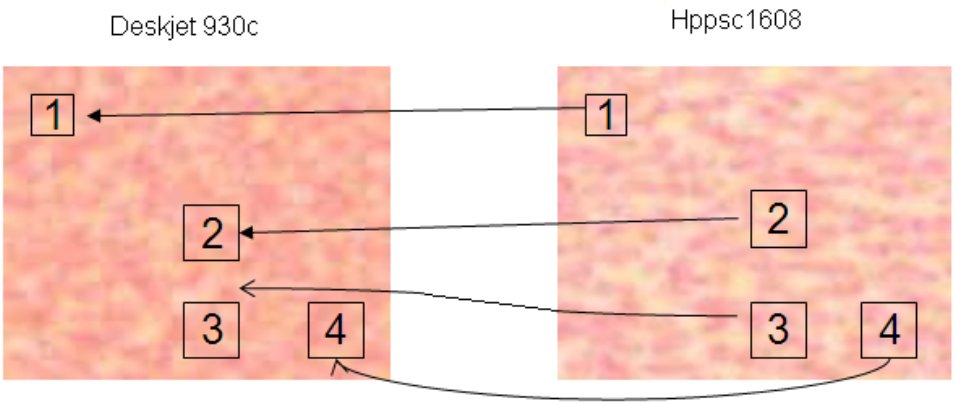


Figure 5.2: Uniform colour regions of images printed on Deskjet930c and Hppsc1608

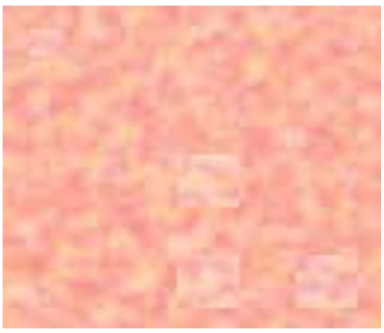


Figure 5.3: Tampered image of Deskjet930c with scattered Hppsc1608 parts

5.2.1 Window-wise analysis of variogram

Window wise analysis involves moving fixed size window over given image for generating variograms and comparing them to inspect any dissimilarities. For example, variograms generated for homogeneous regions of genuine(Deskjet930c) image and tampered image shows an increasing variance of tampered image and as shown in Figure 5.4. The increase in the variance of the tampered image has been observed but this does not reveal whether the image has been tampered or not.

If it is a tampered image then we need to identify the tampered regions. By generating variogram for fixed size non-overlapping windows of sample[93], it is possible to interpret the distinguishable features in the sample which do not belong to the original image. The proposed method of window wise analysis of variogram assists a document examiner in interpreting the altered parts of the document which contains pictures or images. The steps for identifying tampered regions in uniform colour region are explained in Algorithm 5.1.

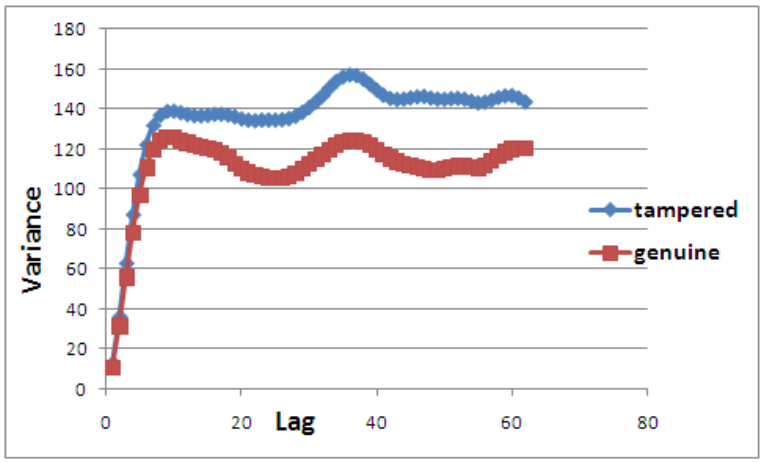


Figure 5.4: Variograms of Deskjet image and tampered image

Once the dissimilar window are identified, these are compared with corresponding windows of the genuine image if available to ensure that identified windows are tampered regions.

Window size

The range of variogram indicates spatial independence at lag equal to range. Thus window of size greater than or equal to the range is essential for capturing the dependence information. The variogram along x-axis of tampered image in Figure 5.5 shows the range is 36 i.e. at the lag/distance 36, the correlation graph reaches zero. Thus, window of size 40 is selected for window wise analysis of image. Tampered sample is the homogeneous/uniform

Algorithm 5.1 FINDTAMPERREGIONS(Sample, d, *scc_t*)

//Algorithm for identification of tampered part of the document//

Input:

Sample: image of size M by M

d: is direction specified in terms of angle

scc_t: is threshold and it is equal to 31

Output:

t: is list of tampered windows

Var: SZ, num, P, SCC

Method:

1: Call VARIOGRAM(sample,d) //It returns V, *sill* S, *Nugget* N and *Range* R//
2: $SZ \leftarrow Max(R,\ M/4)$ //'SZ' is window size and num is number of windows//
3: $num \leftarrow square(floor(M/SZ))$//Where M←size of the sample//
4: Label Sample as non-overlapping windows from window W_1 to W_{num}.
5: **for** $i \leftarrow 1$ *to num* **do**
6: Call VARIOGRAM(W_i, d) //It returns V_i, S_i, where V_i is variogram, S_i is *sill* for W_i//
7: $P_i \leftarrow 1/SZ^2 \sum_{j=1}^{SZ} \sum_{k=1}^{SZ} W_i(x_j, y_k)$ // P_i is picture value for W_i//
8: **end for**
9: $SCC \leftarrow 100 * \sigma(S_i)/Mean(S_i)$ //SCC is sill consistency coefficient, Where σ is standard deviation //
10: **if** $(SCC) > scc\_t$ **then**
11: **for** $i \leftarrow 1$ *to num* **do**
12: $tindex_i \leftarrow S_i/S$
13: **if** $tindex_i > 1$ **then**
14: W_i is tampered window: t(i)←1;
15: **end if**
16: **end for**
17: **else**
18: $t \leftarrow NULL$
19: **end if**
20: Return t

Computational Complexity of FINDTAMPERREGIONS	
step 1	$O(M^3)$
step 5-8 For image of size $M \times M$	$O(M^2)$
Algorithm Complexity: $O(M^3)$	

colour region of 127 by 127. For identifying tampered part in this sample, we adopt window wise analysis where the sample is divided into non overlapping windows of size 40 by 40 and labelled as W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_8 and W_9 which is shown in Figure 5.6.

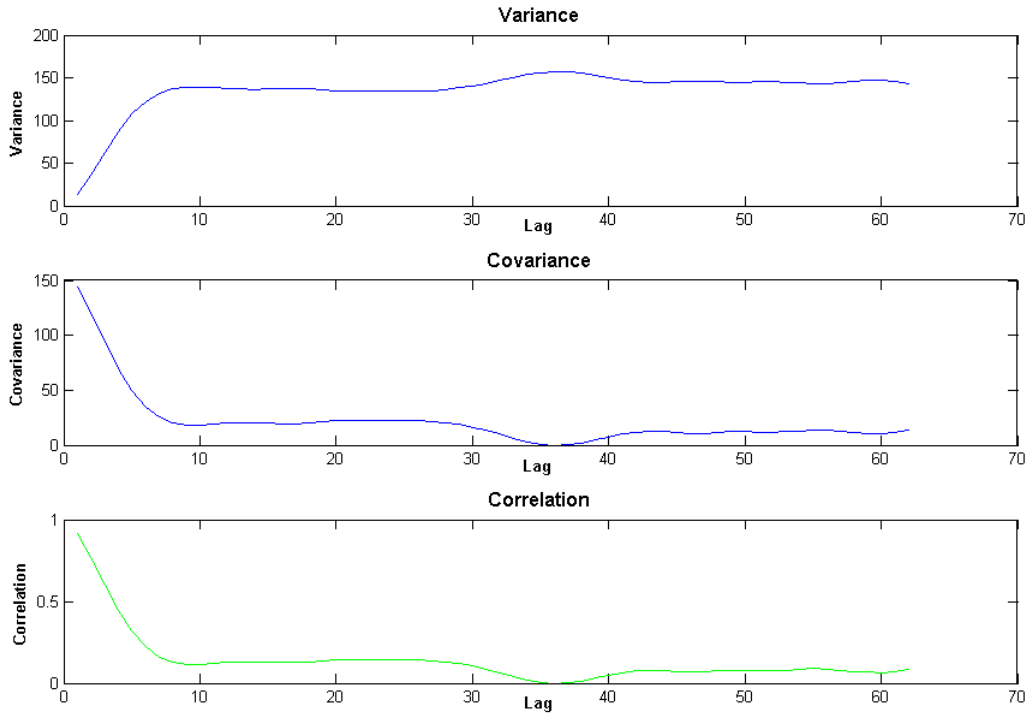


Figure 5.5: Range of tampered image

Window wise variogram and picture value

For each window, generate the variogram V along the selected direction. It can be directional variogram or variogram along x-axis. Window size is fixed based on the *range* of variogram along x-axis. The influence of direction in capturing underlying spatial statistics of sample is discussed in Chapter 3. Selection of direction is an important factor which plays a vital role in identification of tampered parts. Fixing size of window and direction along which variance has to be measured are parameters, which

Figure 5.6: Window of size 40 by 40 labelled on tampered image

influence the accuracy in identification of tampered parts. Compare the variogram V_1 to V_9 generated for windows W_1 to W_9 and shown in the form of graphical representation in Figure 5.7. Parameter of variogram, namely, *sill* for each window is taken as a feature for identification. The intensity of pixels within the window provides the feature to distinguish dissimilar parts in the window. Average intensity of all the pixels lying within the window is referred as picture value. Picture value for windows W_1 to W_9 are P_1 to P_9 respectively.

Plot the *sill* of the variogram in each window of image. S_1, S_5, S_8 and S_9 are *sills* of V_1, V_5, V_8 and V_9. These are slightly higher than the remaining *sills* as observed in Figure 5.7. Figure 5.8 shows comparison of the *sills* of tampered image in each window. Observing Figure 5.9, it is identified that the windows W_1, W_5, W_8 and W_9 have distinguishable picture values compared to the other neighbour windows.

sill Consistency Coefficient(SCC)

The coefficient of variation indicates consistency/stability of the data and it is free from units. It is thought that this coefficient of variation can be aptly adopted for tampered document classification. Hence coefficient of

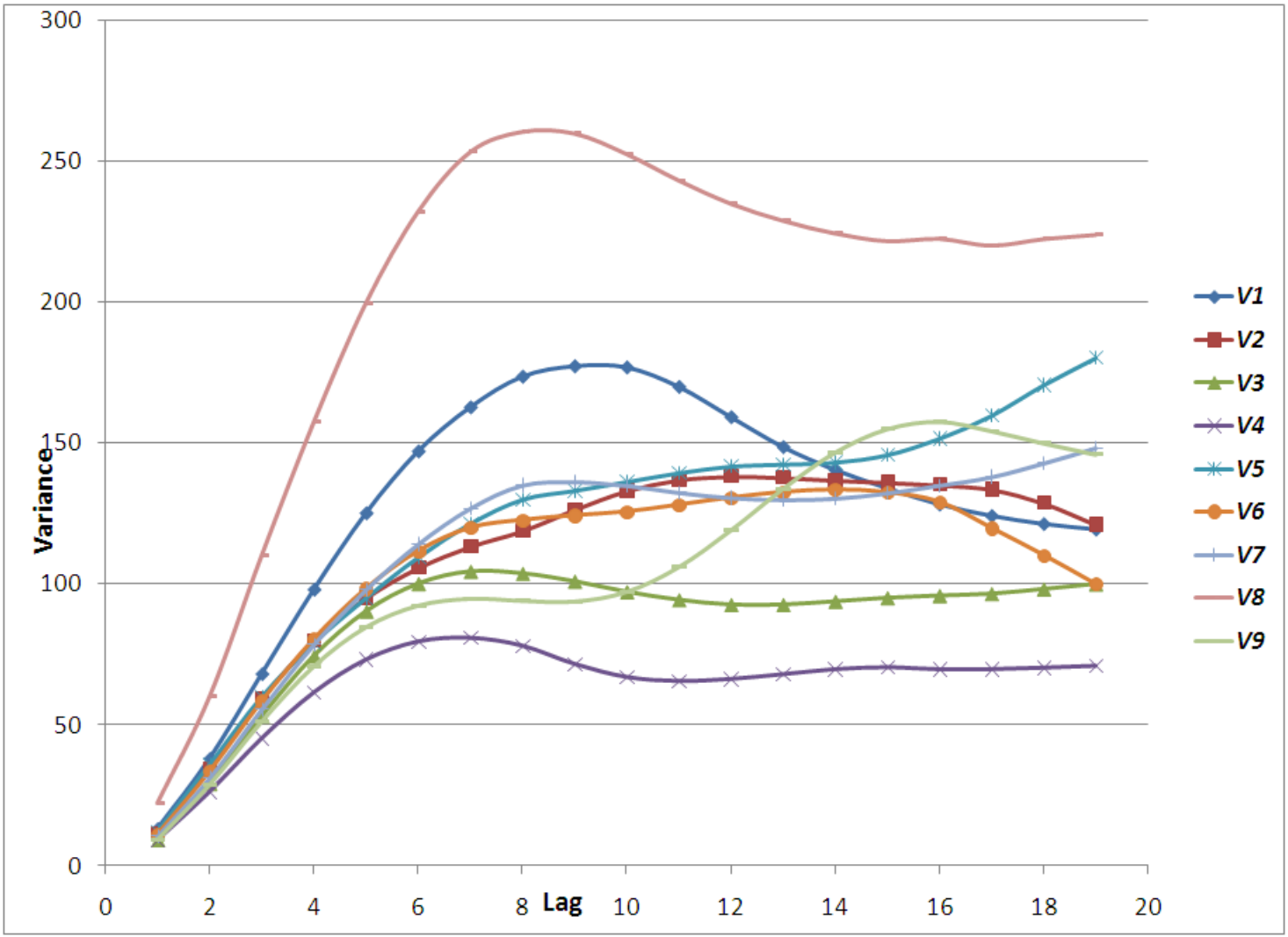


Figure 5.7: Window wise variograms of tampered image

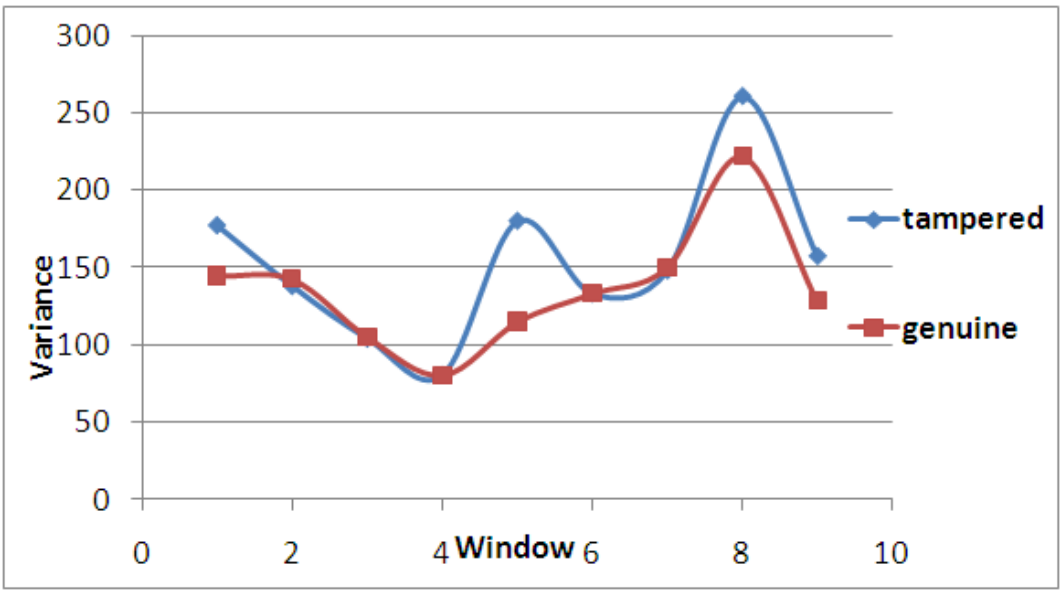


Figure 5.8: Plot of window versus *sill*

variation for the window wise sill values is referred to as Sill Consistency Coefficient(SCC). If the questioned document is identified as tampered with the help of Sill Consistency Coefficient, then one need to identify proper region of tampered data. For the tampered image in Figure 5.3, the sill consistency coefficient is greater than 31. Then it is considered that given image is a tampered image. Thus, one need to compute tamper index t_index to identify the tampered windows. The windows having t_index greater than 1 is considered as tampered. Observing Figure 5.10, the windows W_1, W_5, W_8 and W_9 are showing the tamper index value greater than 1. For the corresponding genuine image(Deskjet930c), the Sill Consistency Coefficient turned out to be less than 31. To ensure that the identified windows are tampered, compare these with the corresponding *sill* and picture values of the genuine image.

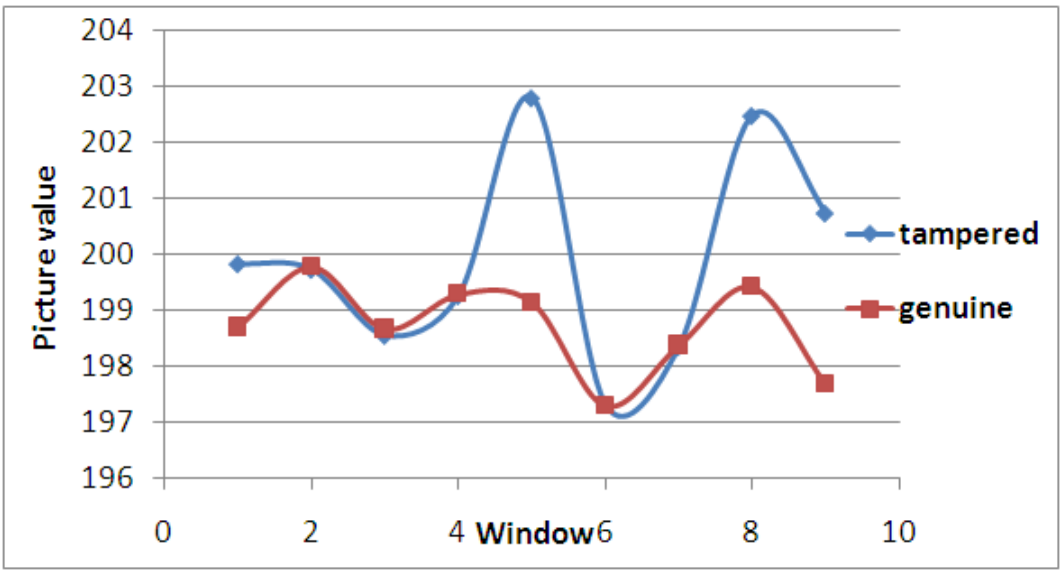


Figure 5.9: Plot of window Vs picture Value

5.3 Identifying mixing print technologies

Often fake documents are created by scanning original documents and reproducing them using printing technology. Sometimes more than one print technology is used in the process of forging the original document. When a printed document is scanned and reproduced using some other printing technique, the reproduced document inherently contains the spatial statistics of both printing techniques. The spatial characteristics of homogeneous/uniform colour region of reproduced document reveals the features of original printed document and reproduced printer characteristics. Hence, fixed size window variogram generated for such tampered documents should not contribute to specific print technology as it inherently contains both features of print technology. While scanning the original document, the original print features are captured by scanners and are reproduced with the print features while taking the print. Variogram of different window size can capture global and local patterns of the document. Adopting variogram of varying size of window will contribute to identifying hierarchical printing techniques contained in the document.

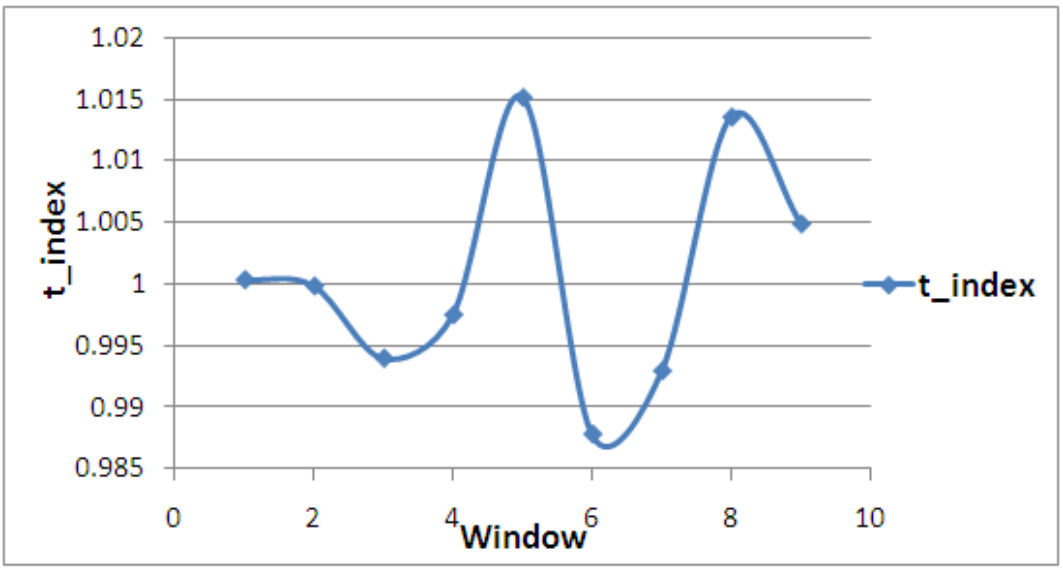


Figure 5.10: Tamper index for windows of tampered image

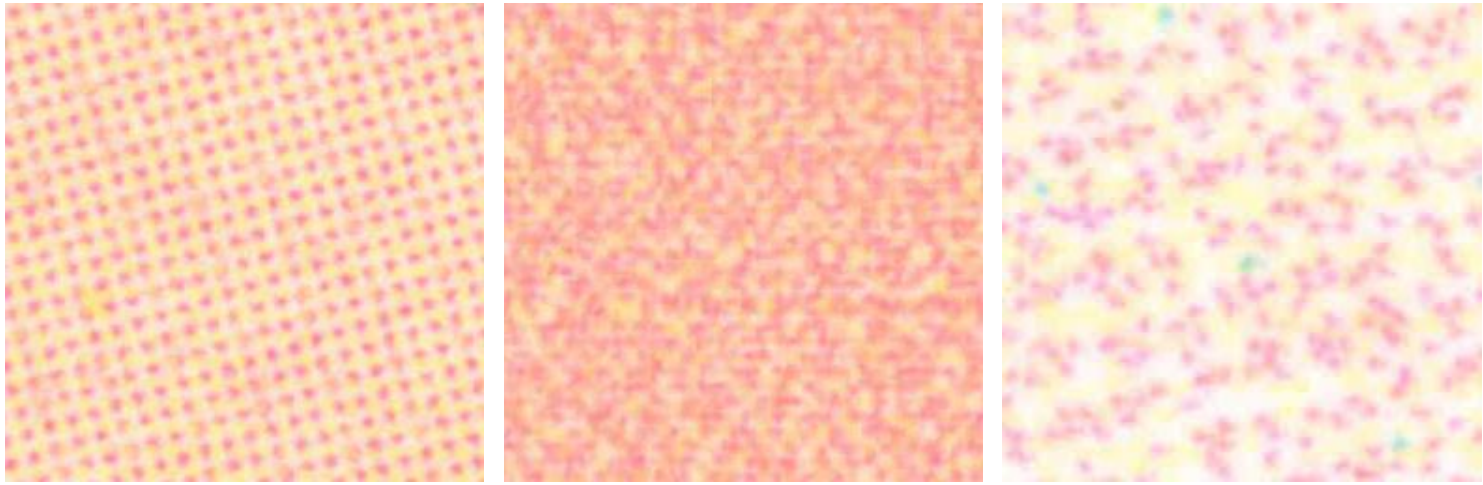
5.3.1 Variogram of mixed print document

In our experiment, uniform colour region of image produced by using laser printer is taken as source document. It is scanned at 2400 dpi and printed on inkjet printer Deskjet840c at 600dpi. Test sample is produced using LaserJet4650 printed image scanned and again printed on Deskjet840c and it is shown in Figure 5.11(c). Variogram generated for the test sample differs depending on the size of the window. If size of window is small, variogram resemble the local pattern, if window size is large variogram resembles the global pattern.

5.3.2 Analysis of window-wise variogram to identify the mixed print technology

The steps for window wise analysis of mixed print sample are explained in Algorithm 5.2. It generates variogram for sample of various window sizes.

Adopting analysis of variogram with varying window size on sample of mixed print such as laser image printed on Deskjet shows the pattern of laser and



(a) Laser4650 sample (b) Deskjet840c sample (c) Mixed print image

Figure 5.11: Sample image of mixed print technology

Algorithm 5.2 FINDMIXEDPRINT(Sample, d)

Input:

Sample: image of size M by M

d: direction specified in terms of angle and d=0 for x-axis

Output:

V_i: variogram for window W_i

S_i: sill for window W_i

N_i: nugget for window W_i

R_i: range for window W_i

Method:

1: W1←image(x,y) for $1 \leq$ x $\leq$ M/4, $1 \leq$ y $\leq$ M/4

2: W2←image(x,y) for $1 \leq$ x $\leq$ M/2, $1 \leq$ y $\leq$ M/2

3: W3←image(x,y) for $1 \leq$ x $\leq$ M, $1 \leq$ y $\leq$ M

4: **for** $i = 1$ *to* 3 **do**

5: Call variogram(W_i,d)

6: **end for**//It returns variogram V_i, sill S_i, nugget N_i, Range R_i for the window of W_i in the sample//

7: Plot variogram of each window

8: Return V_i, S_i, N_i, $Range_i$

Computational Complexity of FINDMIXEDPRINT	
step 4-6	$O(M^3)$
Algorithm Complexity: $O(M^3)$	

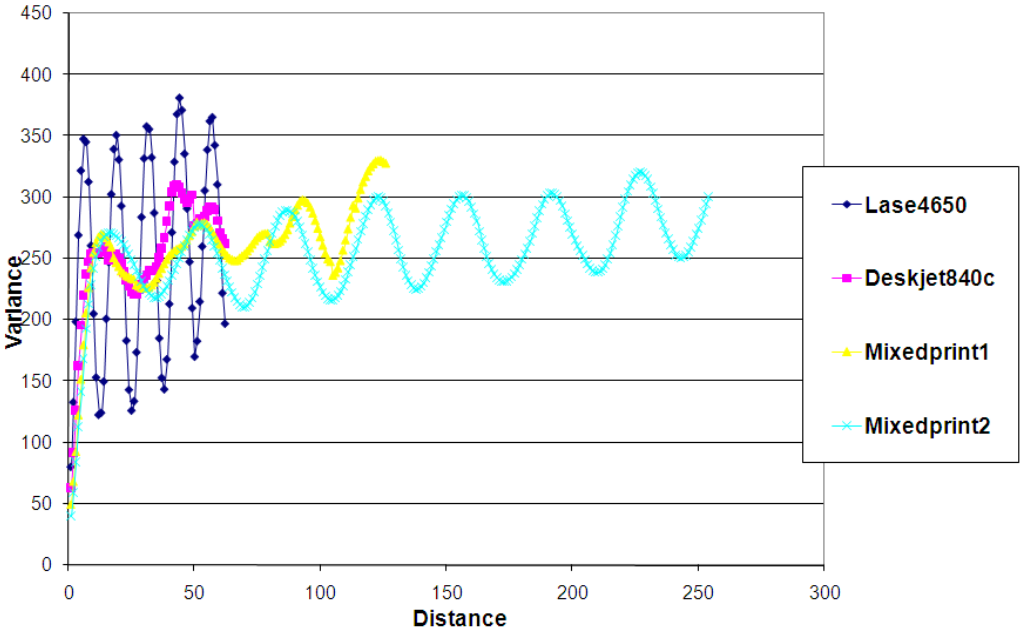


Figure 5.12: Directional variogram 20° degree angle of window size 128 and 256

inkjet print technologies. The variogram results for the mixed print sample (laser image printed on Deskjet) for window size 256 and 512 are demonstrated here. Window size of 256 and for the size 512 are labelled as mixed print 1 and mixed print 2.

The Figure 5.12, 5.13 indicates that the directional variogram varies with window size. In both the Figures variogram corresponds to smaller window size(256) revealed the Deskjet characteristics(local), where as variogram corresponds to large window size(512) exhibits laser printer characteristics(global).

5.3.3 Results

Inkjet and laser combination of printers are selected for preparation of samples to analyse performance of window wise variogram in identifying mixed printing technology. Few samples are prepared using inkjet printers like

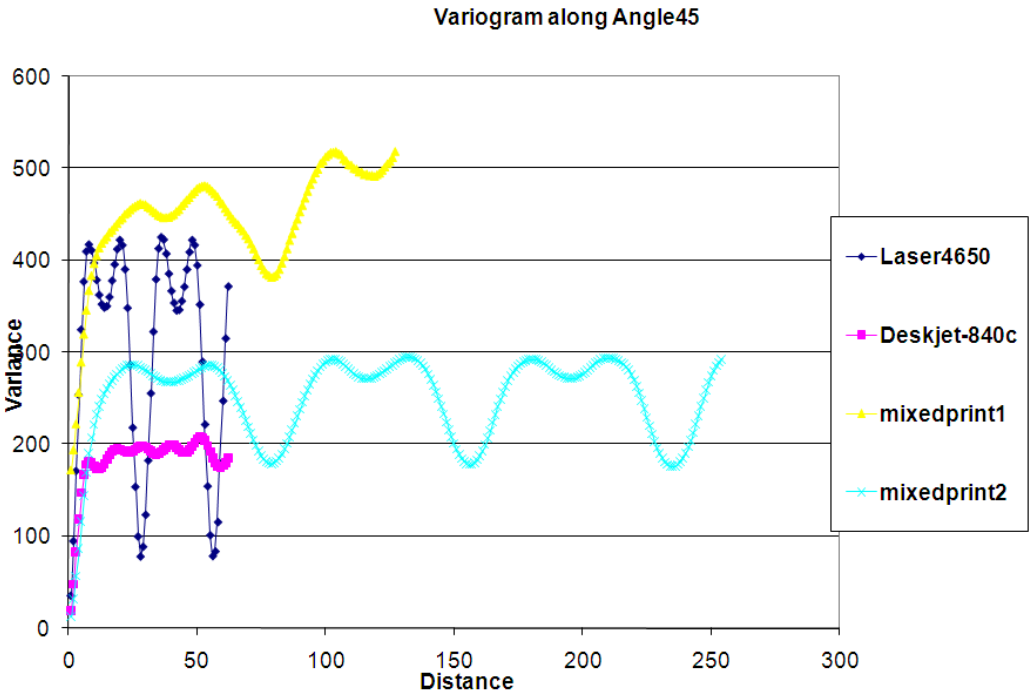


Figure 5.13: Direction variogram 45° angle of window size 128 and 256

Hpofficejet6110, Hpphotosmart3188 printer. These inkjet print images are again printed on laser printers. For example, Hpofficejet print is scanned and printed on Hplaser4550N, Hpofficejet6110 print is scanned and printed on SamsungCLP-510 laser printer. Similarly, Hpphotosmart3188 print is scanned and printed on Hplaser4550N, Hpphotosmart print is scanned and printed on SamsungCLP-510 printer. These samples are all inkjet prints again printed on laser printers. For the above four combinations, samples are generated and the uniform colour region of these samples are analysed employing window wise variogram analysis. Variograms of these gray converted samples for varying window size reveals mixed print combination.

The samples and corresponding variogram (along x-axis) of window size 256, 512, 1024 are labelled with its source printing combinations and shown in Figure 5.14, Figure 5.15, Figure 5.16 and Figure 5.17. In all of the above combinations, local pattern shows periodicity resembling laser print pat-

tern and global pattern shows irregularites in overall trend resembling the inkjetprint. Hence, variogram with varying window sizes uncovers the mixed printing combination in uniform colour regions of an image.

5.4 Observations

The observation from window wise variogram analysis are:

(a) Uniform/fixed window size has potential to classify documents as tampered or not, and to also identify tampered regions. However this works only for uniform regions. One need to develop models or methodology for expressing the variability when non-uniform region is affected due to tampering.

(b) Varying window size analysis has potential to describe hierarchical printing techniques used in producing the questioned document and it also captures the variations in periodic characteristics of printing techniques involved in producing the document. This uniform based method is sufficient in determining hierarchical printing technique identification.

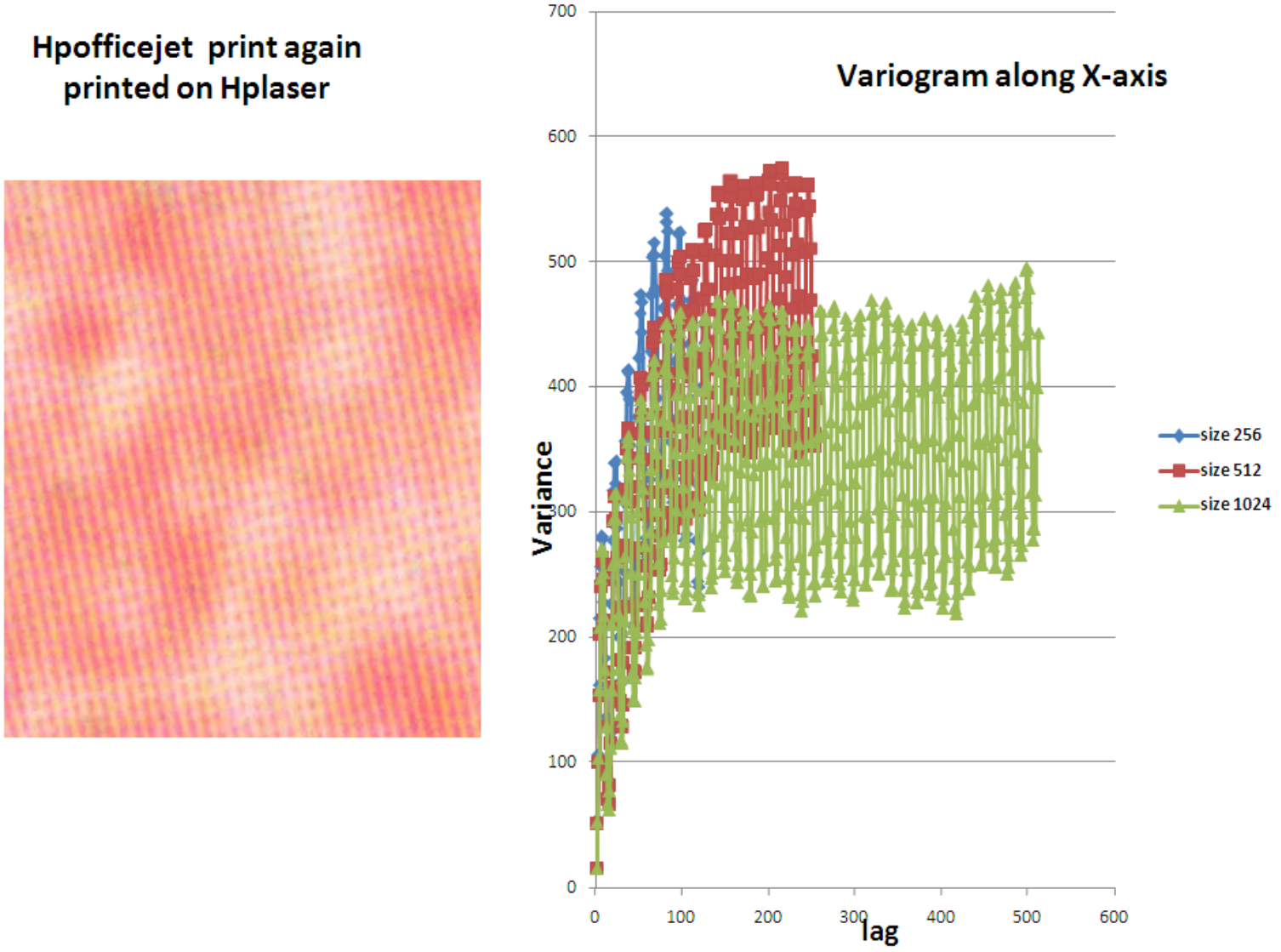


Figure 5.14: Officejet image printed on laser printer and its variogram of size 256, 512 and 1024

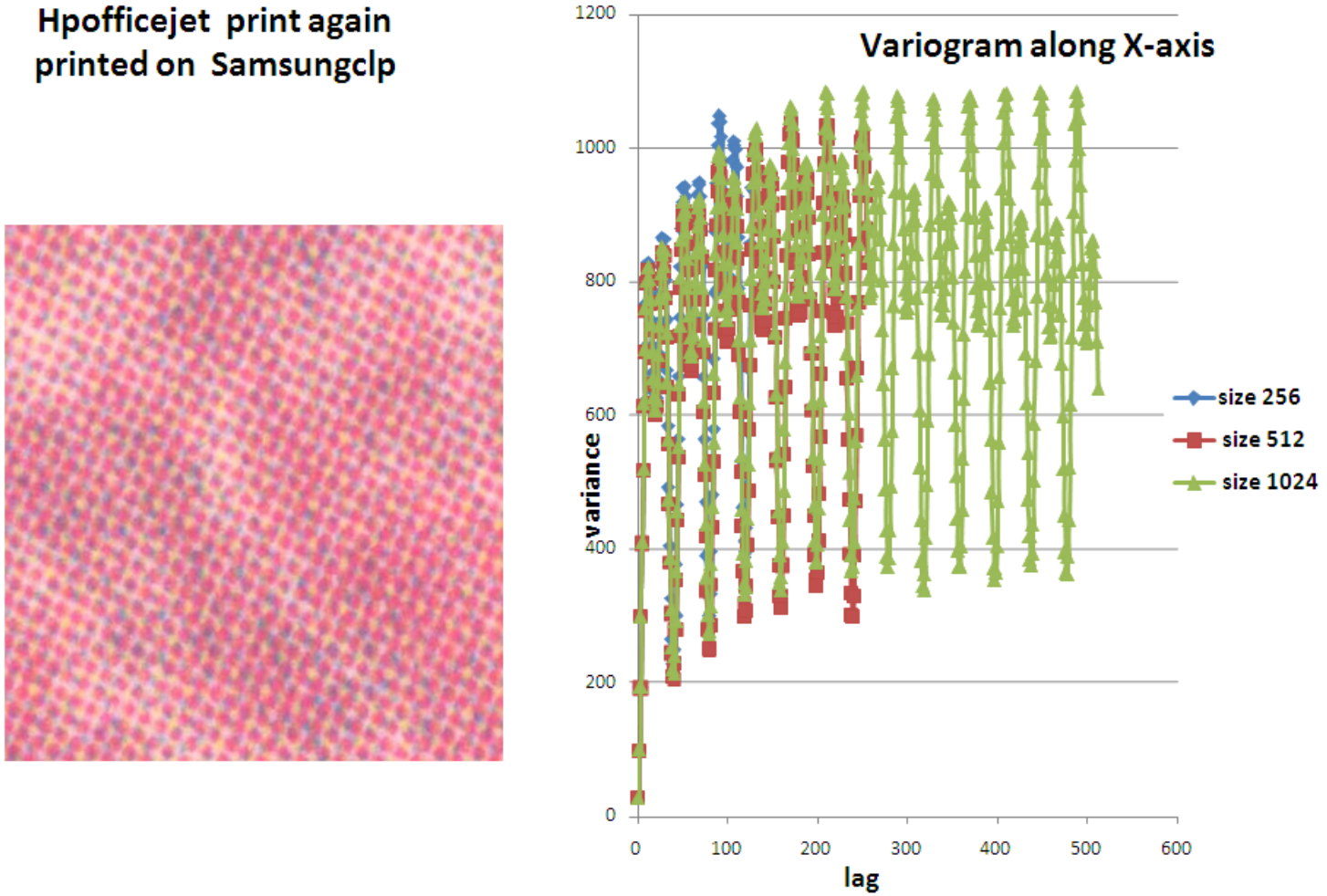


Figure 5.15: Officejet image printed on samsung printer and its variogram of size 256, 512 and 1024

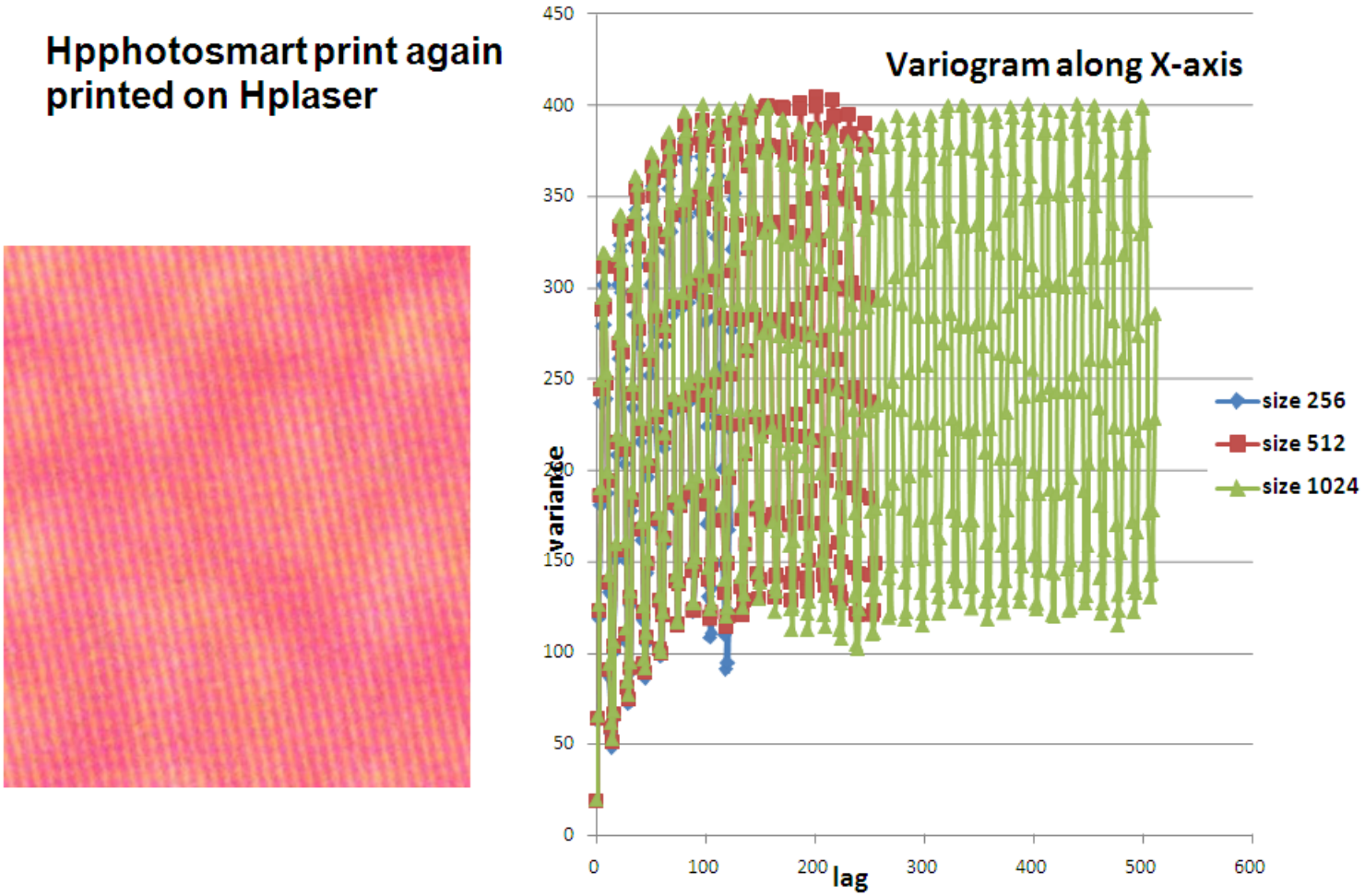


Figure 5.16: Photosmart image printed on Hplaser printer and its variogram of size 256, 512 and 1024

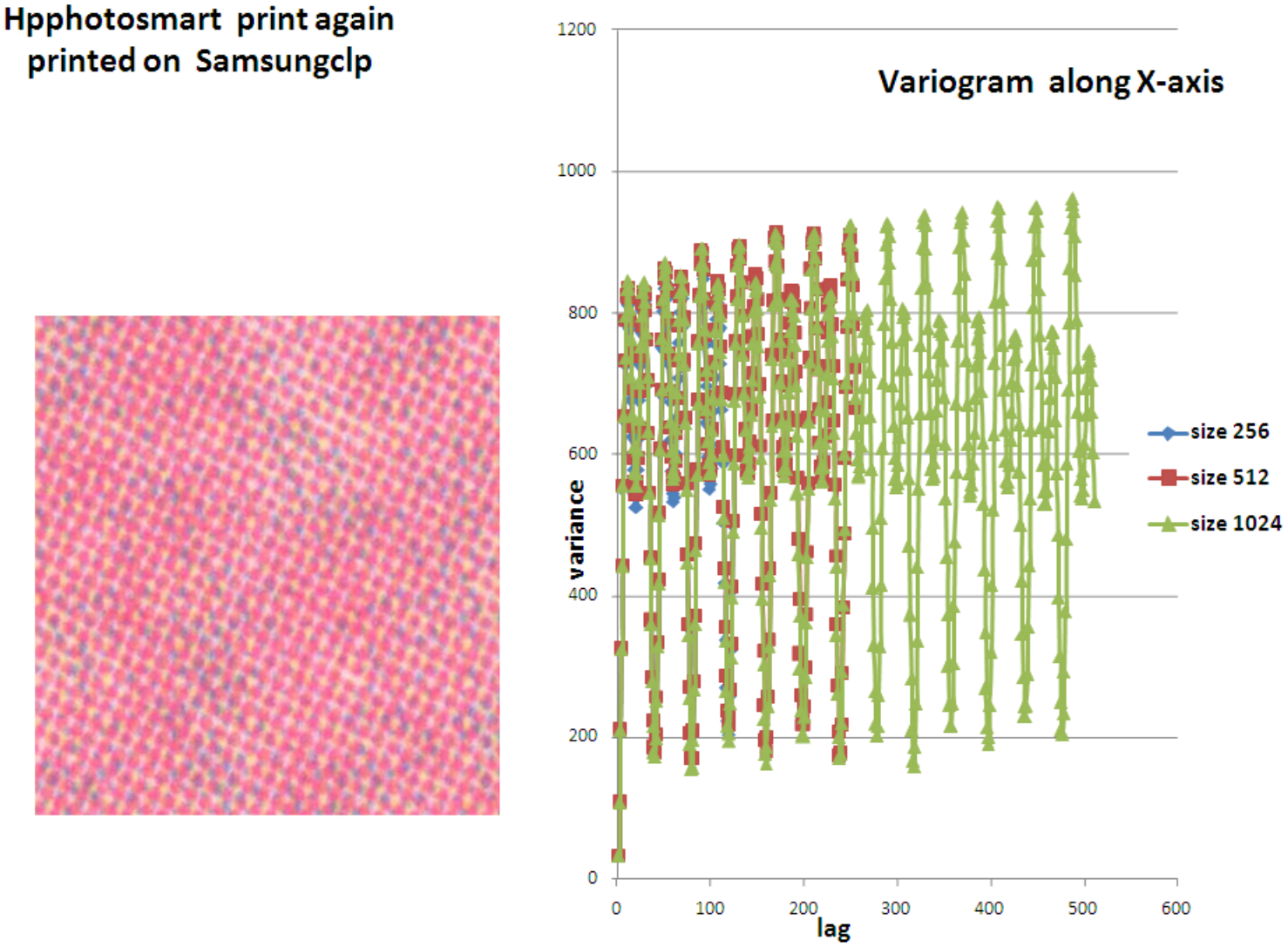


Figure 5.17: Photosmart image printed on Samsung printer and its variogram of size 256, 512 and 1024

Chapter 6

Conclusion and Future Directions

6.1 Conclusion

This thesis presents non-invasive techniques for identification of printing technology in document forensics perspective. It is a contribution to the field of forensic science in identification of source of the document by studying spatial statistics of region/text in the document. The printed document is a heterogeneous combination of pictures, text and various symbols. Region based approaches are based on uniform colour region of pictures in the printed document. As it is difficult to find such region in the printed text, techniques are discussed for identification of print technology based on printed text. The major contribution of this thesis are:

(a) Gaussian Variogram Model for identification of print technology GVMPT, is focusses on dissimilarity measure in homogeneous colour regions of an image. The dissimilarity measure is captured by employing geostatistical tool variogram and abstraction of this can be achieved by producing Gaussian functional forms for the variogram which is named

as Gaussian Variogram Model(GVM). Influence of algorithmic parameters in GVMPT are also demonstrated. Roughset based classification is adopted for identification of predominant attributes and classification of printing technology.

(b) The text based approach for printing technology identification focuses on most frequently used word like 'the' as test sample for characterizing printed text. The novelty of the proposed algorithm is that the selected printed text is modelled as a mixture of three Gaussian models namely text, noise and background. The associated patterns and features of the models are derived using Expectation Maximization(EM) algorithm and few indices are proposed based on these parameters. One of the indices called Print Index(PI) for text is used for basic print technology discrimination. The same is demonstrated on general three letter words like 'and', 'but' etc.,

(c) Histogram based features, skewness and kurtosis, are employed for differentiating printed document from photocopied document. It captures the disturbance that occurs while photocopying printed document. It differentiates inkjet print from its photocopy.

(d) Window wise analysis of variogram based on sliding window protocols is employed for identifying tampered document regions. Variogram with varying window sizes is adopted for identification of combination of print technology and to answer the question: is the document reproduced by scanning the printed document? i.e., inkjet printout scanned on printed on laser printer and vise versa.

Part of this work has been published in the following international conference and journal.

Papers Published

1. **'Gaussian Variogram Model for Printing Technology Identification'**, International Conference on Asian Modeling Symposium, IEEE Computer Society, pp 320-325, 2009.
2. **'A Survey of Image Processing Techniques for Identification of Printing Technology in Document Forensic Perspective'** International Journal of Computer Applications, Special Issue on RTIPPR(1), pp 9-15, 2010.

6.2 Future directions

The work proposed in this thesis is successful for classification of various inkjet and laser jet printers based on homogeneous regions and text. It also differentiate photocopy from inkjet print and identifies tampered parts of the document.

In case of GVMPT or Directional GVMPT we have till now considered single direction is considered for feature set selection and the same is used for classification. This work can be further extended by employing combination of directions which are good enough to model spatial statistics of a pattern. Input image selected for identification is converted to gray level or preprocessed before converting to gray level. Further, it can be extended for colour image identification by adopting multivariate variogram for identification directly rather than converting it to gray scale.

The printed text characterization is based on the Times New Roman font with font size 12pt. This can be experimented for various font type and various sizes. Print Index measure is demonstrated for selected words like 'the' and other three letter word like 'and' and 'but'. Robustness of the

printed text characterization for varying font types, size and for single letter, two, three or n letter words can be tested. Differentiation of laser print out from its photocopy is to be addressed as the laser and photocopy both use electro static printing technique.

Uniform/fixed window size has potential to classify a document as tampered or not. However this works only for uniform regions. One needs to develop models for expressing the variability when non-uniform region is expected to be tampered. Varying window size analysis of variogram for identification of tampered document is based on a specific assumption that if the questioned document is output of mixed print technologies. Hence, it was experimented on synthetic sample with high resolution image (2400 dpi) as there is lack of existing data base of tampered document from forensic point of view. This work can further be extended to real world data with normal resolution.

The work implemented in this thesis is based on documents printed at 600 dpi and scanned at resolution 2400 dpi. The applicability of the proposed methodologies should be tested for the documents digitized at lower resolutions.

Appendix A

Bench mark data set

A.1 Data Preparation

Developing tools for identification of printing techniques requires standard bench mark data sets. There is no standard bench mark database for forensic purpose, in particular for identification of printer characteristics. The present study realises the importance of standard data sets for developing meaningful tools. Hence it aimed at developing schematic methods for compiling datasets.

The availability of identified printers for study purpose even during our research happened to be sparsed due to failures of printers, non availability of spare parts, cost of maintenance, cost effective technological enhancements, growth in quality as well as quantity over generations of the printers.

A printed document in general is a layout of a combination of one or more kinds of content. The content may be text, may be picture; it can be colour or gray. Hence, to develop methodologies for general purpose, one needs to develop standard datasets with text and pictures. The present study developed synthetic text and commonly used images from the web and a typical segment of the documents(questioned) are used for building the data

sets. These documents are printed at 600 dpi on the identified printers and scanned at 2400dpi resolution. These scanned image are treated as benchmark.

Synthetic tampered documents: A original document is scanned and some portions are edited and printed on may be other printer.

Image analysis methods are applied to the sample for developing methodologies to address problems in this thesis. These samples are used to build the models and demonstrate the methodologies developed in the thesis.

(a) Homogeneous samples of size 127 by 127 pixels are used for generating GVMdata. Each sample with its GVM data and labelled with printer id forms training data set.

(b) Three types of text documents are considered for printer identification based on text.

   i. First type of document contains most frequently occurring word 'the', these samples are referred to as Type1 samples.

   ii. Second type of document contains the word 'The', which are referred to as Type 2 samples.

   iii. Third type documents are general text document having some small case three letter words. Any three letter words are referred to as Type 3 samples.

   All these three types of documents having text of Times New Roman font with 12 pt size are printed on various inkjet and laser printers at 600 dpi. The text document is scanned at 2400 dpi to produce image of printed text. Applying minimum bounded rectangle to this scanned image of text and resizing each sample to fixed size forms the database of printed text for training and testing. Type 1, Type 3 sample are resized to [300,400] and Type 2 samples are resized to [300,600]. Resized text samples are used as bench mark data set for feature selection.

(c) The three types of text sample Type 1, Type 2, Type 3 are photocopied and scanned at 2400 dpi.

(d) Few images printed on various printers and the printouts are scanned.

   i. Mixing of scanned images from different printer of a image are used for producing tampered image.

   ii. Scanned image of print is printed on a different printer for producing document with hierarchical mixing print technology.

All the above type of documents are printed using various printers at 600 dpi and scanned at 2400 dpi using HP Scanjet for creating source files for analysis. The respective data sets are divided into two sets, first set is used for building the models(that is training set) and the second for validating the methodology (that is testing set). While building the data sets, various kinds of printers (identified) are commonly used for document preparation and different kinds of documents are also selected for producing standard reference documents for building models and methodologies.

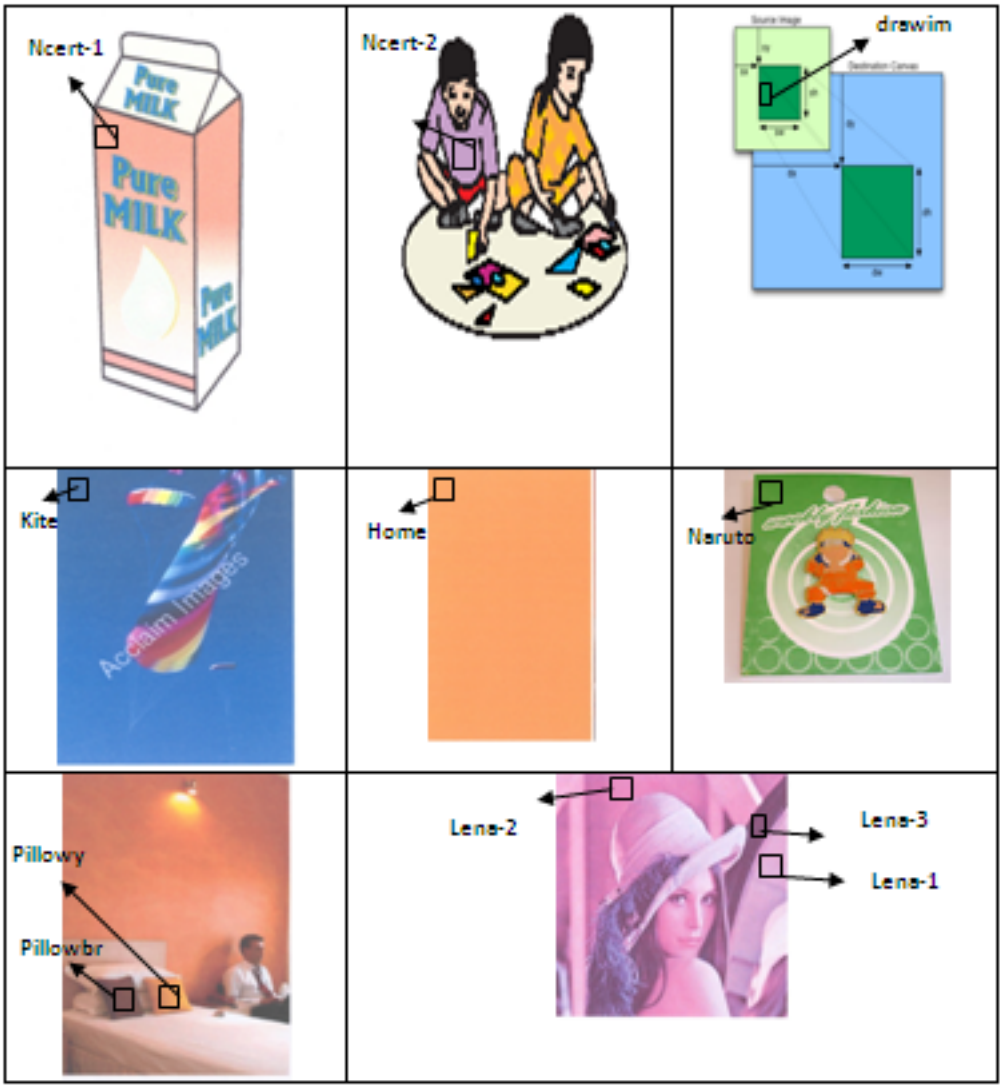


Figure A.1: Selection of samples from various images

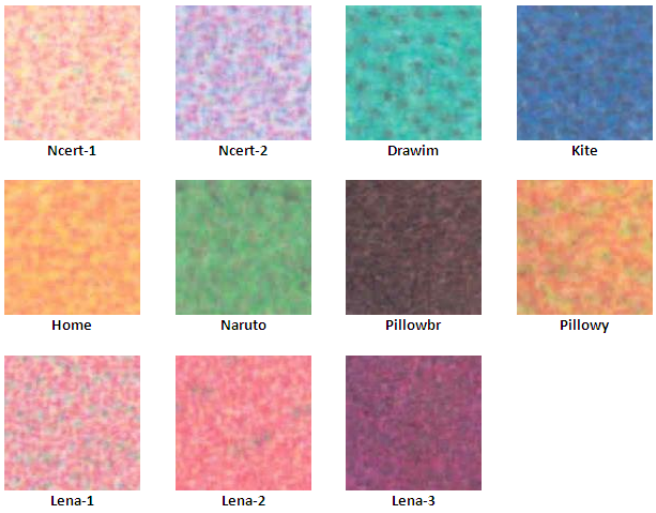


Figure A.2: Samples of printer Photosmart3188 with pid-1

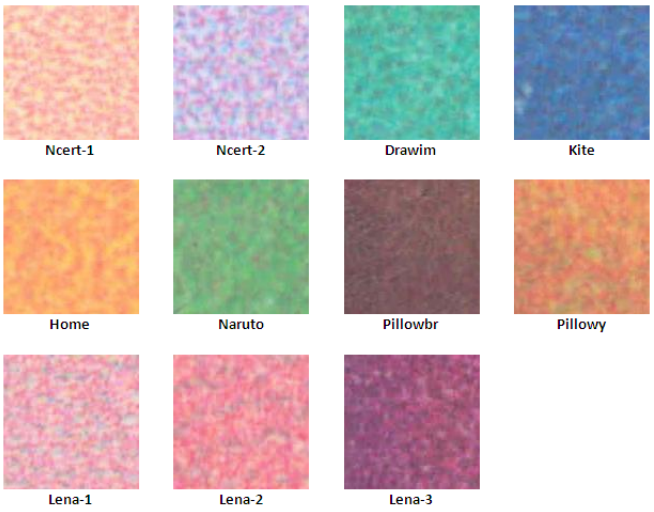


Figure A.3: Samples of printer hppsc1608 with pid-2

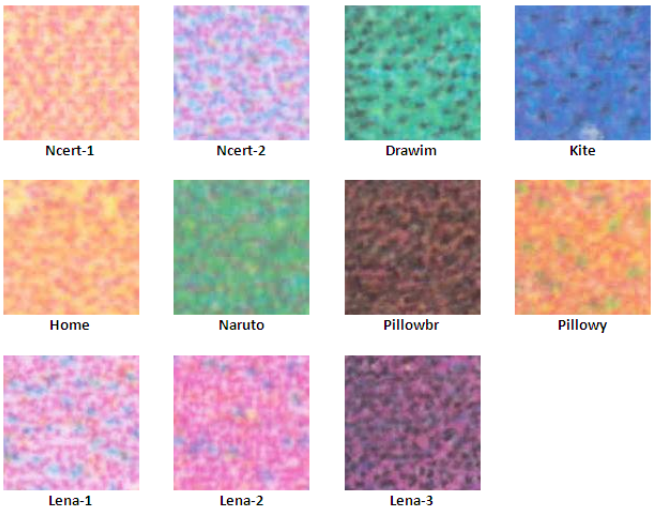


Figure A.4: Samples of printer deskjet840c with pid-3

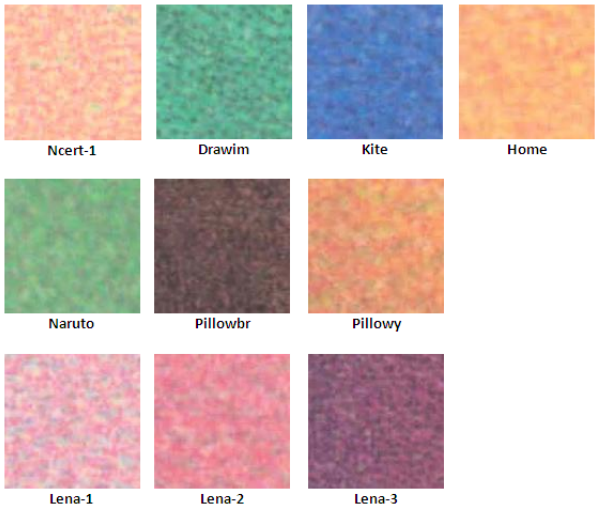


Figure A.5: Samples of printer officejet6110 with pid-4

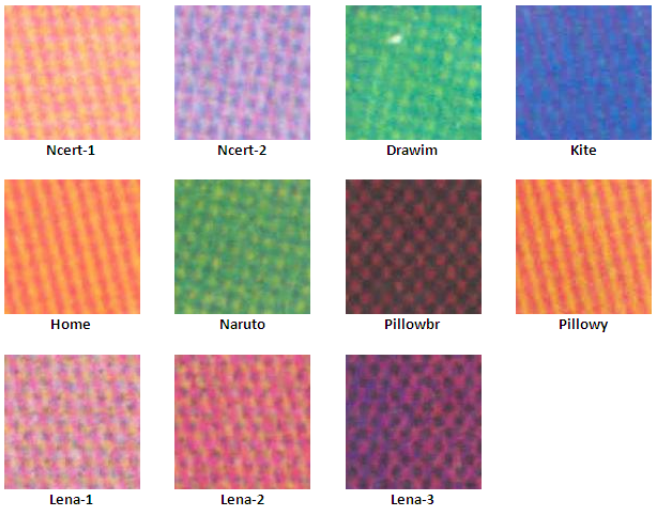


Figure A.6: Samples of printer l4550n with pid-5

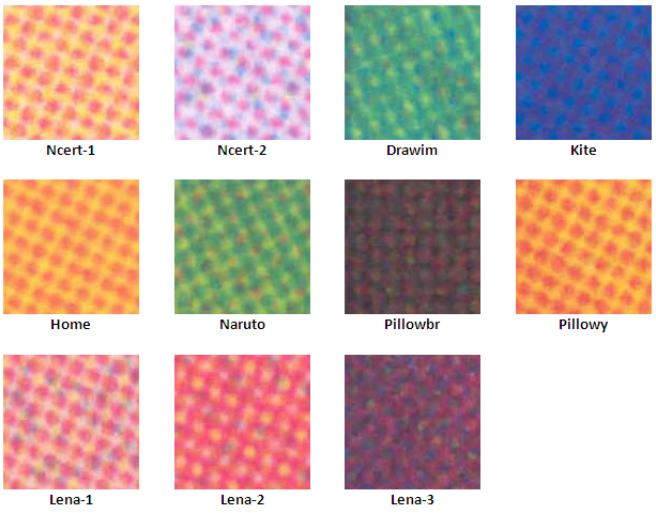


Figure A.7: Samples of printer samsungclp-510 with pid-6

Figure A.8: Selection of samples s1-s29 from various images

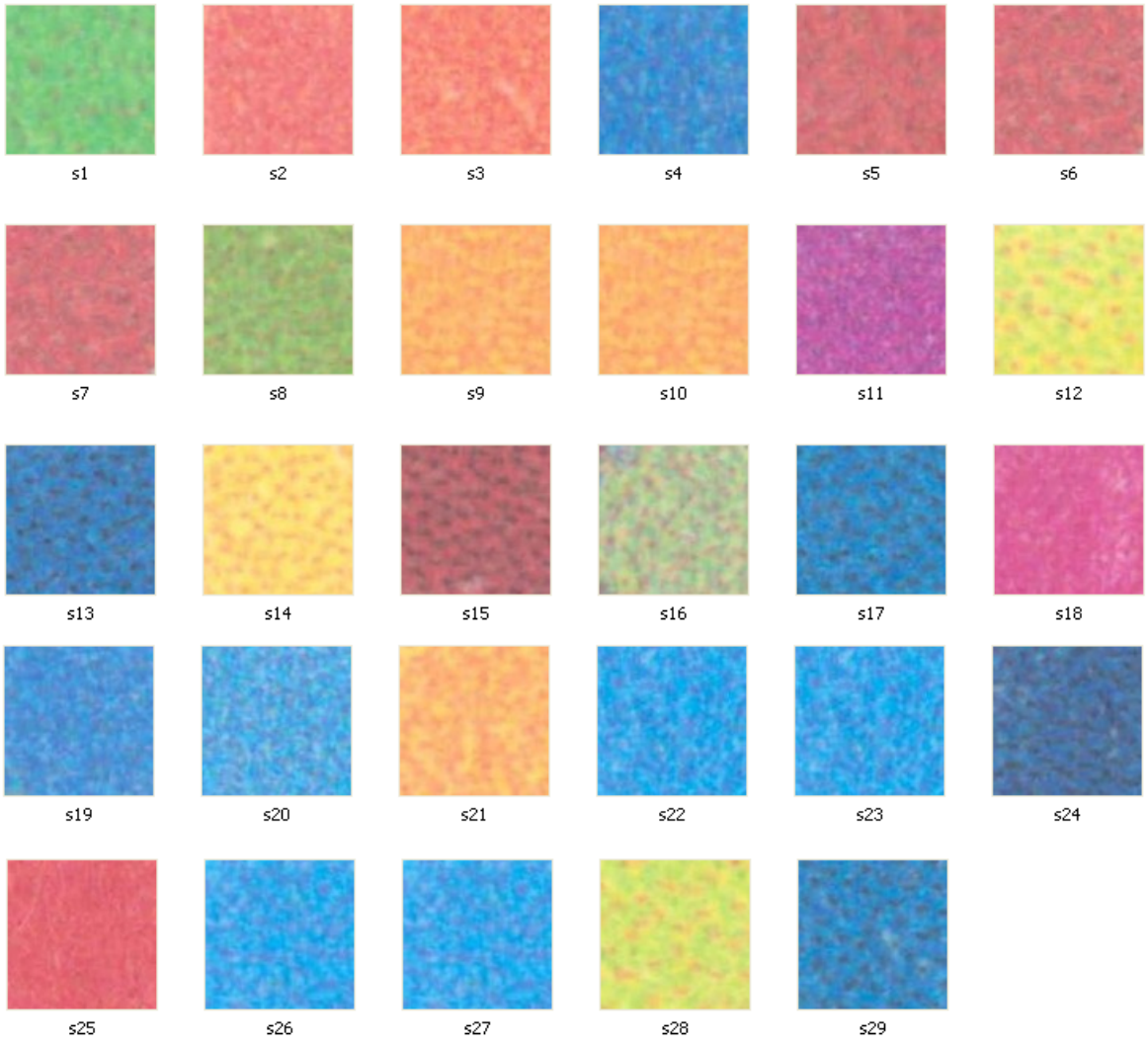


Figure A.9: Samples from Photosmart printer with Pid-1

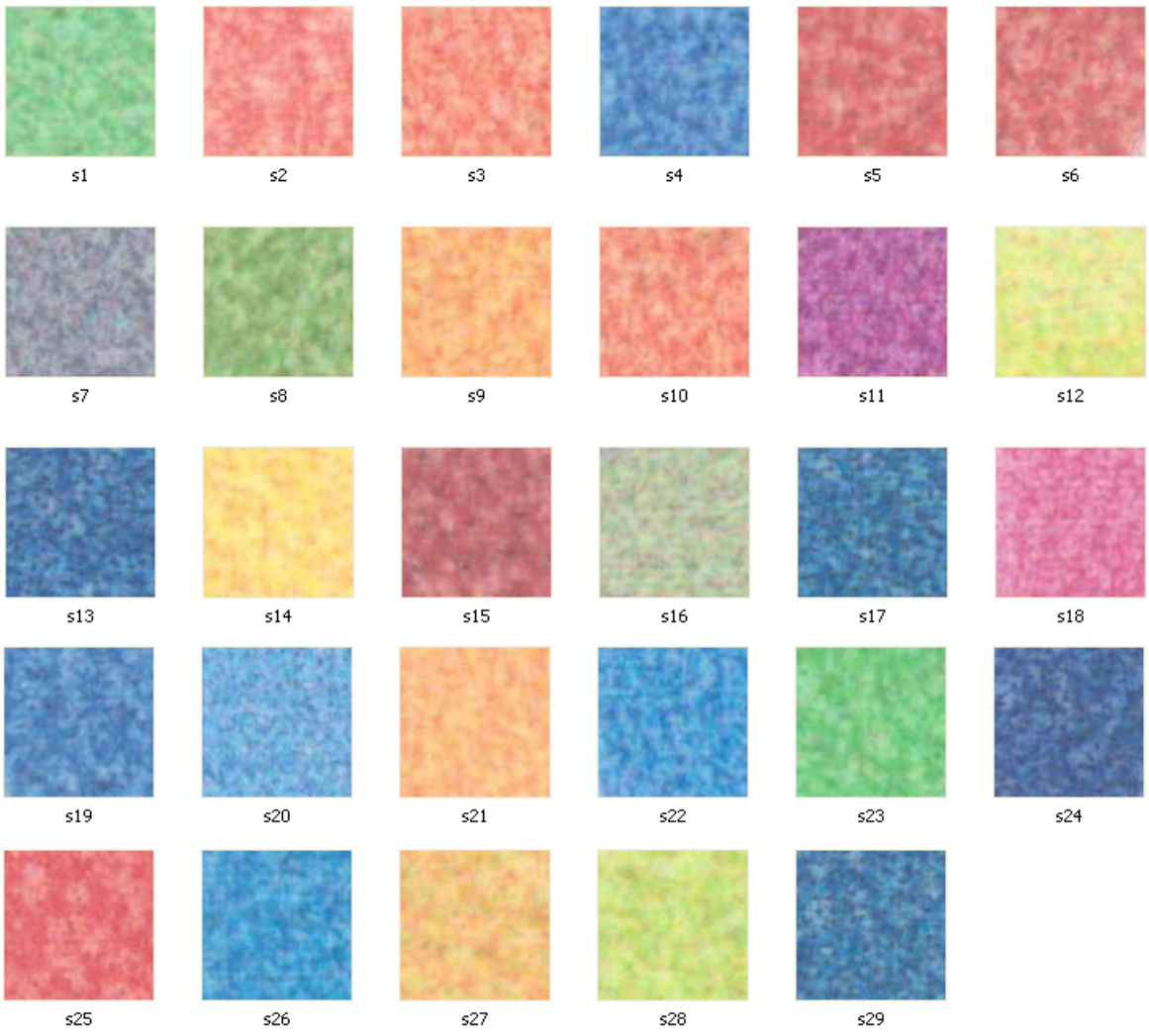


Figure A.10: Samples from Officejet printer with Pid-4

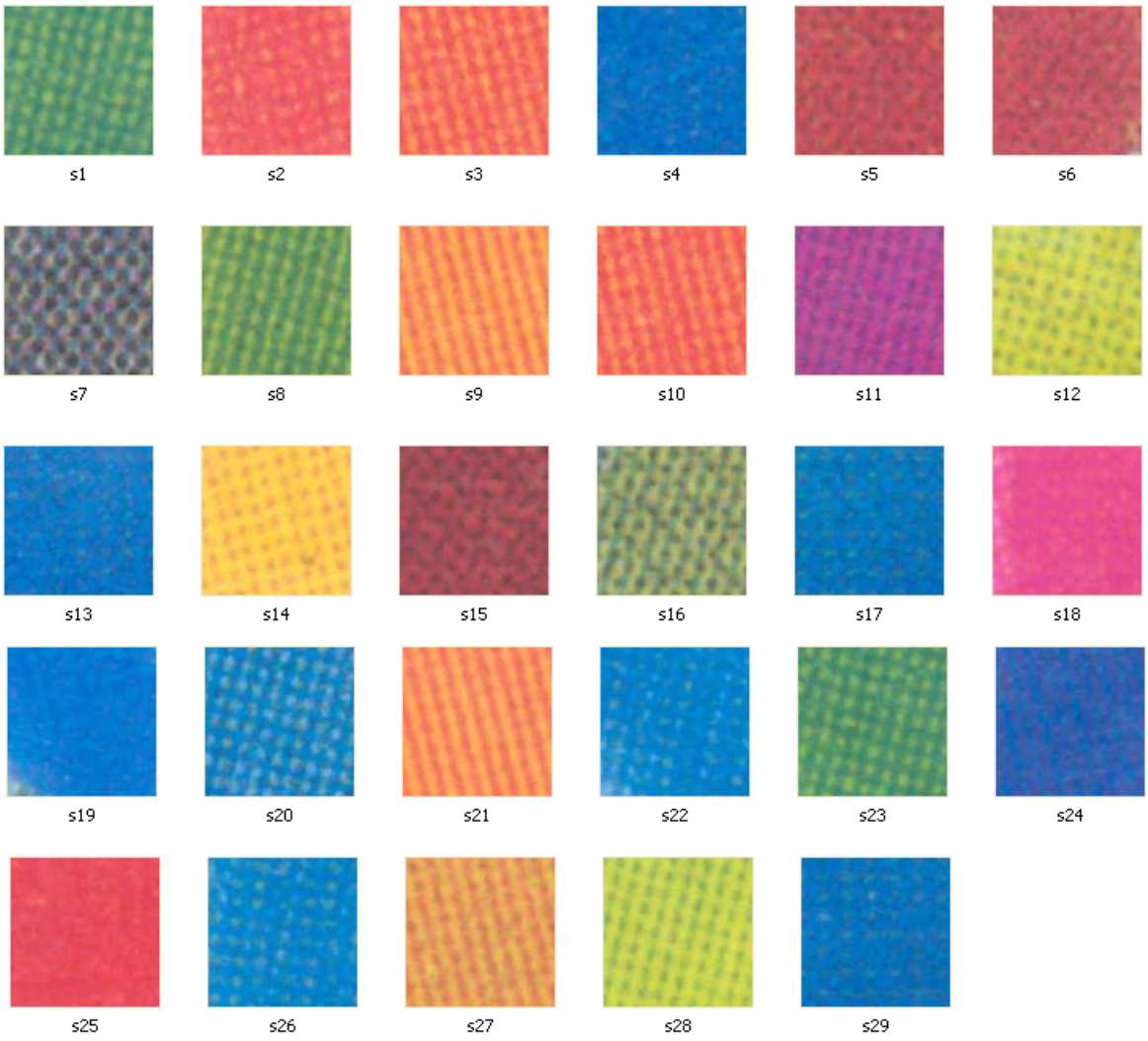


Figure A.11: Samples from Colorlaserjet4550N printer with Pid-5

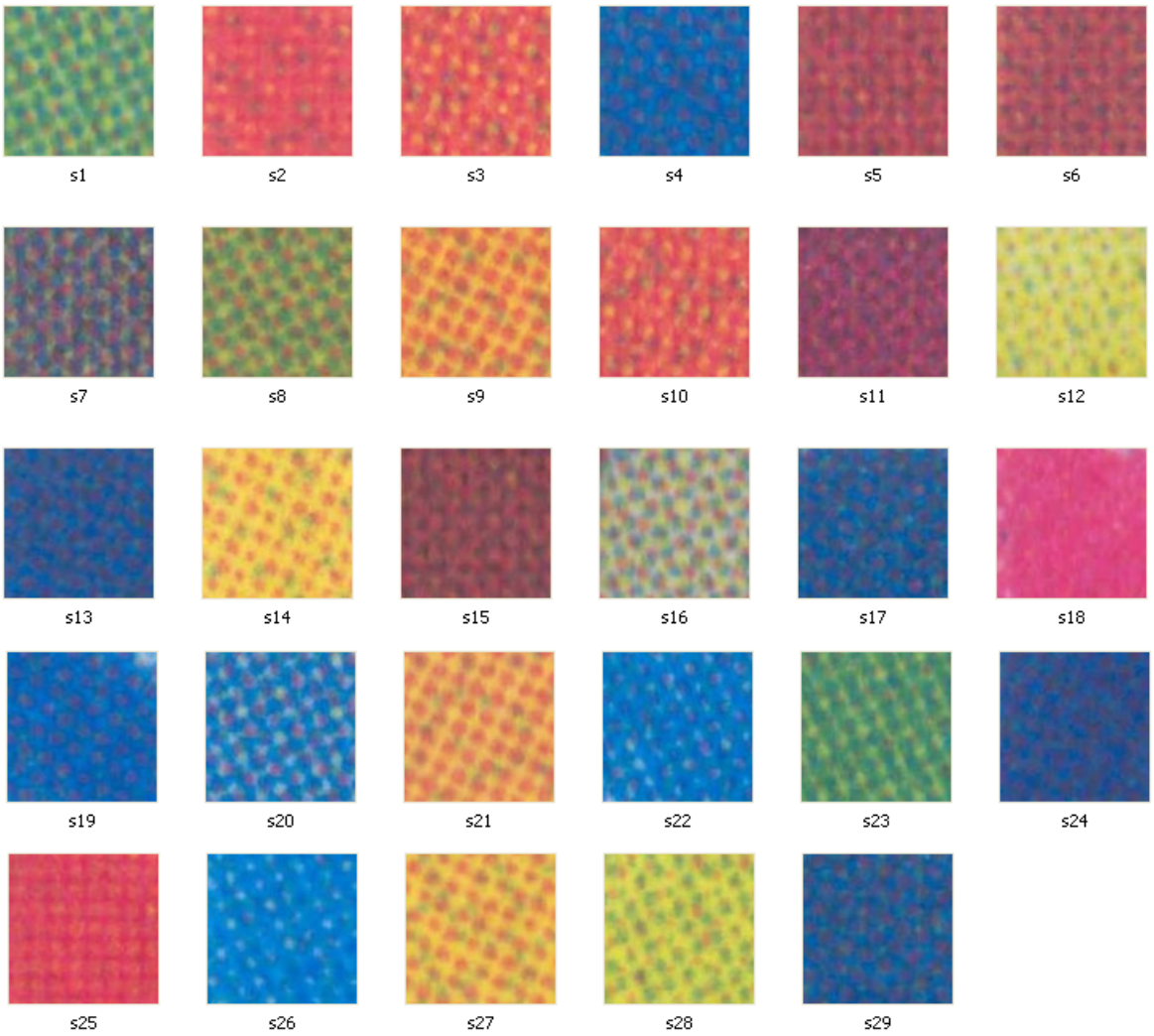


Figure A.12: Samples from SamsungCLP-510 printer with Pid-6

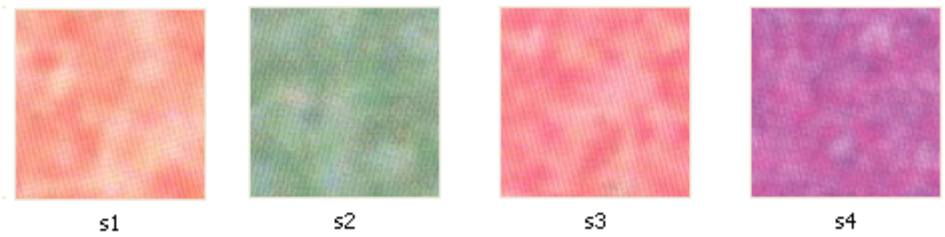


Figure A.13: Samples of Officejet print again printed on Laser printer

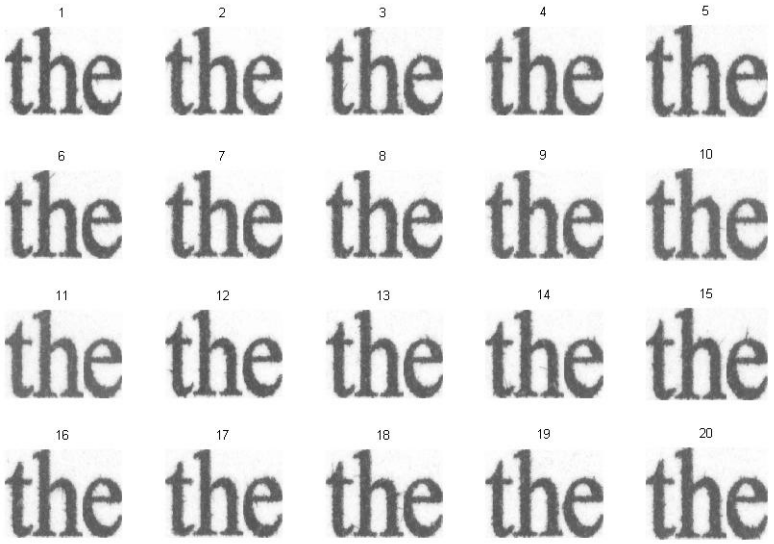


Figure A.14: Type 1 Sample ‘the’ from printer Hppsc1608 with Pid 1

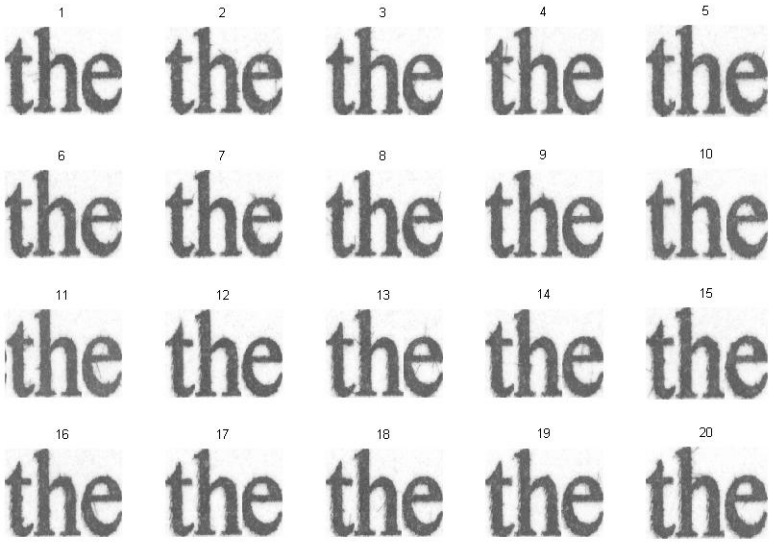


Figure A.15: Type 1 Sample ‘the’ from printer Officejet6110 with Pid 2

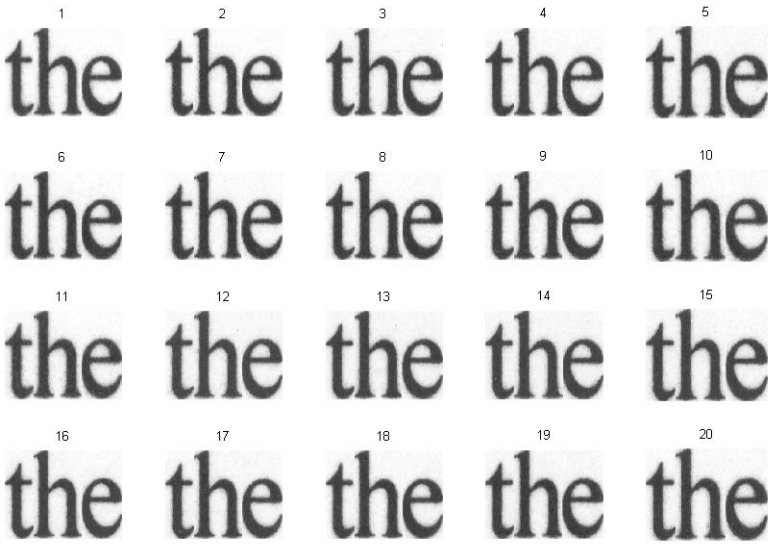


Figure A.16: Type 1 Sample ‘the’ from printer Hplaser4550N with Pid 3

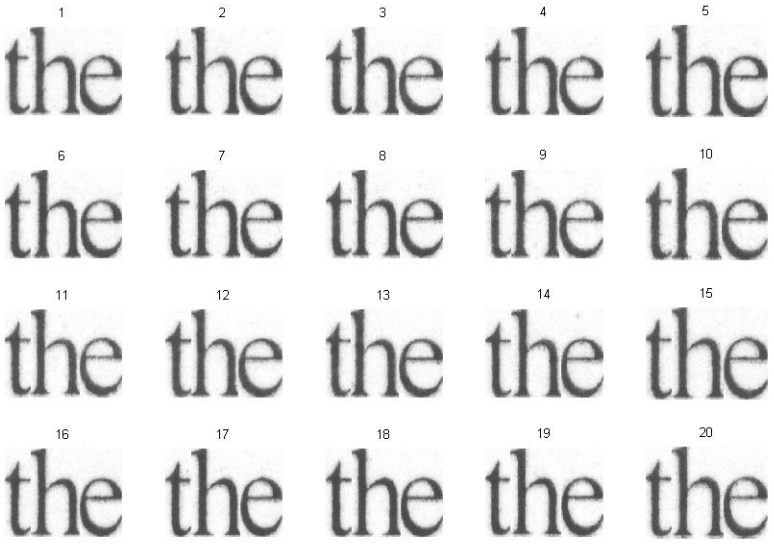


Figure A.17: Type 1 Sample ‘the’ from printer Hplaser1200 with Pid 4

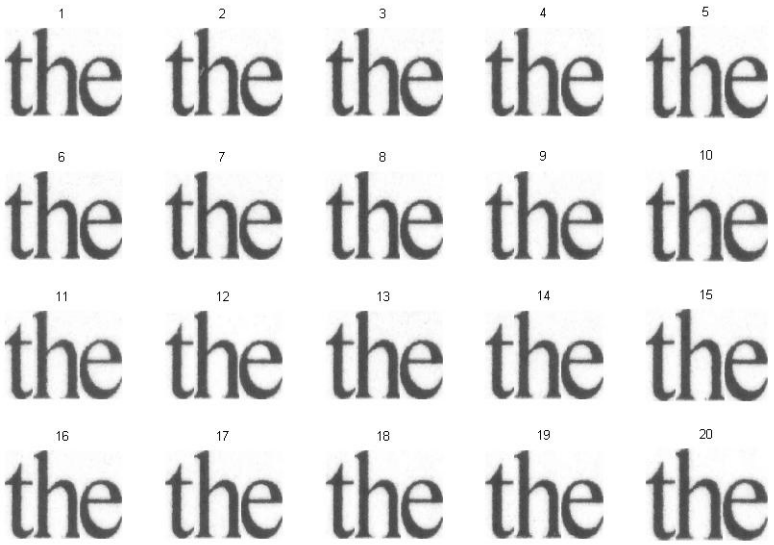


Figure A.18: Type 1 Sample ‘the’ from printer SamsungMl2010 with Pid 5

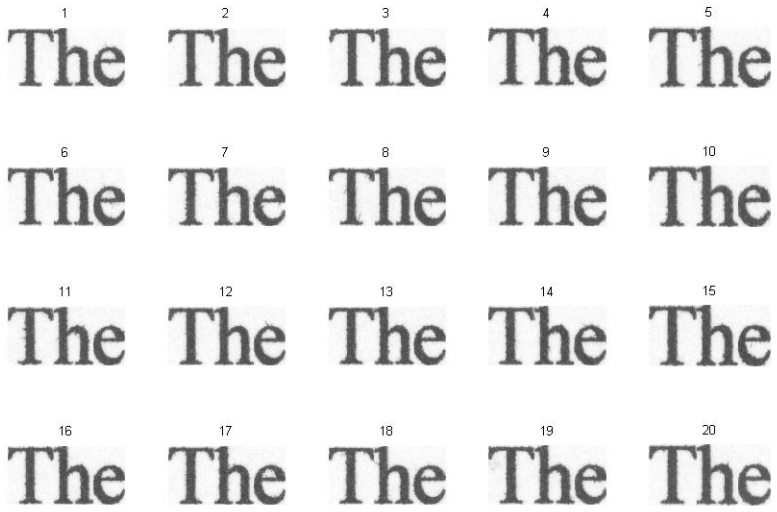


Figure A.19: Type 2 Sample 'The' from printer Hppsc1608 with Pid 1

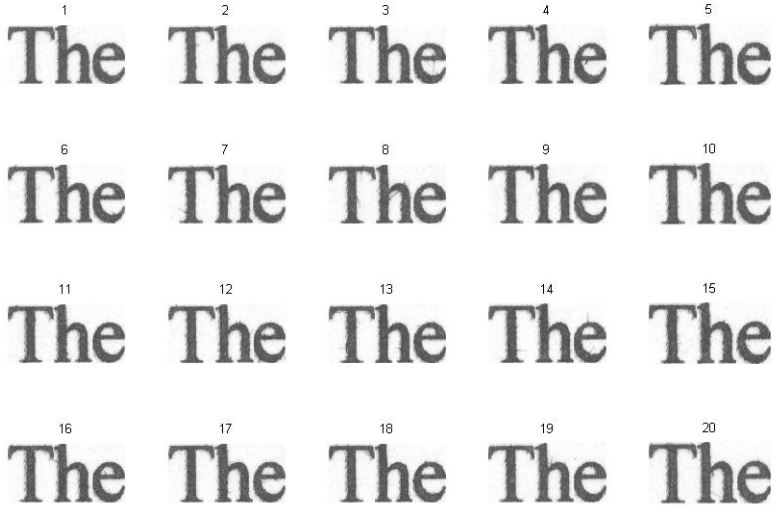


Figure A.20: Type 2 Sample 'The' from printer Officejet6110 with Pid 2

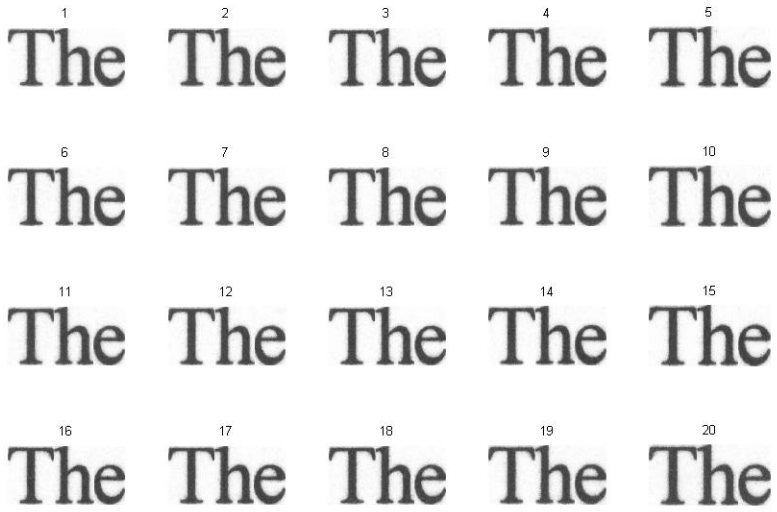


Figure A.21: Type 2 Sample ‘The’ from printer Hplaser4550N with Pid 3

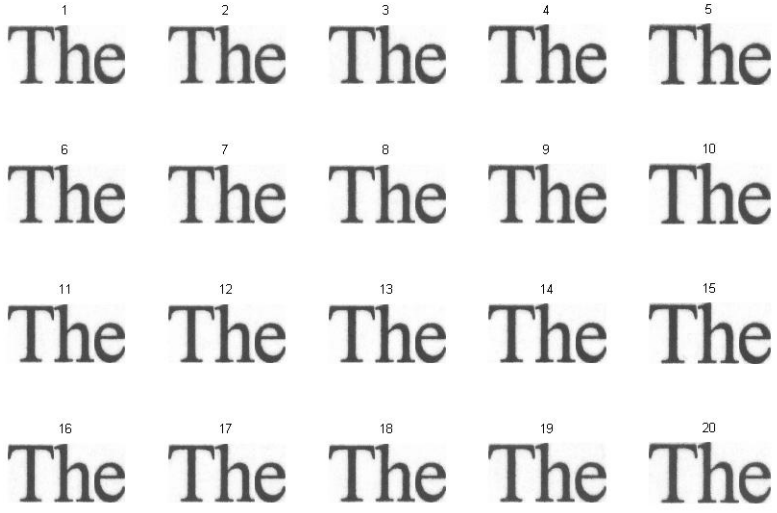


Figure A.22: Type 2 Sample ‘The’ from printer Hplaser1200 with Pid 4

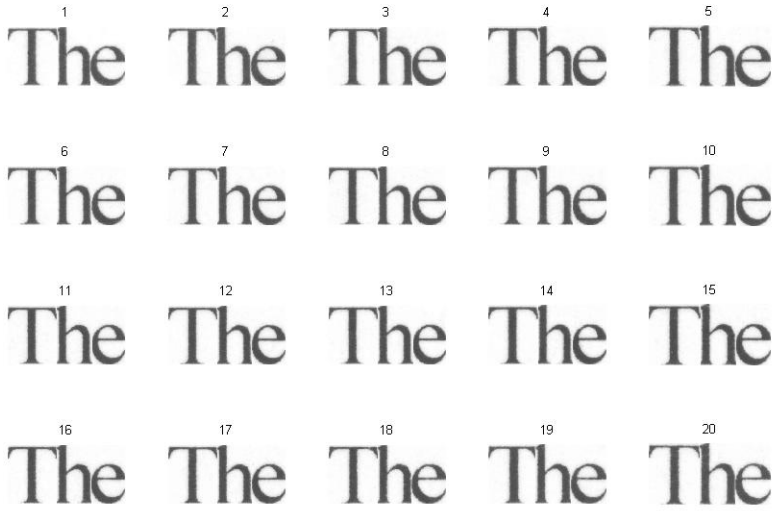


Figure A.23: Type 2 Sample ‘The’ from printer SamsungML2010 with Pid 5

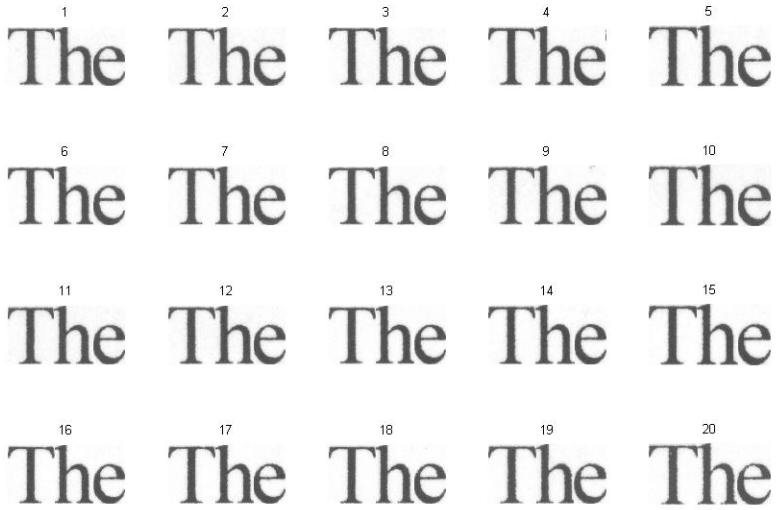


Figure A.24: Type 2 Sample ‘The’ from printer XeroxWorkCentrePe220 with Pid 6

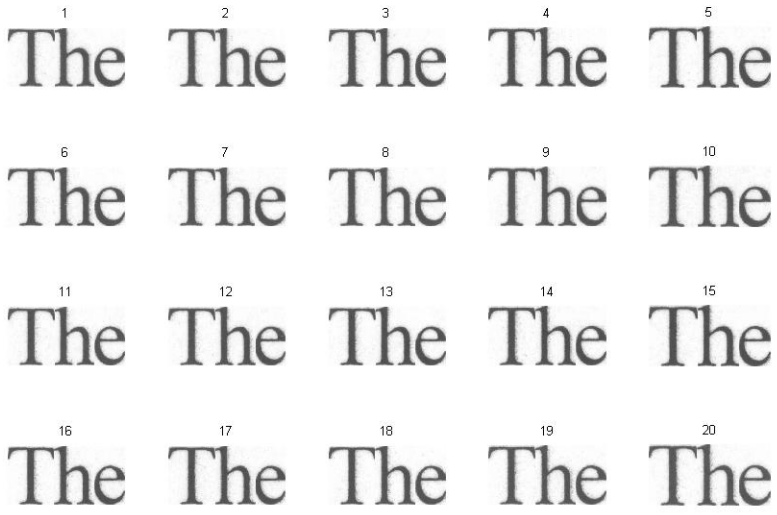


Figure A.25: Type 2 Sample 'The' from printer Hplaser9040 with Pid 7

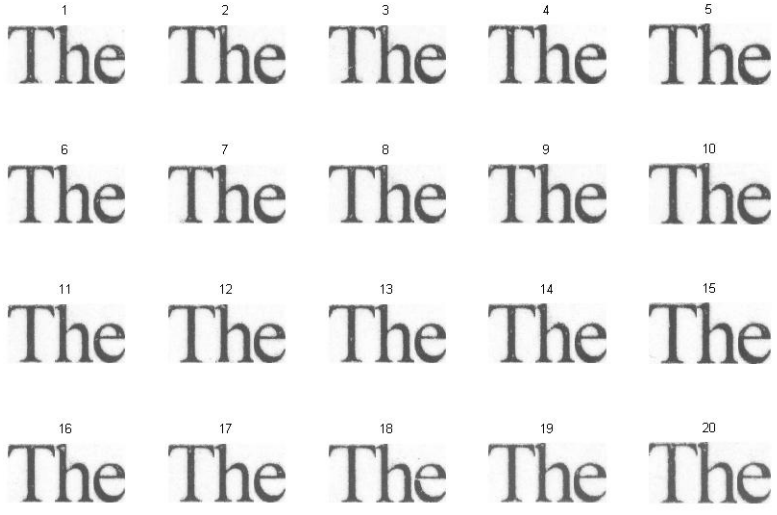


Figure A.26: Type 2 Sample 'The' from printer CannonIR3530 with Pid 8

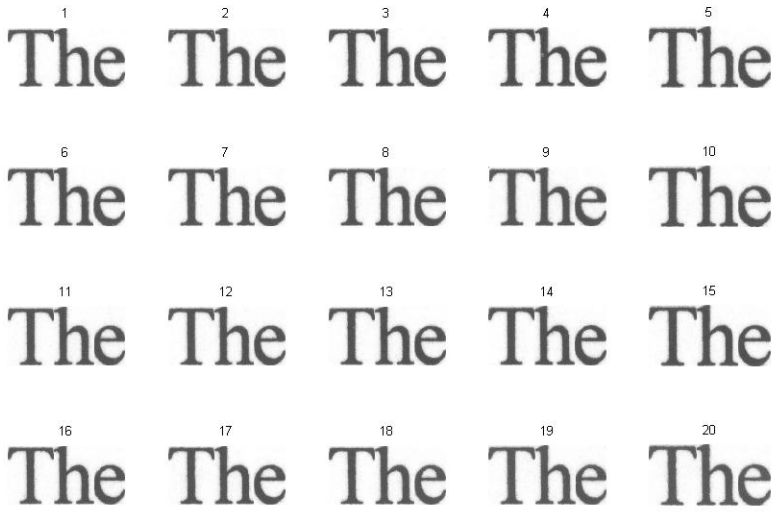


Figure A.27: Type 2 Sample 'The' from printer CannonLBP2900 with Pid 9

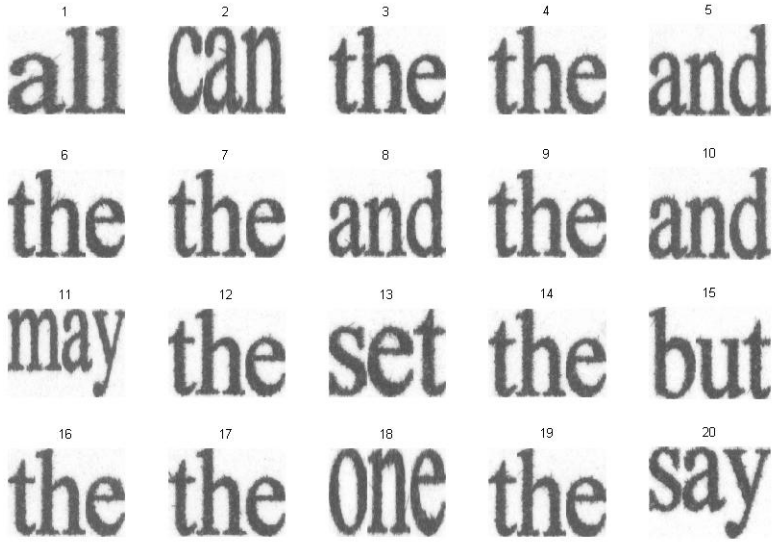


Figure A.28: Type 3 Samples from printer Hppsc1608 with Pid 1

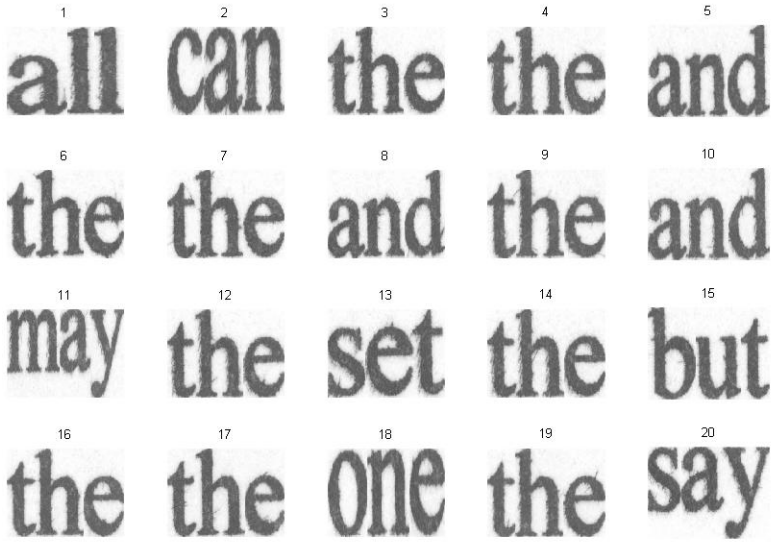


Figure A.29: Type 3 Samples from printer Officejet6110 with Pid 2

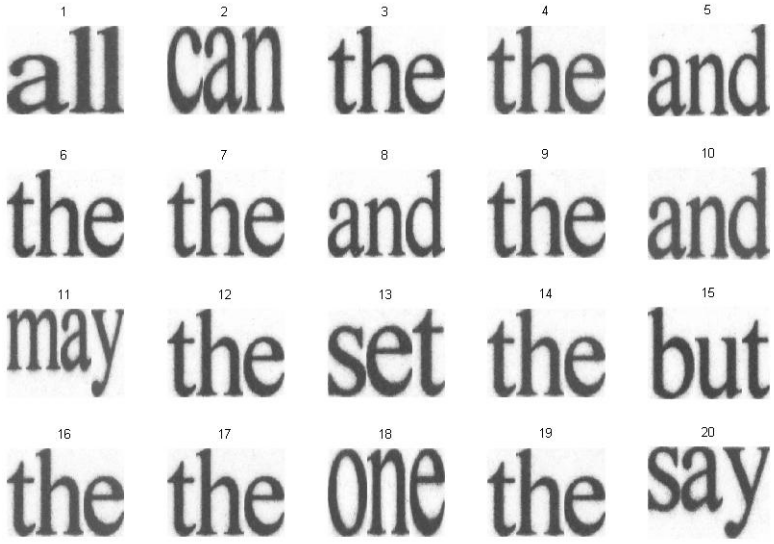


Figure A.30: Type 3 Sample2 from printer Hplaser4550N with Pid 3

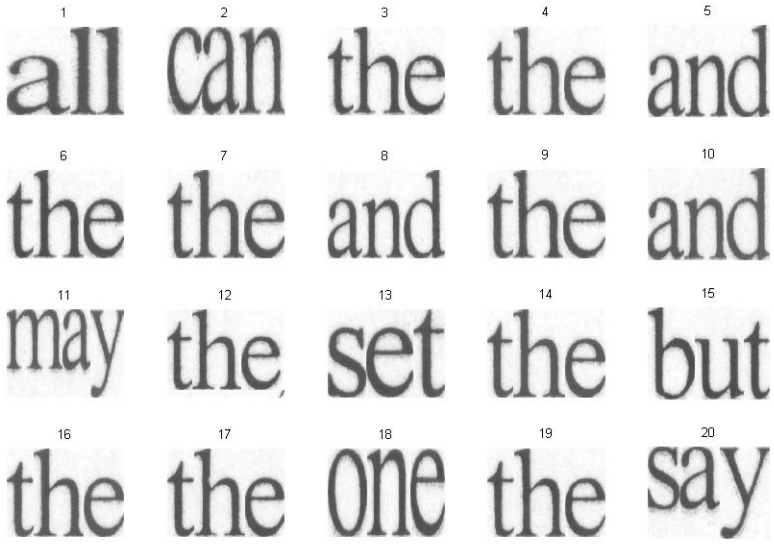


Figure A.31: Type 3 Samples from printer Hplaser1200 with Pid 4

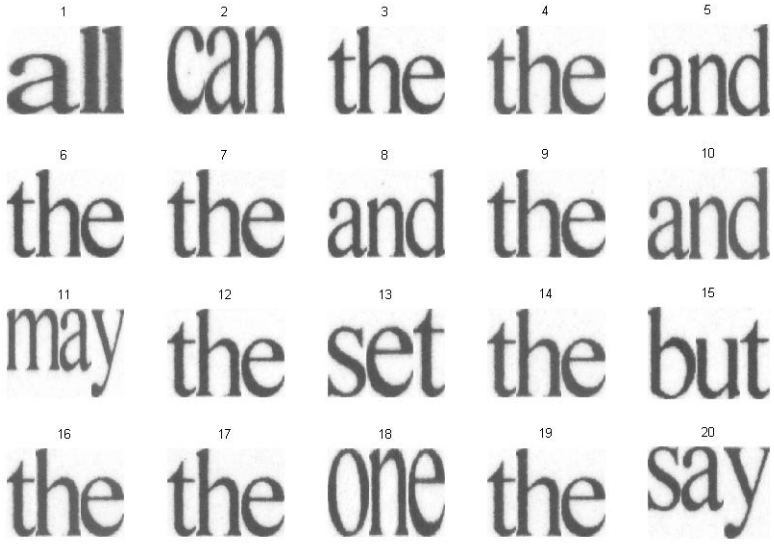


Figure A.32: Type 3 Samples from printer SamsungMl2010 with Pid 5

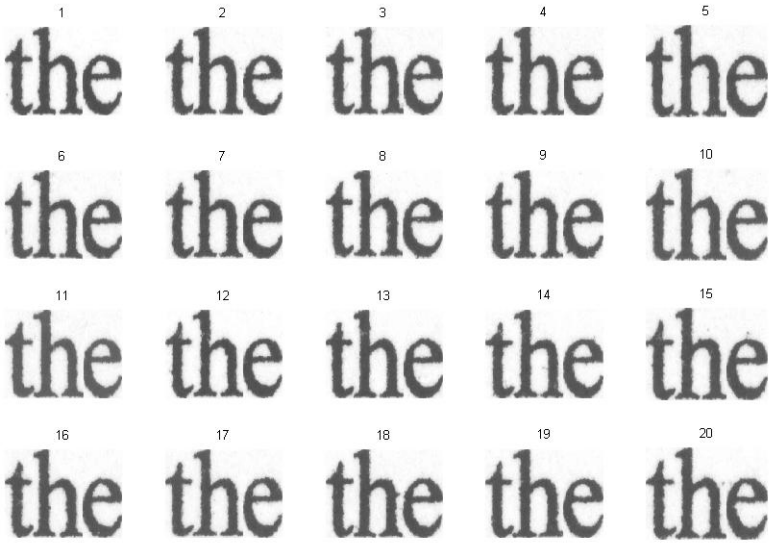


Figure A.33: Photocopy of Type 1 sample from printer Hppsc1608

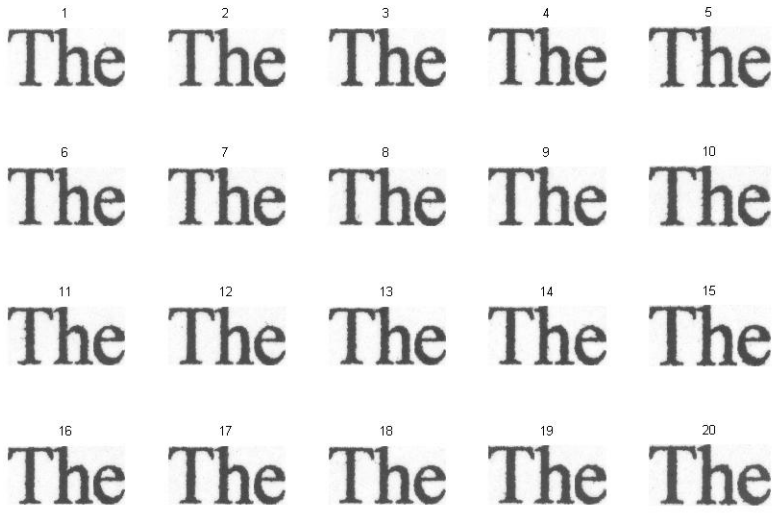


Figure A.34: Photocopy of Type 2 sample from printer Hppsc1608

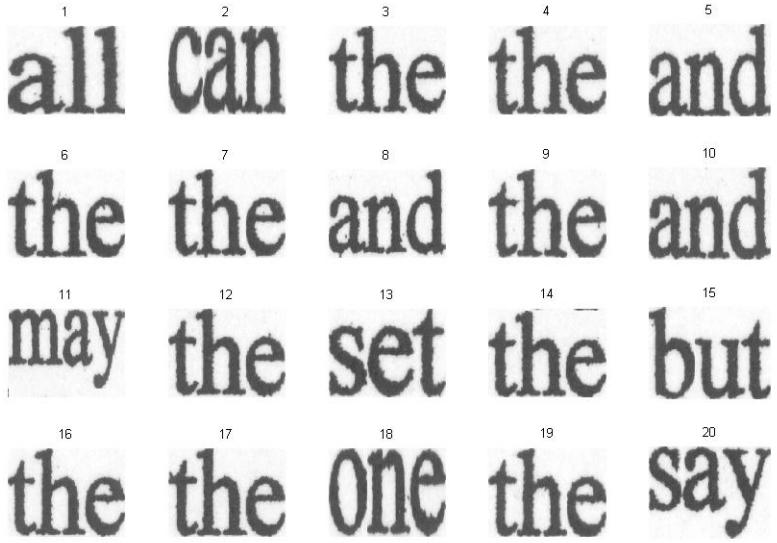


Figure A.35: Photocopy of Type 3 sample from printer Hppsc1608

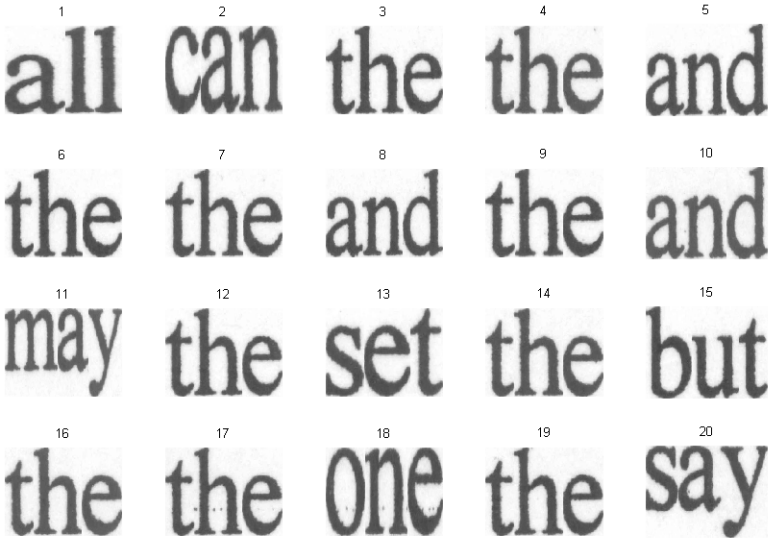


Figure A.36: Photocopy of Type 3 sample from printer Laser4550N

Appendix B

RDT results

B.1 RDT results for Gray DGVM

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 10°			
ANGLE	k	REDUCT	ACCURACY(%)
10	0.1	b1, a5, *nugget*	44.19
10	0.2	b3, a4, b5	51.16
10	0.3	b1, c5, a5, b5	53.49
10	0.4	b3, b1, *sill*, c5	60.47
10	0.5	b3, *nugget*, b1, c5, *sill*, c4	51.16
10	0.6	b3, b1, *nugget*, c5, c4	62.79
10	0.7	b3, b5, *sill*, c3, b4	55.81
10	0.8	*sill*, b4, c5, b3, c3, b1	76.74
10	0.9	*sill*, b3, b1, b5, c5, c4	62.79
10	1	data is not adequate	0

Table B.1: RDT results for Gray DGVMPT data along angle 10°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 15°			
ANGLE	k	REDUCT	ACCURACY(%)
15	0.1	*sill*, c2, *nugget*	39.53
15	0.2	*nugget*, b2, c4	48.84
15	0.3	*sill*, b3, c4, *nugget*	51.16
15	0.4	*nugget*, b2, b3, a5	51.16
15	0.5	*sill*, b2, *nugget*, c2, b4	53.49
15	0.6	b3, *nugget*, c2, c3, a5	67.44
15	0.7	*nugget*, b2, c3, b3, *sill*, a5	55.81
15	0.8	*nugget*, b2, b3, c3, *sill*,b4	55.81
15	0.9	*nugget*, b2, b3, *sill*, b4, c2	67.44
15	1	Data is not adequate	0

Table B.2: RDT results for Gray DGVMPT data along Angle 15°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 20°			
ANGLE	k	REDUCT	ACCURACY(%)
20	0.1	c4, *nugget*, *sill*	44
20	0.2	*nugget*, c5, b5	72
20	0.3	*nugget*, c5, c3	65
20	0.4	*nugget*, b5, c4, c5	58
20	0.5	*nugget*, b5, b3, c4	44
20	0.6	*nugget*, b4, c5, a4, *sill*	65
20	0.7	*nugget*, c5, c4, b5, b3, *sill*	65
20	0.8	*nugget*, b5, b2, a4, c1, c2	72
20	0.9	*nugget*, b4, b5, b2, c2, c3, *sill*	65
20	1	b4, *nugget*, b5, *sill*, c4, c3	65

Table B.3: RDT results for Gray DGVMPT data along angle 20°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 25°			
ANGLE	k	REDUCT	ACCURACY(%)
25	0.1	*sill*, b3	60.47
25	0.2	*sill*, b5, c5	51.16
25	0.3	*sill*, b3, *nugget*, c5	51.16
25	0.4	*sill*, b3, *nugget*, c3	58.14
25	0.5	*sill*, b3, *nugget*, c3, b5	58.14
25	0.6	Data not adequate	0
25	0.7	Data is not adequate	0
25	0.8	Data is not adequate	0
25	0.9	*sill*, b5, b4, c5, c3, b3, a3, *nugget*	46.51
25	1	Data is not adequate	0

Table B.4: RDT results for Gray DGVMPT data angle 25°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 30°			
ANGLE	k	REDUCT	ACCURACY(%)
30	0.1	*sill*, b4, c5	46.51
30	0.2	*nugget*, b2, b4	62.79
30	0.3	*sill*, c2, b4, *nugget*	65.12
30	0.4	*sill*, c4, c3, *nugget*	67.44
30	0.5	*sill*, c2, b3, b4	55.81
30	0.6	*sill*, c2, b2, c3, *nugget*, c5	60.47
30	0.7	*sill*, c2, b2, c5, a3, *nugget*	74.42
30	0.8	*sill*, b5, c2, *nugget*, c4, a5, b4	79.07
30	0.9	*sill*, b3, c2, b4, c3, c5, *nugget*	60.47
30	1	Data is not adequate	0

Table B.5: RDT results for Gray DGVMPT data along Angle 30°

RDT RESULT FOR GRAY DIRECTIONLAL GVMPT ALONG ANGLE 35°			
ANGLE	k	REDUCT	ACCURACY(%)
35	0.1	*nugget*, c1, *sill*	65.12
35	0.2	*sill*, b2, c5	53.49
35	0.3	*nugget*, b2, c5, b5	62.79
35	0.4	*sill*, b2, b5, c4	62.79
35	0.5	*sill*, b2, c2, b3	74.42
35	0.6	*nugget*, b2, b5, c2, c5	74.42
35	0.7	*nugget*, b2, c2, a3, b4	72.09
35	0.8	*nugget*, b2, b5, *sill*, b4	83.72
35	0.9	*nugget*, c2, *sill*, c5, b4, b5	76.74
35	1	*sill*, c2, *nugget*, b5, b3, c5,c4	69.77

Table B.6: RDT results for Gray DGVMPT data along Angle 35°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 40°			
ANGLE	k	REDUCT	ACCURACY(%)
40	0.1	*nugget*, b5, c4	46.51
40	0.2	*nugget*, b5, c3	58.14
40	0.3	*nugget*, c2, c4, c5	58.14
40	0.4	*nugget*, b5, c2, *sill*	62.79
40	0.5	*nugget*, b3, *sill*, b2, c2	60.47
40	0.6	*nugget*, b2, b5, b3, c3	60.47
40	0.7	*nugget*, b4, b3, c2, b1, b5	69.77
40	0.8	*nugget*, b3, *sill*, b2, c2	79.07
40	0.9	*nugget*, b5, b4, c1, b3,c2	60.47
40	1	b3, *nugget*,*sill*, b2, b1, c5, c2	79.07

Table B.7: RDT results for Gray DGVMPT data along angle 40°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANLGE 45°			
ANGLE	k	REDUCT	ACCURACY(%)
45	0.1	b4, *nugget*, *sill*	46.51
45	0.2	b4, *nugget*, b5	41.86
45	0.3	b4, *nugget*, c5, b3	39.53
45	0.4	b4, *nugget*, c5, b1	69.77
45	0.5	c4, b5, *nugget*, c5, b4	58.14
45	0.6	Data is not adequate	0
45	0.7	c4, b5, *sill*, b4, c5, b3, c2, c1	60.47
45	0.8	c4, b5, *sill*, b4, c5, *nugget*, b3, c2	53.49
45	0.9	b4, b5, *nugget*, c4, b1, c5, *sill*, b3	67.44
45	1	data is not adquate	0

Table B.8: RDT results for Gray DGVMPT data along angle 45°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 50°			
ANGLE	k	REDUCT	ACCURACY(%)
50	0.1	*nugget*, b4, *sill*	39.53
50	0.2	*nugget*, b2 ,c4	55.81
50	0.3	*nugget*, b5, c5, *sill*	48.84
50	0.4	*nugget*, b4, *sill*, b5	72.09
50	0.5	*nugget*, b4, *sill*, c3, c5, b5	62.79
50	0.6	*nugget*, b2, c4, c3, c5	65.12
50	0.7	*nugget*, b4, *sill*, c3, b2, c5	69.77
50	0.8	*nugget*, b4, *sill*, b2, c5, c4	62.79
50	0.9	*nugget*, b4, *sill*, b2, c4 ,b5	76.74
50	1	*nugget*, b4, *sill*, b2, c3, c5,c4	67.44

Table B.9: RDT results for Gray DGVMPT data along angle 50°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 55°			
ANGLE	k	REDUCT	ACCURACY(%)
55	0.1	*nugget*, b4, *sill*	48.84
55	0.2	*nugget*, b2, c4	53.49
55	0.3	*nugget*, b5, c5, *sill*	62.79
55	0.4	*nugget*, b4, *sill*, b5	67.44
55	0.5	*nugget*, b4, *sill*, c3, c5, b5	72.09
55	0.6	*nugget*, b2, c4, c3, c5	67.44
55	0.7	*nugget*, b4, *sill*, c3, b2, c5	69.77
55	0.8	*nugget*, b4, *sill*, b2, c5,c4	69.77
55	0.9	*nugget*, b4, *sill*, b2, c4, b5	79.07
55	1	*nugget*, b4, *sill*, b2, c3, c5, c4	74.42

Table B.10: RDT results for Gray DGVMPT data along angle 55°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 60°			
ANGLE	k	REDUCT	ACCURACY(%)
60	0.1	*sill*, c4, *nugget*	58.14
60	0.2	*sill*, c4, c5, *nugget*	58.14
60	0.3	*nugget*, c4, *sill*, c5, a4	65.12
60	0.4	Data is not adequate	0
60	0.5	Data is not adequate	0
60	0.6	Data is not adequate	0
60	0.7	Data is not adequate	0
60	0.8	Data is not adequate	0
60	0.9	Data is not adequate	0
60	1	Data is not adequate	0

Table B.11: RDT results for Gray DGVMPT data along angle 60°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 65°			
ANGLE	k	REDUCT1	ACCURACY(%)
65	0.1	*nugget*, b2, *sill*	41.86
65	0.2	*nugget*, b2, c3, *sill*	48.84
65	0.3	*nugget*, c3, b2, b3	58.14
65	0.4	*nugget*, b3, b1, a3, a2	60.47
65	0.5	*nugget*, b3, c1, b1, c2, b2, c3	58.14
65	0.6	b3, *nugget*, c1, b2, c3, a2, a4	65.12
65	0.7	Data is not adequate	0
65	0.8	*sill*, b3, b1, b2, c1, *nugget*, c2, a4, c3	67.44
65	0.9	Data is not adequate	0
65	1	Data is not adequate	0

Table B.12: RDT results for Gray DGVMPT data along angle 65°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 70°			
ANGLE	k	REDUCT	ACCURACY(%)
70	0.1	*sill*, b2, *nugget*	55.81
70	0.2	*sill*, b3, *nugget*, c3	53.49
70	0.3	*sill*, b3, *nugget*, c2	69.77
70	0.4	*sill*, b3, *nugget*, c2, b2, c3	72.09
70	0.5	Data is not adequate	0
70	0.6	Data is not adequate	0
70	0.7	Data is not adequate	0
70	0.8	Data is not adequate	0
70	0.9	Data is not adequate	0
70	1	Data is not adequate	0

Table B.13: RDT results for Gray DGVMPT data along angle 70°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 75°			
ANGLE	k	REDUCT	ACCURACY(%)
75	0.1	c1, c5, *nugget*	39.53
75	0.2	c1, c5, *nugget*	55.81
75	0.3	c1, *sill*, b1, c5	48.84
75	0.4	c1 , b4, b5, *nugget*, c5	48.84
75	0.5	b1, *nugget*, c1, b4, b5	55.81
75	0.6	b1, *sill*, c1, b4, b5	51.16
75	0.7	b1, *sill*, c1, b4, b5, b2	60.47
75	0.8	b1, *sill*, c1, b5, *nugget*, b4, b2, a1	48.84
75	0.9	b1, *sill*, c1, *nugget*, b4, b2, a1	62.79
75	1	Data is not adequate	0

Table B.14: RDT results for Gray DGVMPT data along angle 75°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 80°			
ANGLE	k	REDUCT	ACCURACY(%)
80	0.1	*nugget*, b3, *sill*	51.16
80	0.2	*nugget*, b5, b3	34.88
80	0.3	b5, *nugget*, b3, *sill*, b4	23.26
80	0.4	b5, *sill*, b4, c3, *nugget*	48.84
80	0.5	b5, b3, *sill*, c5, c4	72.09
80	0.6	b5, *sill*, b3, c4, *nugget*, c3	58.14
80	0.7	Data is not adequate	0
80	0.8	Data is not adequate	0
80	0.9	Data is not adequate	0
80	1	Data is not adequate	0

Table B.15: RDT results for Gray DGVMPT data along angle 80°

RDT RESULT FOR GRAY DIRECTIONAL GVMPT ALONG ANGLE 85°			
ANGLE	k	REDUCT	ACCURACY(%)
85	0.1	*nugget*, *sill*, a5	46.51
85	0.2	*nugget*, b3, a5, *sill*	41.86
85	0.3	*nugget*, b3, *sill*, c3	53.49
85	0.4	Data is not adequate	0
85	0.5	Data is not adequate	0
85	0.6	Data is not adequate	0
85	0.7	Data is not adequate	0
85	0.8	Data is not adequate	0
85	0.9	Data is not adequate	0
85	1	Data is not adequate	0

Table B.16: RDT results for Gray DGVMPT data along angle 85°

B.2 RDT results for Standardised Gray GVM

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 10°			
ANGLE	k	REDUCT	ACCURACY(%)
10	0.1	*nugget*, c5, *sill*	37.21
10	0.2	*sill*, b3, c5, a5	46.51
10	0.3	*sill*, c5, *nugget*, a4, c3	55.81
10	0.4	*sill*, b3, b5, c2, a4	55.81
10	0.5	b5, *nugget*, b2, *sill*, c5, a4	69.77
10	0.6	*sill*, b3, b5, b2, c5, *nugget*, c2	60.47
10	0.7	b5, b3, *sill*, b2, *nugget*, c2	60.47
10	0.8	b5, b3, *sill*, b2, *nugget*, c3, c5, c2	58.14
10	0.9	Data is not adequate	0
10	1	Data is not adequate	0

Table B.17: RDT results for Standardised Gray DGVMPT data along angle 10°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 15°			
ANGLE	k	REDUCT	ACCURACY(%)
15	0.1	*nugget*, c4, *sill*	37.21
15	0.2	b3, *nugget*, b2	41.86
15	0.3	b3, *nugget*, b2, b4	51.16
15	0.4	b3, *nugget*, b2, b4	44.19
15	0.5	b2, *nugget*, b4, c3, c4	51.16
15	0.6	b2, *nugget*, c3, a2, *sill*, c4	53.49
15	0.7	b2, *sill*, b3, b4, c3, *nugget*	53.49
15	0.8	b2, b3, *sill*, *nugget*, c2, b4	67.44
15	0.9	Data is not adequate	0
15	1	Data is not adequate	0

Table B.18: RDT results for Standardised Gray DGVMPT data along angle 15°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 20°			
ANGLE	k	REDUCT	ACCURACY(%)
20	0.1	b4, c1, *nugget*	46
20	0.2	b4, c1, *sill*	58
20	0.3	b4, c1, b1, *nugget*	53
20	0.4	b4, c1, b2, *sill*	58
20	0.5	b4, b3, *nugget*, c5, *sill*	58
20	0.6	b4, b2, *nugget*, c1, b5	74
20	0.7	b4, c1, *nugget*, c2, b1	69
20	0.8	b1, *sill*, b4, b3, c1, *nugget*	67
20	0.9	b1, b4, *sill*, b3, c2, c4	81
20	1	b4, b3, b1, a2, c1, c4, *sill*	63

Table B.19: RDT results for Standardised Gray DGVMPT data along angle 20°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 25°			
ANGLE	k	REDUCT	ACCURACY(%)
25	0.1	b5, b3, *nugget*	37.21
25	0.2	b3, *sill*, b5	48.84
25	0.3	b3, *sill*, b5, *nugget*	60.47
25	0.4	b3, b5, c4, *sill*	51.16
25	0.5	b3, *sill*, b4, b5, *nugget*	62.79
25	0.6	b3, b5, *sill*, c2, *nugget*	62.79
25	0.7	*nugget*, b3, b5, b2, *sill*, a5	55.81
25	0.8	b5, b3, b4, b2 , *nugget*, c3, c5	58.14
25	0.9	b3, b5, *sill*, b4, c3, c2, c5	74.42
25	1	b3, b5, b4, *sill*, a1, c4, c5, c3	69.77

Table B.20: RDT results for Standardised Gray DGVMPT data along angle 25°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 30°			
ANGLE	k	REDUCT	ACCURACY(%)
30	0.1	*nugget*, b2, *sill*	48.84
30	0.2	b3, c1, a3	51.16
30	0.3	b3, *sill*, c1, c4	48.84
30	0.4	b3, *sill*, c1, b4	69.77
30	0.5	b3, *sill*, b1, c3, *nugget*	72.09
30	0.6	b3, *sill*, c1, c2, *nugget*, b4	69.77
30	0.7	b1, *sill*, c3, b3, c1, b4	69.77
30	0.8	b3, *nugget*, c2, c1, b2, *sill*	69.77
30	0.9	b3, b2, *nugget*, c1, c3, *sill*, c4	67.44
30	1	b3, b2, *nugget*, b1, c1, c3,*sill*	72.09

Table B.21: RDT results for Standardised Gray DGVMPT data along angle 30°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 35°			
ANGLE	k	REDUCT	ACCURACY(%)
35	0.1	c3, *nugget*, *sill*	48.84
35	0.2	c3, *nugget*, *sill*	55.81
35	0.3	b2, *nugget*, b3, *sill*	60.47
35	0.4	b3, *sill*, *nugget*, b5	74.42
35	0.5	c3, c2, *nugget*, b5, a5	53.49
35	0.6	b2, c3, *nugget*, *sill*, b5, a5	72.09
35	0.7	b2, *nugget*, b5, b3, c5, *sill*	79.07
35	0.8	c2, *sill*, b5, b3, *nugget*,b2	76.74
35	0.9	b3, b5, c2, *nugget*, b2, *sill*, c3	67.44
35	1	Data is not adequate	0

Table B.22: RDT results for Standardised Gray DGVMPT data along angle 35°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 40°			
ANGLE	k	REDUCT	ACCURACY(%)
40	0.1	*nugget*, c5, *sill*	39.53
40	0.2	b2, c5, *nugget*	62.79
40	0.3	b2, c5, b4, *nugget*	53.49
40	0.4	b2, *sill*, b4, *nugget*	53.49
40	0.5	b3, b2, *nugget*, b5, c5	74.42
40	0.6	b5, *nugget*, b4, c2, c3	65.12
40	0.7	*sill*, b5, b2, b4, *nugget*	65.12
40	0.8	b2 , b4, *sill*, b5, *nugget*, c3	79.07
40	0.9	b2, b4, *nugget*, c5, *sill*, b5	74.42
40	1	b2, b4, b3, *nugget*,c5,*sill*	79.07

Table B.23: RDT results for Standardised Gray DGVMPT data along angle 40°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 45°			
ANGLE	k	REDUCT	ACCURACY(%)
45	0.1	*nugget*, b2, *sill*	41.86
45	0.2	c2, c3, *nugget*	44.19
45	0.3	b2, c3, *sill*, c5	55.81
45	0.4	b1 , c2, *nugget*, c5	51.16
45	0.5	*sill*, b2, b4, c4, *nugget*	55.81
45	0.6	b2, c3,*nugget*, b1, *sill*	65.12
45	0.7	*sill*, b1, b4, c2, *nugget*	62.79
45	0.8	b1, c2,*nugget*, a2, b5, *sill*	67.44
45	0.9	b1, c2, *sill*,c4, c3, b4	74.42
45	1	b2, b3, *sill*, b4, c3, *nugget*, c5	76.74

Table B.24: RDT results for Standardised Gray DGVMPT data along angle 45°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 50°			
ANGLE	k	REDUCT	ACCURACY(%)
50	0.1	a2, b5, *nugget*	34.88
50	0.2	a2, b5, c3	60.47
50	0.3	a2, b5, *nugget*, *sill*	60.47
50	0.4	b5, *nugget*, b2, c5	67.44
50	0.5	c5, *sill*, c2, b1, c3	60.47
50	0.6	b1, *sill*, c3, b2, *nugget*	58.14
50	0.7	b1, *nugget*, b2, c5, c1	76.74
50	0.8	b1, *nugget*, c2, c5, c1, b3, *sill*	67.44
50	0.9	b1, b5, c2, b3, *nugget*, a4	65.12
50	1	b5, *nugget*, b3, b2, b1, a5, c3	65.12

Table B.25: RDT results for Standardised Gray DGVMPT data along angle 50°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 55°			
ANGLE	k	REDUCT	ACCURACY(%)
55	0.1	*nugget*, b1, *sill*	39.53
55	0.2	b1, *nugget*, *sill*	44.19
55	0.3	*sill*, b5, b3, *nugget*	58.14
55	0.4	b1, *nugget*, b5, c1, *sill*	55.81
55	0.5	*nugget*, b1, *sill*, b3, c5	65.12
55	0.6	b5, b3, c1, b1, *nugget*	62.79
55	0.7	b1, *sill*, b5, c1, c5, *nugget*	67.44
55	0.8	Data is not adequate	0
55	0.9	Data is not adequate	0
55	1	Data is not adequate	0

Table B.26: RDT results for Standardised Gray DGVMPT data along angle 55°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 60°			
ANGLE	k	REDUCT	ACCURACY(%)
60	0.1	*nugget*, b3, *sill*	44.19
60	0.2	*nugget*, b3, *sill*	51.16
60	0.3	b3, *nugget*, b1, c2	60.47
60	0.4	b3, *nugget*, b2, b5	48.84
60	0.5	b3,*nugget*,b5,*sill*, b2	60.47
60	0.6	b3, *sill*, b1,*nugget*,c5, c2	69.77
60	0.7	b3, *nugget*, b5,b2,*sill*, c5	62.79
60	0.8	b3, b5, *nugget*, b2,c2,*sill*, c5, a4, b1	62.79
60	0.9	Data is not adequate	0
60	1	Data is not adequate	0

Table B.27: RDT results for Standardised Gray DGVMPT data along angle 60°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 65°			
ANGLE	k	REDUCT	ACCURACY(%)
65	0.1	*nugget*, b5, *sill*	39.53
65	0.2	*nugget*, b5, *sill*	48.84
65	0.3	b2, *nugget*, b5, c4	48.84
65	0.4	b2, *nugget*, b5, c1	62.79
65	0.5	b4, *nugget*, b1,*sill*, c2	60.47
65	0.6	b2, *nugget*, b5, c4, a3, b4	69.77
65	0.7	b2, b5, *nugget*, b1, b4	62.79
65	0.8	b2, b5, *nugget*, b1, *sill*, c5, c2	60.47
65	0.9	b2, b1, *sill*, b5, c5, a2	67.44
65	1	b2, b1, *nugget*, b4, b5, *sill*, a3	67.44

Table B.28: RDT results for Standardised Gray DGVMPT data along angle 65°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 70°			
ANGLE	k	REDUCT	ACCURACY(%)
70	0.1	*nugget*, c4, *sill*	32.56
70	0.2	*nugget*, b5, c4, *sill*	53.49
70	0.3	*nugget*, b3, c4, b5	48.84
70	0.4	b5, *nugget*, *sill*, c4, a5	62.79
70	0.5	b3, *sill*, c4, *nugget*, b2	79.07
70	0.6	b4, *sill*, b5, c3, c5, *nugget*	60.47
70	0.7	*sill*, b5, b3, c4, *nugget*, a5	67.44
70	0.8	b5,*sill*, b4,*nugget*, c4, c3, c2	67.44
70	0.9	Data is not adequate	0
70	1	Data is not adequate	0

Table B.29: RDT results for Standardised Gray DGVMPT data along angle 70°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 75°			
ANGLE	k	REDUCT	ACCURACY(%)
75	0.1	*sill*, b5, *nugget*	34.88
75	0.2	*sill*, b5, *nugget*	48.84
75	0.3	*sill*, b5, *nugget*, c3	44.19
75	0.4	*sill*, b3, *nugget*, b5	55.81
75	0.5	Data is not adequate	0
75	0.6	Data is not adequate	0
75	0.7	Data is not adequate	0
75	0.8	Data is not adequate	0
75	0.9	Data is not adequate	0
75	1	Data is not adequate	0

Table B.30: RDT results for Standardised Gray DGVMPT data along angle 75°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 80°			
ANGLE	k	REDUCT	ACCURACY(%)
80	0.1	b2, *nugget*,c5	34.88
80	0.2	b2, *nugget*, c3	32.56
80	0.3	b2, *nugget*, c3	48.84
80	0.4	b2 ,*nugget*,b3, *sill*, c5	62.79
80	0.5	b2, *sill*, b3, *nugget*, c5	51.16
80	0.6	b2 ,*nugget* ,b3 ,b5, *sill*	55.81
80	0.7	b2, *nugget*, c3, *sill*, b5, c5	67.44
80	0.8	b2, *sill*, b5, c3, *nugget*, b3	67.44
80	0.9	b3, b5, b2, *nugget*, c2, a5	60.47
80	1	b3, b2, b5, *nugget*, c2, c3	69.77

Table B.31: RDT results for Standardised Gray DGVMPT data along angle 80°

RDT RESULT FOR STANDARDISED GRAY GVMPT ALONG ANGLE 85°			
ANGLE	k	REDUCT	ACCURACY(%)
85	0.1	*nugget*, b4, *sill*	27.91
85	0.2	*nugget*, b2, b3	27.91
85	0.3	b2, *nugget*, b5,*sill*	41.86
85	0.4	b2, *nugget*, b3, *sill*	48.84
85	0.5	b3, b2, *nugget*,b5,c4	48.84
85	0.6	b3, b2, *nugget*, b5, c3, *sill*	41.86
85	0.7	b3, *sill*, b4, *nugget*, b5, b2	67.44
85	0.8	*nugget*, b2, b4, *sill*, b3,c3	60.47
85	0.9	b2, *nugget*, b3, b5, c3, b4, *sill*	58.14
85	1	Data is not adequate	0

Table B.32: RDT results for Standardised Gray DGVMPT data along angle 85°

References

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

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

[3] "http://forensicfact.wordpress.com."

[4] M. M. Houck and J. A. Siegel, *Fundamentals of Forensic Science.* Elsevier, 2010.

[5] R. Fridell, *Forensic Science.* Lerners publication company, 2007.

[6] M. Newton, *The Encyclopedia of Crime Scene Investigation.* Checkmark Books, 2008.

[7] P. Kirk, *Crime Investigation, 2nd Edition*, J. I. Thornton, Ed. New York: Wiley and Sons, 1974.

[8] K. M. Koppenhaver, *Forensic Document Examination-Principle and Practice.* Humanpress, 2007.

[9] A. Agarwal, C. Bhagvati, R. K. Jain, and M. S. Rao, "Computer based Decipherment of Obliterations in Questioned Documents," *Proceeding of AICTE-ISTE Second National Conference on Document Analysis and Recognition,*, pp. 1–8, July 2003.

[10] 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*, pp. 660–664, 2005.

[11] V. Conotter, G. Boato, and H. Farid, "Detecting Photo Manipulation on Signs and Billboards," *International Conference on Image Processing*, pp. 1741–1744, 2010.

[12] E. Kee and H. Farid, "Digital Image Authentication from Thumbnails," *SPIE Symposium on Electronic Imaging*, 2010. [Online]. Available: www.cs.dartmouth.edu/farid/publications/spie10a.html

[13] D. E. Bicknell and G. M. Laporte, *Forged and Counterfeit Documents* . Wiley Encyclopedia of Forensic Science, 2009.

[14] "http://www.fosterfreeman.com."

[15] "http://www.docexam.co.uk/esda.html."

[16] Y. Ramadevi, C. R. Rao, and V. Reddy, "Decision Tree Induction using Roughset Theory Comparative Study," *Journal of Theoretical and Applied Information Technology*, pp. 110–114, 2007.

[17] D. E. Ilea and P. F. Whelan, "Color Image segmentation using A Self-Initialization EM algorithm," *Proceedings of Sixth IASTED International Conference*, pp. 417–424, Aug 2006.

[18] S. P. Day and S. Ford, "Hard Evidence from Computers," *European Conference on Security and Detection*, pp. 19–20, 1997.

[19] G. K. Starkweather, "Electronic Color Printing Technology," *IEEE Proceedings of COMPCON '96-41st IEEE International Computer Conference*, pp. 435–439, 1996.

[20] "http://www.tpr.com/electrophotography.htm."

[21] "http://www.certiguide.com."

[22] "http://www.inkjetworkshop.com."

[23] "http://en.wikipedia.org/wiki/Photographic_printing."

[24] G. Unal, G. Sharma, and R. Eschbach, "Efficient Classification of Scanned Media using Spatial Statistics," *International Conference on Image Processing*, pp. 2395–2398, 2004.

[25] C. M. Hains, S. Wang, and K. T. Knox, *Digital Color Halftones:In Digital Color Imaging Hand Book*. G. Sharma Ed,CRC Press, 2003.

[26] M. Umadevi, C. R. Rao, and A. Agarwal, "A Survey of Image Processing Techniques for Identification of Printing Technology in Document Forensic Perspective ," *IJCA,Special Issue on RTIPPR*, pp. 9–15, 2010.

[27] D. Haritha, *Colour Image Processing Techniques for Ink and Toner Analysis in Forensic Document Examination*, University of Hyderabad, 2007.

[28] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Pearson Education Inc, 2002.

[29] A. Hanbury and J. Serra, "Circular Statistics Applied to Colour Images," *Computer VisionCVWW03*, February 2003.

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

[31] H.Tamura, S. Mori, and T. Yamawaki, " Textural Features Corresponding to Visual Perception," *IEEE Transactions on Systems, Man, and Cybernetics, SMC-8*, pp. 460–473, 1978.

[32] K. I. Laws, "Textured Image Segmentation," Ph.D. dissertation, University of Southern California, 1980.

[33] M. Tuceryan and A. K. Jain, *Texture analysis :In The Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing Co, 1998.

[34] M. Sharma and S. Singh, "Evaluation of Texture Methods for Image Analysis," *Seventh Australian and New Zealand Intelligent Information Systems Conference*, pp. 117–121, November 2001.

[35] N. Khanna, A. K. Mikkilineni, A. F. Martone, G. N. Ali, G. T. C. Chiu, J. Allebach, and E. J. Delp, "A Survey of Forensic Characterization

Methods for Physical Devices," *Digital Investigation3s*, pp. s17–s28, 2006.

[36] R. M. Haralick and L. G. Shapiro, *Computer and Robot vision*. Addison-Wesley Publication, 1992.

[37] S. Gooran, *Digital Halftoning*, Department of Science and Technology, Linkoping University, Sweden, 2004.

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

[39] R. Chellappa and S. Chatterjee, "Classification of Textures Using Gaussian Markov Random Fields," *IEEE Transactions on Acoustic, Speech, and Signal Processing, ASSP-33*, pp. 959–963, 1985.

[40] X. Peng, S. Setlur, V. Govindaraju, R. Sitaram, and B. Kiran, "Markov Random Field Based Text Identification from Annotated Machine Printed Documents," *10th International Conference on Document Analysis and Recognition, ICDAR '09*, pp. 431–435, July 2009.

[41] 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*, pp. 1505–1508, 2009.

[42] E. Alpaydin, *The Introduction to Machine Learning*. MIT Press, 2004.

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

[44] 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*, pp. 301–305, Nov 2004.

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

[46] 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*, pp. 35–46, Aug 2008.

[47] W. Deng, Q. Chen, F. Yuan, and Y. Yan, "Printer Identification Based on Distance Transform," *First International Conference on Intelligent Networks and Intelligent Systems, ICINIS*, pp. 565–568, 2008.

[48] Y. Wu, X. Kong, X. You, and Y. Guo, " Printer Forensics Based on Page Document's Geometric Distortion," *ICIP*, pp. 2909–2912, 2009.

[49] C. Y. Wen and C. M. Chou, "Color Image Models and its Applications to Document Examination," *Forensic Science Journal*, pp. 23–32, 2004.

[50] 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*, pp. 497–501, 2007.

[51] 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*, pp. 43–55, 2006.

[52] S. J. Ryu, H. Y. Lee, I. W. Cho, and H. K. Lee, "Document Forgery Detection with SVM Classifier and Image Quality Measures," *PCM '08: Proceedings of the 9th Pacific Rim Conference on Multimedia*, pp. 486–495, 2008.

[53] I. Avcibas, B. Sankur, and K. Sayood, "Statistical Evaluation of Image Quality Measures," *Journal of Electronic Imaging 11(2)*, pp. 206–223, April 2002.

[54] A. M. Eskicioglu and P. S. Fisher, "Image Quality Measures and Their Performance," *IEEE Transaction on Communications, Vol. 43*, pp. 2959–2965, 1995.

[55] P. J. Chiang, A. K. Mikkilineni, R. M. K. 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*, pp. 231–234, 2005.

[56] 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*, pp. 385–396, 2006.

[57] H. Z. Hel-Or, "Copyright Labeling of Printed Images," *Proceedings of IEEE international Conference*, pp. 702–705, 2000.

[58] A. K. Mikkilineni, P. 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*, pp. 65 051U.1–65 051U.11, January 2007.

[59] 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*, pp. 511–515, 2003.

[60] 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.

[61] "http://www.eff.org/issues/printers."

[62] 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*, pp. 75 340A1 – 75 340A12, 2010.

[63] S. Dutta and B. B. Chaudhuri, "Homogenous Region based Color Image Segmentation," *Proceeding of World Congress on Engineering and Computer science*, pp. 1301–1305, october 2009.

[64] S. Ito, M. Yoshioka, S. Omatu, K. Kita, and K. Kugo, "An image segmentation method using histograms and the human characteristics of HSI color space for a scene image," *10th International Symposium on Artificial Life and Robotics*, pp. 6–10, 2006.

[65] D. Mohr and G. Zachmann, "Segmentation of Distinct Homogeneous Color Regions in Images," *Computer Analysis of Images and Patterns*, vol. 4673, pp. 432–440, 2007.

[66] "http://www.goldensoftware.com."

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

[68] 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, pp. 4287–4308, 2004.

[69] T. Cheng and P. Li, "Multivariate Variogram-based Multichannel Image Texture for Image Classification," *IEEE*, pp. 3830–3832, 2005.

[70] A. Hanbury, U. Kandaswamy, and D. A. Adjeroh, "Illumination-Invariant Morphological Texture Classification," *Computational Imaging and Vision*, vol. 30, pp. 377–386, 2005.

[71] A. Jkomulska and K. C. Clarke, "Variogram Derived Measures of Textural Image Classification-Application to Large Scale Vegetation Mapping," *Proceedings of the Third European Conference on Geostatistics for Environmental Applications*, pp. 345–355, 2000.

[72] "Variogram interpretation and modeling," *Mathematical Geology*, 2001.

[73] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[74] L. Polkowski, *Rough Sets:Mathematical Foundations*. Physica-Verlag, 2002.

[75] "http://www.mathworks.com/help/toolbox/images/f8-20792.html."

[76] "http://terpconnect.umd.edu/ toh/spectrum/CurveFitting.html."

[77] L. A. Shalabi, Z. Shaaban, and B. Kasasbeh, "Data Mining: A Preprocessing Engine," *Journal of Computer Science*, pp. 735–739, 2006.

[78] V. Eglin, S. Bres, and H. Emptoz, "Characterization and Classification of Printed Text in a Multiscale Context," *Advances in Pattern Recognition*, 1998.

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

[80] "http://www.world-english.org."

[81] Z. Zhang, C. Chen, J. Sun, and K. L. Chan, "EM Algorithms for Gaussian Mixtures with Split and Merge Operation," *Pattern Recognition*, pp. 1973–1983, Jan 2003.

[82] T. K. Moon, "The Expectation-Maximization Algorithm," *IEEE Signal Processing Magazine*, pp. 47–60, November 1996.

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

[84] "Introduction to Digital Halftoning," *Wasatch Computer Technology, Inc.*, pp. 1–8, 2003.

[85] "http://www.hpl.hp.com/research/isl/halftoning/index.html."

[86] "http://www.cs.umd.edu/ djacobs/CMSC427/Interpolation.pdf."

[87] "http://en.wikipedia.org/wiki/Minimumboundingrectangle."

[88] N. Ye, *The Handbook of Data Mining*. New Jersy: Lawrence Erlbaum Associates Inc, 2003.

[89] "http://en.wikipedia.org/wiki/Cohen's_kappa."

[90] "http://en.wikipedia.org/wiki/Photocopier."

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

[92] "http://forensicdocumentexamination.com."

[93] S. S. Wu, B. Xu, and L. Wang, "Urban Land-use Classification Using Variogram-based Analysis with an Aerial Photograph," *Photogrammetric Engineering Remote Sensing*, vol. 72, pp. 813–822, July 2006.

Synopsis of the thesis

Identification of Printing Techniques: A Document Forensics Approach

Submitted by

Maramreddi Umadevi

(Reg. No. 05MCPC02)

Under the supervision of

Dr. Arun Agarwal and Dr. C. Raghavendra Rao

for the award of the degree of

Doctor of Philosophy

in

Computer Science

Department of Computer and Information Sciences

School of Mathematics and Computer & Information Sciences

University of Hyderabad

Hyderabad - 500046

INDIA

November, 2010

1 Introduction

Forensic science is the application of a broad spectrum of sciences to answer questions of interest to a legal system [1]. The discipline divides neatly into halves, like the term that describes it, the word forensic comes from the Latin adjective forensis meaning " of or before forum" and the word science is the collection of systematic methodologies used to increasingly understand physical world [2]. Forensic science applies various aspects of scientific and technological methods in collection of evidence, reconstruction of crime science and provides scientific explanation of evidence such that it convinces courts, both the parties (accused of crime and accuser) and general public.

Evidence collection is the starting step of forensic investigation. Forensic evidence is sometimes known as hard evidence [3] as it never gets confused, it never forgets and it never lies. Tracing evidence is based on the fact that when two objects come into contact with each other they exchange trace evidence, in its simplest form it says that "Every contact leaves a trace". This is basic principle of forensic science known as Locard's exchange principle [4]. It is formulated by Dr. Edmond Locard and other works explains this principle in this way. "Wherever he steps, wherever he touches, whatever he leaves, even without consciousness, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibres 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."[5]. Evidence collection and its analysis uncover the truth to provide justice.

Within forensic science there are number of individual disciplines. Criminalis-

tics, Forensic anthropology, Computational forensics, Forensic Chemistry, Forensic Botany, Forensic DNA analysis, Forensic Serology, Forensic Toxicology, Digital forensics, Forensic device forensics, Forensic document examinations are some areas, which comes under the Forensic science. Each discipline has its own technological methods for tracing the evidence.

2 Forensic Document Examination

Document Examination popularly known as Questioned Document examination, is a notable part of forensic science that is developed directly from the need of court experts to answer problems regarding documents instead of growing out of established fields of science.

Questioned Documents are documents whose authenticity is in doubt that is documents which are suspected of being fraudulent or whose source is unknown. All the questioned documents may not be fraud documents. A questioned document may be genuine. They may be partially or completely faked by obliterating, erasing or altering the content in the document. A partially fake documents are produced by adding or erasing the content of document by which meaning of the document is changed. Completely faked documents are counterfeit currency, bus tickets, lottery tickets. Documents of less individual value like bus tickets, bus pass can reproduced using simple and cheaper production techniques.

Document forensic has emerged as new field for assisting interpretation of evidence in courts. Hence the document forensics deals with getting evidence from the questioned documents.

2.1 Category of Documents

These questioned documents are generally paper based documents and are classified as handwritten, typed, printed and photocopied documents based on the technology used to produce them.

Hand written documents are produced by individuals using writing instruments and materials. Many styles are used in handwritten documents over centuries. Writing instruments used in hand written documents are pens, pencils, inks and marker. Pens are available in various models like ball point, gel, porous tip pen, roller or fountain pens. Ball point pens marking tip has a small freely rotating ball bearing that rolls the ink onto the paper. Many of these pens use highly viscous inks. Porous tip pen has porous writing point, which spread fluid ink on the paper. The ink tends towards intense colours and these pens deliver heavy and quick drying stroke [6]. These strokes are distinctive and can be distinguished from strokes of ball point and fountain pens. Roller pens are ball point pens which use fluid ink. Differences in inks distinguishes it from ball point pens. Fountain pens are mostly known as nib pointed pens and their writing characteristics varies. The width of the nib point and its degree of flexibility are factors in this variation. Writing ink is liquid used to produce writing on the surface of paper and its colour pigment gives colour to the writing. Inks can be classified as viscous or non-viscous inks.

Typed documents are produced using typing machine. Typed documents varies depending on the typefaces of the machine. Each type machine has type letters particular to that machine. They will have specific physical feature depending on the make of the machine.

Printed documents are produced using various printing mechanism. Printed documents contain features of a printer depending on the specified procedure used by them for placing the marking material on the paper and inks or toners used in that process. They differ in the print pattern, number of drops per dot and

technology used to print like Drop on demand thermal printing e.g., Hp photo REt [7] technology, Laser technology etc. Now a days various printing instrument commercially available in market are inkjet printer, laserjet printer and offset printers. Toners used in the printer can be classified as liquid or powdered toners.

Photocopied documents are produced by copying original printed or handwritten documents by electro photography procedure. Photocopied documents looks similar as printed document to the untrained eye. Photocopiers are machines, which produce documents similar to the original documents. Photocopiers are electrostatic machines like laser printers except in producing grid pattern in document. Most of forged documents like counterfeit currency are produced using colour photocopiers or colour printers.

2.2 Document Forensics and its Scope

The scope of document forensics in different document problems is listed below.

1. Identification of handwritten documents
2. Identification of forged documents
3. Identification of typewriter
4. Deciphering obliterations, alterations and erasures
5. Identification of inks and writing instruments
6. Printer identification of the document
7. Photographic tampering

3 Motivation of Our Work

The technology used to produce documents continues to evolve; the methods used to produce forgeries are ever more sophisticated; the expectations of lawyers and

courts are yet more demanding. When the document's legitimacy is in question, methods are needed to non intrusively analyse distinguishing features of document in order to learn more about its origin. The knowledge of spatial pattern produced by various print technologies are helpful in determining the source device that produced these pattern.

Printed document is spatial distribution of marking material. A printer produces a document to the extent to 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 for tampering to produce fraudulent document, which may be printed by the same or another printer. Hence a forged document will posses composite features of the above processes. Thus printer identification of the questioned document is highly involved and complex process. A printed questioned document examination, by forensic scientist starts with identifying the printer or source from which document has been created. Printed documents contain features of a printer depending on the specified procedure used by them for placing the marking material on the paper.

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

1. Is the document consistent, implying whether the content printed in the document is prepared from a single source?

2. Identification of source printer or printing techniques like ink jet, electrophotography printing etc.

3. Are two documents similar, i.e., printed using same technology or printer?

4. Differentiate between photographic and photorealistic images

Instruments used by document examiner to distinguish genuine document from forged are high resolution microscopes, Electro Static Detection Apparatus(ESDA)

and Video Spectral Comparator(VSC) [8]. ESDA is used to reveal indented impression on paper and it is non destructive technique [9]. VSC is multi spectral 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.

As the current methods and instruments used are expensive in capturing the data as well as in analysing, there is a great need to develop alternative solutions for forensic identification of print technology that is most effective in cost, space and time.

4 Literature Survey

Recent research demonstrate various approaches suggested for discriminating printing techniques. As document is combination of pictures and text, suggested approaches are based on region which may be homogeneous region of picture, text or embedded region.

4.1 Ink and Toner Analysis

As printed document is a spatial distribution of marking material to the paper using printing mechanism, ink and toner analysis in [10], [11], [12], determines the nature of ink/toner and printing process by which ink or toner is transferred to the paper. The absorption characteristics of ink/toner and paper form the basis for classification of inks/toners . Sobel edge detected image captures ink spray

characteristics and help to distinguish inkjet printers and photocopiers. This is a colour image processing technique based on HVS colour space of printed images.

4.2 Techniques Based on Texture of Printed Document

The distribution of marking material in the printed document differs from one printer to other printer based on print technology employed. The source printing technology influences the features of the document. By analysing texture pattern in the document such as whether it contains smoothed or roughed text and direction of texture reveals the source printing mechanism. Various texture models [13] for solving different document problems are addressed [14], [15], [16], [17] here.

In [14], Electro-photography printing fluctuations in the developed toner on the printed page are characterized by statistical modelling of texture. In particular GLCM [18] is employed by considering the most frequently occurring character in English language the "e". Gray level co-occurrence matrix contains overall spatial relation ships among pixels in the image and it is unable to capture features like variability at different distances, periodicity and size of texture. Spatial organization of texture is addressed by geometrical methods and are exploited in [15] by confining the feature, angle of halftone texture for the image in each channel of CMYK. Model based approaches are employed in [16], which models geometric distortion of printed page and it is taken as intrinsic feature of printer for identification of printing technique. Signal processing models like Discrete wavelet analysis [19] is employed for identification of colour laser printers [17]. Feature are extracted from each image are trained using SVM [20] classifier to identify the brand and model of source printer. These characterizes exclusively Electrophotography printers which are known as laser printers.

4.3 Techniques based on region and Text

a. Uniform region based approach: Identification of Electro-photographic printers is based on frequency analysis of banding signal in large mid tone area [21]. 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 percent filled gray level patches. This is based on purely electromechanical fluctuations in laser printers. In [22], features of colour noise and GLCM features of original image are used for identification of colour laser printers.

b. Text based approach: As it is difficult to find large mid tone gray regions in the text, techniques are implemented for identification of print technology based on text. Gray level co-occurrence feature of the most frequently occurring letter 'e' and Gaussian mixture model (GMM) [14], [23] are the techniques used for printer identification. These researches are exclusively for the identification of Electro-photographic printers.

Gray level features are proposed in [24] for discriminating inkjet from laserjet print, which is based on high resolution scanned images, e.g. 3200 dpi. Evaluation of gray level features like perimeter based edge roughness of the text [25] for print technique classification is based on low resolution image. This is developed for high throughput document management system not for forensic perspective.

In [26], Electro-photographic printer identification based on character matching is explained using distance transform measures. It is based on the fact shape of printed character is stable compared to print quality and it presents a macro printing style. Identification and linking fake documents to scanner in [27] proposed new method for identifying fraudulent documents and linking it to colour laserjet printer or colour inkjet printer. Unique colour count and texture feature uniformity and intensity variation are used as parameters for distinguishing fraud documents. In this proposed work they captured images of text using high resolution cameras LEICA MZ 8, LECIA MZ 12.5.

The works mentioned in [14], [23], [24], [25], [26], [27] identifies the print technology based on text only.

Hybrid techniques like Image quality measures[28] are applied to distinguish between genuine and fake document [29]. It has shown low accuracy when fraud documents generated on both laser and inkjet are grouped together for training.

4.4 Techniques based on Embedding Information

Methods of embedding watermark [30] in the printing image taking the advantage of printing process are now discussed. Encoding of the watermark is performed using halftoning techniques which uses set of dither cells. Machine identification code project by Electronic Frontier Foundation [31] identifies presence of pattern of yellow dots in colour laser printouts and these dots reveals information like printer serial number. This is not applicable to all Electro-photographic printers as some printers do not show the presence of these yellow dots. Laser amplitude modulation is used in Electro-photographic printer [32] to embed the information in the text of document. In [33], [34], [35], two strategies are proposed for printer identification. One of those strategies is passive which characterizes the printer by finding the intrinsic features in printed document. Another strategy is active which embeds the information in the printed text by modulating the process parameters in the printing mechanism.

Content Integrity of Printed Documents(CIPDEC) using Error Correction [36] detects any modifications in document without requiring original document for such detection. Error correcting code corrects the document and it reveals the parts of the document that were tampered.

In summary it is seen that literature survey [14], [15], [17], [21], [23], [33], [34], [35] characterizes laser printers only. The embedding techniques [30],[31], [32], [33], [34] [35], [36] are useful for only few group of printers for identification. In [24] discrimination of inkjet from laserjet is based on high resolution scanned image at

3200 dpi. In [11], [12], identification of inkjet , laserjet, photocopy is based on HSV colour space. Hence there is need for forensic techniques which identifies different print technology like inkjet print, laserjet print and photocopy using graylevel data at moderate resolution.

5 Contributions

Document forensics is mostly off line activity but sometimes can be soft real time activity. To give an apt solution the technology has to be built by considering the limitations of equipment at investigation location and time. Some times the data available for analysis will be imprecise as well as partial due to the limited resolution. To address these problems the thesis develops various methods based on computing tools by considering the scanned image of moderate resolution. The contributions can be categorized as shown below.

1. **Identifying print technology of source document based on uniform colour region or homogeneous colour region:** Developed a Gaussian Variogram Model (GVM) model, for identifying the print technology which produced the given document. This method characterizes print technology based on spatial variability. Homogeneous colour region of images are taken as samples for the GVM data generation. The generated GVM data is taken as input to generate Reduct based Decision Tree(RDT) [37], which gives rules to identify the source printer for the given test data. Performance analysis of the model is also presented. Developed method assists the document examiner in finding basic print pattern of printers and it is also helpful in classifying different print technology.

2. **Identifying print technology of source document based on printed text:** We have formulated a Print Index for classification inkjet printed text versus laser jet printed text. This work focuses on frequently used

word like 'the' as test sample for characterizing printed text. The novelty of the proposed approach is that the selected printed text is modelled as mixture of three Gaussian models namely text, noise and background. The associated patterns and features of the models are derived using Expectation Maximization(EM) algorithm [38] and few indices are proposed based on these parameters. One of the indices called Print Index(PI) for text is used for basic print technology discrimination. EM algorithm is also used as dimension reduction technique to characterize printed text.

3. **Differentiation of inkjet print from its photocopy:** Statistical measures skewness and kurtosis of histogram are selected as features for distinguishing inkjet print from its photocopy.

4. **Identification of tampered document:** Questioned document in general may be printed by various printer. Thus identification of tampered document involves

   a. Identification of tampered region Sliding window protocols are proposed and demonstrated for the purpose of identifying tampered document and then the nature of tamper. It involves moving non-overlapping fixed size window for generating variogram to reveal various texture or spatial pattern underlying the sample.

   b. Identification of combination of print technology Questioned document may be document having mixing print technology. Document is reproduced by scanning original document and then reprinting on another printer. In this way the reproduced document contains the features of both previous print as well as the present printing characteristics. Window-wise analysis of variogram with varying window size captures mixed/combination of print pattern existing in the documents. Results for varying window size

also demonstrated.

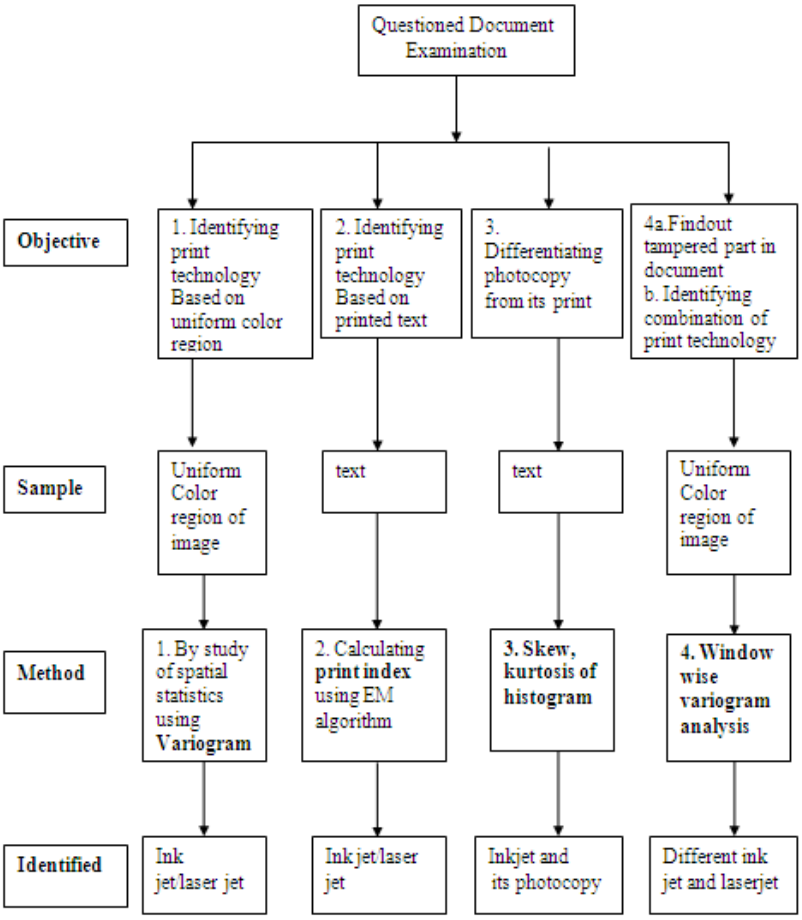


Figure 1: Problems identified and proposed methodology

The above contributions are depicted in Figure 1. Once the region of interest is identified then any one of the above suggested techniques can be applied. The present thesis demonstrates our hybrid system for addressing general problems of the questioned documents which is given in Figure 2. The proof of concept of proposed methodologies has been demonstrated by appropriate experimentation. The following section presents design of experiments and data collection.

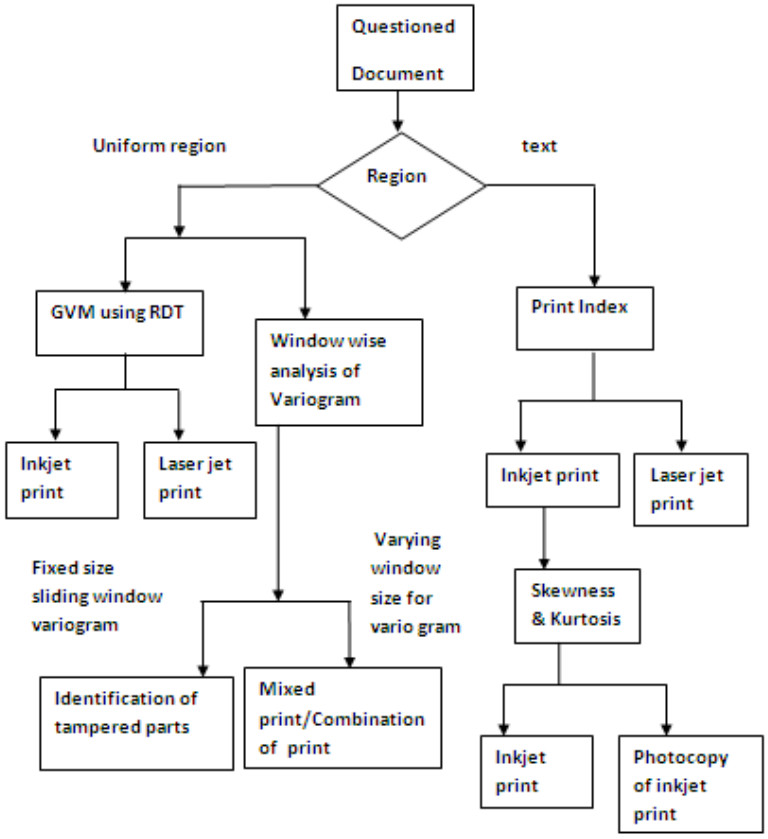


Figure 2: System Overview

5.1 Preparation of Samples

Samples are prepared to build the models and demonstrating the four problems addressed in the thesis.

Type 1: 28 color images are selected and converted to graylevel or normalized graylevel for carrying out region based printer identification.

Type 2: Three types of graylevel text documents are considered for printer identification based on text. 2000 samples are selected from these text documents.

Type 3: Three types of graylevel text documents are considered for discriminating inkjet print from its photocopy.

S.no	Manu-facturer	Model	Print technology
1	HP	Deskjet930c	HP Photo REt III
2	HP	Deskjet840c	Drop-on-demand thermal inkjet
3	HP	Hppsc1608	Drop-on-demand thermal inkjet
4	HP	Hppsc1315	Drop-on-demand thermal inkjet
5	HP	Officejet6110	HP PhotoREt III
6	HP	Photosmart3188	Drop-on demand thermalinkjet
7	HP	Laser 4650	Colourlaserjet
8	HP	Colourlaserjet4550N	Colourlaserjet
9	HP	hplaser1200	Laserjet
10	HP	hplaser 9040	Black and white laser
11	SAMSUNG	CLP-510	Colourlaserjet
12	SAMSUNG	samsungml2010	Laserjet
13	XEROX	xeroxwcpe220	Black and white laser

Table 1: Printers used for producing samples

Type 4: Few images printed on various printers and the printouts are scanned.

a: Mixing of scanned images from different printer of a image are used for producing tampered image.

b: Scanned image of print of a printer is printed on another different printer for producing document with hierarchal mixing print technology.

All the above type of documents are printed using various printers listed in Table 1 at 600 dpi and scanned at 2400 dpi using Hpscanjet for creating source files for analysis. Photocopiers used are listed in Table 2. The respective data sets are divided into two sets, first set is used for building the models (that is training set) and the second for validating the methodology (that is testing set).

S.no	Photocopier
1	Konika Minolta bizhub210
2	Xeroxwcm118

Table 2: Photocopiers used for Producing Type 3 samples

6 Organization of the thesis

The thesis chapters are organized as follows. Chapter 1 introduces Forensic Science, areas of forensic science with emphasis on Document Examination and its

scope. It describes instruments used to prepare document tools used by forensic examiner for examination of documents. It also presents survey of various print technologies used to produce documents. Chapter 2 presents previous work done in identification of printed documents. Literature survey is categorized as techniques based on ink and toner analysis, texture, region based techniques which involves uniform colour region or text and techniques of embedding information. Chapter 3 describes identification of print technology based on uniform colour region of image using Gaussian variogram Models with Rough sets for classifying inkjet versus laserjets print. Chapter 4 explains Identification of print technology based on printed text which characterizes text using Expectation maximization algorithm. Differentiation of inkjet print from photocopy print using the histogram features i.e., skew and kurtosis are developed and demonstrated. Chapter 5 addresses window wise variogram analysis for interpretation of the tampered documents and identification of tampered parts in the document. Chapter 6 concludes the thesis with future directions.

Acknowledgments

This research work is supported by Government Examiner of Questioned Documents(GEQD), Hyderabad, under the Fellowship of Directorate of Forensic Science Ref.No. 87(1)/GEH/JRF/SRF for the period of Dt: 28-02-2005 to 27-02-2009.

7 Papers Published

1. **'Gaussian Variogram Model for Printing Technology Identification'**, International Conference on Asian Modeling Symposium, pp 320-325, 2009.

2. **'A Survey of Image Processing Techniques for Identification of Printing Technology in Document Forensic Perspective'** IJCA, Special Issue on RTIPPR(1), pp 9-15, 2010.

References

[1] http://forensicfact.wordpress.com.

[2] M. M. Houck and J. A. Siegel. *Fundamentals of Forensic Science*. Elsevier, 2010.

[3] R. Fridell. *Forensic Science*. Lerners publication company, 2007.

[4] M. Newton. *The Encyclopedia of Crime Scene Investigation*. Checkmark Books, 2008.

[5] P. Kirk. *Crime Investigation, 2nd Edition*. Wiley and Sons, New York, 1974.

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

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

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

[9] http://www.docexam.co.uk/esda.html.

[10] D. Haritha. *Colour Image Processing Techniques for Ink and Toner Analysis in Forensic Document Examination*. Ph. D. Thesis, University of Hyderabad, 2007.

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

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

[13] M. Tuceryan and A. K. Jain. *Texture analysis :In The Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing Co, 1998.

[14] 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. In *Proceedings of the SPIE International Conference on Security, Volume 5681*, pages 430–440, Mar 2005.

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

[16] Y. Wu, X. Kong, X. You, and Y. Guo. Printer Forensics Based on Page Document's Geometric Distortion. In *ICIP*, pages 2909–2912, 2009.

[17] 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 . In *ICIP*, pages 1505–1508, 2009.

[18] R. Haralick and L. G. Shapiro. *Computer and Robot vision*. Addison-Wesley Publication, 1992.

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

[20] E. Alpaydin. *The Introduction to Machine Learning*. MIT Press, 2004.

[21] N. Khanna, A. K. Mikkilineni, A. F. Martone, G. N. Ali, G. T. C. Chiu, J. Allebach, and E. J. Delp. A Survey of Forensic Characterization Methods for Physical Devices. In *Digital Investigation3s*, pages s17–s28, 2006.

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

[23] G. N. Ali, P. J. Chiang, A. K. Mikkilineni, Chiu G.T.-C, E. J. Delp, and J. P. Allebach. 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, Volume 20*, pages 301–305, Nov 2004.

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

[25] C. Schulze, M. Schreyer, A. Stahl, and T. 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.

[26] W. Deng, Q. Chen, F. Yuan, and Y.Yan. Printer Identification Based on Distance Transform. In *First International Conference on Intelligent Networks and Intelligent Systems, ICINIS*, pages 565–568, 2008.

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

[28] I. Avcibas, B. Sankur, and K. Sayood. Statistical Evaluation of Image Quality Measures. In *Journal of Electronic Imaging 11(2)*, pages 206–223, April 2002.

[29] S. J. Ryu, H. Y. Lee, I. W. Cho, and H. K. Lee. Document Forgery Detection with SVM Classifier and Image Quality Measures. In *PCM '08: Proceedings of the 9th Pacific Rim Conference on Multimedia*, pages 486–495, 2008.

[30] H. Z. Hel-Or. Copyright Labelling Of Printed Images. In *Proceedings of IEEE international Conference*, pages 702–705, 2000.

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

[32] 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. In *Proceedings of SPIE International Conference on Security , Stegnography and*

Watermarking of multimedia contents IX, Vol 6505, pages 65051U.1–65051U.11, January 2007.

[33] 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. In *Proc. IST's NIP21: International Conference on Digital Printing Technologies, vol. 21*, pages 231–234, 2005.

[34] 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. In *Proc. SPIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents VIII, Vol. 6072*, pages 385–396, 2006.

[35] 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. In *Proc. IST's NIP19: International Conference on Digital Printing Technologies,Vol.19*, pages 511–515, 2003.

[36] Y. S. subramaniam, B. Narayanan, K. Viswanathan, and K. Anjaneyulu. Detecting Modifications in Paper Documents: A Coding Approach. In *Document Recognition and Retrieval XVII, Proc. of SPIE-IST Electronic Imaging, SPIE Vol. 7534*, pages 75340A1 – 75340A12, 2010.

[37] Y. Ramadevi, C. R. Rao, and V. Reddy. Decision Tree Induction using Roughset Theory Comparative Study. In *Journal of Theoretical and Applied Information Technology*, pages 110–114, 2007.

[38] D. E. Ilea and P. F. Whelan. Color Image segmentation using A Self-Initialization EM algorithm. In *Proceeding of Sixth IASTED International Conference*, pages 417–424, Aug 2006.