Intelligent Printing Technique Recognition and Photocopy Detection Using Digital Image Analysis for Forensic Document Examination

Diploma Thesis

Marco Schreyer

submitted to the

Department of Computer Science IV
Prof. Dr.-Ing. Wolfgang Effelsberg
Faculty of Mathematics and Computer Science
University of Mannheim

and

Image Understanding and Pattern Recognition Research Group
German Research Center for Artifical Intelligence

November 11^{th}, 2008

Supervisor: Dipl.-Ing. Christian Schulze

Hiermit versichere ich, die vorliegende Diplomarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen die aus den Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mannheim, 11. November 2008

(Marco Schreyer)

Danksagung

Den Menschen, die mich während der Entstehung dieser Arbeit unterstützten:

- Prof. Dr.-Ing. Wolfgang Effelsberg, der mir seitens der Universität Mannheim diese Diplomarbeit in Kooperation mit dem Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) ermöglichte.
- Prof. Dr. Thomas Breuel, für die Gelegenheit diese Diplomarbeit im Forschungsbereich "Bildverstehen und Mustererkennung" des DFKI zu schreiben.
- Dr. Armin Stahl und Dr. Daniel Keysers, die mich mit ihrer Weitsicht und Begeisterung zu dieser Thematik motivierten.
- Dipl.-Ing. Christian Schulze, für die Fülle an Denkanstößen, Gesprächen und vielen Hilfestellungen die diese Diplomarbeit erst ermöglichten.
- Den Wissenschaftlern, Angestellten und Diplomanden des Forschungsbereichs "Bildverstehen und Mustererkennung" die es für mich zu einem besonderen Erlebnis machten Teil dieses Teams zu sein.
- Dipl.-Psych. Elisabeth Hoffmann, die mich und diese Diplomarbeit um die Perspektive einer Schriftsachverständigen und mit Fachliteratur bereicherte.
- Damian, der mir als Freund mit unermüdlicher Unterstützung zur Seite stand.
- Andreas, der mir trotz großer Entfernung eine besondere Hilfe bei der Korrektur dieser Diplomarbeit war.
- Monika und meinem Bruder Patrick.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Thesis Objective 5

1.3 Thesis Constraints 6

1.4 Thesis Outline 7

1.5 Notation 8

2 State of the Art in Forensic Document Examination **9**

2.1 Forensic Document Examination 9

2.2 Characteristics of Forensic Document Examination 13

2.3 Related Work 16

3 Printing Processes and their Characteristics **21**

3.1 Printing Technologies 21

3.2 Printing Characteristics 28

4 Intelligent Forensic Document Examination **35**

4.1 Pattern Classification Systems 35

4.2 Machine Learning Strategies 37

4.3 Pattern Classifier Design 38

4.4 Process for Intelligent Forensic Document Examination 39

5 Document Preprocessing and Character Segmentation **43**

5.1 Connected Components 43

5.2 Document Binarization 44

5.3 Document Segmentation 47

6 Feature Identification and Extraction **51**

6.1 Preliminary Document Image Examination 51

6.2 Features for Printing Technique Recognition 54

7 Classification Results and Hypotheses Evaluation **75**
7.1 Experimental Setup 75
7.2 Performance Measures 77
7.3 Parameter Optimization 79
7.4 Evaluation of Global Document Examination 80
7.5 Evaluation of Local Document Examination 84

8 Conclusion and Future Work **93**
8.1 Future Work 95

A The Evolution in Document Forensics **97**
A.1 The Technological Evolution 97
A.2 The Statutory Evolution 98
A.3 Concluding Observations 99

B Pattern Classification Techniques **101**
B.1 No Free Lunch Theorem 101
B.2 Decision Trees 102
B.3 Artificial Neural Networks 106
B.4 Support Vector Machines 112

C Experimental Results **119**
C.1 Global Feature Evaluation Results 119
C.2 Local Feature Evaluation Results 120

D Template Document **123**

Bibliography **125**

List of Figures

1.1 The Fraud Triangle 2
1.2 Photocopy warning sign 3
1.3 Exemplary printing technique variations 4
1.4 High resolution printing technique comparison 5
1.5 'Document Verification Suite' Architecture 7

2.1 Overview of forensic document examination techniques. 11

3.1 Taxonomy of different printing technologies 22
3.2 The inkjet printing process 24
3.3 The electrophotgraphic printing process 25
3.4 Surface plots of inkjet printed characters 28
3.5 Typical inkjet printing defects 29
3.6 Surface plots of electrophotographic printed characters 30
3.7 Photocopy image degradation 32
3.8 Photocopy document image noise 33
3.9 Typical traces of physical printer properties 34

4.1 Pattern classification process 36
4.2 Classifier design cycle 38
4.3 Intelligent Printing technique recognition process 40

5.1 Graylevel histograms of distinct document classes 45
5.2 Document image binarization 47
5.3 Connected component detection 49
5.4 Connected component extraction 50

6.1 Representative document image extracts 52

6.2 Gaussian Filter based denoising 55
6.3 Median filter based denoising 56
6.4 Gradient based printing technique comparison 59
6.5 The first 64 basis functions of the Discrete Cosine Transformation (DCT). . 61
6.6 Exemplary Discrete Cosine Transformation 62
6.7 Frequency domain printing technique comparison 64
6.8 Horizontal and vertical frequency subband analysis 65
6.9 Exemplary Multiresolution Analysis (MRA) 70

7.1 Feature parameter optimization results 79
7.2 Decision tree global feature evaluation 80
7.3 Multilayer Perceptron global feature evaluation 82
7.4 Support Vector Machine global feature evaluation 83
7.5 Local feature evaluation mean accuracy rates of laser and inkjet printed documents 85
7.6 Accuracy rate distribution of local feature evaluation on the basis of laser and inkjet printed documents 86
7.7 Local feature evaluation mean accuracy rates of all document classes 87
7.8 Accuracy rate distribution of local feature evaluation on the basis of all document classes 88
7.9 Local feature evaluation mean accuracy rates of laser printed and photocopied documents 89
7.10 Accuracy rate distribution of local feature evaluation on the basis of laser printed and photocopied documents 90

B.1 Exemplary decision tree and decision boundaries 103
B.2 Exemplary perceptron and logistic sigmoid functions 107
B.3 Exemplary multilayer perceptron and error surface 111
B.4 Linear support vector classification 113
B.5 Non-linear support vector classification 115

D.1 'Grünert' letter 124

List of Tables

1.1 Symbols used throughout this work. 8

5.1 Parameters used in document character segmentation 50

7.1 C4.5 Decision Tree global features evaluation results 81

7.2 Multilayer Perceptron global features evaluation results 82

7.3 Support Vector Machine global features evaluation results 83

7.4 Exemplary confusion matrix of three class based local feature evaluation . . 88

8.1 Summary of global feature evaluation 94

B.1 Common kernel functions used in support vector classification 116

C.1 Global results obtained by C4.5 Decision Tree 119

C.2 Global results obtained by Multilayer Perceptron 120

C.3 Global results obtained by Support Vector Machine 120

C.4 Local accuracy results with C4.5 Decision Tree on all document types . . . 120

C.5 Local results with Multilayer Perceptron on all document types 121

C.6 Local results with C4.5 Decision Tree laser prints and photocopies 121

C.7 Local results with Multilayer Perceptron laser prints and photocopies 121

C.8 Local results with C4.5 Decision Tree laser and inkjet prints 122

C.9 Local results with Multilayer Perceptron laser and inkjet prints 122

1. Introduction

1.1 Motivation

The history of expressing mankinds thoughts in printed form is equivalent to the story of changes in technology. With almost every major technological advance there has been a corresponding change in the way we record and store our ideas. Within 40 years mankind has gone from typewriters to the digital age [1]. In this context the advent of digital printing and imaging technologies had a tremendous impact on the way we generate, publish and store information nowadays.

This technological progress, as more and more applicable, is not only used for legitimate purposes but also for illegal activities. Recent cases reported to the American Society of Questioned Document Examiners (ASQDE) revealed the increasing involvement of modern printing technologies in the production of counterfeited banknotes [2, 3] and forged documents [4, 5]. According to a survey issued by the Association of Certified Fraud Examiners (ACFE) United States companies lose 650 billion dollars each year due to fraud caused by counterfeiting. Especially, document fraud is making up more than two-thirds of that[1].

In response to the accounting scandals such as Enron, WorldCom and Adelphia the American Institute of Certified Public Accountants (AICPA) introduced the concept of the so called "Fraud Triangle" [6]. The Fraud Triangle as shown in Figure 1.1 is describing three major factors enabling people to commit fraud, namely "Incentive/Pressure", "Attitude/Rationalization" and "Opportunity" as explained in the following:

[1] The Association of Certified Fraud Examiners's "2006 Report to the Nation on Occupational Fraud and Abuse" is available online at `http://www.acfe.com/documents/2006-rttn.pdf`

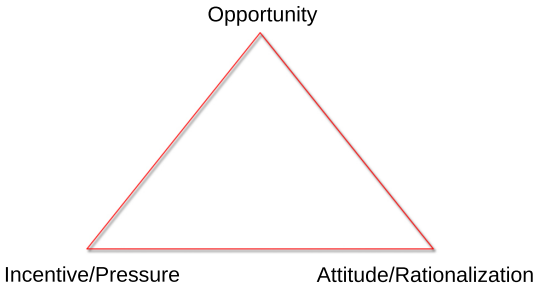


Figure 1.1: The causal factors of fraud “Incentive/Pressure”, “Attitude/Rationalization” and “Opportunity” illustrated in the “Fraud Triangle”. At least one of the three factors causes and is involved in every situation of fraud.

- **Incentive/Pressure** describes what causes a person to commit fraud. Most of the time, pressure results from a significant financial problem or need.
- **Attitude/Rationalization** depicts the mindset of the fraudster justifying fraud i.e. the ethical principles or values.
- **Opportunity** specifies the situation that enables fraud to occur i.e. caused by the circumstance that controls are weak or nonexistent or the fraudulent activities are difficult to detect.

According to the AICPA these factors play a vital role in nearly every situation of fraud. Cendrowski et al.[7] stated that, breaking the Fraud Triangle implies that at least one of the elements in the triangle has to be confined in order to reduce the likelihood of fraudulent activities.

Observing the constant development in digital imaging and printing technologies that are available for domestic use, it can be agreed that the creation of fraudulent documents is never as comfortable as today. Subsequently, the factor ”Opportunity“ described above as a major reason provoking document fraud is exceeding. Using these technologies, it is possible to create high quality forgeries or altered documents within short timeframes and without noteworthy effort. Figure 1.2 shows a warning sign found at the ”Image Understanding and Pattern Recognition (IUPR) Research Group“ lab photocopier prohibiting to photocopy sensitive documents like money, passports or cheques.

The genuineness of documents has always been a critical issue in history. In ancient times documents were sealed by monarchs and dignitaries using unique signet rings to prevent forgery [8]. In analogy much effort has been undertaken since the introduction of digital printing technologies to maintain the authenticity of documents and protect them against various counterfeiting attempts. The well known techniques are, for example, the usage of particular paper types, holographic images, specific inks and printing technologies or

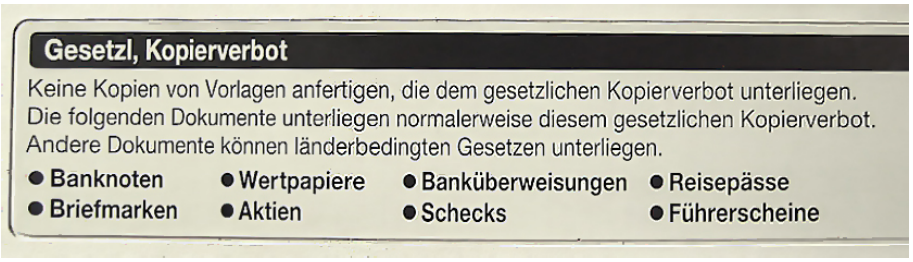


Figure 1.2: Warning sign found at the Image Understanding and Pattern Recognition Group lab photocopier prohibiting to photocopy sensitive documents, for example, money, passports or cheques.

physical or chemical signatures [9]. Nevertheless, the usage of such sophisticated security features is often limited to particular federal or financial documents like passports, banknotes or highly confidential contracts.

There are two reasons for this, which concern organizational as well as financial aspects. Firstly, within applications that handle governmental or business processes these techniques are impractical. Since in such processes typically different parties are involved issuing different document types that are printed on multiple devices. Secondly, the majority of the above mentioned techniques are enhancements of a document's paper or printing process. Hence, most of these techniques either require special equipment or materials to embed security features and are therefore simply too expensive for the average consumer [10]. As a result there is an urgent need for novel document examination techniques.

On the other hand the detection of documents involved in fraud is becoming increasingly difficult. Especially in the case of banks, insurance companies and tax authorities, processing several thousand documents issued by a high number of invoicing parties each day. Observing, for example, the high amount of processed bills related to payments and assuming that only a small percentage of these are forged or manipulated it is easy to imagine that quite some disprofit could be prevented with the use of authenticity verification systems. In such scenarios there is an urgent need for intelligent methods to determine if a processed document is genuine or manipulated.

As a result, computer-assisted and fully automatic computer-based document analysis and visualization systems are needed to examine fraud and money laundry [11, 12]. Typical examples of forged documents are annual tax declarations often including joined voucher bundles from different invoicing parties. Another example are medical doctor bills sent by patients after a medical or physical treatment to their health insurance company.

In the detection of fraudulent documents important insights in an examination process can be obtained by answering the questions: How was the document at hand created? Is the document an original or a photocopy? The ability to investigate documents for consistence in the used printing technology can be a first useful observation in making the decision if a given document is genuine or faked. The detection if specific document regions are printed

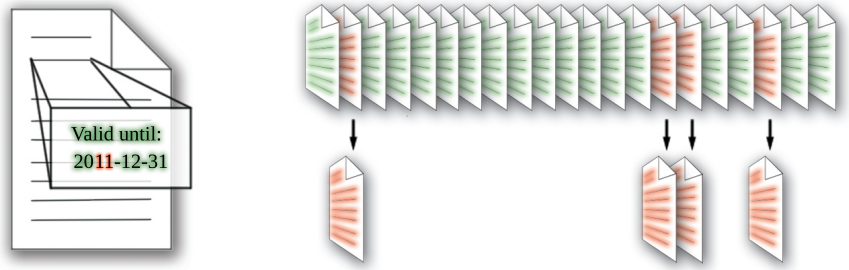


Figure 1.3: Exemplary printing technique variations: (l) document inherent and (r) time inherent printing variation. Both variations could potentially be a first evidence of document tampering and counterfeiting.

using the same non impact printing technique, namely laser or inkjet, is an essential piece of information. Time inherent as well as document inherent variations in the detected printing technology are of particular interest in an investigation of questioned documents and can be valuable trace to detect potential forgeries. Both types of variation will be explained in the following:

- **Document Inherent Printing Technique Variation**
  Depicts the observation of a sudden or random change in the printing technology of a documents printed content. Observed changes can be a trace to malicious document alterations e.g. addition of new content or the cut-and-paste of content from another document. The questions therefore to answer are: Had more than one technology been involved in the creation of the document? And if so, are there comprehensible reasons for this variation? This is especially of interest in the case of sensitive content regions as shown left (l) in Figure 1.3 like billing amounts, account informations, dates, person names or addresses. In this thesis this examination will be referred to as *Local-Document-Examination*.

- **Time Inherent Printing Technique Variation**
  Describes the observation of a sudden or random deflection in the printing technology of a known document type or class within a certain period of time without comprehensible reason. Printing technique comparison of a specific document type over a certain time period is offering the ability to expose counterfeiting activities as illustrated right (r) in Figure 1.3. This approach is in particular useful in the case of large insurance companies or governmental organizations which handle large document quantities originating from similar sources. In this thesis this examination will be referred to as *Global-Document-Examination*.

If a document or a series of documents appears to be inconsistent regarding to the detected printing technology it might be an evidence of forgery. Therefore, a so called ”red

Figure 1.4: High resolution printing technique comparison illustrated by the printed character 'a' scanned at 4800*dpi*: (l) template character, (m) laser and (r) inkjet printed result.

flag“ could be raised requiring a second level inspection to proof the suspected documents authenticity.

1.2 Thesis Objective

Figure 1.4 illustrates a high resolution scan of an inkjet and laser printed character “a” extracted from a regular business correspondence letter. Both printing technologies leave a unique printing signature on the surface of the printing paper. These signatures and its corresponding pattern can be used to recognize the printing technique applied in the creation of the document.

An opportunity to resolve the problem statement outlined above lies therefore in the appliance of techniques derived from digital image processing and statistical pattern recognition. However, the usage of these techniques in forensic document examination is relatively new [13]. In awareness of this situation an application named “Document Verification Suite” is currently in development by the “Image Understanding and Pattern Recognition (IUPR) Research Group” at the “German Research Center for Artificial Intelligence”. An outline of the proposed application architecture and its components is presented in Figure 1.5.

The objective of this diploma thesis is to study the feasibility of detecting potential forgeries by techniques derived from image processing and statistical pattern recognition. Therefore, a potential system to be developed should be able to recognize a document's underlying printing technology and detect documents created by xerographic printing processes known as photocopies. This task is of particular challenge since laser printers and photocopiers are extremely similar in operation.

Derived from this objective the following four hypotheses should be evaluated:

`[Hypothesis 1]`: **Printing technique recognition is achievable at document level.**
`[Hypothesis 2]`: **Photocopy detection is achievable at document level.**
`[Hypothesis 3]`: **Printing technique recognition is achievable at character level.**
`[Hypothesis 4]`: **Photocopy detection is achievable at document character level.**

In order to achieve this, the following sub-tasks have to be successfully resolved:

1. Study of photocopy and printing processes and their characteristics.
2. Identification and extraction of the characters within a document.
3. Extraction of appropriate features for printing technique recognition.
4. Classification of a document characters underlying printing technology.
5. Appropriate visualization of the classification results.

The suitability of the proposed idea to counterfeit document detection is underpinned by forensic examinations on how document forgeries in general are created these days. Only a very small contingent of forgers are professionally equipped with an expensive laboratory of advanced printing technology, for example, offset printing. *"Instead, the vast majority of forgeries are created on ordinary personal computers using commercially available scanners and printers"* [14].

1.3 Thesis Constraints

The "Document Verification Suite" should be employable in scenarios where high throughput document management is required. Therefore, several constraints for the evaluated and developed approaches have to be considered. These are resulting from the application scenario itself as well as from technical document processing limitations and are described in the following:

- **Document Characteristics**

  The system should be able to detect document printing technique variations on the receiver side of document based communications. Incoming documents could be send by a high amount of different document issuing parties. Typically, the receiving side in such a scenario has no influence on particular document characteristics. Therefore, no assumptions concerning document specific properties should be made. The system should provide the ability to cope with different paper, ink, and toner types in a reliable manner.

- **Document Information**

  Another constraint derives from the fact that high throughput scanning devices are in general limited to a maximum scanning resolution of 400dpi nowadays. This limitation is caused by an intention to reduce scanning time as well as the allocated document data storage. Consequently, the developed method should provide the ability to classify a documents underlying printing technology at low resolutions.

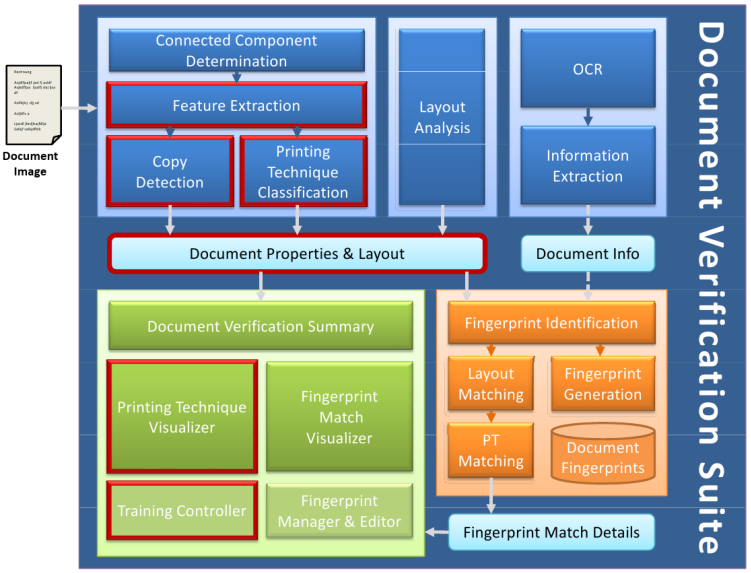


Figure 1.5: Draft version of the proposed "Document Verification Suite" application architecture currently under development by the German Research Center of Artificial Intelligence. Red framed components are within the context of this thesis.

- **Document Capturing**

  Nowadays, illumination technologies regularly applied in the capturing of suspected documents also include the invisible light spectrum i.e. infrared or ultraviolet illumination or more advanced methods like x-ray radioscopy. However, the starting point of examination techniques developed in this thesis should be a document image captured by a commercially available scanning device. Therefore, no assumptions concerning the appliance of sophisticated document capturing technologies should be made.

1.4 Thesis Outline

The thesis at hand is divided into the following three parts:

1. **Theoretical Fundamentals and Intelligent Document Examination**
   Following this general introduction an outline of the state of the art in forensic document examination as well as the characteristics of different printing techniques is presented in Chapters 2 and 3 respectively. In Chapter 4 these fundamentals as well as techniques derived from statistical pattern recognition are merged to introduce a new system for intelligent document examination.

2. **Document Preprocessing and Feature Description**
   In Chapter 5 the document preprocessing steps required for intelligent document examination are described. Subsequently, in Chapter 6 the from the documents

extracted features are introduced for the purpose of printing technique recognition and photocopy detection.

3. **Hypothesis Evaluation and Conclusion**
   Finally, in Chapter 7 the in Section 1.2 formulated hypotheses are evaluated based on the introduced system and features. The concluding evaluation results and potential future work is stated in Chapter 8.

1.5 Notation

This section provides a reference of the symbols and their corresponding meanings used throughout this work.

Table 1.1: Symbols used throughout this work.

Symbol	Description
D	Set of document images
d_i	Single document image, $d_i \in D$, $i = 1...n$
$\|D\|$	Number of document images in D
C	Set of printing technique classes
c_i	Single printing technique class, $c_i \in C$, $i = 1...n$
$\|C\|$	Number of printing technique classes in C
X	Set of feature vectors
x_i	Single feature vector, $x_i \in X$, $i = 1...n$
$\|X\|$	Number of feature vectors in X

2. State of the Art in Forensic Document Examination

"Forensic document examiners are confronted on a daily basis with questions like by whom or what device a document was created, what changes have occurred since its original production, and is the document as old as it purports to be" [1]. The textbooks of Hilton [15], Ellen [13], Nickell [16], Kelly and Lindblom [1] offer excellent overviews of the state of the art in the techniques applied to questioned documents. However, a review of these textbooks revealed a shortage of new examination approaches derived from digital image processing and pattern recognition. This absence is caused by the circumstance, that the appliance of these scientific disciplines to the domain of document examination is relatively new [17].

In this Chapter a general overview of actual as well as the emerging digital techniques which are utilized in questioned document examinations will be given. Furthermore, for the purpose of categorization of this thesis, several characterstics are outlined which are of particular importance within the investigation of machine generated documents. Finally, the related work is reviewed and classified according to the respective examination objectives.

2.1 Forensic Document Examination

Forensic Document examiners have commonly been referred to as handwriting experts. However, the field of forensic document examination is much broader than the examination of handwriting. *"This discipline involves the complete scientific study of documents for any visible, microscopic, or spectral evidence that can assist in answering legal questions that arise in criminal, civil, and administrative matters."* [18]. Therefore, a document investigation case also covers ink and paper analyzes, identification of office machines, in-

dented writing, charred documents, fracture matches and document dating[1].

2.1.1 Questioned Documents

According to the Encyclopædia Britannica[2] the term document is defined as *"an original or official paper relied on as the basis, proof or support of something"* or in a more general perspective *"a writing conveying information"*. Prior to the start of an investigation the forensic document examiner usually receives the suspected documents that are associated with particular questions.

The daily casework of the forensic examiner is faced with documents were not everything about a document is accepted for what it appears to be. A document is labeled a *questioned document* if its authenticity is in doubt. Questioned document my be genuine, partially faked by obliterating, erasing or altering the original information, or completely faked as, for example, in counterfeit currency and lottery tickets, and blank educational certificates [17].

2.1.2 Examination Techniques and Practices

"Forensic sciences are challenged by the fact that only tiny pieces of evidence are hidden in a mostly chaotic environment" [12]. Subsequently, a forensic investigation is often a multi-facet approach regardless of the number of documents involved in an examination. Even if an examination is limited to a single page, the document examiner is likely to investigate various artefacts. The techniques used in an examination, are depending to a large extent on the document nature, for example, the investigation of signatures, paper type, ink type, or accidental marks. Consequently, a variety of sophisticated techniques have been developed by the forensic science community including destructive and nondestructive examinations.

Nowadays, according to LaPorte [19] and Ellen [13], three major categories of investigation methodologies can be distinguished in document examination namely physical, optical and chemical. Due to the observed technological progress of digital investigation methods and their increasing involvement in document forensics as described in Appendix A a fourth category named *digital examination* was added as illustrated in Figure 2.1. Since these techniques covering full textbooks it should be appreciated that the following introduced categories are merely exemplary embodiments and should not be understood in an all encompassing manner.

[1] The interested reader may be surprised why the discipline graphology is not listed in this context. By definition, graphology means the personality assessment of suspects by handwriting examination [13]. In the United States as well as in Germany, forensic document examiners refuse the practice of graphology and courts deny evidence based on graphological assumptions.

[2] Websource: Encyclopædia Britannica, `http://www.britannica.com`, as of August 29, 2008

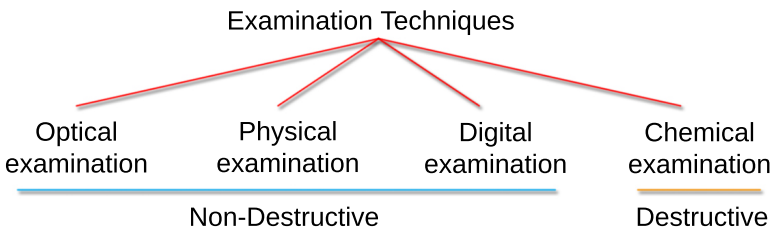


Figure 2.1: Overview of forensic document examination techniques.

Optical Examination

Quite often an optical inspection alone can be sufficient to reveal that the document at hand is not genuine. An optical examination might consist either of macroscopy (the scrutiny of things visible to the naked eye) or microscopy (an investigation by microscope).

Macroscopic examination may be conducted by reflected, oblique or transmitted light using an ordinary magnifying class. Reflected light examination exposes the absorbance of light on the document surface and can be used to date the document. Furthermore, ink types can be distinguished according to their rate of light reflection. Oblique light striking the surface of the document from one side at low angle can reveal shadows produced by surface irregularities, for example, erasures, indentations or embossment. Examination of a document by transmitted light involves the illumination the document from behind by placing it on a light table. Such examination facilitates identifying the type of paper, watermarks and detecting erasures or alterations [13].

Microscopic examination in contrast is used to achieve higher magnification in document work. Usually a low-power stereoscopic or stereo microscopes to inspect a document directly and laboratory microscopes for certain specific analyzes are used.

Stereoscopic microscopes are especially useful for document work due to their use of comparatively low magnification. Stereomicroscopic examination is often applied to pen type identification used to produce a given writing, for example, roller pen or fountain pen. As a result valuable information about the used pen like stroke density, thickness, outlines, uniformity, used pen pressure, additional ink, and the writing direction can be gained. Thus, stereomicroscopic examination is also an appropriate tool for many other writing features, including erasures, corrections, alterations and sequence of pens stroke or distinguising paper types [16].

Physical Examination

Visible light represents only one band of the electromagnetic spectrum. Forensic document examiners utilize also other bands in an investigation like ultraviolet light and infrared radiation. These techniques also including laser technology are referred to as spectral techniques.

Ultraviolet rays have wavelengths shorter than those of the visible light. Applying ultraviolet light to the surface of a document the light is absorbed by some substances and its energy transformed and radiated back in light of different colors. Using this interesting phenomenon, referred to as fluoresce, stains on documents, alterations on cheques, or secret writing in a letter can be detected. Furthermore, according to their rate of fluorescence different inks, paper types and erasures can be distinguished [13].

Infrared wavelengths are longer than those of visible rays and can be used to analyze similar document artefacts as using ultraviolet light. In addition infrared is offering the ability to detect undersketched, damaged document content. Furthermore, this type of light can be used to visualize and extract the content of unaccessible documents like unopened letters.

A more recent spectral tool applied in the examination of documents is laser technology. Unlike ordinary light, which radiates uncontrolled in all dimensions, laser is beamed so that all its waves are parallel as well as in phase. Laser technology can be applied to nearly the whole spectrum of light [13]. Therefore, it is used in combination with other spectral and non spectral techniques to achieve a higher precision of examination. One of the most extensively used forensic applications of laser has been the detection of a documents latent fingerprints and impressions but it is also often used for ink discrimination purposes.

Chemical Examination

While the so far introduced methods are *nondestructive*, chemical document examination methods are referred to as destructive since they may cause little but acceptable damage to the document under investigation. Provoked by their destructive manner chemical examinations are not the first choice in an examination, especially in the case of highly valuable or ancient documents. Among these are so called *spot tests* in which special chemical reactants are applied to tiny fragments or artefacts of a document [13].

A technique used in the discrimination of ink types is referred to as *Thin-layer chromatography*. In which ink samples taken from one or more documents are compared onto a silica-gel chromatogram sheet. According to run of the ink substrate onto the sheet the different dye compostions of the ink can be distinguished.

Also different toner types can be distinguished due to their composition of organic and inorganic elements, for example, the degree of iron, calcium, sulfur or silicon. Therefore, pyrolysis gas chromatography is applied to a set of toner samples that are heated to emit small toner molecules into an carrier gas i.e. nitrogen. In a subsequent step these molecules can than be isolated from the carrier gas and their individual amount measured.

Nowadays, another technique often applied to develop special document artefacts is the usage of different chemical solutions. For example, a mixture of ninhydrin with ethyl alcohol can be used to develop latent fingerprints or ammonium hydrosulfate can be applied to restore bleached old iron based ink.

Digital Examination

A variety of modern image processing techniques can be utilized to investigate different document properties on digital bases. Digital image processing applications like Photoshop™ or Corel Draw™ are more and more used by forensic document examiners. Because of their ability to even perform complex investigation tasks like reconstruction of destroyed document content in a convenient manner.

These applications utilize tools and filters which can, for example, be used to detect accidental marks by increasing a documents contrast or the magnification and coloration of specific document regions. Herbertson [20] introduced these kind of techniques in his book to analyze handwritten signatures or to determine a documents underlying printing technology. More sophisticated image processing features like graylevel coocurrence matrices or edge descriptors can be utilized to determine the contents textural or edge characteristics of a suspected document.

A new approach in this context is the appliance of classification techniques derived from statistical pattern recognition. These methodologies used in combination with machine learning approaches is offering the ability to learn document properties like the arrangement of a document components or signatures. These learned document models can in a later step be used to compare documents to detect significant alterations. A more detailed introduction about so far developed approaches in digital document examination will be given in Section 2.3.

2.2 Characteristics of Forensic Document Examination

As already described, forensic document examination can be understand as a multifacet approach based to a large extent on the suspected features of the investigated document. Every examination, however, exhibits more general characterstics allowing the forensic document examiner to categorize and explain the performed investigation procedure. In the following a subset of the most important document examination characteristics will be presented. Concluding the within this thesis developed technique will be categorized according the introduced characteristics.

2.2.1 Model Based and Generic Document Examination

All document examination problems fall into two basic groups. One being those that require specimens or a priori knowledge from an individual source to reach a conclusion. The other group is investigates the questioned document exclusively without any prior knowledge [1]. According to Lampert et al. [14] there are two ways the problem of counterfeit documents can be accessed: (1) model-based or (2) generically.

The *model-based* approach requires pre-knowledge on characteristic features to be checked and then searches specifically for them. Often the documents to be checked are already

created having the possibility of such an examination in mind and therefore an extrinsic document signature is added to the document. The validity of the security features can then be checked later, either by the human eye or by the usage of special devices in a straightforward manner.

Model based systems have the drawback that only those documents can be checked, for which a model of the document layout or the security feature is available, for example, from a database. Many document classes, like invoices, are exhibiting a great variety of characteristics therefore the creation of database for all derived document models is impractical or even impossible [14].

An alternative, to avoid this drawback, is the examination of questioned documents in a *generic* way. Using this approach a general selection of features is extracted from a document and the decision if a document is genuine or forged is based only on a class membership of a document features and their statistical information.

Generic counterfeit detection systems show a larger rate of error than model-based due to the limited pre-knowledge, but have the advantage of being applicable for a wider class of documents. Typically, forensic document examination casework is not exclusively dominated by one investigation approach it is often an interaction of model-based and generic investigations.

2.2.2 Class and Individual Document Characteristics

All document elements are valuable traces to identify a document's source or author, but the most unusual ones have the greatest value. *"In general the examination of a questioned document is directed towards the discovery of those elements that become its identifying attributes or characteristics"* [1]. Especially in the case of printing technique or device identification of a particular document, forensic document examiners utilize their training and databases to assist in identifying class and individualizing document features.

The examination strategy usually progresses from the general to the specific. A document is exhibiting many identifying characteristics that are common to a larger document group from different sources. These features are referred to as *class characteristics*, for example, printing technology, paper type or document layout.

Class characteristics are of special importance in the classification of printing devices for a variety of reasons. Knowledge of the class characteristics provides a forensic document examiner with the basis from which individual characteristics can be recognized. Therefore, the determination of a documents class characteristics is often a starting point within an examination procedure [18].

Individual characteristics are highly unique document elements because they can be rarely found on other documents, for example, toner trash marks, stapler perforation or erasure marks. This type of characteristics can result from wear, damage and defects. These

features are often providing evidence on which forensic document examiners can draw a link of a specific device to a particular document.

In the perspective of a forensic document examiner a questioned document can be the product of a combination of several materials put together with common instruments such as pen, pencil, typewriter, and/or various printing processes. Each of these materials and instruments leaves an individual signature on the document in question. Therefore, an investigation of the particular signature combination can yield valuable insights to personalize and identify the documents source and history [4].

2.2.3 Extrinsic and Intrinsic Document Signatures

The judgement wether a document is genuine or forged is based on the evaluation of the document's signature. The document signature exhibits a specific feature or the sum of features, allowing to classify the document. Two types of document signatures can be distinguished namely intrinsic and extrinsic signatures. Since Mikkilineni et al. [21] applied a similar distinction in the discrimination of different printer types it will be used here in a more general sense for document examination.

A documents *extrinsic signature* is generated by actively modifying the documents creation or the addition of features, for example, for security purposes. According to Zhu et al. [22] four major classes can be distinguished in the efforts to apply extrinsic document features, namely the use of special material, fingerprints, digital methods, and digital cryptography. As a result extrinsic document signatures can be created by applying security features like watermarks [23], holograms [24] or a special papers [25] to a document. Another way to increase document security extrinsicly can be achieved by modifying the process parameters of the printing mechanism to encode identifying information such as for example the printers serial number or date of printing [26].

In contrast a document's *intrinsic signature* is exclusively generated by the pure printing process and characterized by the physical characteristics of a particular printer, model or manufacturerers products. Therefore, neither an intentional modification of the printing process nor the document itself before or after printing has been undertaken. Documents exhibiting only intrinsic signatures are lacking special security features. For the investigation of such documents an understanding of the printer's mechanism and the underlying printing technology is required.

2.2.4 Contextual Classification of Thesis Objectives

In summary, the aim within this diploma thesis is the evaluation and development of recognition approaches to detect photocopies and a given document's underlying printing technology. Since these characteristics stand independent of a specific document they are categorized as *class characteristic* within this thesis. The approaches proposed in the

following chapters can be classified as *generic* since no preliminary computed document model is consulted in the examination. Furthermore, according to the above explained characteristics the printing technology is recognized based on the *intrinsic* printing signature of a document.

2.3 Related Work

Forensic disciplines are currently in a transformational stage [27]. An outline of this transformation process is given in Appendix 2. In comparison to traditional document forensic approaches, only a small number of publications related to digital document investigation exists nowadays. Furthermore, since the examination of machine printed documents is a multifacet approach, the reviewed publications in the context of this thesis are covering a variety of topics.

A brief outline of the related work encompassing the topics of examination like ink and toner types, paper type, accidental marks, print quality assessment and other printing device characteristics will be presented in the following:

Printing Quality

As illustrated in Section 3.2, distinctive information in the classification of a document's underlying printing technology can be gained by an assessment of the printing quality. Oliver et al. [28] proposed several digital quality metrics incorporated into the ImageXpert™ printing quality assessment system. They determined the line width, line raggedness, line overspray, dot roundness, dot perimeter and the number of satellite drops of printed characters scanned at high resolution. In doing so they could also successfully classify printed characters obtained from one offset device, three inkjet and three xerographic devices.

Lampert et al. [14] also assessed the print quality of laser and inkjet printed documents. They proposed texture features based on graylevel coocurrence and local binary map histograms as well as edge roughness features measuring line edge roughness, correlation and area difference for printed document characters. Printouts of eight laser printers and five inkjet printers were scanned at a resolution of 3200 dpi and classified using a support vector machine. An accuracy rate of 94.8% was achieved in their evaluation. In [29] Schulze et al. examined the features proposed by Lampert et al. for the purpose of high-throughput document management systems. Furthermore, they supplemented the features by two additional edge roughness features as well as a grayvalue distribution feature. In their extensive evaluation they successfully distinguished laser and inkjet printed documents at low resolution.

Mikkilineni et al. [30, 21] traced printing devices by extracting 22 graylevel coocurrence-occurrence features from the printed letter "e". Their method was based on printouts

created by ten printers scanned at a resolution of 2400 dpi. The extracted features were classified using a 5-Nearest-Neighbor classifier applying a majority vote within the feature space. With their approach they were able to classify nine of ten printers correctly. In a subsequent publication [10] a support vector machine was used and all ten printers were classified correctly. Furthermore, they were able to show that slight changes in font size, font type and paper types had no significant influence on the classification performance.

The systems outlined by Tchan [31] captured printed documents with a camera and differentiates them by measuring edge sharpness, surface roughness and image contrast. The discriminatory ability of the features was tested on two laser printed and one photocopied document containing squares and circles but no characters. According to the extracted feature values presented within the publication all three documents could be classified correctly.

Color Discrimination

Caused by the reduction in price of color laser printers in recent years, the dimension color is added to the document feature space and is more and more recognized within the forensic science community. In [17] Dasari and Bhagvati demonstrated the capability to determine different printing substrates and therefore printing techniques by analyzing the HSV color space representation of a document. Therefore, the Hue and Saturation histograms obtained from documents, scanned with a resolution of 1200 dpi, and printed by six color inkjet printers, four color laser printers and two color laser copiers were obtained and evaluated. Their presented histograms underpined their aim to discriminate between the printing technologies.

Printing Characteristics

The physical characteristics of printing devices can besides the printing signature, also leave distinctive fingerprints on printed documents. As recently shown by Akao et al. [32, 33] the investigation of spur gears, holding and passing the paper through the printing device, can also be used to link questioned documents to suspected printers. Therefore, the pitch and mutual distance of spur marks found on documents were compared to the parameters of already known printing devices.

Another approach proposed by Mikkilineni et al. [26] used banding frequency of the printing signature to identify electrophotographic printers. Banding is a print quality defect caused by electromechanical fluctuations in the printer mechanism (mostly from gear backlash) resulting in an non-uniform line spacing on the printed page. To measure the mentioned effect, test documents containing large midtone areas were scanned at a resolution of 2400 dpi. Obtaining the banding frequency from the printed patterns they were able to distinguish different printers. In [34] Ali et al. enhanced this method using Principal

Component Analysis (PCA)[3] to reduce the dimensions of the feature space. Furthermore, the utilize Gaussian Mixture Models (GMM)[4] and a decision tree for classification purposes and showed the ability to classify printouts obtained from five test printers correctly.

Layout Distortion

Gupta et al. [35] showed that the distortion created due to introduced imperfection by the printing device can be facilitated to detect fraudulent documents. They compared the original document and questioned documents according to their overall similarity and their similarity in coarse area, for example, printed character edges. The approach was tested measuring the distortion created by fake documents of two different Hewlett Packard printers. Both printouts showed a significant distortion in comparison to the original and therefore could be detected as counterfeit.

A method also based on image distortion was presented by Beusekom et al. [8]. They measured a documents layout deviation within the non-variable document parts (i.e. the documents header and footer) caused by reprinting it after altering the content. Their results revealed that connected component based document signature allows to estimate the likelihood of a document to be an original or not. All of the 12 altered test documents could be detected successfully according to the derived signature.

Protection Patterns

Another cutting edge approach proposed by Tweedy [36] and Li et al. [3, 37] are yellow dotted protection patterns distributed on documents printed by color laser printers that are nearly invisible for the unaided human eye. It was demonstrated that these distinctive dotted patterns are directly related to the printer serial number and printing date and time of the document and could be used for tracing a particular laser printer. So far the keys to decode these distinctive protection patterns are only accessible to governmental authorities for to purpose of tracking counterfeiting activities[5].

Zhu et al. [22] used the non-repeatable randomness existing in a printing process for authentication of printed paper documents. A unique print signature of a printed protection pattern consisting of four symmetrically aligned dots, was printed to documents and examined. This was done by measuring the Euclidean distance between the dot center and dot the perimeters at different angles. Their experimental results revealed that all forged

[3] Principal component analysis (PCA) is a vector space transform used to reduce multidimensional data sets to lower dimensions for analysis.

[4] Gaussian Mixture Models are used to estimate the probability density function of certain observation in the feature space.

[5] Obtained from the "Electronic Frontier Foundation" (EEF) at `http://www.eff.org/wp/investigating-machine-identification-code-technology-color-laser-printers`, as of August 23, 2008

documents were detected.

Scanner Identification

Prior to printing a photocopied document it has to be captured by the photocopiers associated scanning device. Several imperfections are introduced to a document image during the scanning phase referred to as scanning noise. Gou et al. [38] showed that scanner brands and models can be effectively identified by the examination of their corresponding image noise. They utilized several filtering methods, octave subband wavelet analysis as well as a neighborhood prediction model. Within their experiments they utilized support vector classification and achieved an accuracy rate of 96% in using 25 training documents.

In this context another interesting approach was proposed by Khanna et al. [39]. In their work they obtained a scanner's two dimensional reference noise pattern. This was achieved by denoising the reference images using an anisotropic local polynomial estimator and subtracting the denoised result from the original document. The obtained noise reference patterns where then compared to the noise pattern of a randomly given image. Using this approach they were able to classify the images of four distinct scanners with 96% accuracy.

Scanner-Printer Identification

In [35] Gupta et al. presented a structured methodology to detect the scanner-printer combination used in the creation of tampered documents. Their approach is based on the investigation of document image imperfections introduced by the scanning as well as the printing device. Therefore, the overall similarity as well as the similarity in coarse document areas between the original and the tampered document was measured. Furthermore, the standard deviation and average saturation of the image noise were obtained using mean, median and Gaussian filters. Their experimental results showed that the scanner and printer used in the generation process of the faked document could be identified successfully.

3. Printing Processes and their Characteristics

Current printing technologies are based on a variety of inventions. The discoveries made in the engineering sciences, information technology, physics and chemistry have in particular influenced the development of nowadays printing technologies. Therefore before a forensic document examiner can properly examine computer generated documents, she must have a sufficient understanding of the technology involved. *"A working knowledge of the various classes of typewriters and their distinguishing features allows to differentate one technology from the other"* [1]. The following chapter will attempt to touch on the most important aspects concerning printing processes as well as their printed image characteristics. Both aspects have significant relevance in the forensic examination of computer generated documents.

3.1 Printing Technologies

Printing in a more general perspective can be seen as a complex *"reproduction process in which printing ink is applied to a printing substrate in order to transmit information in a repeatable form using an image carrying media (e.g., a printing plate)"* [40]. An overview of printing technologies is given in Figure 3.1. According to Dasari and Bhagvati [17], there are two primary characteristics in the analysis of printing methods: the nature of ink used, and the process by which the ink is transferred to the paper. A major distinction is made between technologies requiring a printing master for ink transference referred to as *conventional printing*, and so-called *non impact printing* technologies that do not requiring a printing master. Both classes are described in the following:

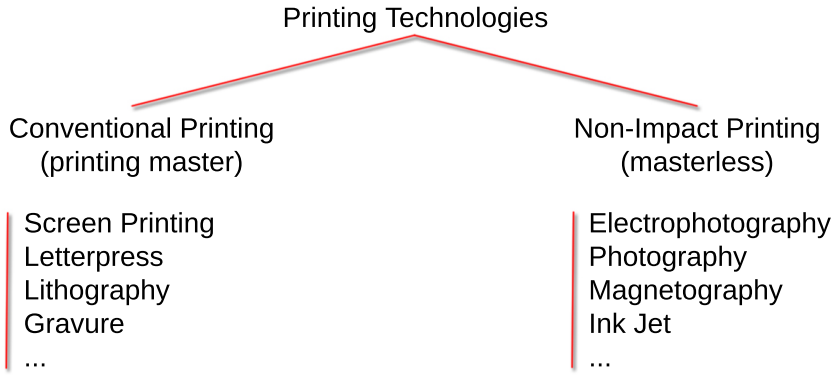


Figure 3.1: Taxonomy of different printing technologies according to Kipphan [40]. A distinction is made between technologies that touches the printing substrate to crate an image (Conventional printing) and those that do not (Non-impact printing).

Conventional Printing

Conventional printing technologies are denoted by the fact that they are using an mechanism that touches the printing substrate to create an image. The information to be printed is transfered by a partial surface transfer of ink or toner to the document. This is achieved in most of the cases utilizing an image carrying printing master, for example, printing plate. After the information transfer, the document exhibits regions containing transferred ink also referred to as image elements and regions without ink referred to as non-image elements. Conventional printing technologies are, for example, lithography (offset printing), gravure, letterpress and dot matrix printing.

Non-impact Printing

Printing technologies, not requiring a stable, physical and fixed image carrier are referred to as non-impact[1]. These printing technologies do not touch the printing substrate while creating an image. The most well-known non-impact technologies are *inkjet* and *laser printing*. In inkjet printing tiny drops are sprayed directly onto the paper whereas laser printing toner particles are attracted by electric charge.

The examination of documents produced by non-impact printing technology has evolved significantly since Chester Carlson's discovery of the electrophotgraphic process in 1938. Nowadays, the electrophotgraphic process is the underlying technology of the majority of laser printers and photocopiers. Therefore, inkjet printing and the electrophotgraphic process are the two most prevalent non-impact printing technologies. These systems can be found in many households and offices because they are capable of reproducing documents with a rapid and high quality output at affordable prices [19].

[1]The term non-impact is derived from early digitally controlled printing systems were printouts were often generated using dot matrix printers. Typefaces for these matrix printers were controlled electronically and the information transferred to the paper via typeforming pins impacting an ink ribbon [40].

Resulting from this evolution of non-impact printing technologies the forensic document examiner will in many cases encounter evidence generated with non-impact printing devices. According to Doherty *"now more than ever the forensic document examiner should be equipped with tools to accurately evaluate these technologies"* [41]. A basic understanding of these technologies is crucial for the investigation of non-impact printed documents. Hence, the underlying technological principles of non-impact printing technologies and their characteristics will be outlined within the subsequent sections.

3.1.1 Inkjet Printing

As already mentioned one of the most frequently non impact printing method used nowadays is the inkjet technology. In the following section the basic technological principles of inkjet printing will be described. Additionally, some regular characteristics of inkjet printed signatures will be introduced and elaborated in greater detail.

Inkjet printers utilize a printhead to emit tiny ink droplets onto the printing paper. As the paper is drawn through the printer, the printhead moves back-and-forth horizontally and usually transfers the ink directly to the paper. The ink deposition is digitally controlled by the printers inherent firmware. Applying inkjet printing the ink is sprayed onto the paper in a way that multiple gradients of dots accumulate to form an image with variable color tones. As illustrated in Figure 3.2 inkjet processes can be classified as *continuous inkjet* and *drop on demand inkjet*.

Continuous Inkjet

The *continuous inkjet* technology generates a constant stream of small ink droplets, which is charged according to the content to be printed and controlled electronically. The charged droplets are deflected by a subsequent electric field, while the uncharged ones flow onto the paper. This is indicating that the imaging signal for charging the droplets corresponds to the negative of the printed image. As a result continuous ink jet printing usually feeds only a small fraction of the droplet stream onto the printing paper. The larger amount of droplets is collected by the printing system.

Drop On Demand Inkjet

With so-called *drop on demand inkjet* technology, a droplet is only generated if required by the printed image. The most important drop on demand technologies are *thermal inkjet* and *piezo inkjet* printing. Thermal inkjet (also known as *bubblejet*) is generating ink droplets by heating a localized vaporization of the liquid in an inkjet chamber. While using piezo inkjet technology the ink is formed and catapulted out of a nozzle by mechanically deforming the inkjet chamber. This deformation is caused by an electronic signal which is send to the piezoelectric crystal of the chamber wall. Due to physical constraints, the

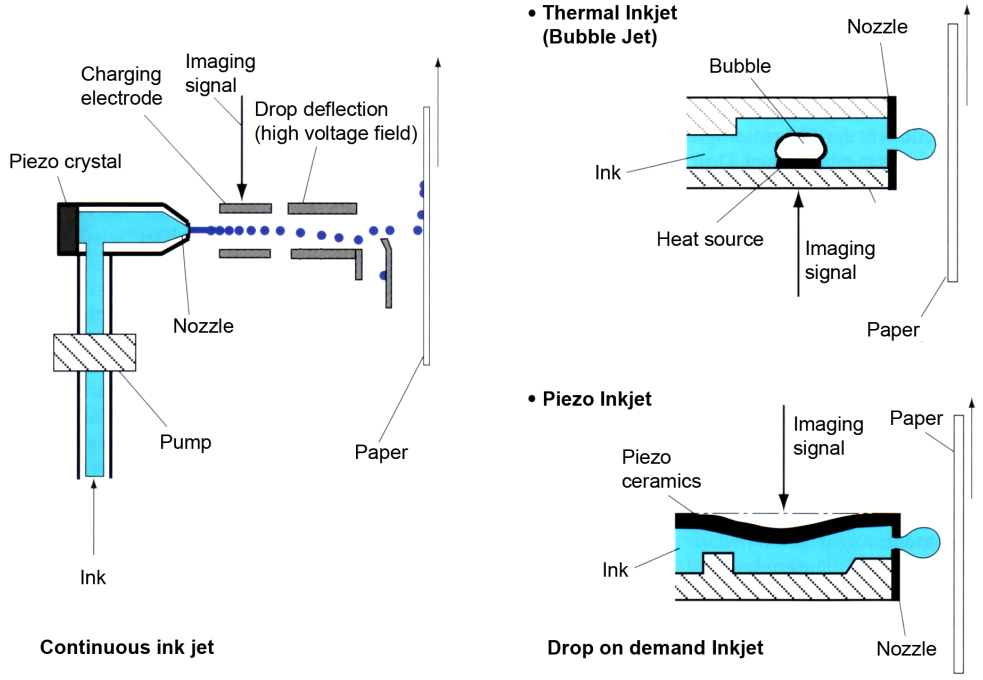


Figure 3.2: The inkjet printing process: (l) basic elements of continuous inkjet printing and (r) basic elements of drop on demand inkjet printing as illustrated by Kipphan [40].

rate of droplet frequency is lower for the thermal generated droplets in comparison to the one generated using piezo technology.

The ink used for inkjet printing is usually in liquid form. An alternative, are hot-melting inks which are liquefied by heating. The ink is sprayed onto the printing paper where it solidifies after cooling.

3.1.2 Electrophotographic Printing

The *electrophotgraphic process* also known as *xerography* was derived from the Greek words for "dry" and "writing". Today the huge majority of laser printers and photocopiers are based on the electrophotographic technology. Electrophotography is based on two natural phenomena: materials of opposite electrical charge attract each other and special materials become better conductors of electricity when exposed to light[2].

The underlying concept of electrophotgraphic printing is the creation of a visible image using an electrostatic latent image created by surface charge patterns on a photoconductive surface. In difference to inkjet printing technology (outlined in Section 3.1.1) where the image is written directly to the paper, in electrophotgraphic printing the image is first written as a pattern of charge on a photoreceptor. The final visible images consist of fine particles called toner.

[2]Websource: `http://www.imaging.org/resources/web\_tutorials/xerography.cfm`, as of June 10, 2008

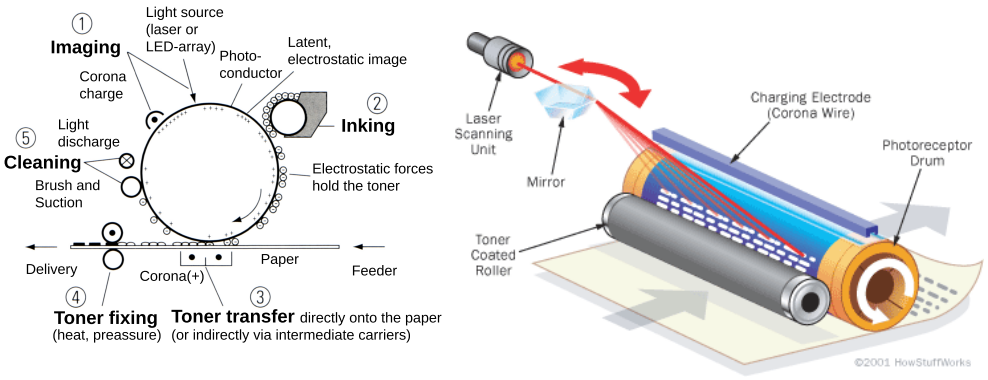


Figure 3.3: The electrophotographic printing process: (l) basic elements of electrophotgraphic printing as illustrated by Kipphan [40] (r) charging of the photoconductor drum [4].

As illustrated in Figure 3.3, the electrophotographic printing process can be subdivided into the following five basic steps [40]:

1. Imaging/Developing

Imaging is carried out by charging a photoconductive surface (creating a homogeneous charged surface) with subsequent imaging via a controlled light source. The charged print image on the photoconducter drum corresponds to the positioning of the light signals. The areas of the drum that are exposed to the light source become conductive and hence discharged. The area of the drum not exposed to the light source remain negatively charged. As a result a latent electrical image on the surface of the drum is written. Imaging in electrophotography can be done both by laser light or by an array of light-emitting diodes. As a result, according to Kipphan [40], the frequently-used term laser printer instead of the term electrophotgraphic printer is misleading.

2. Inking

Special inks which may be powder or liquid toner containing the colorant in the form of pigments are used for electrophotograpic processes. Both types can vary in structure according to their composition. The ink is the fundamental decisive element for the impression. Inking is done via systems which transfer fine powder particles, without contact to the photoconducter drum. The toner charge is configured in such a way that the charged areas of the photoconducter surface attract the toner. As shown in Figure 3.3, imaging is done with a negative image because the positive charges have been discharged by exposure to light. After inking, the latent image on the photoconducter drum becomes visible at regions were toner is applied.

3. Transfer/Printing

The toner may be transferred directly onto the paper, although in some cases it may also be transferred via intermediate systems, in the form of a drum or a belt. As shown in Figure 3.3 the transfer mostly is taking place directly from the photoconducter drum to the substrate. To transfer the charged toner particles from the drum surface to the paper, electrostatic forces are used supported by contact pressure between the drum surface and the paper.

4. Fixing/Fusing

A fixing unit is required to anchor the toner particles on the paper and create a stable print image. This is usually designed in a way that melting and consequent anchoring of the toner on the paper is achieved by the application of heat and contact pressure. Therefore in most of the investigated printers the toner is melted and bonded to the paper by high-heat and high-pressure rollers.

5. Cleaning

As shown in Figure 3.3, residual charges and individual particles remain on the drum after the print image has been transferred from the photoconductor drum to the paper. To prepare the drum for printing the next image, both mechanical and electrical cleaning of the surface is necessary. The mechanical cleaning, which removes particles of the toner, can be done by brushes and/or rubber blades. The electrical cleaning (neutralizing) is accomplished by a homogeneous illumination of the surface. After electrical neutralization the surface is free from toner particles. The photoconducter drum is now charged again with an subsequent image via the corona as described in step one.

3.1.3 Photocopying

The ability of contemporary photocopiers to create fast, inexpensive document reproductions has lead to their widespread use at every level of business. Modern photocopy devices also can operate as printers and faxes in which case the digital image is obtained from a computer or phone. As already mentioned these devices work the same basic way as laser printers. Stand-alone photocopy devices mostly use an indirect electrostatic process. However, in comparison to laser printing photocopying comprises the two following significant differences:

1. Image Source

The most obvious difference is the image source: In photocopying the image has to be scanned prior to its printing, while a laser printer receives the image of a document directly in digital format. Nowadays, nearly all modern copiers capture the image digitally during

the document scan. A typical photocopy scenario requires the placement of a picture or document onto the scanner platen where it is scanned and converted into a data file.

The scanning device found in the majority of photocopiers is usually composed of a glass pane, under which a bright light source is installed (often xenon or cold cathode fluorescent lamp) illuminating the pane, and a moving charge-coupled device (CCD) array. The CCD consists of a collection of tiny light-sensitive diodes, which convert photons (light) into electrons (electric charge). The brighter the light that hits a single photodiode, the greater the charge that will accumulate at the diode. The device containing the linear array of CCD photodiodes is also referred to as scanhead.

Usually a scanning process is started by placing a document on the glass pane and closing of the photocopiers cover. During the scanning procedure the scanhead is moved slowly across the document. By sweeping the scanhead across the paper the lamp is used to illuminate the document. The reflectance image created by the illumination of the document using the lamp is absorbed by a construction of angled mirrors. Where the last mirror reflects the image onto a lens which focuses the image through a filter on the CCD array. The motion of the scanhead divides the image into microscopic rows and columns by measuring how much light, or the lack thereof, reflects from each individual intersection of the rows and columns [42]. In addition color photocopiers typically contain three rows (arrays) of sensors equipped with red, green, and blue filters. The charge collected during the scanning process by each photodiode is proportional to the reflectance of the specific area of the paper. The amount of reflectance is recorded as a dot, or picture element (pixel). After the scanner collected the information from all the dots, it compiles the result into a digital file on the device.

2. Image Development

Another major difference between a photocopy device and a laser printer is how the electrostatic image is developed. When a photocopier bounces light off a piece of paper, the light reflects back onto the photoreceptor from the white areas but is absorbed by the dark areas. In this process, the background is discharged, while the electrostatic image retains a positive charge. This method is called *write-white*.

In contrast for the majority of laser printers, this process is reversed: The laser discharges the content of the electrostatic image and leaves the background positively charged. In a printer, this *write-black* system is easier to implement than a write-white system, and in general produces better print quality results[5].

[5]Websource: `http://home.howstuffworks.com/photocopier.htm`, as of June 30, 2008

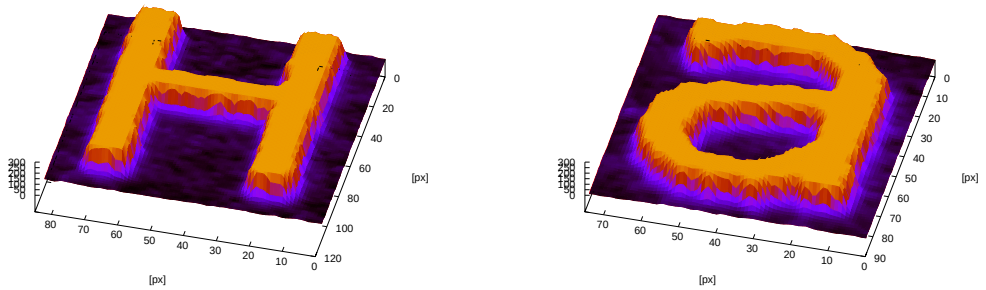


Figure 3.4: Exemplary surface plots of the inkjet printed characters ”H” and “a” scanned with a resolution of 800dpi. A light blurring and an increased roughness at the character edges is observable.

3.2 Printing Characteristics

Having presented the differences in the printing technologies now the focus is turned to the possible defects of the latter. Examination of these defects is the main source of evidence in document forensics. As stated by La Porte in [19] *“there are a number of physical defects that the forensic document examiner should remember when examining printed documents to determine forensic discriminations”*. In 1941 Osborn and Osborn [43] already emphasized, *“a document may have any one of twenty or more defects that are not seen until they are looked for. Some of these things are obvious when pointed out, while others to be seen and correctly interpreted must be explained and illustrated*[6]*”*. The detection and analysis of these defects also provides a valuable source of information usable in printing techniques recognition.

The identification of defects specific to particular printing technique, also referred to as *document image degradations* in the context of Optical Character Recognition (OCR), show the ability to provide first insights in a documents examination process. This especially holds in the investigation of questioned document were in many cases a certain printing technique can be recognized by its printing defects and degradations. Throughout this work document image defects and degradations are understood as *”every sort of less-than-ideal properties“* or *”the departure of an ideal version“* as defined by Biard [44].

According to Smith [45] the defects occurring frequently in scanned document images can be categorized into:

1. Printing paper defects, for example, yellowing, wrinkles or coffee stains.
2. Printing process defects, for example, toner dropout, bleeding or scatter.
3. Document scanning defects, for example, sensor sensitivity noise or skew.

[6] Albert S. Osbornand Albert D. Osborn, co-founders of the American Society of Questioned Document Examiners (ASQDE) published the year before the formal founding.

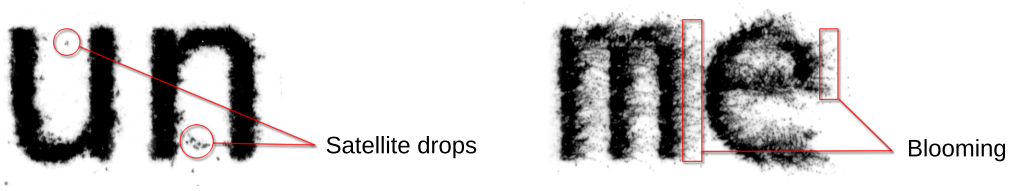


Figure 3.5: Typical inkjet printing defects: (l) satellite drops induced by unclean ink ejection and (r) ink blooming/bleeding resulting from inappropriate ink absorption of the paper.

The subsequent subsections will attempt to describe the most important aspects concerning printing processes and their defects which are relevant in the forensic examination of computer generated documents. The focus of the following explanations is laid on defects caused by the printing process as well as document scanning[7].

3.2.1 Inkjet Printing Characteristics

Investigating the printing signature of inkjet printed characters, as illustrated in Figure 3.4, some characteristic features can be observed. Document images created by inkjet technologies are in general characterized by a stronger blurring and roughness of character edges especially when compared to laser printing. Further, a more irregular ink distribution at printed areas resulting in a weaker reflectance of the printed characters can be recognized. Both observations are provoked by the effects described in the following:

Blooming/Bleeding

Blooming and Bleeding is caused by the absorption of ink into the paper and diffusing further than the preferred dot size. When ink is distributed and migrates along the paper fibers it can create a spider web effect resulting in an uneven and unsharp print of the character edges. The result is often a poor print quality evident to the unaided eye as shown in Figure 3.5. This phenomenon occurs in particular when the applied ink is incompatible with the used paper type, when printing in fast mode, or when printing in unfavorable conditions such as high temperature or high humidity.

Mottling/Cockling

Another characteristic referred to as mottling occurs when the print density or color of an image is uneven because of an inaccurate dot assignment or dot density. The results often observed are heavily inked areas followed by weakly inked areas leading to unsteady

[7]However, to also obtain a more elaborated understanding about the impact caused by paper properties (e.g. paper moisture, roughness and porosity and printing) to the printed document image the interested reader should be referred to the work of Borch and Svendson [46].

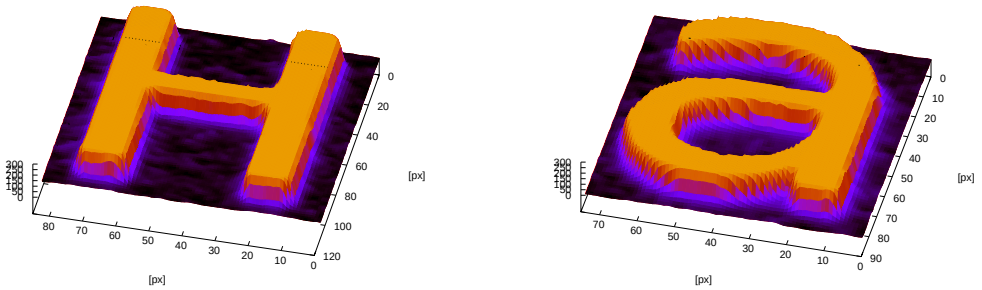


Figure 3.6: Exemplary surface plots of the electrophotographic printed characters ”H” and “a” scanned with a resolution of 800*dpi*. A sharp transition between character and non-character regions is observable.

and uneven halftone pattern of the printed character image. Cockling characterizes the wrinkling or puckering from the deposition of too much ink or nonuniform drying of the ink on the substrate. These occurrences are often found on documents where ink and substrate are not compatible leading to an irregular perception of the printed area.

Satellite Drops

Satellite drops are tiny droplets of ink that separate from the main drop as the ink is being ejected as illustrated in Figure 3.5. Each drop of ink can have one or more satellite drops that follow it or land nearby. General observations indicate that they will gather at the edges of printed material.

Looking at the position of the satellites on a questioned document can reveal the printhead directionality. Typically, when satellites are formed from the main drop, they land ahead of the drop in the direction of the printhead. Thus, if satellites appear to the right of the printed area, it indicates that the printhead was moving to the right. When satellites appear on one side of the print on every line it can be assumed that the printer utilizes a unidirectional printhead. If satellites are positioned on opposite sides from line to line, then this is an indicative if bidirectional printing. LaPorte [19] observed that inkjet printed documents from Epson^{TM} printers, for example, have very few, if any observable satellites. This might be due to the piezoelectric technology of the printhead.

3.2.2 Electrophotographic Printing Characteristics

Investigating images of printed characters created by electrophotographic printing, as illustrated in Figure 3.6, in comparison to inkjet printing, an improved print quality can be recognized. Since the majority of electrophotographic printing devices using dry toner particles for image creation, the diffusion of toner into the printing paper fibers is diminished [46]. Therefore, electrophotgraphic printed characters are characterized by sharp

transitions between toner to non toner areas resulting in unblurred character edges. In addition, originating from the reduced diffusion of toner into the printing substrate and the precise image development at the photoconductor drum a more consistent distribution of toner particles on printed document areas is achieved. As a result the halftone regions of laser printed documents are perceived as more homogeneous and reflective than in ink jet printed documents.

In addition cases occur were electrophotographic created documents are perceived as slightly darker in comparison to inkjet printed documents. This is often caused by a distribution of toner scatter as described in the following:

Toner Scatter

The random distribution of fine toner particles on the surface of a printed document is referred to as toner scatter or background noise. Toner scatter can be observed on unprinted areas diffusing over the entire document or accumulating at specific document regions i.e. document corners or letter edges. This noise follows from the contrast difference between paper and printed content [47]. Usually, toner particles are originating from an improper installed toner cartridge. A second reason creating this defect is a dirty or worn transfer roller, which results in improper biased voltages that effects the electrophotgraphic image developing process. Toner scatter can also be caused by a bad fuser assembly (especially on fusers containing ceramic heating elements). In this case residue toner particles of recently processed documents are not removed correctly after printing.

3.2.3 Photocopy Characteristics

The electrophotographic printing process as described in the proceeding sections is the underlying concept of the majority of todays photocopiers and laser printers. Therefore, photocopiers and laser printers are extremely similar in operation. Subsequently, even for the experienced forensic document examiner, it is difficult to determine whether the document at hand is photocopied or laser printed [1].

As already described the process to photocopy a document consists of at least two distinct phases, namely the scanning of the template document and the final printing of the obtained content. Due to technical or physical limitations of a photocopier device a relatively small amount of the documents original information will be lost or changed. These *device introduced imperfections* as described by Gupta et al. [35] can provide valuable traces in the identification of photocopies. In the previous Sections 3.2.2 and 3.2.1 imperfections introduced by document printing have already been discussed. In the following, the document image degradations caused by document scanning as determined by Smith and Qiu [45, 48] are presented:

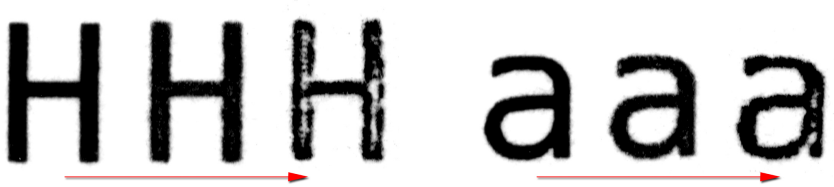


Figure 3.7: Document image degradation at several photocopy generations (creating a photocopy of a photocopy). Progressive degradations as well as a blurring at the character edges is observable from one photocopy generation to the next.

Edge Blurring

Character edges are characterized by a sharp transition of black to white document image content. Photocopying a document and observing the obtained character images, in comparison to the original document, a blurring or smoothing of the character edges can be recognized. This effect can be traced back to light diffusion reflected by the scanning sensor during the document scanning procedure. Usually, it is desired for each scanned pixel value in the document image to be composed entirely of the amount of light reflected from the central sensor location. However, scanned pixel values are in addition often influenced by the papers image brightness and the light reflected within the sensors neighborhood.

Corner Degradation

A similar effect can be observed at character corners. In the case of corners the photocopying process is having an impact from several directions. Therefore, corner regions are in particular sensitive to photocopying. As a result, sharp corner tips will be rounded which is referred to as corner erosion. Corner erosion can be recognized on exterior black to white character corners as well as on interior white to black character corners.

To photocopy a document several times and therefore repeating the above explained degradations will result in poor character images which is also shown in Figure 3.7. This effect can yield to a filling or the break up of characters and is especially a critical issue in OCR were document recognition rates can fall abruptly when image quality degrades even slightly.

Image Noise

As described in Section 3.1.3 the majority of todays photocopiers is using a moving CCD imaging sensor array to record the image information of the to be photocopied document. Several kinds of noise can occur when photoelectrons are created by the CCD sensor array induced by imaging sensor imperfections and the discrete and random nature of photoelectrons [38].

These kinds of noise can be categorized into two classes, namely random noise and pattern noise. Random noise can be traced back to array defects that cause irregular image

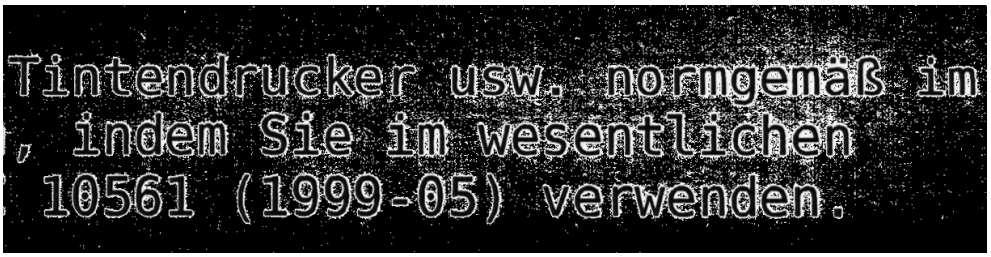


Figure 3.8: Exemplary document image noise caused by the photocopying process that is frequently observable at photocopied document images.

deviations caused by e.g. hot point defects, dead pixels, and array defects. In contrast pattern noise is not changing significantly from image to image and is caused by dark signal non-uniformity and photoresponse non-uniformity [39]. Dark signal non-uniformity represents the pixel to pixel variations in the CCD voltage under the circumstance that no light is incident on the CCD sensors [38]. Photoresponse non-uniformity models the variation in pixel response and can be observed under fixed-light intensity. Both noise pattern types are linked to imperfections created during the manufacturing process e.g. variations between the detectors or thickness in coating. A comprehensive overview of the noise types described above can be found by Holst [49].

Another type of noise recently investigated by Dirik et al. [50] in the context of camera identification is sensor dust noise. Sensor dust on images can occur if dust and moisture which are attracted by the electrostatic fields of the imaging sensor is causing a dust pattern in front of the CCD sensor surface. This especially holds for photocopiers in which fine toner scatter and dust particles can effect the CCD array during a ordinary copy process or if the scanner platen is opened due to maintenance. In the case of heavily utilized photocopiers, dust and microscratches are detectable at the front side of the scanner platen leading to a similar effect. In general this type of noise is visually not very significant but is transfered to the photocopied document as shown in Figure 3.8.

3.2.4 General Printing Defects

In the following general printing defects that could not exclusively traced back to a specific printing technique are outlined.

Banding and Lining

Banding is defined as uniform, regularly occurring, density variations or voids in a given color. These appear along the axis of printhead movement and are evidenced by dark or light linear stripes that occur in the print areas e.g. the absence of color component can result in dark or light areas.

Lining is another term that is used to describe a missing area, or line of ink or toner. Both of these are more noticeable in image prints, but are also observed in text documents. The

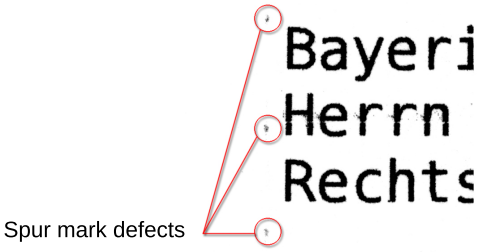


Figure 3.9: Typical printer traces left at a document image that are caused by the physical properties of the printer. The image above shows exemplary spur marks induced by the gears of the printing device.

occurance of both banding or lining provide the possibility to link multiple documents to each other or in certain instances, to link documents to a particular printing device.

Physical Markings

There are a number of mechanical components of inkjet printers and electrophotgraphic printers, such as star wheels and pinch rollers, which may cause microscopic or macroscopic marks on documents after they have been printed. In some instances, the developed markings can directly be associated to various parts within the printer. Broken or misaligned components can result in visible (e.g. paper tears) and latent individual defects as shown in Figure 3.9.

The proceeding sections gave an overview of non-impact printing characteristics and defects. However, it remains important to mention that the defects presented within this chapter should only be considered as a starting point for further investigations. Most of the presented defects are derived from research undertaken by La Porte [19] on inkjet printing and a recent publication of Hewlett-Packard™[8] that lists common laser printers defects.

[8] A comprehensive list of Hewlett-Packard laser printer defects that should at least 90% of the most common image defects can be found online: http://www.printertechs.com/tech/print-defects/print-defects-index.php.

4. Intelligent Forensic Document Examination

As stated in Chapter 1, the objective of this thesis is to develop an intelligent examination technique for the automatic detection of photocopies and and a given document's underlying printing technology. To achieve this objective, the central idea is the enhancement of traditional document investigation methods, as presented in Chapter 2, by techniques derived from digital image processing and pattern recognition.

In this chapter the outline of a general pattern classification system will be explained. Furthermore, the question, how computer systems become able to learn, for example, solving an arbitrary classification problem, will be addressed. Finally, the gained insights will be transferred to the design of a novel process for intelligent photocopy detection and printing technique recognition.

4.1 Pattern Classification Systems

According to Watanabe [51] a *pattern* is defined as *"a vaguely defined entity"* representing the *"opposite of chaos"*. A pattern could therefore be any observation or group of observations derived from an object, for example a fingerprint image, a human face or, as in the context of this thesis a scanned document image. In general patterns are represented by several measurements derived from an object referred to as *features*. The composition of features, referred to as *feature vector*, defines a points in an multidimensional feature space. The task of *pattern classification systems* can be understood as the process to classify patterns into distinct categories [52]. Therefore either a priori knowledge or any other statistical pattern information is utilized.

According to Duda et al. [52] each pattern classification system can be interpreted as a process consisting of three basic steps namely preprocessing, feature extraction and clas-

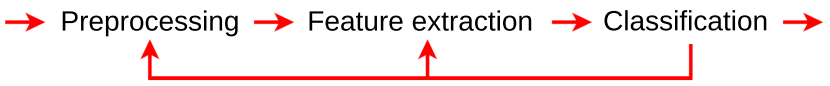


Figure 4.1: General pattern classification process comprised of the three process steps preprocessing, feature extraction and classification according by Duda et al. [52].

sification. Figure 4.1 illustrates the interconnection of these steps.

1. Preprocessing

Within the first basic operation of a pattern classification system general or distinct features from the object of interest are captured. This is achieved using some kind of transducer e.g. utilizing a camera in the case of optical data or a microphone in the case of measuring an audio signal. In general these measurements are associated with a certain amount of noise. The role of preprocessing therefore is to segment the desired information from noise and irrelevant data.

2. Feature Extraction

The goal of feature extraction is to characterize an arbitrary observation by a set of individual features. Furthermore, feature extraction is the determination of an appropriate feature subspace exhibiting a smaller dimensionality in comparison to the initial feature space. The values of these features should as similar as possible to observations of the same class. Simultaneously, the feature values should be as discriminant as possible for observations corresponding to different classes. A finall applied dimension reduction leads to an compact feature vector representing the observation.

3. Classification

Within the classification step a classifier assigns the extracted feature vector to a particular class. Therefore, the feature vector is applied to a probability distribution model that is iteratively learned by the classifier. Based on the learned model the classifier calculates a posterior probability of class membership for each of the classes. Finally, a class membership prediction is made based on the probabilities.

Depending on the obtained classification results, the feedback path allows the designer of the classification system to optimize the preprocessing and feature extraction/selection

strategies. Therefore, designing a pattern classification system becomes an iterative process.

4.2 Machine Learning Strategies

The question of how a system learns is another important aspect in designing a pattern classification system. As described in the preceding Chapter 4.1 feature vector classification is achieved based on a learned model. The design and development of techniques that allow computer to learn those models is referred to as *machine learning*. In general machine learning relies on the acquisition of different types of knowledge and can therefore range from trivial memorization of experience, to the creation of entire scientific theories [53].

Consulting the corresponding literature [52, 54, 55] in general machine learning is achieved according to three essential learning strategies:

1. Supervised Machine Learning

Scenarios in which a set of *ground truth* training samples is provided are referred to as supervised learning problems [54]. Where the term ground truth denotes a set of correctly labeled samples. The objective of supervised learning is to learn a reliable model from the discriminative distribution of the ground truth sample's class labels. The learned model is than utilized as classifier to predict the class of so far unseen samples based on their feature and their distribution in the feature space [56].

2. Unsupervised Machine Learning

Scenarios in which a set of training samples is provided without any corresponding class label are referred to as unsupervised learning [56]. The objective of unsupervised learning is the discovery of clusters of similar samples or the determination and/or the probability density distribution for the given unlabeled samples [54]. Applying unsupervised learning techniques is driven by the idea to discover unknown, but useful information about the input samples distribution within the feature space [57].

3. Reinforcement Machine Learning

Reinforcement learning, is referred to as learning with a critic [52]. In contrast to supervised learning only a tentative feedback is given by the teacher whether the predicted class is right or wrong. Therefore, no degree of certainty is provided of how wrong it is specifically. The objective of this type of learning strategy is to find a balance between exploration and exploitation and is often applied in the context autonomous agents.

Having introduced the different learning strategies it can be concluded that the strategy to be chosen depends to a large extent on the problem scenario in which the system will

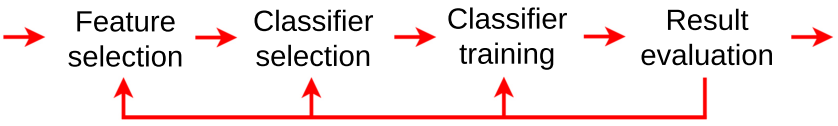


Figure 4.2: General classifier design process according to Duda et al. [52]. The evaluation of classification results my call for a recursive repetition of various steps till satisfactory results are obtained.

be employed. Furthermore, it can be concluded that the characteristics of the problem scenario have an impact on the design of a pattern classification system.

As already mentioned in Chapter 1.2 the purpose of this thesis is the prediction of an unseen document's class in terms of photocopy detection or printing technique recognition. Therefore, the objective is to learn a discriminant model from a set of ground truth training samples containing several documents that are photocopied or printed by different printing technologies. In a subsequent step the learned model should then be applied to predict the underlying printing technology of so far unseen documents. Judging from the above introduced machine learning strategies this can be generalized to a supervised machine learning scenario. The design of a supervised pattern classifier, that is needed in such a scenario, will be introduced in the subsequent section.

4.3 Pattern Classifier Design

As already described, supervised pattern classification is concerned with the classification of patterns based on a priori knowledge extracted from these or similar patterns. In statistical pattern recognition, these patterns are represented by a set of features and are obtained from the to be classified pattern, e.g. a document image as in our case. These features compose a d-dimensional feature vector X in a d-dimensional feature space $\mathbb{R}^d$. Machine Learning is than achieved by establishing so called *decision boundaries* to partition the feature vectors in the feature space. The learned partition of the feature space, referred to as *model*, can than be utilized to predict so far unknown feature vectors.

This process of learning, is conducted by a *classifier*, that provides the ability to learn a partitioning of the feature space. The design of an appropriate classifier usually consists of two central phases: (1) a training and (2) a testing phase. An overview of the classifier design process is illustrated in Figure 4.2. Subsequently, both phases are described:

Classifier Training

Within the training phase the classifier is trained to partition a feature space $V \in \mathbb{R}^d$ into disjunct subspaces $c_1, ..., c_n, \quad = C$ according to class labels of a set of training samples. Suppose there are given a set of $|D|$ random document images. Furthermore, each document image $d_i \in D$ is associated with a pair $[x_i, c_i]$: the feature vector $c_i \in V$ extracted from the document image and its associated class label $c_i \in C$. Let $[x_i, c_i]$ be the set of ground truth labeled samples given by a trusted source [58]. The aim of the training phase is to learn a feature space separating hyperplane referred to as discriminant function g:

$$\begin{aligned} g : \mathbb{R}^d &\rightarrow \mathbb{R} \\ x_i &\mapsto c_i \end{aligned} \tag{4.1}$$

that maps the feature vector x_i of a training sample to its corresponding class label c_i.

Classifier Testing

Finally, during the evaluation phase the learned discriminant function g is than utilized to predict the class label of unseen samples represented by x_i. Therefore, the single features vector values $x_i^1, x_i^2, ..., x_n^n$ are assumed to have a probability density function conditioned on a particular pattern class c_i. Finally, x_i is then classified as belonging to a particular class c_i if it can be viewed as an observation drawn randomly from the class likelyhood probability density function $p(x_i|c_i)$ [52]. An intuitive decision rule can be derived e.g. from the simplified Bayes Rule [59] which assigns an unseen feature vector x_i to class c_i in favor of c_j if:

$$P(c_i|x) > P(c_j|x), \text{ where } i \neq j. \tag{4.2}$$

Based on the evaluation results a decision is made wether satisfactory results are obtained or the repetition with a modified set of features or classifier is aimed. Therefore, according Duda et al. [52] the design of an appropriate classifier for a given classification task usually entails the repetition of a number of different activities: feature choice, classifier choice, classifier training and classifier evaluation.

4.4 Process for Intelligent Forensic Document Examination

Within this section the question will be addressed of: *"How the concepts introduced above can be transfered to solve the problem of classifying documents according to their underlying printing technology?"*

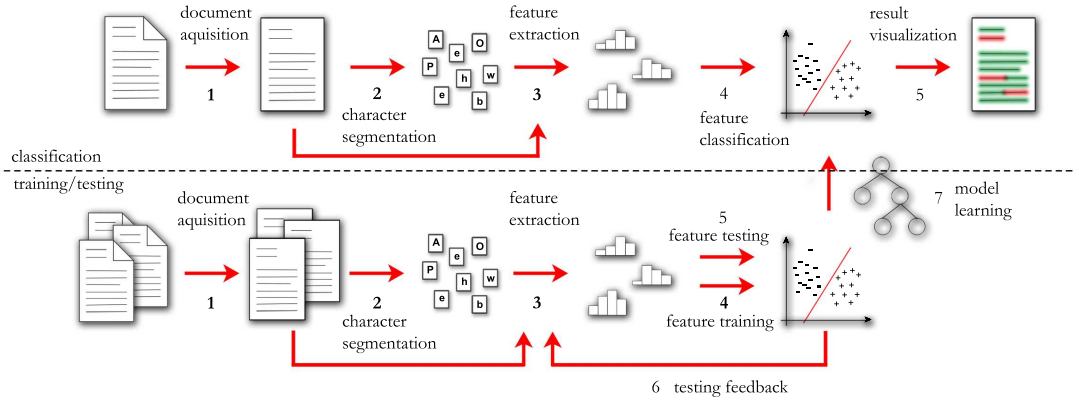


Figure 4.3: Process for intelligent printing technique recognition based on the elementary process steps of each pattern classification system in conjunction with supervised machine learning.

The proposed system architecture for intelligent forensic document examination, that exhibits the desired characteristics of supervised machine learning, is illustrated in Figure 4.3. It can be observed that architecture is divided into two phases: (1) a *training/testing phase* and (2) a *classification phase*. Within the training/testing phase the model for printing technique recognition is learned by the classifier while in the classification phase the learned model is deployed to classify unseen documents. The adapted architecture incorporates also the introduced process steps of general pattern classification systems ranging from the acquisition and preprocessing of the document to the final visualization of the classification result. The distinct process steps to achieve intelligent document examination are:

1. Document Acquisition

In the first step ground truth document samples are collected for the classifier learning procedure as well as unseen document's to be classified. To achieve printing technology recognition using digital image processing techniques it is necessary to obtain a documents information in digital form. Therefore, the document images are captured via a scanning device for further processing.

2. Document Character Segmentation

As already mentioned the objective of this thesis is the prediction of a documents underlying printing technology at (1) a global (document) level as well as on (2) a local (character) level. To perform local classification on the basis of single characters they have to be segmented within the document image. The extracted character images are then utilized for the purpose of further feature extractions. For this purpose the document images are binarized and region labeling is applied for segmentation to extract the document characters. Both methods will be described in Chapter 5.

3. Feature Selection and Extraction

As stated in Chapter 4.1 to classify a documents underlying printing technique a set of discriminant features have to be obtained from both the document image as well as from its corresponding character images. However, this feature extraction is associated with different objectives for both the training/testing and classification phase.

During the training/testing phase the goal is to discover and extract document features comprising the most discriminative behavior. As described in Section 4.2 designing a classification system is an iterative approach in which several features combinations are learned and evaluated. The final outcome of the training/testing phase is one particular set of features as well as their corresponding parameters yielding the highest classification accuracy. This set of features is selected for classification within the final system.

In contrast, the objective of the classification phase is the classification of so far unseen documents. Therefore, the according to the learning phase most discriminative features are extracted for further printing technology prediction.

An overview of the proposed features for intelligent printing technique classification and their evaluation results are described in Chapter 6 and 7.

4. Model Learning and Classification

During the next step of the printing technique recognition process the features obtained from document and character images are being classified and in dependence of the phase utilized for model learning. Again, a difference in the objectives for training/testing and the classification can be determined.

Within the learning process one or several machine learning techniques are applied. As described in Section 4.3 these techniques provide the ability to learn a classification model by processing the extracted ground truth data.

During the classification process the learned model is then utilized to classify so far unseen documents based on their extracted feature values. The class of a character or document is than predicted according to the learned model. Subsequently, a decision about the documents creation process is made by the system.

An outline of the applied machine learning techniques used within this thesis, namely Decision Tree, Multilayer Perceptron and Support Vector Machine classification will be presented in Appendix B.

5. Result Visualization

Finally, to give a system user an immediate impression of the classification results they are illustrated in a graphical manner. For this purpose the segmented document characters are colored according to their predicted printing technology. Figure 4.3 shows the visualized

results of an exemplary classified document.

The system architecture, explained above, is the outline of the work presented in the subsequent thesis chapters. Within the respective chapters the reader will find detailed explanations of particular ideas and techniques concerning the single process steps.

5. Document Preprocessing and Character Segmentation

Prior to the start of intelligent forensic document examination several document preprocessing steps are necessary. Especially on the level of single characters they have to be segmented and extracted from the document image. To achieve this a given document image is binarized using an adaptive global thresholding method. In a second step the region labeling algorithm is applied to the binarized document image in order to label and obtain the document's characters. Both steps will be described in the subsequent chapter.

5.1 Connected Components

Since digital images are two-dimensional discrete signals they are in general represented as brightness functions of two spatial variables. A digital image $I(x, y)$ can therefore be represented by one or more two-dimensional arrays of luminance values comprised of $M \in \mathbb{N}$ rows and $N \in \mathbb{N}$ columns where $x \in \{0, ..., M-1\}$ and $y \in \{0, ..., N-1\}$. The image origin is given by $(x, y) = (0, 0)$ and the next value along the first image row (column) is represented as $(x, y) = (0, 1)$ $((x, y) = (1, 0))$. Mathematically a digital image is than defined according to the following mapping:

$$\begin{aligned} I : \mathbb{N} \times \mathbb{N} &\rightarrow \mathbb{N} \\ (x, y) &\mapsto I(x, y), \end{aligned} \tag{5.1}$$

where each of these mapping elements $\{(x, y) | x \in M \wedge y \in N\}$ is referred to as pixel and the value of each pixel is given by $p(x, y)$.

Pixel Neighborhood

A given pixel at coordinates $p(u, v)$ is surrounded by horizontal and vertical neighboring pixels whose coordinates are given by:

$$p_1(x+1, y),\ p_2(x-1, y),\ p_3(x, y+1),\ p_4(x, y-1). \tag{5.2}$$

This set of pixels is referred to as *4-neighbors* of p and will be denoted by $N_4(p)$ in the following. Furthermore, the pixel $p(u, v)$ is surrounded by four diagonal neighboring pixels having the coordinates:

$$p_5(x+1, y+1),\ p_6(x+1, y-1),\ p_7(x-1, y+1),\ p_8(x-1, y-1). \tag{5.3}$$

This set of pixels in conjunction with the 4-neighbors build the so called *8-neighbors* of p, denoted by $N_8(p)$ in the following. While, each of the neighboring pixels is a unit distance from the center pixel $p(x, y)$.

Pixel Path

A path of pixels between two pixels $path(p, q)$ starting from pixel p with coordinates (x_0, y_0) to pixel q with coordinates (x_n, y_n) is a sequence containing difference pixels with coordinates as given by:

$$p_0(x_0, y_0),\ p_1(x_1, y_1), ...,\ p_i(x_i, y_i), ...,\ p_n(x_n, y_n). \tag{5.4}$$

Where any two pixels $p_i(x_i, y_i)$ and $p_{i+1}(x_{i+1}, y_{i+1}),\ i = 0, ..., n$ of the path are adjacent which indicates that $p_{i+1} \in N_8(p_i)$.

Let S be a subset of pixels obtained from a random image. Then two pixels p and q are called connected within the subset S if there exists a path $path(p, q)$ between them consisting solely of pixels in S. Choosing an arbitrary pixel p in S, the set of pixels in S that are connected to p is called a *connected component* [60].

5.2 Document Binarization

To classify the characters within an arbitrary document the characters have to be segmented and extracted from the document image. This can be achieved by separating a documents foreground (characters) from its background (non-character area). In the field of OCR and document layout analysis, where the objective is to separate character pixels from non-character pixels so called image thresholding techniques are utilized. The result

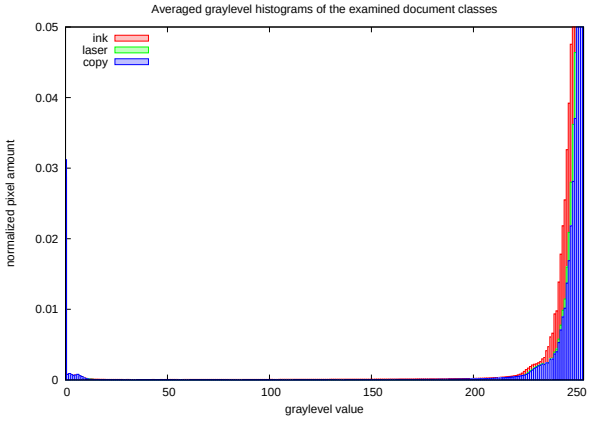


Figure 5.1: Normalized graylevel histograms for each of the examined printing techniques. The illustrated histograms are obtained by averaging the histograms of a set of documents for each document class. Due to the bimodal histogram shape the segmentation of character and non-character pixels can be achieved by global thresholding.

of thresholding is a binary image. In binary images one pixel state determines the image foreground objects and the complementary pixel state will correspond to the image background. This thresholding process is usually referred to as *binarization*.

Various approaches have been developed that estimate a proper threshold based on the structure of the underlying image histogram. A comprehensive overview is given by Sezgin and Senkur in [61].

When applying binarization to a given greylevel document image $I_G(x, y)$ every image pixel $p(x, y) \in I_G$ is tested against a thresholding function T denoted by:

$$T = T[x, y, l(x, y), I_G(x, y)] \tag{5.5}$$

where $l(x, y)$ denotes some local property of the pixel $p(x, y)$ [60]. A graylevel document image $I_G : \{0, .., M-1\} \times \{0, .., N-1\} \rightarrow \{0, .., 255\}$ can then be transformed into a binary version $I_B(x, y) : \{0, .., M-1\} \times \{0, .., N-1\} \rightarrow \{0, 1\}$ applying the thresholding function T to all pixels in $I_G(x, y)$:

$$I_B = \begin{cases} 0, & \text{if } I_G(x, y) \geq \mathrm{T} \\ 1, & \text{else.} \end{cases} \tag{5.6}$$

The foreground pixels become the pixels exhibiting a luminance value less than the threshold T and the background become the pixels exhibiting a luminance value above the threshold [61]. If T is determined by the whole image $I_G(x, y)$ the thresholding technique is referred to as *global thresholding*. The thresholding technique is named local if it also depends on some local properties $l(x, y)$, for example, the average gray level within the neighborhood of $p(x, y)$.

The underlying assumption of global thresholding is the existence of a clear intensity difference between character pixels and non-character pixels. This holds especially for regular document images captured via a scanning device. Figure 5.1 illustrates the averaged graylevel histogram of several exemplary scanned document images for each of the examined document classes. It can be observed that due to the bimodal histogram shape, the separation of character and non-character pixels can be achieved by global thresholding.

One method which is in particular suitable for thresholding of scanned document images because of its simplicity and good results was proposed by Otsu [62]. Using this thresholding approach, the pixels in I_G are divided according to a global threshold T into two pixel clusters C_0 and C_1 labeled as character and non-character pixels. The Otsu's methods basic idea is to find a threshold that minimizes the weighted sums of the inner class variances of those clusters as calculated by:

$$\begin{aligned} \omega_0(t) &= |\{(x,y)|I_G(x,y) < t\}|, \\ \omega_1(t) &= |\{(x,y)|I_G(x,y) \geq t\}|, \\ \mu_0(t) &= \frac{1}{\omega_0} \sum_{I_G(x,y)<t} I_G(x,y), \\ \mu_1(t) &= \frac{1}{\omega_1} \sum_{I_G(x,y)\geq t} I_G(x,y). \end{aligned} \tag{5.7}$$

This turns out to be the same as maximizing the between class variance of the two clusters. An optimal threshold t_{opt} is then determined by a maximization of the following quality criterion:

$$Q(t) = \omega_0(t)\omega_1(t)(\mu_0(t) - \mu_1(t))^2. \tag{5.8}$$

with:

$$t_{opt} = \underset{t=0,..,255}{\arg\max}\, Q(t). \tag{5.9}$$

Maximizing the quality function results in: (1) a clear separation of the two document pixel clusters through maximization of the cluster mean squared distances $\mu_0(t)$ and $\mu_1(t)$, (2) a preservation of the resulting binary image's information by a maximization of the document image entropy given $\omega_0(t)\omega_1(t)$. Since Otsu's method can be applied directly on gray level histograms and each pixel needs only be visited once, this approach is very fast once the histogram has been computed.

Figure 5.2 depict an exemplary extract of a document image and its binarized version obtained utilizing the Otsu binarization approach.

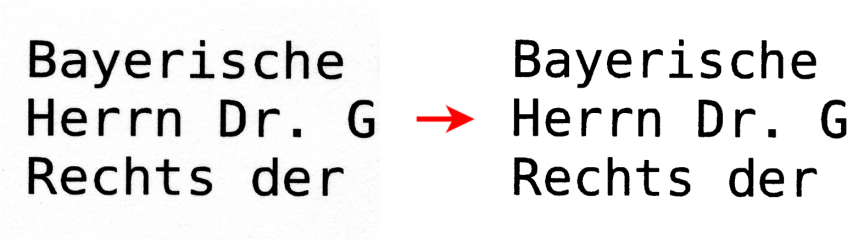


Figure 5.2: Exemplary extract of a binarized document image obtained utilizing Otsu's [62] binarization approach: (l) original document image (r) binarized result.

5.3 Document Segmentation

The subsequent task after document image binarization is the identification and extraction of the document characters. Considering a binarized document image, a character contained within the image can be interpreted as connected component of black pixels. Consequently, the identification of characters can be generalized to the recognition a binarized document's connected components.

5.3.1 Connected Component Labeling

One well-developed image segmentation technique to extract connected components is the region growing algorithm as described by Gonzalez and Woods [60]. This algorithm is based on the assumption that neighboring pixels within the same region are characterized by similar intensity values. The idea is therefore to label pixels with the same or similar intensities values as belonging to the same region according to a given homogeneity criterion.

Another approach similar to the region growing algorithm is region labeling as described by Burger and Burge [63]. The main advantage of the region labeling algorithm is its reduced storage allocation. As a result it is in particular suitable for large amounts of image data. Therefore it is used for the detection of connected components within this thesis.

The region labeling algorithm is a non recursive technique composed of two distinct sweeps through the to be processed image data: (1) a preliminary region labeling and (2) a subsequent collision resolving sweep. Finally, a label image I_L is obtained as defined by:

$$I_L : \mathbb{D}_{I_B} \rightarrow \mathbb{N}. \tag{5.10}$$

Applying the region labeling algorithm two pixels p_i and p_j are marked with the same label if they are connected within the binarized document image I_B via a 8-neighborhood connected path of foreground pixels as defined by:

$$I_L(p_i) = I_L(p_j) \leftrightarrow path(p_i, p_j), \tag{5.11}$$

where $i \neq j$.

1. Preliminary Region Labeling

During the first sweep the binary image I_B is processed and every binary image foreground pixel marked by $I_B(x, y) = 1$ will be labeled in I_L. In order to achieve this the labels corresponding to a subset of the 8-neighborhood within $I_L(x, y)$ are considered. The considered set N_{RL} of neighboring pixels is given by:

$$\{p_L^1(x-1, y-1), p_L^2(x, y-1), p_L^3(x+1, y-1), p_L^4(x-1, y)\} \in I_L. \tag{5.12}$$

Let L be the set of distinct foreground pixel labels l already applied to one of the four neighboring pixels in N_{RL}. Furthermore, let $|L|$ depict the total number of labels and L_1 the first label found in the neighborhood in N_{RL} of an arbitrary pixel $p_L(x, y)$. Labeling a so far unlabeled pixel in I_L, three cases are distinguished, depending on the pixel values in the binarized image I_B and its neighboring pixels in N_{RL}:

$$I_L(x, y) = \begin{cases} 0, & \text{if } I_B(x, y) = 0 \\ \text{new label}, & \text{if } I_B(x, y) = 1 \wedge |L| = 0 \\ L_1, & \text{if } I_B(x, y) = 1 \wedge |L| \geq 1 \end{cases} \tag{5.13}$$

While in the first case the pixel $I_B(x, y)$ of the binary image is recognized as background pixel and therefore the label 0 is assigned to its corresponding pixel within the label image $I_L(x, y)$. In the second case the pixel $I_B(x, y)$ within the binary image is recognized as foreground pixel surrounded by neighboring pixels that are all labeled as background. Therefore, a new label is assigned to the pixel since potentially a new region (connected component) is found. In the third case the pixel $I_B(u, v)$ is also recognized as foreground pixel but has at least one neighboring pixel which is belonging to an already recognized region. In this case the neighboring pixel labels have to be investigated in more detail. If only one of the neighboring pixels N_{RL} is labeled meaning the pixel belongs to an already identified region the label of this neighboring pixel is assigned to $I_B(x, y)$.

A special situation occurs if the pixel $I_B(x, y)$ is surrounded by more than one differently labeled neighboring pixels indicated by $|L| > 1$. This case is referred to as *region collision*. A region collision describes the situtation that two or more regions are preliminary labeled as distinct regions but in fact become a single region through the so far unlabeled pixel $I_B(x, y)$. This circumstance is not immediately solved in the first sweep of the region labeling algorithm but is remembered to be resolved in the second sweep. An efficient remembrance of such collisions and their involved labels can be achieved using dynamic data structures like the Union-Find structure described in Cormen et al. [64].

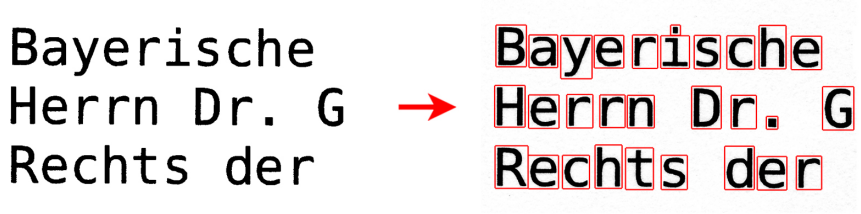


Figure 5.3: Connected components of a document detected applying the region labeling algorithm: (l) binarized document image (r) detected connected components.

2. Dissolving of Collisions

After the first sweep, all region collisions and their involved labels are memorized in the Union-Find structure. In the second sweep these label collisions are resolved by merging the connected regions of label equivalences in the way that Equation 5.11 is satisfied. Once Equation 5.11 is fullfilled, the set of the documents image connected components can be obtained and is expressed by:

$$C_i := \{p | I_L(p) = i\}, \tag{5.14}$$

where the connected component C_0 is referred to as image background. All other connected components are represented by $\{C_i | i > 0\}$.

5.3.2 Connected Components Representation

For each of the above recognized connected components a seed pixel as $p_i \in C_i$ will be stored as well as its corresponding *bounding box*. A bounding box is referred to as the smallest rectangular box of image pixels containing all pixels of a connected component and is defined in the following way:

$$B_i := [x_0^i, x_1^i, y_0^i, y_1^i] \tag{5.15}$$

where

$$\begin{aligned} x_0^i &:= min\{x | x \in C_i(x, y)\} \\ x_1^i &:= max\{x | x \in C_i(x, y)\} \\ y_0^i &:= min\{y | y \in C_i(x, y)\} \\ y_1^i &:= max\{y | y \in C_i(x, y)\} \end{aligned} \tag{5.16}$$

Figure 5.3 illustrates an exemplary outcome applying the region labeling algorithm to a binarized document image. The detected connected components are highlighted by their red marked bounding boxes. Finally, for the purpose of single character classification the

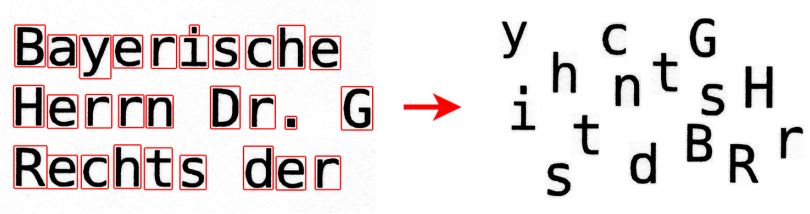


Figure 5.4: Extracted bounding boxes detected within the document image utilizing the region labeling algorithm: (l) detected bounding boxes (r) extracted character images.

detected bounding box images are extracted from the original document as illustrated in Figure 5.4.

However, as already mentioned in Chapter 3, laser printed and photocopied documents are characterized by small amounts of toner scatter distributed randomly over the document image. To avoid the extraction of toner scatter as well as other noise artefacts a threshold is applied defining the minimum area of pixels a bounding box should exhibit to be considered as document character. In Table 5.1 an overview of the optimal minimum area and their corresponding scanning resolution is presented. The optimal minimum areas are obtained by comparing the amount of characters within a set of several document images and the obtained amount of extracted character images. The minimum pixel area is considered for specific resolution as optimal if a low divergence of both amounts could be observed.

Table 5.1: Document scan resolutions and their corresponding connected component extraction parameters: (1) area size and (2) padding.

dpi	Min. Area [pixel]	Padding [pixel]
100	6	1
200	12	2
300	20	3
400	28	4
800	56	5

Furthermore, since connected component detection is performed on the binarized document images the simple extraction of corresponding bounding boxes from the original image causes a loss of important character edge information. To preserve the unique edge characteristics for each of the extracted document characters the detected bounding boxes are padded. The padding is added to the bounding box corner pixels presented in Equation 5.15. Furthermore, the applied padding values in pixel as well as their corresponding resolution is presented in Table 5.1.

6. Feature Identification and Extraction

As already mentioned in Chapter 4 the objective of pattern recognition based classification systems is the identification of a set of features, derived from an observation, as a member of a predefined class. The effectiveness of the featureset is determined based on how well feature vectors from different classes can be distinguished. To express it in the words of Ross: *"The more relevant features at your disposal, the better your decision will be"* [65]. To achieve high classification accuracy the aim is to develop and choose those features that show a high discriminative behavior.

In this chapter, the set of features utilized within this work for the detection of photocopies and the recognition of a documents underlying printing technology will be presented. A distinction between global and local document features is made. Furthermore, the proposed features are developed with the particular focus to distinguish the different document classes at low scan resolution.

6.1 Preliminary Document Image Examination

Prior to the beginning of feature development, several scanned document images that have been created by each of the investigated techniques are examined. The aim of this examination is the detection of class characteristics corresponding exclusively to a particular document class. In a subsequent step the observed characteristics are used to serve as the foundation for feature selection and development.

Figure 6.1 illustrates representative extracts of inkjet printed, laser printed and photocopied document images scanned with a resolution of $2400dpi$. A thresholded version as well as the corresponding gradient map of the extracts are added for the purpose of highlighting the characteristic differences. It will be shown that these characteristics offer the possibility to serve as valuable fingerprints in the determination of photocopies or a specific document's printing technique.

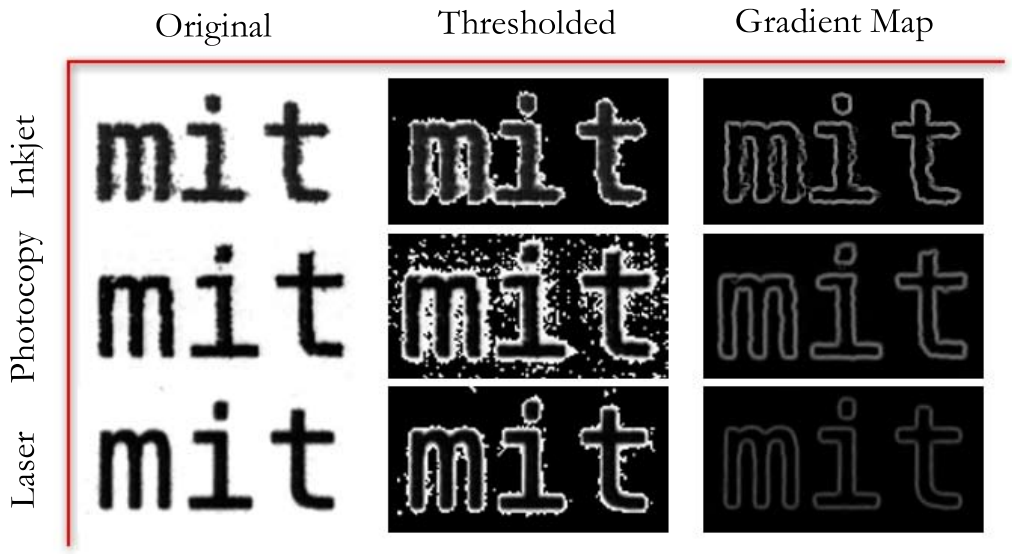


Figure 6.1: Representative document image extracts of an inkjet printed, laser printed and photocopied document scanned with a resolution of 2400dpi. A high amount of image noise can be observed within the document images obtained by laser printed and photocopied documents. Furthermore, inkjet printed and photocopied character images comprise a high amount of character edge degradation in comparison to laser printed characters.

Observing the exemplary extracts within Figure 6.1, four discriminative characteristics can be identified: (1) image noise and artifacts, (2) character edge roughness, (3) character edge contrast and (4) uniformity of printed character area. In the following, the recognized characteristics and their causes will be elaborated in more detail:

1. Image Noise and Artefacts

The thresholded images, illustrated in Figure 6.1, show that the document images of all printing techniques are affected by noise. However, comparing photocopied and laser printed document images a significant difference in the degree of noise is observable. Whereas the inkjet printed image can be recognized as nearly noise free. This observations can be traced back to the effects described in Chapter 3.2: Laser printed documents are affected by noise effects occurring during the printing process, for example, toner scatter. In addition character images of photocopied documents are affected by scanning noise caused by imaging sensor defects or dust specks.

In general it can be assumed:

$$noise_{photocopy} > noise_{laser} > noise_{ink}. \tag{6.1}$$

2. Edge Sharpness and Contrast

In the case of scanned document images edge sharpness can be interpreted as the degree of intensity change at a particular image region. The gradient between two image pixels is a measure of the highest image intensity change occurring at that particular pixel position. A high intensity change is represented by a high gradient value. High gradients are illustrated by a brighter pixel appearance in the gradient maps shown in Figure 6.1. Comparing the gradient maps it becomes obvious that laser printed document images are characterized by sharp transitions between character and non character areas. In contrast, both the photocopied as well as the inkjet printed document images show a tendency to smoother and blurred character edges. This observation can be traced back to printing substrate diffusion in the case of inkjet printed images. In the case of photocopies this effect is caused by th light diffusion during scanning of the template document. In Chapter 3.2 both effects have been described in detail.

In general it can be assumed:

$$sharpness_{laser} > sharpness_{photocopy} > sharpness_{ink}. \tag{6.2}$$

3. Edge Roughness and Corner Degradation

Edge roughness and corner degradation denotes the divergence of the printed character shape from the original template character shape. Considering again the gradient maps presented in Figure 6.1, different degrees of character shape degradation can be observed. A high degree of edge roughness can be observed in the case of inkjet printing. In contrast, the edge roughness of laser printed characters is perceived as less pronounced. Similarly to edge sharpness, edge roughness is determined by several factors including the printer resolution, dot placement accuracy, rendering algorithms and the interaction between the colorant and the paper [66].

In general it can be assumed:

$$roughness_{ink} > roughness_{photocopy} > roughness_{laser}. \tag{6.3}$$

4. Uniformity of Printed Character Area

Comparing the printed areas of the character images, as illustrated in Figure 6.1, the printed area of the laser printed document is perceived as most uniform. In contrast, the printed areas of inkjet and photocopied document images show a greater variety in intermediate graylevel values and are therefore perceived more non-uniform. Both observations can be traced back to the printing process characteristics described in Chapter 3.2 like printing substrate diffusion in the case of inkjet printing or the defects caused by scanning illumination in the case of photocopied images.

In general it can be assumed:

$$homogenity_{laser} > homogenity_{photocopy} > homogenity_{ink}. \tag{6.4}$$

6.2 Features for Printing Technique Recognition

Based on the observations determined in the previous Section 6.1 the features applied for printing technique recognition within this thesis will be described. To comply with the in Section 1.1 described approaches of Global- and Local-Document-Examination a distinction is made between global and local document features. Global features are extracted from the entire document image. The local features are extracted from the character images that are obtained by the preprocessing steps described in Chapter 5.

Furthermore, this distinction originates from the observation that especially in the case of document images scanned at low resolutions the discriminative information contained within single character images is limited to a certain extent. As a result it can become difficult to distinguish the printing technique at character level without information obtained from the entire document. The basic idea of global document features is therefore the extraction of discriminant document information which is beyond the spatial scope of the segmented document characters. Additionally, in terms of Local-Document-Examination a preclassification at document level can be used to improve the subsequent classification at character level.

6.2.1 Global Document Features

As illustrated in Figure 6.1 different printing technologies can be distinguished by the amount of noise and character degradation evident within their corresponding document images. In this context four distinct examination approaches have been developed namely:

1. Document Image Difference Analysis,
2. Document Image Gradient Analysis,
3. Document Image Frequency Analysis and
4. Document Image Multiresolution Wavelet Analysis.

The global document features obtained by the developed approaches are designed to capture the above explained discriminative characteristics.

Figure 6.2: Exemplary (l) photocopied document image extract and (r) the obtained denoised result filtered with a 5×5 Gaussian filter mask with $\sigma = 2$.

6.2.1.1 Document Image Difference Analysis

According to Gonzalez and Woods [60] the degradation within an image can be interperted as an additive component to the original image. However, especially in the case of documents originating from distinct sources the exact degradation terms are often unknown. Therefore, the straightforward substraction of noise and character blurring contained within a document image is often not feasible. Nevertheless, to obtain a document's image noise and degradation an approach proposed by Lukas et al. [67], originally developed to detect digital image forgeries, is adapted for the purpose of document examination.

Let I be a scanned document image of size $M \times N$ pixels. Let I_{noise} be the image degradation corresponding to the original image and $I_{denoised}$ be the denoised image obtained by applying denoising filter to I. The image noise I_{noise} can than be obtained by pixel-wise subtraction:

$$I_{noise} = I - I_{denoised}. \tag{6.5}$$

To reveal a denoised image version $I_{denoised}$ two linear and a non-linear filtering techniques are applied. This is done to capture different kinds of document image noise and degradations as proposed by Guo et al. [38] in the context of scanner identification. The applied filters are: (1) the mean filter, (2) the gaussian filter and (3) the median filter.

1. Mean Filter

The purpose of *mean filtering* is the extraction of high-frequency noise and degradation [38]. The mean graylevel of each pixel value $p(x, y) \in I$ is calculated by a convolution of $p(x, y)$ with a mean filter according to:

$$I^{mean}_{denoised}(x, y) = \frac{1}{|A||B|} \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x - i, y - j), \tag{6.6}$$

where $|A|$ depicts the filter size in x-direction and $|B|$ in y-direction. Equation 6.6 calculates the mean of the pixel graylevel values contained in the filter mask neighborhood.

Figure 6.3: Exemplary (l) photocopied document image extract and (r) the obtained denoised result filtered with a 3×3 median filter mask.

Mean filtering results in an image with reduced sharp image transitions referring to high image frequencies. Since only the low frequencies remain in the image mean filtering is also referred to as lowpass filtering [60].

2. Gaussian Filter

Another linear lowpass filtering method applied for the purpose of high-frequency degradation detection is *Gaussian filtering*. Applying Gaussian filtering the filter mask is modeled as isotropic gaussian curve. The Gaussian filtering formula for a pixel-wise convolution with a Gaussian filter mask is given by:

$$I^{gauss}_{denoised}(x,y) = \frac{1}{|A||B|} \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x-i, y-j)\ e^{-\frac{i^2+j^2}{2\sigma^2}}, \tag{6.7}$$

where $|A|$ depicts the filter size in x-direction and $|B|$ in y-direction. As seen in Figure 6.1 gaussian degradation can be observed in document images often due to sensor noise. Therefore, Gaussian filtering is of particular interest for detecting a documents printing technique. Figure 6.2 illustrates (l) an exemplary photocopied document image extract and the (r) obtained denoised result filtered with a 5×5 Gaussian filter mask with $\sigma = 2$.

3. Median Filter

The above presented linear filter techniques are associated with the main disadvantage that they also strongly affect contours and lines within a document image. Applying nonlinear filters like the *median filter* these effects can be avoided to a certain degree. The median pixel value ξ is calculated by halving a set of pixels in a way that half of the values in the set are less than or equal to ξ, and half are greater than or equal to ξ [60]. The median filter in a neighborhood of size $M \times N$ can than be expressed by:

$$I^{median}_{denoised}(x,y) = median\{p(x,y) |\ p(x,y) \in I, \forall x \in -a...a \wedge y \in -b...b\}, \tag{6.8}$$

where $a = (M-1)/2$ and $b = (N-1)/2$.

According to Gonzalez and Woods [60] median filtering provides excellent noise-reduction capabilities for random noise types. Furthermore, median filtering is effective in the presence of so called *"salt and pepper noise"* as observed in the thresholded images within Figure 6.1. In Figure 6.3 (l) an exemplary photocopied document image extract and (r) the obtained denoised result filtered with a 3×3 median filter mask is illustrated.

6.2.1.2 Extracted Image Difference Features

Utilizing each of the above presented methods the image degradation I_{noise} is calculated according to Equation 6.5. To obtain the degradation and noise, the mean μ and standard deviation σ are obtained from the noise image I_{noise} as given by:

$$\mu(I_{noise}) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I_{noise}(x,y), \tag{6.9}$$

$$\sigma(I_{noise}) = \Big(\frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (I_{noise}(x,y) - \mu(I_{noise}))^2 \Big)^{\frac{1}{2}}. \tag{6.10}$$

To measure the distance between I and $I_{denoised}$ the correlation ϱ and the mean squared error MSE are calculated according to Equation 6.11 and Equation 6.12.

$$\varrho(I, I_{denoised}) = \frac{Cov(I, I_{denoised})}{\sigma(I)\ \sigma(I_{denoised})}, \tag{6.11}$$

$$MSE(I, I_{denoised}) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (I(x,y) - I_{denoised}(x,y))^2. \tag{6.12}$$

6.2.2 Document Image Gradient Analysis

The analysis of the image gradients is another approach to examine the amount of noise containt within a document image as well as the edge blurring of the document's characters. The gradient at a certain pixel can be interpreted as the pixels derivative since it depicts the maximum rate of image magnitude change in a certain image direction [60].

Document gradient analysis is in particular suitable in the context of printing technique analysis. Since minor intensity variations can usually be traced back to document image noise and high intensity variations are in general associated with sharp document characters edges. As described preliminary in Section 6.1 the characteristics of these intensity variations are essential for document class discrimination.

In [31] Tchan et al. showed that gradient filters can be utilized to discriminate between different printing techniques. Within this thesis the idea of gradient filters is enhanced to obtain a document's gradient histogram. It will be shown that statistical information obtained from gradient histograms provides the ability to discriminate printing technologies.

6.2.2.1 Gradient Filters

To obtain the gradients of an arbitrary document image, a gradient filter mask is applied. Mathematically, a gradient filter calculates the image derivatives in horizontal $\frac{\partial I}{\partial x}(x,y)$ and vertical $\frac{\partial I}{\partial y}(x,y)$ direction for a given image I. This results in gradient images G_x and G_y for both directions as obtained by:

$$G_x(x,y) = [\frac{\partial I}{\partial x}(x,y)], \tag{6.13}$$

$$G_y(x,y) = [\frac{\partial I}{\partial y}(x,y)]. \tag{6.14}$$

Both derivatives form a vector whose magnitude gives the maximum rate of increase per unit distance within the original image I. The gradient magnitude at a certain pixel is calculated by the magnitude of the derivatives in both directions as given by:

$$\nabla I(x,y) = (G_x(x,y)^2 + G_y(x,y)^2)^{\frac{1}{2}}. \tag{6.15}$$

As a result the gradient map $\nabla I(x,y)$ of a document image is obtained containing the locations and magnitudes of a document image's intensity changes. Two common filters for gradient calculation are the so called Prewitt and Sobel gradient filters, which are applied for the purpose of document examination. It has to be mentioned that there exist also gradient methods that calculate also the diagonal derivatives in gradient image calculation, like the Frei-Chen gradient filter. However, since the discriminative blurring effects are in particular observable at vertical and horizontal transitions of character and non-character areas the experimental results showed that only the horizontal and vertical character gradients are of interest.

1. Prewitt Gradient Filter

The gradient applying the *Prewitt filter* is calculated by convolving the image with a small and separable filter mask. Since the Prewitt filter masks are integer valued they are computational inexpensive and especially suitable for rapid document examination. The 3×3 Prewitt filter masks in horizontal $w_x^{Prewitt}$ and vertical $w_y^{Prewitt}$ direction are given by:

$$w_x^{Prewitt} = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix}, \quad w_y^{Prewitt} = \begin{pmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}. \tag{6.16}$$

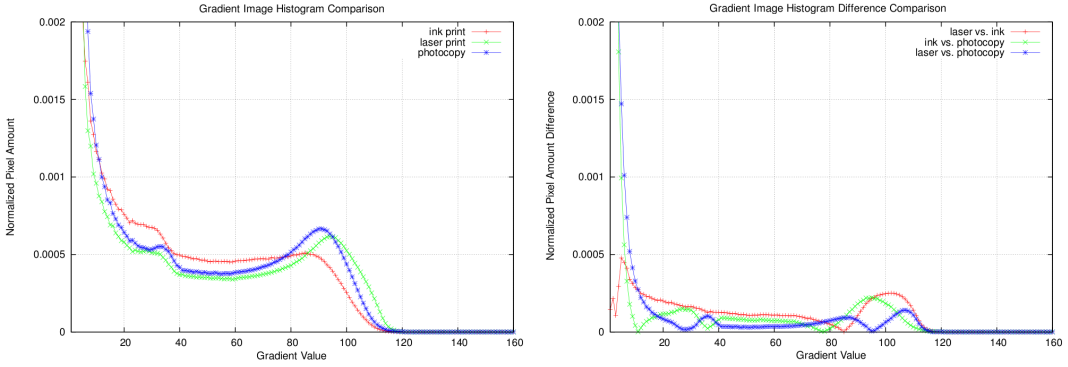


Figure 6.4: Printing technique comparison based on the averaged gradient histograms obtained from document images: (l) average gradient histograms corresponding to specific printing technique (r) printing technique gradient histogram differences. The gradient histograms are obtained by applying 3×3 Prewitt filter masks in horizontal and vertical directions.

2. Sobel Gradient Filter

Another gradient filtering technique are the appliance of Sobel filter masks. The *Sobel filter* is characterized by a slightly superior noise suppression and is therefore more suitable for character edge examination. The gradient image map is again calculated via convolution of the original document image I with a 3×3 horizontal and vertical mask w_x^{Sobel} and w_y^{Sobel} given by:

$$w_x^{Sobel} = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad w_y^{Sobel} = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}. \tag{6.17}$$

After obtaining a documents image gradient map $\nabla I(x, y)$ in a subsequent step its histogram $\nabla H(i), i = 0, ..., 255$ is calculated. Figure 6.4 illustrates the averaged gradient histogram plots obtained by set of documents for each of the examined document classes. The gradient histograms are obtained by applying Prewitt filter masks in horizontal and vertical directions. To illustrate the discriminative behavior, the gradient histogram differences obtained by pairwise substraction of the averaged histograms for each document class are shown in the left subimage.

Considering the gradient histograms for each printing technique a discrepancy could be observed in the gradient histogram values of $\nabla H(i)$ ranging from $i \in [1, 40]$ and $i \in [80, 120]$. The difference within the gradient histogram values ranging form $i \in [1, 40]$ can be traced back to several sources like background scatter, missing uniformity of the printed character areas as well as to a certain degree to the blurring at the documents character edges. In this context Figure 6.4 already revealed that inkjet printed as well as photocopied document images are characterized by a superior amount of small gradients.

Observing the distribution and the amount of gradient histogram values $\nabla H(i)$ in the interval $i \in [80, 120]$ the printing techniques impact on the document character edges becomes evident. Inkjet printed documents are characterized by a comparatively low amount of high gradient values due to the already described ink diffusion effect. Another important property can be observed comparing laser printed and photocopied document images. As already described, laser printed document scans are characterized by a sharp transition between character and non-character areas. This characteristic is in particular evident comparing the gradient histogram values of both printing techniques ranging from $i \in [100, 115]$. In contrast, the photocopied document images are showing a high amount of gradient values within the interval $i \in [80, 100]$. This observation can be traced back to the character edge degradation and therefore the loss of edge sharpness induced by photocopying documents as described in Section 3.2.3.

6.2.2.2 Gradient Analysis Features

To obtain the discriminative characteristics mentioned above, mean and standard deviation of the gradient histogram $\nabla H(i)$ are calculated as features for printing technique recognition. This is done for both gradient histogram value ranges, $i \in [1, 40]$ and $i \in [80, 120]$, as given by:

$$\mu(\nabla H(i)) = \frac{1}{N} \sum_{i}^{N} \nabla H(i), \tag{6.18}$$

$$\sigma(\nabla H(i)) = \Big(\frac{1}{N} \sum_{i}^{N} \nabla H(i) - \mu(\nabla H(i) \Big)^{\frac{1}{2}}. \tag{6.19}$$

6.2.3 Document Image Frequency Analysis

Both examination approaches introduced so far are based on the representation of a document image in the spatial domain. As described in the analysis of document image gradients, printing technique discrimination can be achieved examining the transition characteristics between character and non character areas.

Observing the concept of frequency in this context, as the rate at which image pixel intensities change, it becomes obvious that frequency analysis can be utilized for character edge and document noise analysis. High spatial frequencies are denoted by large graylevel alterations within a small image area. In contrast, low spatial frequencies are denoted by large areas of nearly constant graylevel values.

In awareness of this coherency, document images created by the different printing technologies are examined in the frequency domain. Therefore, a document image representation

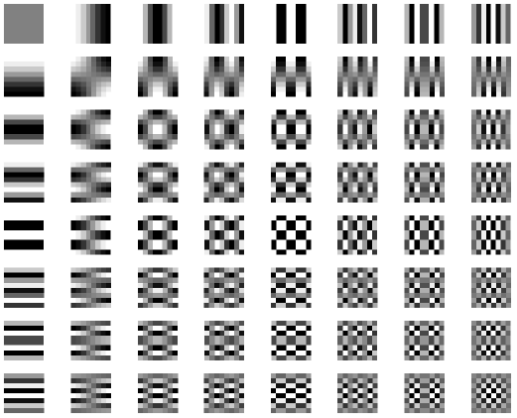


Figure 6.5: The first 64 basis functions of the Discrete Cosine Transformation (DCT).

within the frequency domain has to be obtained. A very popular frequency transformation technique in imaging and video compression is the *Discrete Cosine Transformation (DCT)* which was first applied in 1974 by Ahmed et al. [68]. In the context of document examination this transformation is used for another purpose. According to Gonzalez and Woods [60], the DCT transform can be interpreted as a "mathematical" prism that splits a given image into its various cosine frequency components.

6.2.3.1 Discrete Cosine Transformation (DCT)

The DCT transforms a given image from the spatial domain into another space called the frequency domain. Therefore, the DCT applies a base transformation to the image to express the image information in terms of cosine functions oscillating at different frequencies. Figure 6.5 illustrates the first 64 basis functions of the DCT.

Let I be an arbitrary document image of size $N \times N$ in the spatial domain, its discrete cosine transformed representation in the frequency domain $F(u, v)$ can be obtained by the following transformation:

$$F(u,v) = \frac{2}{N} C(u)C(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(u,v) cos(\frac{(2x+1)u\pi}{2N}) cos(\frac{(2y+1)v\pi}{2N}), \tag{6.20}$$

where x and y are the spatial coordinates of the image I and u and v are the coordinates in the frequency domain of the discrete cosine transformed image $F(u, v)$.

To receive a truly orthogonal transformation the terms $C(u)$ and $C(v)$ are calculated by:

$$C(z) = \begin{cases} \frac{1}{\sqrt{2}}, & z = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{6.21}$$

As shown in Equation 6.20 the values of the original image $I(x, y)$ are multiplied by cosines of distinct frequencies. The value range of $F(u, v)$ is called the frequency domain because u and v are determining the frequencies of the transform.

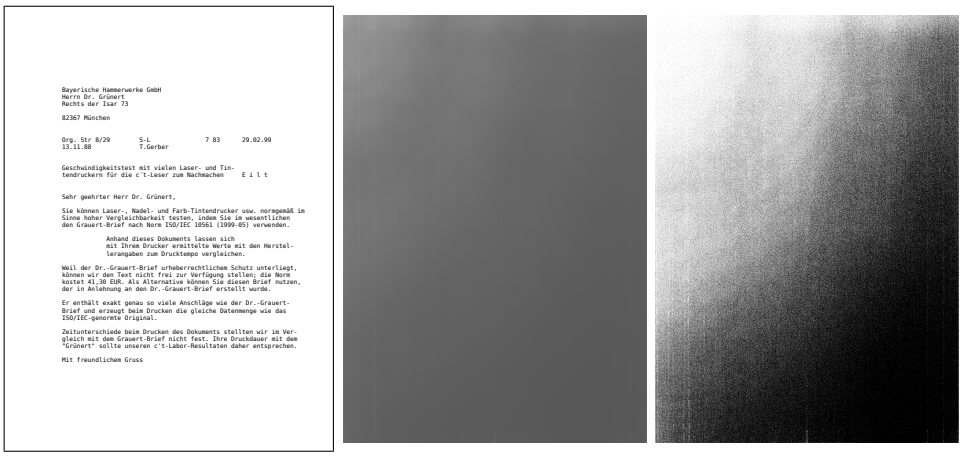


Figure 6.6: Exemplary Discrete Cosine Transformation (DCT): (l) document image in the spatial domain (m) corresponding DCT transformed frequency spectrum in the frequency domain (r) binarized spectrum using Otsu's method [62].

Each of the $N \times N$ terms in $F(u, v)$ are called frequency coefficients of the transform. The frequency coefficients contain important information about the distribution of the frequency basis functions within the original image. The larger the value of a frequency basis coefficient is, the higher the proportional occurrence of its corresponding frequency.

For visualization purposes the calculated frequency spectrum is scaled according to Equation 6.22:

$$F'(u, v) = \frac{F(0, 0)}{255} \log(1 + |F(u, v)|), \tag{6.22}$$

where $F(0, 0)$ defines the origin of the transformed image comprising the highest amount of image energy. For fast calculation of the DCT coefficients the "FFTW3" library is used as outlined in [69].

6.2.3.2 Frequency and Spatial Domain Relationship

For the purpose of printing technique examination, some important relations between the frequency components and the spatial characteristics of the original image are evident. In a general perspective the concept of frequencies defines the rate of intensity change from pixel to pixel. Therefore, a relationship between the intensity variations in the original image and the frequency values obtained by the DCT exists.

Figure 6.6 illustrates a standard office document scan as well as its corresponding normalized frequency domain transformation result. High coefficient values $F'(u, v)$ are marked by brighter pixels within the transformation image. To enhance the perceivability of the coefficient strength as well as their distribution a binary version of the spectrum is presented using Otsu's [62] global thresholding method.

The transformed image can be interpreted as follows: The origin or so called Direct Current (DC) coefficient given by $F(0,0)$ is located at the transformed images upper left corner. It corresponds to the constant element of the original image and depicts the average graylevel in I. In general the transformation image origin contains most of the originals image energy and therefore exhibits the highest coefficient value.

The DCT transformed spectrum and its binarized version in Figure 6.6 show a smooth transition from brighter to darker image values. Moving further away from the transformation image origin relates to coefficients of higher frequencies and therefore to increased intensity variation in the original image. The coefficients of these frequencies, also referred to as Alternating Current (AC) coefficients, are correlated to the amount of image details.

High values in an increased distance from the transformation image origin correspond in particular to edges of image objects that are characterized by abrupt graylevel changes [60]. The thresholded frequency image presented in Figure 6.6 shows that most of the document images energy is concentrated in the upper left corner and result in the lower frequency components of the image. The spectra also show no visible coefficient values beyond a certain distance from the origin, which marks the maximal intensity change in the original image.

6.2.3.3 Comparison of Printing Technique Frequency Spectrums

Frequency spectrum comparison is realized by first calculating the average frequency spectrum based on a set of sample documents created for each printing technique. Subsequently, the average frequency spectrums are examined in a pairwise manner. For each pair, of printing techniques the average image frequency coefficient values corresponding to the same frequencies are compared. Finally, the frequencies in the comparison image are colored according to the technique exhibiting the superior frequency coefficient value.

Figure 6.7 illustrates the results of the pairwise printing technique comparison (superior laser printing coefficients are marked as red, superior inkjet printing coefficients are marked as green and superior photocopy coefficients are marked as blue). Therefore, the left image displays the obtained result of comparing the average laser printed spectrum versus the average spectrum of the photocopied documents. The center image illustrates the average laser printed spectrum versus the inkjet printed and the left image illustrates the average inkjet printed spectrum versus the average spectrum of the photocopied documents.

The comparison images in Figure 6.7 reveal clear differences between the compared printing techniques. Especially, comparing the average DCT coefficients spectrum of laser printed documents in comparison to both other printing technologies is showing a radial symmetric pattern. For the comparison of the inkjet and photocopied document spectra the high frequency pattern is less evident but still recognizable.

As mentioned preliminary, this observation can be traced back to the relationship between the spatial and frequency domain. In comparison to photocopied and inkjet printed doc-

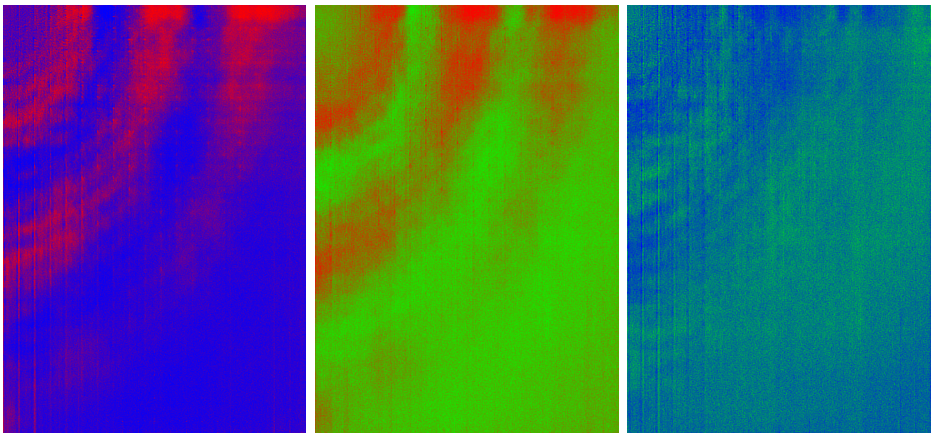


Figure 6.7: Comparison of printing technique classes in the frequency domain based on the average difference spectrum calculated for each pair of techniques: (l) laser printing [red] vs. photocopy [blue], (m) laser printing [red] vs. inkjet printing [green] and (r) inkjet printing [green] vs. photocopy [blue].

uments, laser printed characters are characterized by sharp transitions between character and non-character areas in general. Therefore, this property is in particular evident in specific DCT coefficient values.

From the comparison images in Figure 6.7 it can also be recognized that starting at the spectra origin the DCT coefficients corresponding to horizontal and vertical frequencies especially show a discriminant behavior. This observation can be traced back to the properties of latin fonts where an unproportionally high amount of sharp transitions between character and non-character areas occur at horizontal and vertical dimensions. However, sharp horizontal and vertical transitions are in particular fragile to edge blurring induced by inkjet printing and photocopying. As a result these printing techniques show a tendency to lower coefficient values for the high frequencies of the horizontal and vertical DCT spectrum.

6.2.3.4 DCT Frequency Analysis Features

Based on the comparison images presented in Figure 6.7 and the mentioned printing technique characteristics it can be assumed that different printing techniques are distinguishable according to their frequency spectrum. This especially holds for frequency subband coefficients corresponding to vertical and horizontal image intensity variations in the spatial domain.

A two step approach is applied to determine the DCT coefficient distribution and their strength within the frequency spectrum subbands which are derived from an arbitrary document image. First, the DCT coefficients of a particular subband were extracted from the frequency spectrum and in a subsequent step statistical features were obtained from

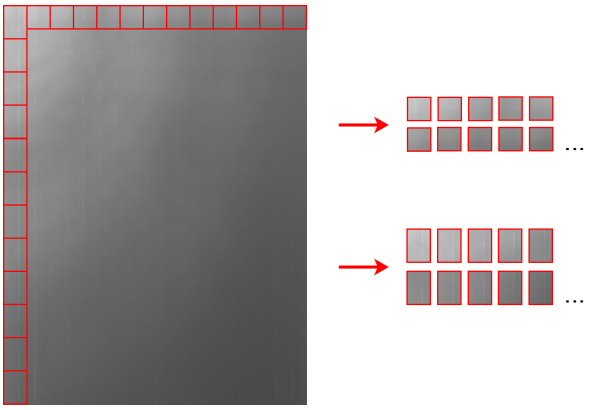


Figure 6.8: Horizontal and vertical frequency subband coefficient detection and extraction: (l) Detected and horizontal and vertical frequency subband coefficients (r) extracted frequency subband coefficients.

those subband coefficients:

Frequency Subband Extraction

As obvious in Figure 6.7, a high discriminative character is in particular evident in the horizontal and vertical frequency subbands of the spectra. Therefore, these frequency subband coefficients are obtained from each document's spectrum by extracting a set of horizontal and vertical subband boxes as illustrated in Figure 6.8. To also capture the frequency distortion caused by the aspect ratio of the document images the subband boxes aspect rate is adapted to $\frac{1}{\sqrt{2}}$.

Statistical Feature Extraction

Let k be the number of frequency subband boxes box_i obtained from a documents frequency spectrum $F(u, v)$ as given in Equation 6.22. To obtain a documents unique horizontal and vertical frequency subband pattern the mean and standard deviation of the coefficients is calculated according to Equations 6.23 and 6.24 for each subband box_i:

$$\mu(box_i) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} F(m,n), \tag{6.23}$$

$$\sigma(box_i) = \left(\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (F(m,n) - \mu)^2 \right)^{\frac{1}{2}}, \tag{6.24}$$

where $m, n \in box_i$ and M, N indicate the size of the box_i.

6.2.4 Document Image Multiresolution Wavelet Analysis

Considering the DCT presented in the previous Section 6.2.3, it can be recognized that its application is associated with two major disadvantages: (1) the documents entire spa-

tial information is lost within the transformed image and (2) the basis functions of the transformations are restricted to cosine base functions.

Applying a similar transformation technique referred to as *Wavelet Transformation*, the image is transformed using small wave functions, called *wavelets* as base [60]. Having the ability to choose different types of small transformation functions the Wavelet Transformation becomes suitable for the purpose of document examination. Since wavelets can be chosen which correspond to character edge characteristics of certain printing techniques.

The foundations of the Wavelet Transformation were laid by Haar in 1910 in which he proposed a new approach for signal analysis [70]. In comparison to other transformation techniques the Wavelet Transformation is a relatively recent approach. Since it was first utilized 1984 by Morlet and Grossmann [71] in relation to signal decomposition.

6.2.4.1 Discrete Wavelet Transformation (DWT)

Applying the *Discrete Wavelet Transformation (DWT)* to an image the basis of the transformed image could be any function ψ that fullfills the two following permissibility conditions:

$$0 < 2\pi \int \frac{|\Psi(u)|^2}{|u|} du < \infty, \tag{6.25}$$

$$\int \psi(x) dx = 0, \tag{6.26}$$

where $\Psi(u)$ depicts the Fourier Transformation[1] of the function.

Equation 6.25 demands the compactness of the function indicating that its corresponding frequency spectrum has to be restricted to certain frequency interval. The condition given by Equation 6.26 demands the oscillation of the function since its integral has to be zero [73]. A function that fullfills both permissibility conditions is than referred to as wavelet.

In the context of Discrete Wavelet Transformation a wavelet depicts a template function also named *mother wavelet*. The main purpose of the mother wavelet is to serve as source function for the generation of the *daughter wavelets* which are translated and scaled versions of the mother wavelet.

[1] Similar to the Cosine Transformation the Fourier Transformation decomposes a signal into sine and cosine functions oscillating at different frequencies. A detailed introduction of Fourier Transformation can be found by Kammler [72].

A definition of a general mother wavelet is given by:

$$\psi_{\tau,s}(x) = 2^{\frac{s}{2}}\psi(2^s x - \tau), \tag{6.27}$$

where $\tau, s \in \mathbb{Z}$.

It can be observed from Equation 6.27 that the mother wavelet function ψ exhibits two parameters, a scale factor s and a translational value τ. Utilizing these parameters a set of daughter wavelets oscillating at different frequencies are derived. It can be shown that the mother wavelet is an orthonormal basis within the Lebesgue space of functions $L^2(R)$.

Since the initial introduction of the Wavelet Transformation a variety of wavelet functions have been developed. The most prominent one, originating from Haar's initial work, is the Haar Wavelet as defined by:

$$\psi(x) = \begin{cases} 1, & 0 \leq x < 0.5 \\ -1, & 0.5 \leq x < 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{6.28}$$

Another prominent wavelet family fullfilling the above mentioned permissibility conditions are the Daubechies 1-20 wavelets developed by Daubechies [74]. The Daubiches wavelets are characterized by a good localization in the spatial and the frequency domain.

Using one of the mother wavelet functions mentioned above an arbitrary discrete signal $h(x), x = 1, ..., N$ can be decomposed into a multitude of different scaled daughter wavelets. This transformation referred to as Discrete Wavelet Transformation is defined according to:

$$W_{\psi,h}(\tau, s) = \frac{1}{\sqrt{|s|}} \sum_{x=1}^{N} h(x)\psi(\frac{x-\tau}{s}), \tag{6.29}$$

Each of the $W_{\psi,h}(\tau, s)$ terms is called a wavelet coefficient of the transform and depicts the proportional occurance of its corresponding daughter wavelet within $h(x)$. Choosing a relatively small scale factor, the resulting daughter wavelet is more receptive to high frequencies in $h(x)$ while for larger scale factors it becomes more receptive to low frequencies.

6.2.4.2 Multiresolution DWT Analysis

In digital image processing the DWT is often utilized for noise and edge detection applying so called *Multiresolution Analysis (MRA)* to an image [75]. Therefore, a given image is continuously filtered with a wavelet of increasing scale. The filter results obtained by an increasing wavelet scale exhibit the coarse structures within the original image since these structures correspond to lower frequencies. Filtering in contrast the image with a

low scaled wavelet results in a filtered image containing the fine structures of the image corresponding to high frequencies.

Considering the explained properties of the MRA it becomes obvious that this technique could also be utilized for the purpose of printing technique recognition. This originates from the assumption that image noise and sharp character edges are reflected within the wavelet coefficients obtained from low scaled wavelet transformations. In comparison, information about blurred character edges are reflected in the wavelet coefficients obtained by filtering the image with high scaled wavelets.

Mallat demonstrated in [76] that a wavelet based MRA of a given signal could be achieved by filtering the signal with a qualified pair of highpass and lowpass filters. Where the highpass-filter h_0 corresponds to an arbitrary wavelet function and the lowpass-filter h_1 corresponds to a scaling function which is orthogonal to the wavelet function. A detailed description of the criteria to be fullfilled by a qualified filter pair in order to become suitable for MRA can be found in Burke-Hubbard [77].

To perform MRA of an arbitrary signal the scaling function is utilized to approximate a given image at different resolutions. At every signal resolution scale a two step approach is applied to the signal consisting of (1) filtering and subsequent (2) subsampling. Both steps of MRA will be explained within the context of document analysis:

Image Filtering

Within the first step of MRA, both the scaling function and the wavelet function are applied to the document image. As already mentioned, the scaling function serves as lowpass filter h_1 and therefore captures only the low frequencies of the image. Applying the scaling function to an image therefore accompanies with the loss of fine image details in the filtered result. However, the idea of MRA is to capture exactly the lost image detail by applying a suitable wavelet function. This is achieved since the wavelet serves as the highpass filter h_0 of the MRA and is orthogonal to the scaling function. The fine image details lost in the appliance of the scaling function are captured by the highpass-filtered image I_{high}. As a result the original document image I is decomposed into its high frequency contents I_{high} and its low frequency contents I_{low} at every resolution scale s.

Image Subsampling

After filtering the resulting lowpass- and highpass-filtered images are subsampled. Subsampling is achieved by removing every second pixel value from both filtered images I_{high}^{s+1} and I_{low}^{s+1}. The image filtering step described above is than recursively applied to the subsampled lowpass filtered image I_{low}^{s+1} as shown in Equation 6.30:

$$I^s = I_{high}^{s+1} \oplus I_{low}^{s+1}. \tag{6.30}$$

As a result a repeating decomposition $\oplus$ of the image into high and low frequencies is achieved at lower resolution scale. As shown by Mallat [78] this process of filtering and subsampling equals an increase of the scale parameter of the the wavelet function given in Equation 6.29.

Following the described process of filtering and subsampling a signal is successively decomposed until the size of the subsampled signal becomes smaller than the mask of the applied filters. While the last lowpass-filtered image corresponds to the average frequency within the image. Mallat and Mayer proved that using their MRA approach with a qualified pair of highpass and lowpass filters no signal information will be lost [78]. Since the original signal is orthogonally decomposed into signal approximations and signal details at every scale of resolution. This technique is also referred to as the *Fast Wavelet Transformation (FWT).*

As shown by Equation 6.29 the DWT is defined for one dimensional signals. In order to apply MRA also to two dimensional signals, like digital images, the filters of the MRA are successively applied to the image rows and columns. At every decomposition stage the corresponding highpass- and lowpass filter can be applied in four filtering combinations to the original image I^s. This results in four decomposition images for each scale:

1. A *Low-Lowpass* filtered image I_1^{s+1} in which pixel rows and pixel columns are filtered with the lowpass-filter. I_1^{s+1} contains low frequencies in both image directions within I^s.

2. A *High-Lowpass* filtered image I_2^{s+1} in which pixel rows are filtered with the highpass filter and pixel columns are filtered with the lowpass filter. I_2^{s+1} contains the high horizontal frequencies within I^s.

3. A *Low-Highpass* filtered image I_3^{s+1} in which pixel rows are filtered with a lowpass filter and image columns are filtered with a highpass filter. I_3^{s+1} contains the high vertical frequencies within I^s.

4. A *High-Highpass* filtered image I_4^{s+1} in which the pixel rows as well as the pixel columns are filtered with a highpass filter. I_4^{s+1} contains high frequencies in both image directions within I^s.

while S defines the resolution scale $S = 1, ..., 4$ of the applied filter combination.

Figure 6.9 illustrates the results of MRA applied to an exemplary character combination "ba" which was scanned at a scan resolution of $800dpi$. High frequency coefficients at each scale of the High-Lowpass, Low-Highpass and High-Highpass filtered MRA images are presented by darker pixel values. Observing the decomposition images for each scale it can be recognized that high frequency information is in particular evident at the character

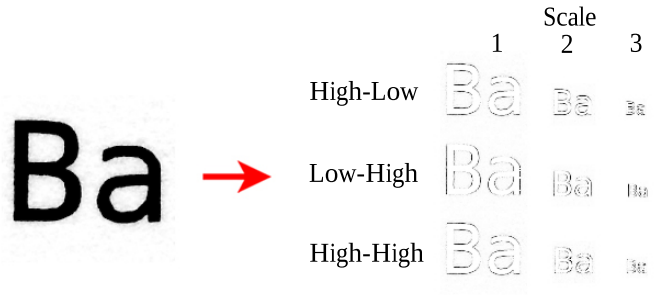


Figure 6.9: Multiresolution Analysis (MRA) applied to an exemplary character combination "ba" which was scanned at a resolution of 800*dpi*. High frequency coefficients at each scale of the High-Lowpass, Low-Highpass and High-Highpass filtered MRA images are presented by darker pixel values. Observing the decomposition images for each scale it can be recognized that high frequency information is in particular evident at the character edges of the original image.

edges of the original image.

6.2.4.3 DWT Frequency Analysis Features

For the purpose of printing technique recognition at each scale of MRA the mean and standard deviation is obtained from each of the three Highpass-filtered decomposition images I_2, I_3 and I_4 at each scale of decomposition:

$$\mu(I_i^s) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I_i^s(m,n), \tag{6.31}$$

$$\sigma(I_i^s) = \left(\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} (I_i^s(m,n) - \mu)^2 \right)^{\frac{1}{2}}. \tag{6.32}$$

As a result statistical information about the high frequency as well as the low frequency document image content is obtained at distinct frequency subbands.

6.2.5 Local Character Features

As described preliminary in Section 1.1, the basic idea of local document features is the detection of photocopies and the recognition of printing technique at a document's character level. This Local-Document-Examination allows the examination of printing technique consistence within a document.

Consulting the related work, that is reviewed in Chapter 2.3, several features for printing technique recognition at character level have been proposed within the work of Lampert et al. [14], Schulze et al. [29] and Mikkilineni et al. [30]. Especially the features proposed by Lampert et al. and Mikkilineni et al. show a high degree of similarity. Both utilized graylevel co-occurrence matrices in their work.

Therefore, for the purpose of document character classification, the features of Lampert et al. [14] and Schulze et al. [29] are evaluated. In the following a brief outline of these features will be presented.

6.2.5.1 Area Difference

Due to the obvious edge degeneration effect, described in Section 6.1, of inkjet printed and photocopied characters, their edge transition contains more intermediate graylevel values. The area difference feature exploits this fact and therefore calculates a measure for the amount of intermediate graylevel values found at the edge of a given character. This is achieved by performing two different binarizations onto the character image. One binarization with the Otsu threshold [62] described, in Chapter 5.2, and the other with an Otsu threshhold differing by δ value. The area difference D_A between both binarizations is than obtained by:

$$D_A = \frac{|A_{otsu} - A_{otsu+\delta}|}{A_{otsu}}. \tag{6.33}$$

6.2.5.2 Correlation Coefficient

The preliminary evaluation of the differences in edge sharpness and roughness is also the basic idea of the correlation coefficient feature. Similar to the area difference an Otsu binarized image is created. Since only the character contour is the *Region Of Interest (ROI)*, an edge image is created by dilating the original image with a circular mask. The correlation coefficient C_C is then obtained from the edge image A and its binarization B.

$$C_C = \frac{\sum_{i,j}(A[i,j] - \overline{A})(B[i,j] - \overline{B})}{\sqrt{\sum_{i,j}(A[i,j] - \overline{A})^2}\sqrt{\sum_{i,j}(B[i,j] - \overline{B})^2}}, \tag{6.34}$$

where $\overline{A}$ and $\overline{B}$ depict the mean of A and B inside the ROI.

6.2.5.3 Graylevel Co-Occurrence Features

In order to describe also the texture of the extracted character images, co-occurrence matrices [79] are proposed by Lampert et al. [14]. The co-occurrence matrices are obtained from two different transformations of the grayscale image and its original. The transformations applied are:

- **Gaussian Filter**
  The first transformation is obtained from blurring the original image with a Gaussian filter.

- **Local Binary Maps**
  For the second transformation the *rotation invariant local binary map* as proposed by [80] is used.

As a result, a 2D co-occurrence histogram $p[i,j]$ for each of the transformed images is created. The calculated histogram therefore states how often a graylevel value i within the ROI in image I occurs in combination with value j in the transformed image J. Four features, namely *contrast, correlation, energy and homogeneity* are then extracted from each of the histograms, according to:

$$contrast = \sum_{i,j} |i-j|^2 p[i,j], \tag{6.35}$$

$$correlation = \frac{\sum_{i,j}(i-\mu_x)(j-\mu_j)p[i,j]}{\sigma_x \sigma_y}, \tag{6.36}$$

$$energy = \sum_{i,j} p^2[i,j], \tag{6.37}$$

$$homogeneity = \sum_{i,j} \frac{p[i,j]}{1+|i-j|}, \tag{6.38}$$

with

$$\mu_x = \sum_{i,j} ip[i,j] \qquad \mu_y = \sum_{i,j} jp[i,j],$$

$$\sigma_x = \sqrt{\sum_{i,j}(i-\mu_x)^2 p[i,j]} \qquad \sigma_y = \sqrt{\sum_{i,j}(j-\mu_y)^2 p[i,j]}.$$

6.2.5.4 Perimeter Based Edge Roughness

Another approach for measuring the roughness of printed characters proposed by Schulze et al. [29] is the perimeter comparison of the binarized and a smoothed binarized character image. For the binarization the first valley next to the lowest gray level found in the histogram of the original image is chosen as global threshhold. A character image binarized with threshhold T depicts the perimeter p_b. After applying smoothing with a median filter, the smoothed perimeter p_s can be obtained. The perimeter based edge roughness is than calculated based on the difference between both parameters:

$$R_{PBE} = \frac{p_b - p_s}{p_s}. \tag{6.39}$$

6.2.5.5 Distance Map Based Edge Roughness

A second feature for the purpose of edge roughness examination compares the character perimeter size of the binarized character image I_b and its smoothed version I_s. Therefore, this feature relates edge pixel locations via distance mapping. In a first step the distance map is initialized with the values taken from the smoothed binary image. Within a second step the distances are propagated by filling all entries of the distance map with the minimal distance to the nearest edge pixels of I_s as given by

$$DIST = \min\{d|d = \sqrt{(x-m)^2 + (y-n)^2}\}, \tag{6.40}$$

where $(x, y) \in I_b$ and $(m, n) \in I_s$. As a third step, the distance map histogram is obtained and the histograms *mean, sample standard deviation, maximal* and *relative* distance are calculated to form a feature vector exhibiting relative distance defined as:

$$DIST_{rel} = \frac{\sum\limits_{x \in Edge} dist_{map}(x)}{|Edge|}, \tag{6.41}$$

with *Edge* the set of edge pixels, and the maximal distance:

$$DIST_{max} = \max_{d \in dist_{map}} \{d - \overline{dist_{map}}\}$$

6.2.5.6 Gray Value Distribution on Printed Area

As stated in the preliminary explanations of this chapter the printed regions uniformity can be used for printing technique discrimination. Therefore, a mask for the printed area is constructed by Min-Max threshholding applied to the regions containing black pixels. Afterwards a gray value histogram is extracted from the masked image. To characterize the obtained histogram, linear regression is applied to the grayvalue histogram while the regression lines parameters a,b are used as feature values. The parameters a,b are calculated by:

$$b = \frac{\sum\limits_{i=0}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=0}^{n}(x_i - \overline{x})^2} \tag{6.42}$$

$$a = \overline{y} - b\overline{x} \tag{6.43}$$

with

$$\overline{x} = \sum_{i=0}^{n} x_i, \qquad \overline{y} = \sum_{i=0}^{n} y_i, \qquad n = 255.$$

Having introduced the different types of local and global features in a subsequent step the extracted features are applied to distinct machine learning approaches for the purpose of intelligent document examination as mentioned in Chapter 4.

7. Classification Results and Hypotheses Evaluation

This chapter documents the evaluation of the four initial hypothesis. Therefore, the global and local features as proposed in Chapter 6 are applied to distinct machine learning techniques as described in Chapter 4. To ensure comparability of the results, the evaluation is undertaken based on a representative ground truth document database. Furthermore, the obtained classification results are compared by different performance measures depending on the evaluated feature type.

7.1 Experimental Setup

Before presenting the performed hypotheses analysis, the experimental setup is outlined.

7.1.1 Ground Truth Data Generation

As described in Section 4.2 to train and evaluate a model learned by a classifier as well as its corresponding set of features, ground truth data is inevitable. In a first step well known document image databases are reviewed for the purpose of document printing technique recognition:

1. UW English Document Image Database I - III [81, 82, 83],
2. Medical Article Records System (MARS) [84],
3. MediaTeam Oulu Document Database [85],
4. Google 1000 Books Project.

However, none of the reviewed databases is currently providing a document annotation of the printing technique used for its creation. As a result, the necessity emerged to create a new document image database annotated with the needed ground truth information.

An important aspect in terms of ground truth generation is the selection or creation of a suited test-document. In German speaking countries a document called the "Grauert" letter, implementing the DIN-ISO 10561 standard, is used for the test of printing devices. The "Grünert" letter, which is derived from this document, yields the same results in printer tests. Because of its high similarity in layout and content to regular written business letters, the 'Grünert' letter[1], is used as template for ground truth database creation. Another advantage of the Grünert letter is its free of charge online availability[2]. An exemplary printout of the Grünert letter can be found in Appendix D.

The created document database, comprises 49 different laser printouts, 14 different inkjet printouts as well as 46 photocopied documents. The variety of printing and photocopier device manufacturers exhibits all major brands typically present in (home) office environments, for example Hewlett-Packard™, Epson™, Canon™and Ricoh™. To create a realistic evaluation scenario half of the photocopied documents are generated using laser printed templates while the other half is based on inkjet printed templates. Furthermore, exclusively photocopies of the first generation are contained in the ground truth database implying that no copy of an already copied document is applied. This is emphasized since first generation photocopies are in particular difficult to distinguish from their printed templates.

7.1.2 Document Scanning

In order to find a proper sampling frequency all documents are scanned at resolutions of $100dpi$, $200dpi$, $300dpi$, $400dpi$, and $800dpi$. For scan resolutions up to $400dpi$ a "Fujitsu 4860®" high speed scan device is used. This scan device is especially designed for high throughput scan procedures in which huge amounts of documents have to be processed in a limited time interval. For the purpose of saving scan time and storage space the maximal scan resolution of the "Fujitsu 4860®" is limited to $400dpi$. To evaluate feature performance at a higher resolution and independently from a particular scan device the scans of $800dpi$ are obtained from the "EPSON 4180®" high resolution scan device. This scan device is designed for domestic use and can therefore be found in regular (home) office environments. No special filters or preprocessing techniques are applied during document scans. All document images are scanned as 8 bit grayscale and stored in the TIFF dataformat to avoid further information loss.

[1] The 'Grünert' letter exhibits the following characteristics, fonttype *'Courier New' normal*, fontsize *12 pt*, lineheight *12 pt*

[2] Websource: `http://www.heise.de/ct/testbilder/gruenert-brief/`, as of October 27, 2008

7.1.3 Document Feature Classification

For the purpose of feature evaluation the question arose if one of the actual supervised machine learning techniques can reach the general superiority compared to all the other techniques. In 1995 and 1997 Wolpert and Macready [86, 87] showed that for static and time dependent search and optimization problems *"the average performance of any pair of algorithms across all possible problems is identical"*. This insight was postulated in the so called *"No-Free-Lunch Theorem"*.

In [88] Wolpert transfered the No-Free-Lunch Theorem to supervised learning. Assuming a noise free scenario one can not state the general superiority of a specific supervised learning algorithm in every classification task. Or as stated in the words of Wolpert, *"for any two learning algorithms, there are just as many situations in which algorithm one is superior to algorithm two as vice verca"* [88]. This implies that it is desireable to compare the appliance of different supervised learning techniques for a given classification task.

To comply with the No Free Lunch Theorem, the subsequent evaluation is performed on the basis of three well known supervised machine learning techniques: (1) C4.5 Decision Tree, (2) Multilayer Perceptron and a (3) Support Vector Machine. Both the C4.5 Decision Tree [89] and the Multilayer Perceptron [90] are derived from the Weka collection of machine learning algorithms as described in Witten and Eibe [91]. For the purpose of support vector classification the LibSVM library by Chang and Lin [92] is applied.

7.2 Performance Measures

In terms of feature and classifier evaluation a set of performance measures is used within the global and local feature evaluation experiments.

7.2.1 Global Performance Measures

To evaluate the classification capability of the extracted features without losing the generalization ability of the learned model *stratified cross validation* is applied for global feature evaluation.

Stratified Cross Validation

In stratified N-fold cross validation a given feature data set D is splitted into S distinct feature data subsets, D_s , $s = 1, \ldots, S$. While each of the resulting data subsets contains the fraction of $\frac{S-1}{S}$ training samples. In contrast to randomly selecting data samples for each subset D_s in stratified sampling the data samples are selected maintaining the original class probability distribution within D. In a subsequent step the resulting feature data subsets D_s are divided to form a testing data set and $S-1$ training data sets. This procedure is applied S times in a way that each of the S subsets is used once as test set

and the other $S-1$ subsets are put together to form the training set.

Cross Validation Accuracy

Let D be the set of scanned document images and $d_i,\ i = 1, \ldots, N$ depicts an arbitrary document within D. For each cross validation trail $s \in S$, the accuracy acc_s given by the percentage of correct classified documents within the actual training set of trail s is calculated. Finally, to calculate the global classification performance, the average accuracy across all S trials is computed as the overall result of the stratified S-fold cross validation as given by:

$$\mu(acc_s) = \frac{1}{S} \sum_{s=1}^{S} acc_s. \tag{7.1}$$

7.2.2 Local Performance Measures

To evaluate the classification performance of local features the ground truth document set is divided into a training set D_{tr} and testing set D_{ts}. Three performance measures are obtained for each local feature evaluation experiments: (1) the train accuracy, (2) the test accuracy and (3) the test standard deviation.

Train Accuracy

To estimate the generalization ability of the by the classifier learned model, the training accuracy is calculated for each classifier training procedure. This is achieved by S-fold cross validation using the obtained feature vectors from the documents within the training set D_{tr}. The final training accuracy for local feature evaluation is then calculated as the average accuracy across all S cross validation trials according to Equation 7.1.

Testing Accuracy and Standard Deviation

To obtain the local feature classification performance the average amount of correct classified characters over all test documents D_{ts} is calculated. Let $c_l^i,\ l = 1, ..., T$ be the set of extracted characters of an arbitrary document $d_i \in D_{tr}$. Furthermore, let $|c_l^i|_{corr}$ define the percentage of correct classified characters within d_i. The test accuracy referring to the mean amount of correct classified characters per document is than obtained by:

$$\mu(D_{ts}) = \frac{1}{N} \sum_{i=1}^{N} |c_l^i|_{corr} \tag{7.2}$$

and the standard deviation of a test experiment is obtained by:

$$\sigma(D_{ts}) = \Big(\frac{1}{N} \sum_{i=1}^{N} (|c_j^l|_{corr} - \mu)^2 \Big)^{\frac{1}{2}} \tag{7.3}$$

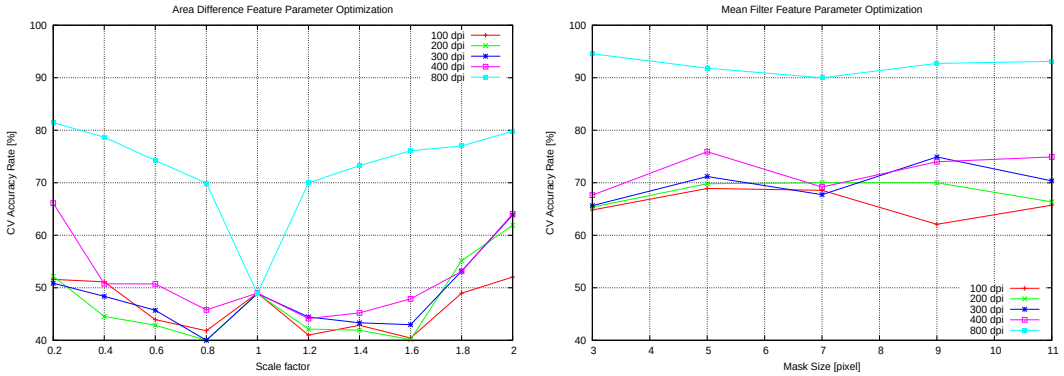


Figure 7.1: Feature parameter optimization results obtained by coarse grid searching the feature parameter space of the (l) Area Difference feature and the (r) Mean Filtering feature. A high dependence of classification result and an eligible feature parameter selection can be recognized observing the Area Difference feature. The Mean Filtering feature exhibits a higher robustness to a change in feature parameters.

7.3 Parameter Optimization

To gain a high classification performance it is essential to find the optimal parameters of the features and classifiers used in performance evaluation. Since the ground truth data is scanned at five resolutions and applied in each case to seven global and local features utilizing three different classifiers a manual test of different parameter combinations is not practicable.

The majority of the evaluated features exhibit a low dimensional parameter space. Therefore, the best parameter combination for each feature is obtained by a coarse grid search of the respective feature parameter space. To measure the results of each parameter setting the accuracy rate of cross validation served as performance indicator. To achieve a higher reliability the final parameter values are selected from a parameter space accuracy surface region where the local maximum only slightly exceeds the local average accuracy.

Additionally, the C4.5 decision tree algorithm with tree pruning is chosen as classifier for feature parameter optimization. The advantages of the C4.5 decision tree is the high classification speed making this algorithm suitable for parameter optimization. Another important property of the C4.5 decision tree is that the learned model can be visualized and therefore be checked if unnecessary branches have been created.

In terms of global feature parameter optimization the optimization experiments are undertaken based on all 109 ground truth documents. A parameter search for local feature optimization over the entire ground truth data set is highly time consuming. Consequently, the ground truth dataset is reduced to a smaller dataset containing eight documents for each of the examined printing techniques. This results in approximately 8800 feature vectors obtained for each class to optimize local feature parameters.

The parameter optimization experiments revealed the sensitivity of certain features to parameter changes. Figure 7.1 shows the parameter spaces of (l) the Area Difference

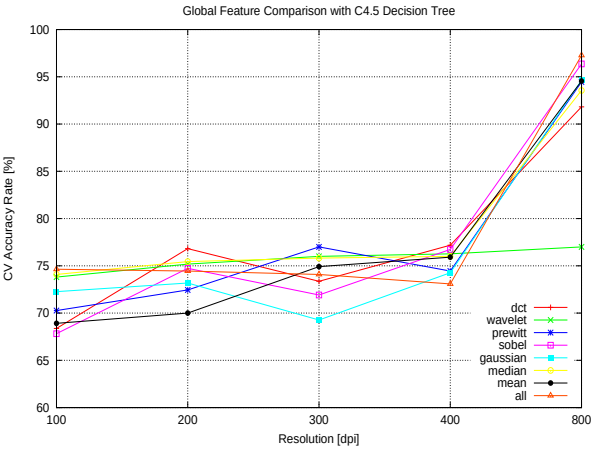


Figure 7.2: Global feature evaluation accuracy results obtained by decision tree classification performing 10-fold stratified sampled cross validation.

feature and (r) the Mean Filtering feature. In the case of the Area Difference feature a high variance in classification accuracy can be observed for different parameter combinations and evaluated resolutions. The Mean Filtering feature is showing a more robust behavior in terms of parameter optimization.

In a subsequent step the optimized features are utilized for coarse grid searching the parameter spaces of the distinct classifiers. Where the experimental setup for global and local classifier optimization remains unchanged. Again the results are obtained by measuring the cross validation accuracy rate of each parameter setting. Finally, the classifier parameters are selected yielding a local classification accuracy maximum.

7.4 Evaluation of Global Document Examination

In order to evaluate `[Hypothesis 1]` and `[Hypothesis 2]` the proposed global feature their classification performance is measured based on the entire set of 109 ground truth documents. This implies that a three class classification task had to be solved successfully by the applied classifiers. In the following, a comparative analysis of the different global features and the classifiers is presented.

7.4.1 Decision Tree Classification Results

C4.5 Decision tree classification evaluation is performed on postpruned trees. The classifier parameter optimization revealed that good classification results are obtainable by setting the confidence threshold for tree pruning to 25% and the minimum number of instances per leaf to 2. The respective cross validation results for each feature are illustrated in Figure 7.2 and Table 7.1.

It can be seen that for low scan resolutions of $100dpi$ to $400dpi$ the cross validation accuracy $\mu(acc_s)$ of the evaluated features ranges from 67% to 77%. Within this low resolution

Table 7.1: Accuracy results obtained by C4.5 Decision Tree classification of printing techniques applying the global frequency domain features and global spatial domain features. All presented results are given in percent [%].

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	68.37	**76.82**	73.36	**77.18**	91.82
Wavelet MRA	73.81	75.18	76.00	76.27	77.00
Gaussian Filter	72.27	73.18	69.27	74.27	94.63
Mean Filter	68.91	70.00	74.91	75.91	94.55
Median Filter	74.09	75.45	75.82	76.00	93.55
Gradient Prewitt	70.27	72.45	**77.00**	74.45	94.45
Gradient Sobel	67.82	74.73	71.91	76.73	96.36
All Spatial	**74.64**	74.45	74.09	73.09	**97.27**

range a homogen performance for all applied features is observable. Slightly superior results are obtained by the Mean Filter, the DCT Coefficients Analysis and the Wavelet MRA. In the case of 800*dpi* high resolution scans cross validation accuracy rates between 77% and 97% are achieved. The highest cross validation accuracy of 97.27% is obtained by the appliance of all spatial features while the best single feature result is obtained by Sobel Gradient Analysis yielding 96.39% cross validation accuracy.

7.4.2 Multilayer Perceptron Classification Results

Multilayer Perceptron (MLP) classification is performed using sigmoid perceptron functions for propagation. The performed parameter optimization shows that optimized classification results in terms of classification time and cross validation accuracy are obtained by training the MLP with 500 training epochs and setting the backpropagation learning rate to 0.3. The obtained experimental results applying MLP cross validation are illustrated in Figure 7.3 and Table 7.2.

Comparing the cross validation results obtained for low scan resolutions within the range from 100*dpi* to 400*dpi* it can be observed that high classification results are obtained at scan resolutions of 400*dpi*. Especially, the DCT Coefficients Analysis feature outperformes the other features reaching a cross validation accuracy $\mu(acc_s)$ of 90.09%. High classification results are also obtained by applying the gradient features yielding an accuracy of 83.18% for Prewitt Gradient Analysis and 82.18% for Sobel Gradient Analysis. In the case of 800*dpi* scan resolution the highest accuracy of 98.18% is received by Sobel Gradient Analysis

7.4.3 Support Vector Classification Results

Classification experiments using Support Vector Machine (SVM) classification are based on a radial-basis kernel function using optimized parameters. The optimal parameters for

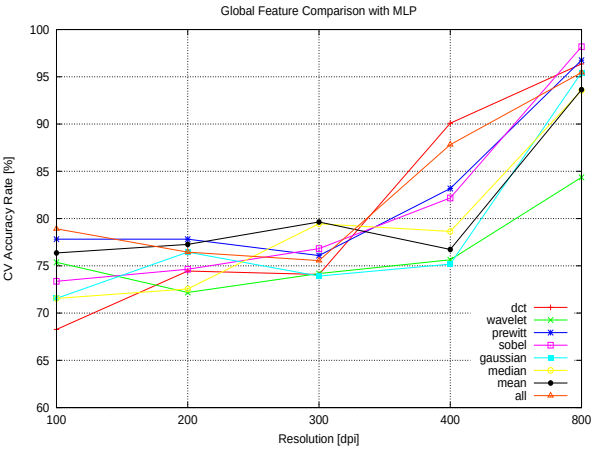


Figure 7.3: Global feature evaluation accuracy results obtained by multilayer perceptron classification performing 10-fold stratified sampled cross validation.

Table 7.2: Accuracy results obtained by Multilayer Perceptron classification of printing techniques applying the global frequency domain features and global spatial domain features. All presented results are given in percent [%].

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	68.27	74.45	74.09	**90.09**	96.36
Wavelet MRA	75.37	72.18	74.18	76.64	84.36
Gaussian Filter	71.55	76.45	73.91	75.18	95.45
Mean Filter	76.36	77.27	**79.64**	76.73	93.64
Median Filter	71.55	72.55	79.45	78.64	93.55
Gradient Prewitt	77.82	**77.82**	76.09	83.18	96.75
Gradient Sobel	73.36	74.64	76.82	82.18	**98.18**
All Spatial	**78.91**	76.45	75.55	87.82	95.45

C and γ are obtained within the intervals $C = [2^{-5}; 2^{15}]$, $\gamma = [2^{-15}; 2^{3}]$ for each of the scanned resolutions. The obtained cross validation results are illustrated in Figure 7.4 and Table 7.3.

Within the resolution range of 100*dpi* and 400*dpi* high cross validation accuracy rates $\mu(acc_s)$ are achieved by the DCT Coefficients Analysis and both Gradient Analysis features. Therefore, the results obtained at a scan resolution 400*dpi* are notable in particular. An accuracy of 92.92% is achieved by the DCT feature while an accuracy of 91.57% is obtained by combining all spatial features. The evaluation of 800*dpi* high resolution scans reveals that DCT Coefficients Analysis yield the highest classification result of all evaluated features by exhibiting an accuracy of 99.08%. Additionally, high results at 800*dpi* scan resolution are obtained by Prewitt and Sobel Gradient Analysis leading to 98.17% and 98.07% respectively.

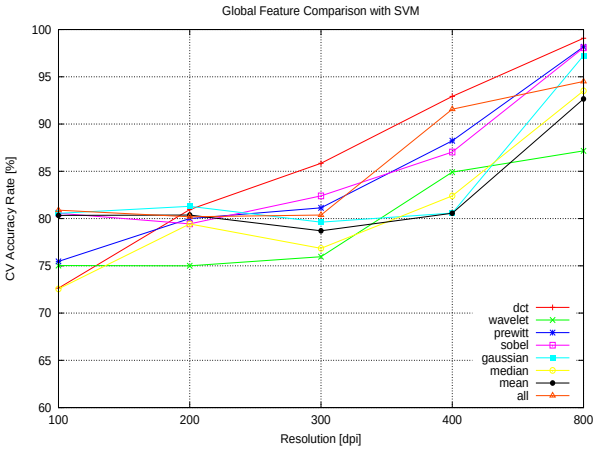


Figure 7.4: Global feature evaluation accuracy results obtained by support vector machine classification performing 10-fold stratified sampled cross validation.

Table 7.3: Accuracy results obtained by Support Vector Machine classification of printing techniques applying the global frequency domain features and global spatial domain features. All presented results are given in percent [%].

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	72.64	80.95	**85.85**	**92.92**	**99.08**
Wavelet MRA	75.03	75.00	75.96	84.91	87.17
Gaussian Filter	80.55	**81.30**	79.62	80.58	97.24
Mean Filter	80.33	80.37	78.70	80.56	92.66
Median Filter	72.55	79.43	76.85	82.40	93.51
Gradient Prewitt	75.47	80.00	81.13	88.23	97.17
Gradient Sobel	80.56	79.44	82.40	87.04	98.07
All Spatial	**80.89**	80.19	80.37	91.57	94.50

7.4.4 Concluding Observations

Having presented the classification results of global document features the following concluding statements are derived:

- The initial `[Hypothesis 1]` and `[Hypothesis 2]` can be confirmed according to the results obtained from global document feature evaluation presented above. Printing technique recognition and photocopy detection can be realized on the basis of collecting discriminative information from the entire document scan.

- In terms of feature comparison, high cross validation accuracy rates are achieved for DCT Coefficients and Gradient analysis. Comparing both best performing techniques, slightly better results are obtained for document analysis within the frequency domain. This can be ascribed to the unique properties of the frequency domain in which the discriminative properties like high frequency noise and high frequency edge

sharpness are mapped to particular frequency subbands. Therefore, the amount and characteristics of those discriminative document properties are obtained more precisely.

- The superiority of document frequency analysis is also underpinned by the results obtained using the Wavelet MRA feature. Utilizing MLP and SVM classification high classification results are already obtained at low MRA scales. However, due to the time constraints of this thesis only a preliminary study of the Wavelet MRA feature is undertaken. In trying other suitable wavelet types and families exists a major source of improvement. The so far tested wavelets encompass the Haar Wavelet, the Daubiches 1-5 Wavelets as well as the 1-4 Coiflets. Furthermore, an increase of MRA scales should be considered.

- In terms of classification results it becomes evident that MLP and SVM classification outperforms the results obtained by C4.5 Decision Tree classification. This especially holds for the results obtained at low scan resolutions ranging from $100dpi$ at $400dpi$. Within this range more complex decision boundaries are learned by the MLP and the SVM while the created C4.5 Decision Tree exhibits a strong pruning. To prevent this circumstance a finer parameter search has to be performed in terms of optimizing decision tree classification.

- The experimental comparison of different scan resolutions revealed that printing technique recognition and photocopy detection is possible at low and high scan resolutions. This is in particular true for documents scanned at high resolutions of $800dpi$. However, also remarkable results could are obtained utilizing MLP and SVM classification for documents scanned with $400dpi$.

7.5 Evaluation of Local Document Examination

Both initial thesis `[Hypothesis 3]` and `[Hypothesis 4]` are evaluated based on three distinct experimental setups. In the first setup local features are extracted from the laser and inkjet printed document scans within the created ground truth database. Therefore, to evaluate `[Hypothesis 3]` a two class classification task had to be solved by the applied classifiers.

The two subsequent setups are designed to evaluate `[Hypothesis 4]`. In the second setup the local classification performance on all document classes is evaluated and in the third setup the evaluation is performed on laser printed and photocopied documents. This leads to a three class classification task for the second experimental setup and a two class classification task for the third experimental setup.

The applied training sets for local feature evaluation contains 75% of each document class within in the ground truth database. The learned model is then tested on the remaining

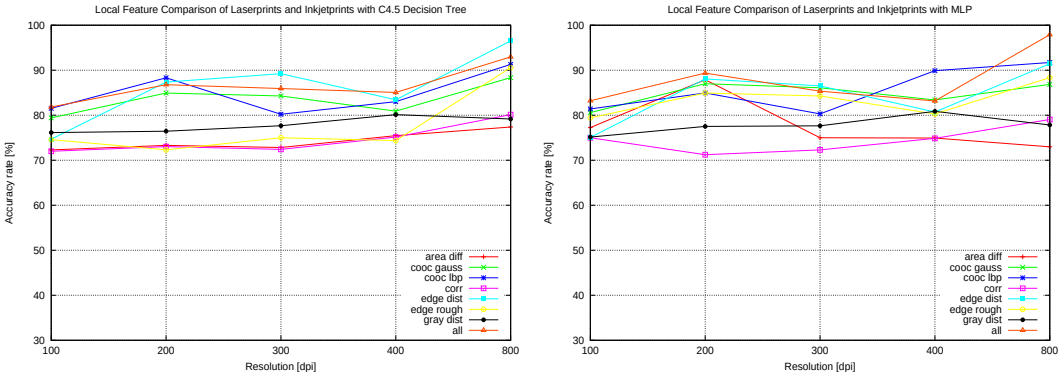


Figure 7.5: Local feature evaluation mean accuracy rates obtained on the basis of all laser and inkjet printed documents within the ground truth database. The classification results are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

25% of document images. In terms of measuring the classification performance, mean and standard deviation are obtained as defined in Section 7.2.2. To evaluate the discriminative ability of the features C4.5 Decision Tree and Multilayer Perceptron classification is performed.

7.5.1 Classification of Laser and Inkjet Printed Documents

For the purpose of evaluating `[Hypothesis 3]` the ground truth database is subsampled to laser and inkjet printed documents while 46 documents (37 laser and 9 inkjet printouts) are used for classifier training and 17 documents (12 laser and 5 inkjet printouts) for classifier testing.

Figure 7.5 shows the mean testing accuracy rates $\mu(D_{ts})$ as obtained by each feature and classifier at different scales of scan resolution. More detailed classification results can be found in Appendix C. Both classifiers show constantly high mean testing accuracy rates within the low resolution range. Superior classification results are achieved for the documents scanned at a high scan resolution of 800dpi.

For both classifiers high accuracy rates are obtained using all local features. Slightly higher classification results are obtained by C4.5 Decision Tree classification leading to a mean testing accuracy rate of 85.06% for 400dpi and 92.96% at 800dpi.

In terms of single feature evaluation it can be observed that the highest classification results are obtained by the Distance Map Based Edge Roughness feature and the Graylevel-Coocurrence features. The highest mean testing accuracy within the low resolution range of 100dpi to 400dpi is achieved by the Graylevel-Coocurrence Feature using Local Binary Pattern and MLP classification yielding an accuracy of 89.90% at 400dpi. For scans of 800dpi the highest classification result of 96.52% is obtained by the Distance Map Based Edge Roughness feature in combination with C4.5 Decision Tree classification

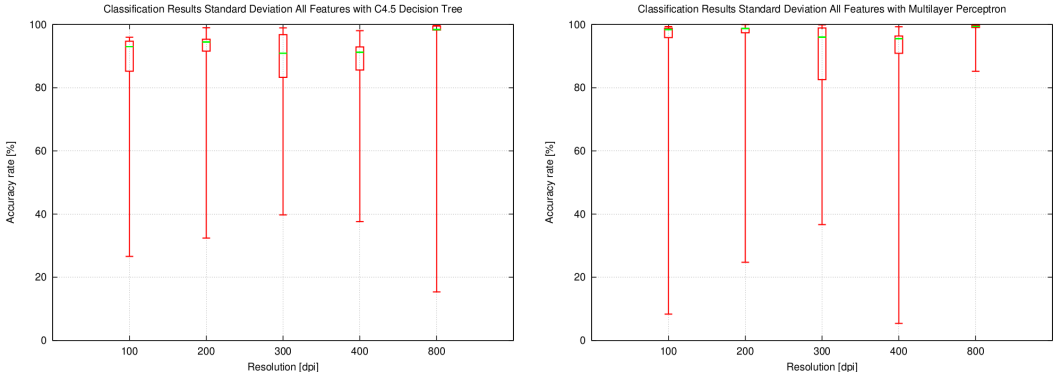


Figure 7.6: Distribution of mean accuracy rates obtained by local features evaluation on the basis of all laser and inkjet printed documents within the ground truth database. The classification results standard deviation is presented in terms of quartile accuracy box plots that are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

In Figure 7.6 the quartile accuracy box plots obtained by the evaluation experiments for the combination of all local features are shown. The accuracy box plots illustrate the distribution of the obtained mean training accuracy rates and reflect therefore the standard deviation $\sigma(D_{ts})$ of the testing experiments. The green bar depicts the median of the obtained result and the box depicts range of the lower quartile (cuts off the lowest 25% results) to the upper quartile (cuts off highest 25% results, or the lowest 75%).

It can be observed that the mean accuracy rates presented above are associated with high standard deviations for scan resolutions ranging from $100dpi$ to $400dpi$. Utilizing Multilayer Perceptron classification a standard deviation of 20.46 is obtained for $300dpi$ and 24.53 for $400dpi$. Furthermore, a high amount of outlying results is observable within the range of $100dpi$ to $400dpi$.

7.5.2 Concluding Observations

Observing the local feature evaluation results represented above the following concluding statements are derived:

- The local feature evaluation presented above confirms `[Hypothesis 3]` of printing technique recognition at document character level. Even high mean accuracy rates are obtained based on documents scanned at low resolutions.
- However, the obtained evaluation classification results exhibit a high standard deviation in the case of low resolution document scans ranging from $100dpi$ to $400dpi$. This derives from the circumstance that the class predictions of characters within a certain document exhibit only low variations. Meaning, that in the case of character missclassification in the majority of cases all document characters of a document are

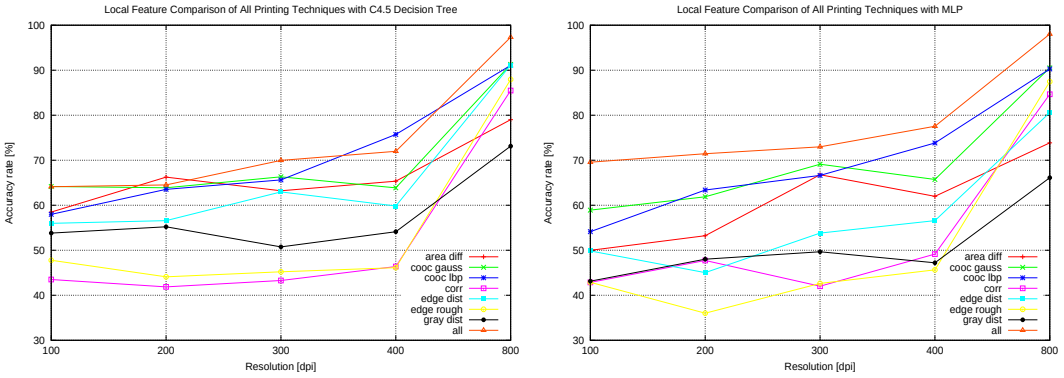


Figure 7.7: Local feature evaluation mean accuracy rates obtained on the basis of all documents within the ground truth database. The classification results are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

missclassified. As a result several outlying mean test accuracy rates are obtained. These outlying test mean accuracy rates could be traced back to high quality inkjet printouts.

7.5.3 Classification of All Document Classes

For the purpose of evaluating `[Hypothesis 4]` the ground truth database is splitted into 80 documents (37 laser printouts, 9 inkjet printouts and 34 photocopies) used for classifier training and 29 documents (12 laser printouts, 5 inkjet printouts and 12 photocopies) used for classifier testing.

Figure 7.7 illustrates the mean training accuracy rates $\mu(D_{ts})$ of the classifiers obtained by single feature evaluation as well as for all features. For both classifiers an increase of the mean training accuracy rates in correlation with a higher scan resolution is observable. The more detailed classification results are given in Appendix C.

The highest accuracy rates are obtained using all local features. This holds for the results achieved at low scan resolutions within the range of $100dpi$ to $400dpi$ as well as high resolution scans of $800dpi$. Utilizing MLP classification a mean training accuracy rate of 77.55% is achieved for $400dpi$ and 98.03% for $800dpi$.

Single feature evaluation shows that superior classification results are obtained from the Graylevel-Coocurrence features and the Area Difference feature. Applying the Graylevel-Coocurrence with Local Binary Pattern Feature and MLP classification, a mean training accuracy of 73.83% is obtained at $400dpi$ and 90.29% at $800dpi$.

Observing the quartile accuracy box plots presented in Figure 7.8 it can be recognized that the mean accuracy rates presented above are associated with high standard deviations $\sigma(D_{ts}$ for scan resolutions ranging from $100dpi$ to $400dpi$. In the case of applying all local features which yield the highest mean accuracy rates still a standard deviation of 26.24 is

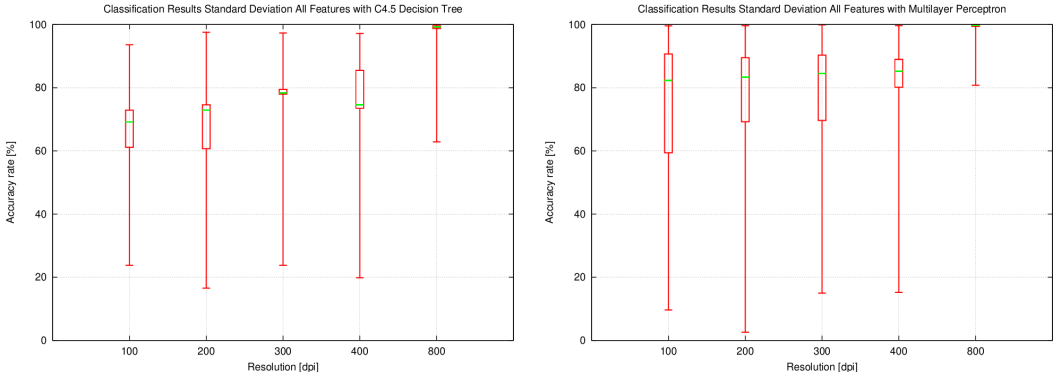


Figure 7.8: Distribution of mean accuracy rates obtained by local features evaluation on the basis of all documents within the ground truth database. The classification results standard deviation is presented in terms of quartile accuracy box plots that are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

obtained for 300*dpi* and 22.11 for 400*dpi* using MLP classification. Several investigative experiments had been executed to identify the reason for this high standard deviation.

Table 7.4 shows the confusion matrix for one of these experiments. They have been performed on a subset of documents containing all three document classes. The ground truth documents are scanned with a resolution of 400*dpi*. Each column of the confusion matrix represents the instances of the actual character class, while each row represents the instances of the predicted character class. It can be observed that a high amount of inkjet printed characters are falsely classified as photocopied characters. Also a certain amount of photocopied characters are recognized as laser printed characters. Several further experiments confirmed this observation.

Table 7.4: Exemplary confusion matrix obtained from an experiment based on document subset that contains all three document classes. It can be observed that a high amount of inkjet printed characters are falsely classified as photocopied characters.

		True Class		
		Ink	Laser	Copy
Pred. Class	Ink	6625	1163	1390
	Laser	1281	15851	**4123**
	Copy	**1460**	3777	15879

It is discovered that beside the difficulty to discriminate between laser printed and photocopied characters the discrimination between inkjet and photocopied characters depicts a second major source of missclassification. This is an observation of particular interest since both document classes have been created by different printing techniques. Further analysis revealed that in both cases of missclassification the typical printing technique characteristics of the template document are also transferred to the respective photocopied document

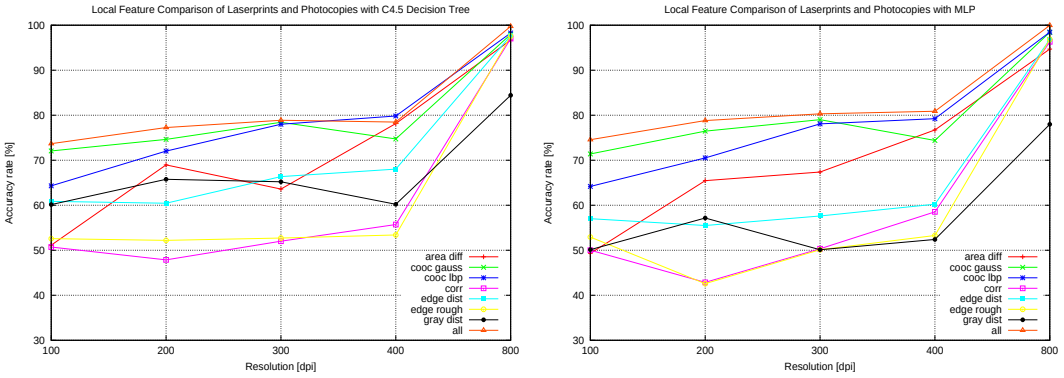


Figure 7.9: Local feature evaluation mean accuracy rates obtained on the basis of all laser printed and photocopied documents within the ground truth database. The classification results are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

images. Therefore, the first generation photocopies show a high similarity to their template documents and are difficult to classify at character level.

To verify this discovery an additional experimental setup is designed. The three class problem is reduced to a two class problem excluding the conflicting document class of inkjet printed documents.

7.5.4 Classification of Laser Printed and Photocopied Documents

The third local feature evaluation is performed based on laser printed and photocopied documents. To achieve this the ground truth database is subsampled containing 71 documents (37 laser printouts and 34 photocopies) used for classifier training and 24 documents (12 laser printouts and 12 photocopies) for classifier testing.

In Figure 7.9 the mean training accuracy rates $\mu(D_{ts})$ are illustrated as obtained by each classifier. More detailed classification results are given in Appendix C. Both classifiers show an increase of mean training accuracy rates with rising scan resolution.

Similar to the previous experiments the highest accuracy rates are obtained by the appliance of all local features an the usage of MLP classification. A mean training accuracy rate of 80.89 is achieved for a scan resolution of 400dpi and 99.97% for a scan resolution of 800dpi.

Single feature evaluation reveals that the Graylevel-Coocurrence features as well as the Area Difference feature outperforms all other features. Applying the Graylevel-Coocurrence Feature with Gaussian Filter yields a high a mean training accuracy at low scan resolutions for both classifiers. The highest single feature results are obtained by the Graylevel-Coocurrence Feature with Local Binary Pattern. Using this feature a mean training accuracy of 79.24% at 400dpi and 98.46% at 800dpi is achieved.

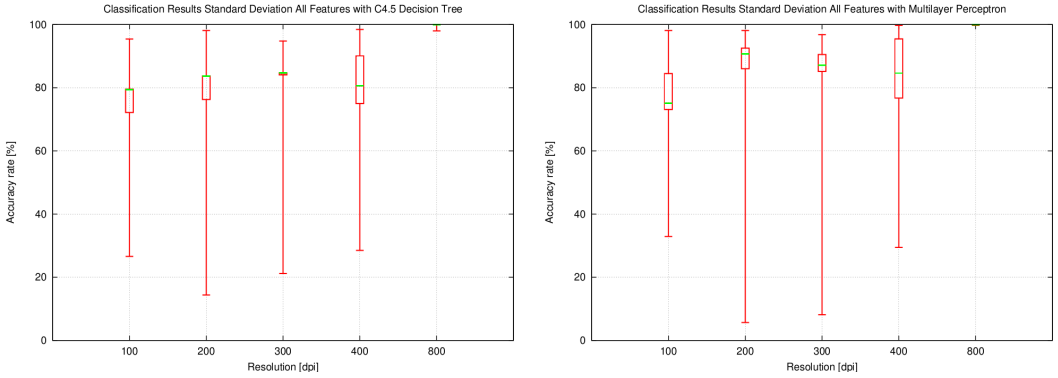


Figure 7.10: Distribution of mean accuracy rates obtained by local features evaluation on the basis of all laser printed and photocopied documents within the ground truth database. The classification results standard deviation is presented in terms of quartile accuracy box plots that are obtained at several scan resolutions utilizing (l) C4.5 Decision Tree classification and (r) Multilayer Perceptron (MLP) classification.

In comparison to results obtained by the classification of all document classes the quartile accuracy box plots presented in Figure 7.10 show a reduced standard deviation $\sigma(D_{ts})$. This especially holds for the resolution range of $100dpi$ to $300dpi$ as well as the high resolution scans of $800dpi$. Utilizing MLP classification a standard deviation of 20.79 is obtained for $300dpi$ and 18.24 for $400dpi$.

7.5.5 Concluding Observations

Observing the local feature evaluation results represented above the following concluding statements are derived:

- According to the local feature evaluation presented above `[Hypothesis 4]` is only partly verifiable. The performed evaluation shows insufficient classification results for the detection of photocopies at character level for low scan resolutions. However, in the case of high resolution scans the majority of the documents is classified correctly. The main sources of missclassification at low resolution originate from (1) a large amount of laser printed characters that are falsely classified as photocopied characters and (2) a large amount of inkjet printed characters that are falsely classified as photocopied characters.

- The large amount of missclassified laser printed characters originates from the high similarity of first generation photocopies to the template documents used in their creation. In the case of laser printed documents this is obvious since both document classes are created by an electrophotographic printing process. In the case of high quality photocopies derived from inkjet printed template documents the same effect could be observed. Since the examination of the misclassified photocopies derived

from inkjet printed templates showed that the typical inkjet printing characteristics are still observable on the photocopied documents. Although, both documents are created by distinct printing processes. Therefore, the discrimination of both document classes at character level is difficult at low scan resolutions.

- In terms of the high amount of falsely classified inkjet printed character images another interesting observation is made. The majority of these misclassifications can be traced back to the imperfections that are introduced by the photocopy scan procedure. In general photocopied document images exhibit a small amount of blurring at the character edges. In the case of inkjet printed document images a small amount of ink diffusion into the printing paper can be recognized at the character edges. Both effects exhibit a small degree of character edge degeneration. This edge degeneration shows a high similarity within the extracted character images especially for low scan resolutions.

- The experiments based on a ground truth database reduced to two photocopied and laser printed documents underpinned the derived observations. While in the evaluation higher mean test accuracy rates are obtained. In this scenario the applied classifiers are able to successfully separate the photocopies derived from laser printed templates. Since the photocopied character images obtained from inkjet printed documents exhibit a low similarity to the character images extracted from laser printed documents.

- In terms of standard deviation a similar observation to the first experiment covering laser and inkjet printed documents is made. Most of the outlying test mean accuracy rates can be traced back to the high similarity of photocopied documents to their templates and high quality inkjet printouts.

8. Conclusion and Future Work

In this thesis a novel document examination approach for the intelligent detection of photocopies and a document's underlying printing technology was developed. This is motivated by the increasing demand by companies, like banks and insurance companies or governmental organisations, for reliable methods to determine if a processed document is genuine or manipulated. The obtained evaluation results demonstrate that this demand can be meet by enhancing traditional document investigation methods with techniques derived from digital image processing and pattern recognition.

Several photocopied, laser and inkjet printed documents have been examined to detect class characteristics corresponding exclusively to a particular document class. The examination leads to the conclusion that these document classes exhibit different degrees of:

- Image Noise and Artefacts,
- Edge Sharpness and Contrast,
- Edge Roughness and Degradation,
- Uniformity of Printed Character Areas.

Based on this discovery a set of existing and newly developed features for photocopy and printing technique recognition was utilized to evaluate the four initial hypotheses. The evaluation was performed based on a representative ground truth document database containing photocopied as well as laser and inkjet printed documents. Furthermore, Decision Tree, Multilayer Perceptron and Support Vector Machine classification was applied at different document scanning resolutions.

The performed evaluation leads to the following conclusions for each hypothesis:

`[Hypothesis 1]`: **Printing technique recognition is achievable at document level.**
`[Hypothesis 2]`: **Photocopy detection is achievable at document level.**

The evaluation results reveal that both hypotheses can be confirmed. Because printing technique recognition and photocopy detection can be realized by the extraction of discriminative information distributed over the entire scanned document image. High cross validation accuracy rates are achieved applying the three different classifiers in combination with a novel feature set:

- Document Image Difference Analysis,
- Document Image Gradient Analysis,
- Document Image Frequency Analysis and
- Document Image Multiresolution Wavelet Analysis.

A comparative study of both best performing techniques, namely Gradient and Frequency Analysis, demonstrated that document analysis within the frequency domain yield higher classification results. Utilizing MLP and SVM classification for documents scanned with $400dpi$ and $800dpi$ remarkable classification results are obtained as illustrated in Table 8.1.

Table 8.1: Cross validation classification results obtained by document image Gradient and Frequency Analysis. All presented results are given in percent [%].

	Gradient Analysis		Frequency Analysis	
Classifier	400 dpi	800 dpi	400 dpi	800 dpi
MLP	83.18	96.75	90.09	96.36
SVM	88.23	98.07	92.92	99.08

`[Hypothesis 3]`: **Printing technique recognition is achievable at character level.**

The experimental evaluation confirms the hypothesis. High mean accuracy rates are obtained even for document characters scanned at low resolutions. However, also some outlying test results are obtained. These outlying test results can be traced back to high quality inkjet printouts that exhibit a high similarity to laser printed documents.

`[Hypothesis 4]`: **Photocopy detection is achievable at character level.**

Based on the evaluation results this hypothesis has to be partly rejected. Insufficient classification results are obtained for character images of low scanning resolution. However, in the case of high resolution scans the majority of the documents is classified correctly.

The main sources of missclassification at low resolution originate from:

1. The similarity of high quality first generation photocopies to their respective template documents.
2. The high similarity in character edge degeneration of photocopied and inkjet printed character images.

Both sources lead to low mean accuracy rates and a high standard deviation for the obtained classification results.

8.1 Future Work

The techniques developed within this thesis for forensic document examination show a high potential for new applications in the field of document fraud and counterfeit detection. However, several areas of future work have been identified concerning the improvement of the actual techniques itself but also new research directions of the field:

- **Global and Local Feature Fusion**
  The actual system could be improved by a fusion of global and local features in terms of classifying single document characters. Therefore, it has to be evaluated of how the different features and feature types have to be weighted within the classification. In this context also a boosting of different classifiers should be considered to derive several predictions for a document or single document characters. First attempts of feature fusion have already been developed within the actual system but could not be evaluated due to the time constraints of this thesis.

- **Unsupervised Learning and Clustering**
  A very interesting new field of research in document forensics is given by the shift of the applied machine learning strategy to unsupervised learning. A new idea would be to cluster the features obtained from the characters within a single document. An analysis of the obtained clusters could be performed in a subsequent step to detect potential outliers. The detected outlying features can then be ascribed to document characters that show a high difference in printing technique compared to the majority of the document characters.

- **Feature Robustness**
  As described in Chapter 7 the presented experimental results were obtained on the basis of ground truth documents. These documents are created by several printouts and photocopies derived from the 'Grünert' letter. To obtain a comprehensive impression about the robustness of the so far developed features several further experiments should be carried out on the basis of documents containing a different

geometrical layout, font size and type. In this case the evaluation of local features would be of particular interest since character edge degeneration effects are influenced to a large extent by the characteristics (ascenders, descenders and serifs) of the documents font type.

- **Parameter Optimization**
  Another source of improvement of the current system lies within the feature and classifier parameter optimization procedure. In Chapter 7 it was shown that in order to reduce the computation time of parameter optimization the search for optimal parameters was splitted. In a first step coarse grid searching was carried out within the feature parameter space. In a second step the classifier parameter space was searched utilizing the optimized features. This is regarded as a non-optimal approach since the parameter optimization procedure should be done simultaneously for features and parameters. This issue could be resolved in acceptable computation time by the appliance of more advanced searching strategies like genetic algorithms.

A. The Evolution in Document Forensics

A paradigm shift can be recognized observing the recent developments in computer science in alliance with forensic document examination [27]. As more and more obvious, the monolithic standing of both scientific disciplines will be outdated by the increasing demand for digital methods in document examination capable to create valuable benefits. To underpin this observation and illustrate the derived advantages a brief analysis of the ongoing process will be given within this Appendix.

A.1 The Technological Evolution

Within the history of modern forensic document examination[1] an impressive toolbox of methods and techniques was developed to examine the genuineness of questioned documents (as outlined in Section 2.1.2). In comparison the use of image processing techniques in forensic document examination is relatively new [17].

Computer-based trace analysis, although promising research has been done, is rarely applied in daily forensic casework [12] but gaining momentum nowadays. In 2002 Gary Herbertson former chief of the Federal Bureau of Investigation's document examination lab published the first book on document examination and the usage of digital image processing techniques, in which he stated, *"now that computers are ubiquitous in office environments, document examiners should be using digital image processing as the standard method for capturing, examining and presenting questioned documents"* [20].

Since digital image processing technologies have now become the method of choice in the fraudulent reproduction of many official documents, these methods are more and more used by forensic document examiners in the opposite way to investigate questioned documents. However, *Computational Forensic* referred to as the appliance of techniques derived from

[1] According to Nickell [16] the first significant modern text that attempted a thourough scientific approach to questioned documents, including chemical test for alteration detection was E.E. Hagan's *Disputed Handwriting* published in 1894.

computer science to the field of the traditional document forensic casework is still in its beginnings [12]. As an emerging interdisciplinary research domain it is offering a variety of advantages in an examiners daily casework. These advantages can be summarized according to Herbertson [20] as following: increased speed of examination steps, support within the presentation of examination results, the ability of non-destructive examination steps, artless storage of intermediate examination, and subsequently the opportunity for more cost efficient examinations.

Another major advantage is resulting from the opportunity of automated document examination: forensic document investigations in the past have been largely based on manual inspections by experts to discern writing habits and other characteristics [93]. Nowadays, with the usage of digital image analysis, automated methods can be developed for the examination of handwritings [94, 95], printing devices [30] or printing technique recognition [29].

A.2 The Statutory Evolution

According to Saks *"little more than a decade ago, forensic document examiners compared pairs of handwriting and document marks, and testified in court whether they matched or not. The foundations of the asserted expertise was rarely questioned"*[27]. However, initiated by reports of erroneous forensic identification, courts begun more and more to question the core assumptions of forensic sciences.

In 1993 the judicature began to question scientific considerations of forensic methodologies. Accordingly, in the case *Daubert v. Merrel Dow Pharmaceuticals* [509 U.S. 579 (1993)], the U.S. Supreme Court introduced a new standard for the admissibility of scientific evidence. Referring to the *Daubert* standard, scientific testimony must be shown to stand on a reliable scientific foundation. As an effect the *Daubert* test lowered the threshold for admission of sound cutting-edge science and raised the threshold for expertise that lacks a scientific foundation, like forensic document examination, at that time.

In the case *United States v. Starzecpyzel* [880 F. Supp. 1027 (S.D.N.Y 1995)] it became obvious how the Daubert standard had changed judicial views. A federal district court concluded that *"despite the existence of a certification program, professional journals and other trappings of science"*, forensic document examination cannot, after Daubert, be regarded as *"scientific knowledge"*.

However, in this case the testimony given by the forensic document examiner was not excluded. The court reasoned that handwriting identification did not have to reach the Daubert standard because it is applied only to scientific evidence, and therefore handwriting identification was referred to as a *"technical skill"* [27]. Nevertheless, the forensic document community was stunned by labeling one of the oldest forensic disciplines in this manner.

In the following years, Kam, a professor in engineering and computer science, played an initial role in the proficiency testing of forensic document examination. By researching the published literature Kam and his team were surprised to find an absence of scientific studies. Therefore, they were highly sceptical of the so far published statistics and emphazised the need for controlled studies. As a result Kam published four studies in which he compared the skills of document examiners and laypersons in identifying handwriting in 1994, 1997, 1998 and 2001 [96, 97, 98, 99]. All studies revealed a significant superiority of the results achieved by document examiners in comparison to the results achieved by laypersons.

In 2002 the case *United States v. Prime* [220F. 2d 1203] also included a Daubert hearing of forensic document examination. In the final analysis it was concluded that the techniques practiced by document examiners satisfied all the Daubert factors and therefore were reliable and admissible [1].

A.3 Concluding Observations

In a sensitive disciplines like forensic science the alliance with computer science is offering the ability to create a winning situation for both disciplines. According to Franke and Shrihari [17], *"new scientific disciplines, approaches and studies in "Computational Forensics" need to be peer-reviewed and published for the purpose of discussion, consequent general acceptance, and rejection by the scientific community"* [12]. This is especially of importance for the scientific examination of questioned documents in the post "Daubert era", where the comparatively small community of forensic document examiners can hardly foster scientific bases for their methods without support of other scientific disciplines like computer science [27].

On the other hand forensic examination of documents as stated by Dasri and Bhagvati [17] is fast emerging as a challenging field of research. Especially, since the profileration of fake and questioned documents through the use of computer-based technologies. Therefore, with the advent of digital document examination an tremendously exciting field of research is currently opening within Computational Forensics for computer scientists.

B. Pattern Classification Techniques

This appendix introduces the classifiers used in the experiments presented in Chapter 7 for the purpose of feature evaluation. The classifiers are C4.5 Decision Tree, Multilayer Perceptron and Support Vector Machine. As already described idea of pattern recognition is the design and development of techniques that allow computers to construct approximations or models of artifical or real world circumstances. To create such models usually techniques referred as machine learning are applied to learn decision boundaries or discriminant functions from a given set of observed patterns [100]. Therefore, the question aroused if one of the actual techniques can aim the general superiority to all the other techniques.

B.1 No Free Lunch Theorem

In this context in 1995 and 1997 Wolpert and Macready [86, 87] showed that for static and time dependent search and optimization problems "the average performance of any pair of algorithms across all possible problems is identicial". This insight was postulated in the so called *''No-Free-Lunch Theorems"*. In [88] Wolpert transfered the "No-Free-Lunch Theorems" to supervised learning. Assuming a noise free scenario,

- let D be the training set consisting of n ordered pairs of input and output values $\{d_t, c_t\} : 1 \leq i \leq n$, and
- g be the true discriminant function that labels the "true" or "target" input output relationship between the d_t and c_t, furthermore
- let h be a hypothesis or the learning algorithms guess about g derived in response to D, and
- *Cost* representing the absolute off-training-set "cost" or "generalization errors" associated with a particular g and h.

The expected off-set-training error $E(Cost|D)$ of a given learning algorithm $P(h|D)$ can than be written by:

$$E(Cost|D) = \sum_{h,f} E(h, f, D)P(h|D)P(f|D) \tag{B.1}$$

while the off-set-training error $E(h, f, D)$ of a given hypothesis h is expressed by:

$$Er(h, f, d) = \sum_{x \notin d_t} P(x)[1 - \delta(f(x), h(x))] / \sum_{x \notin d_t} P(x), \tag{B.2}$$

and δ depicts the Kronecker-Delta. Wolpert showed that for any two learning algorithms i and j, the estimated generalization errors uniformly averaged over all target functions $P(f)$ do not differ independently of the sampling distribution $P(x)$. This insight is formulated in the following Equation B.3:

$$E_i(Cost|d) - E_j(Cost|d) = 0, \tag{B.3}$$

where $i \neq j$. As a result one can not state the general superiority of a specific supervised learning algorithm in every classification task. Or as stated in the words of Wolpert, *"for any two learning algorithms, there are just as many situations in which algortihm one is superior to algorithm two as vice verca [88]"*. This implied that it is desireable to examine the appliance of different supervised learning techniques to a given classification task.

To comply with the "No Free Lunch Theorem", three supervised machine learning techniques were applied and evaluated for the purpose of printing technique classification. These techniques are the C4.5 Decision Trees [89], the Weka Multilayer Perceptron [90] and the LibSVM Support Vector Machine [92]. Since the theory of these classifiers usually covers full textbooks an outline of their major principles as used within this work will be presented in the following sections. To obtain an all encompassing introduction the reader is kindly referred to the textbooks of Duda et al. [52], Bishop [54] and Mitchell [55].

B.2 Decision Trees

Decision trees are data mining techniques evolved from traditional statistical disciplines like linear regression. Using decision trees for machine learning tasks accompanies with the main advantage that their created results could be interpreted and communicated very well in symbolic and visual terms [101]. Another benefit of using decision trees are their robustness i.e. they accommodate the incooperation of missing feature values.

As shown in figure B.1 decision trees are acyclic graphs comprised of leafs and decision nodes. Every leaf indicates a single class c_i and each decision node defines a test carried out on a single feature vector attribute value $x_i^i \in x_i$ of the input data. Every leaf node defines an according to the tree structure optimal decision that is marked by a classvalue.

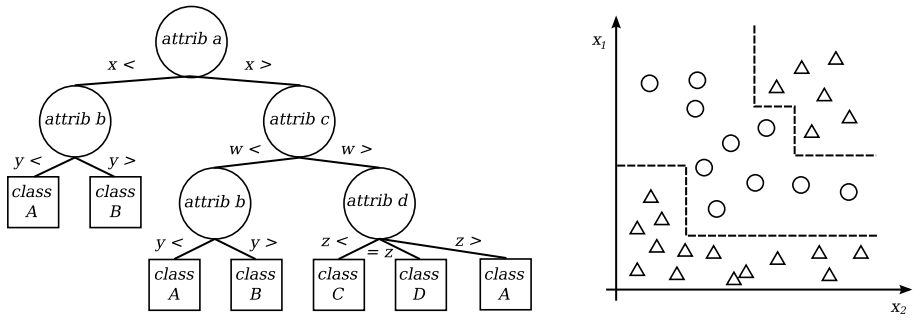


Figure B.1: Decision tree classification: (l) exemplary decision tree, (r) decision boundaries in the feature space created through decision tree building.

A decision tree can than be applied to classify a given input feature vector x_i by starting at the tree root moving downwards until a leaf node is reached.

Nowadays, several approaches are used to create decision trees like the CHi-squared Automatic Interaction Detector (CHAID) [102] or the Classification and Regression Tree (CART) [103]. Another well known and extensively used algorithm to create decision trees is the so called C4.5 developed by Quinlan [89] in 1993. A significant benefit of the C4.5 algorithm is its ability to handle also continues attribute values.

B.2.1 Measuring Uncertainty

Decision tree classification in general is driven by the idea that each data attribute x_i^i of a feature vector x_i can be used to make a decision. Building a decision tree is concerned with two central questions: How to split the training data efficiently? When to stop further splitting of the training data?

The basic idea of decision tree classification is to continuesly split the training data into distinct subsets $D = D_1 \cup D_2 \cup ...D_n$. This is done in a way that the uncertainty about the corresponding class labels c_i of the features within the created distinct data subsets will be reduced as maximum as possible.

As a result a function needs to be designed which measures the uncertainty about the class labels of a given set of sample feature vectors D_1 according to their class labels. Let n be the number of distinct classes c_i and $p_i(x_i) \geq 0$ the probability of feature vector x_i belonging to c_i. The uncertainty function should exhibit the following properties for all i:

1. global maximum at point $(\frac{1}{i}, \frac{1}{i}, ..., \frac{1}{i})$,
2. local minimas at the points $(1, 0, ..., 0), (0, 1, ..., 0), ..., (0, 0, ..., 1)$,
3. is symmetric over $p_1, ..., p_n$.

The needed function can be found in Shannon's model of communication theory [104]. In this model, the concept of information entropy as a measure of uncertainty was introduced.

Applying his idea to the task of decision tree creation, the uncertainty about the class labels of set of feature vectors can be determined by calculating the entropy over all probability distributions $P = (p_1, p_2, ..., p_n)$ according to:

$$Entropy(P) = -\sum_{i=1}^{n} p_i log_2 p_i, \text{ where } 0\ log_2\ 0\ :=\ 0, \tag{B.4}$$

.

It can be shown that the concept of entropy is fullfilling the desired properties stated above. The measurement of entropy to Shannon is the underlying concept used in the C4.5 algorithm for building decision trees.

B.2.2 Treebuilding

Considerung the measurement of uncertainty as given by Equation B.4, the disjunct term $1 - Entropy(P)$ can be interpreted as the certainty or information content of an arbitrary random variable. Using these observations and applying them to an arbitrary disjunct amount of training data subdivided by a feature attribute A, the information gain $InformationGain(D, A)$ obtained by the subdivision, is formulated by:

$$\begin{aligned} InformationGain(D, A) &= Entropy(D) - Entropy(D|A), \\ &= Entropy(D) - \sum_{i=1}^{n} \frac{|D_i|}{|D|} Entropy(D_i). \end{aligned} \tag{B.5}$$

Equation B.5 expresses the gained amount of information derived from the difference between the subdivided data weighted entropies and the entropy of the entire Data. For each training data attribute A the normalized information gain is examined when choosing A for splitting the data into $D_1 \cup D_2 \cup ...D_n$. Subsequently, the attribute exhibiting the highest information gain is chosen as tree node according to which the data is splitted. The same procedure is than recursively applied to the resulting splitted data sets.

It can be shown that using the concept of information gain as given by Equation B.5 attributes that exhibits a wide value margin are privileged for training data set splitting. To avoid this observation the calculated information is normalized by a split information term expressed by:

$$SplitInformation(D, A) = -\sum_{i=1}^{n} (\frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|}). \tag{B.6}$$

The normalized information gain ratio used in the final decision of splitting the data can than be obtained by:

$$GainRatio(D, A) = -\frac{InformationGain(D, A)}{SplitInformation(D, A)}. \tag{B.7}$$

The C4.5 algorithm is based on a divide and conquer strategy. There are two cases in which the creation of the tree terminates [89]:

1. A pure subtree meaning that all left training samples belong to a single class c_i. Therefore, a leaf node is created identifying class c_i

2. A subtree is reached in which further tree splitting delivers no additional information gain. In that case also a leaf node is created and its corresponding classvalue is set to the most frequent class determined at the leafs parent.

One advantage of the C4.5 algorithm is the ability to classify also continues data. Therefore, in the case of continues attributes a threshold value $\Theta_{D,A} := A > v$ is set for all possible attribute values $v \in A$. In a subsequent step this threshold value is optimized to determine the maximum information gain that could be obtained by splitting the attribute as formulated in the following Equation B.8:

$$\Theta_{D,A} = \underset{v}{\operatorname{argmax}}\{GainRatio(D, A > v)\}. \tag{B.8}$$

.

B.2.3 Treepruning

According to the "Occam's razor principle" it is desireable having different classifiers yielding to similar classification performances to choose the classifier that introduces the fewest assumptions and postulates the fewest entities. This holds also for the creation of decision trees. Considering the C4.5 algorithm, cases can occur in which a non optimal tree is created in terms of size and classification performance. In order to achieve a higher degree of generalization and more simplified tree structures several parts of the tree are cutted. This technique is is referred to as treepruning.

In [105] Quinlan proposed a heuristic tree pruning approach based on statistical conclusions which referred to as pessimistic treepruning. The approach, applied in the C4.5 algorithm, utilizes the so called postpruning technique in which a tree is pruned after its creation. Inner nodes are replaced by subtrees (subtree-raising) on the basis of an estimated error rate.

During the pruning procedure the amount of training samples reaching each decision tree node is considered. This is done based on the assumption that every inner tree node could potentially be replaced by the leaf node that is reached by the majority of training samples. Subsequently, the decision wether to replace or not an inner node is made. This decision is based on the inner nodes estimated pessimistic error e_{inner} and the weighted estimated pessimistic error of its corresponding leaf nodes e_{leaf}. If $e_{inner} > e_{leaf}$ the inner node will be replaced by its majority leaf node otherwise the leaf nodes are cutted.

To estimate the pessimistic error occurring at each node the decision tree classification is interpreted as a Bernoulli-Process. Let $|E_i|$ be the number of incorrect classified training samples d_i and $|D_i|$ be the total number of all classified samples reaching node i. The error rate observed at node i is then calculated by $r_i = N_i/|D_i|$ while a_i depicts the actual rate of missclassified training samples. As a result the missclassifications error confidence interval upper bound is than calculated by [91]:

$$|\frac{r_i - a_i}{\sqrt{a_i(1-a_i)/|N_i|}} > z_i| = c, \tag{B.9}$$

where c names the confidence value.

The C4.5 algorithm uses a pessimistic error estimation e_i of the actual missclassification rates a_i upper bound by setting c to 0.25 as default value. This leads to value of 0.69 for z_i since a normal distribution is assumed. Resolving Equation B.9 to z_i the pessimistic error e_i of node i can be obtained by [89]:

$$e_i = \frac{f_i + \frac{z_i^2}{2N_i} + z_i\sqrt{\frac{f_i}{N_i} - \frac{f_i^2}{N_i} + \frac{z_i^2}{4N_i^2}}}{1 + \frac{z_i^2}{N_i}}. \tag{B.10}$$

If desired the confidence value could also be set to a value lower than 0.25 yielding to a more drastic pruning of the tree.

Concluding, it has to be mentioned that the error estimation given by Equation B.10 is based on heuristic assumptions lacking proper a statistical foundation [106]. However, underpinned by its good practical results the C4.5 algorithm is beside these critics one of the most used approaches in the induction of decision trees nowadays [91].

B.3 Artificial Neural Networks

Artificial neural networks are a machine learning method derived from the microstructure of the human brain and therefore can be interpreted according to Jain et al. [107] as *"as massively parallel computing systems consisting of a large number of interconnected simple processors"*. Nowadays, neural networks are applied extensively to various pattern recognition tasks starting from simple classification to more sophisticated time series prediction. Furthermore, they possess the ability to learn complex nonlinear input-output relationships through sequential training procedures.

One particular and commonly used class of neural networks is the *Multilayer Perceptron (MLP)* in which learning is carried out through so called backpropagation of error. In the following the basic elements and concepts of Multilayer Perceptrons will be explained.

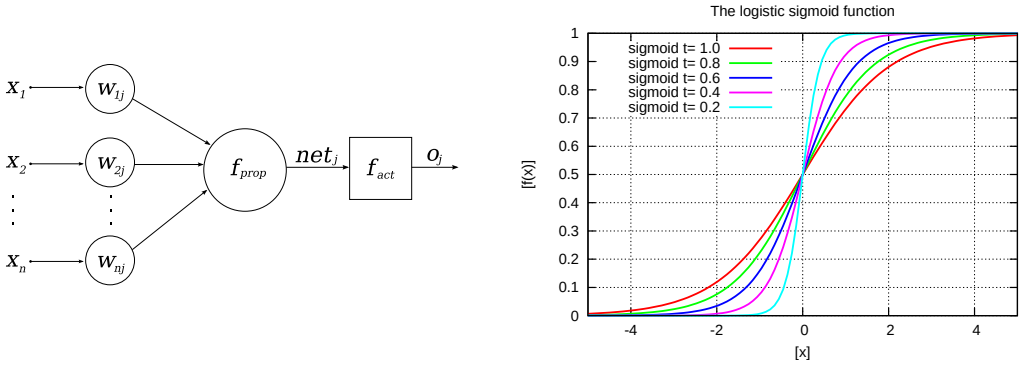


Figure B.2: Simple perceptron classification: (l) perceptron structure comprised of a propagation and an activation function f_{prop} and f_{act}, (r) exemplary plots of logistic sigmoid function for different temperature parameters t commonly applied as activation function within perceptrons.

The Perceptron

In 1958 Frank Rosenblatt [108] introduced the perceptron as the simplest form of a neural network. In general, a perceptron can be interpreted as a linear discriminant function that classifies a given input feature vector into two classes. Furthermore, a perceptron is comprised of a single neuron with several adjustable weights. Figure B.2 shows the structure of an exemplary neuron. As illustrated, each neuron consists of the following three basic elements:

1. A set of connecting or input links also referred to as *synapses*, while each link is associated with a weight $w_0, w_1, ..., w_n$.
2. A *propagation function* f_{prop} that calculates the linear combination of the input signals and their respective weights.
3. An *activation function* f_{act} that indicates the state of a neurons activation. Typically the output is limited to the interval $[0; 1]$ or $[-1; 1]$ respectively.

The linear combination calculated by perceptrons propagation function f_{prop} given a arbitrary n-dimensional input vector $x_i = (x_i^1, x_i^2, ..., x_n^m)$ and a n-dimensional weight vector $w_i = (w_i^1, w_i^2, ..., w_n^m) \in W$ is expressed by:

$$\begin{aligned} f_{prop} &: X \times W \rightarrow \mathbb{R} \\ f_{prop}(x_i, w_i) &= \sum_{j=1}^{m} (x_i^j w_i^j) + b, \qquad \text{(B.11)} \\ &= net_i, \qquad \text{(B.12)} \end{aligned}$$

where b is the so called *bias* of the perceptron. To achieve linear classification the result of the propagation function is applied to the activation function f_{act} of the neuron. In principle every functions which is strictly increasing monotonically over a certain interval can be

used as activation function. A simple discrimination between two classes can be achieved using e.g. the McCulloch-Pitts model [109] which uses as simple bilevel thresholding:

$$f_{act}^{thres}(net_i) = \begin{cases} 1, & \text{if } f_{act}(net_i) \geq \Theta \\ -1, & \text{if } f_{act}(net_i) < \Theta \end{cases}, \tag{B.13}$$

where Θ denotes the perceptrons threshold value. Considering Equation B.13 it becomes obvious that the bias parameter b of the propagation function applied in Equation B.11 can be used to alter the spatial postion of the perceptrons decision boundary.

Depending on the perceptron learning rule it can become essential that the activation function f_{act} should exhibit the property of being differentiable over a specific interval. Often used activation functions are the tangents hyperbolicus or the logistic sigmoid function. The logistic sigmoid function, is expressed by:

$$f_{act}^{sigmoid}(x,t) = \frac{1}{1+e^{\frac{-x}{t}}}, \tag{B.14}$$

were $x \in [0;1]$. It has the comfortable property that its derivative can be computed in a straightforward manner as shown in Equation B.15:

$$\begin{aligned} f_{act}^{'sigmoid}(x,t) &= (\frac{1}{1+e^{\frac{-x}{t}}})', \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{\frac{-x}{t}}}, \\ &= g(x)(1-g(x)). \end{aligned} \tag{B.15}$$

This quality is of particular importance in perceptron learning and will be explained in detail the next section.

B.3.1 Perceptron Learning

The basic question, so far unanswered is, how to choose the perceptrons weights $w_1, w_2, ..., w_n$ yielding to a high number of correct classifications. The answer therefore is that after random selection of initial weights they have to be adapted by the perceptron through an iterative learning process. The general perceptron learning rule also referred to as Delta or Widrow-Hoff rule is calculating the difference between the expected t_i and the actual output o_i of the perceptron. Assuming a linear propagation function, the generalized Delta rule can be expressed according to Mitchell [55] by:

$$\begin{aligned} w_{i,p+1} &= w_{i,p} + \Delta w_{i,p}, \\ &= w_{i,p} + \eta(t_i - o_i)x_i, \\ &= w_{i,p} + \eta\delta_p x_i. \end{aligned} \tag{B.16}$$

where x_i corresponds to i-th input feature vector and p defines the number of applied training epochs. Furthermore, $\eta \in [0;1]$ defines the learning rate of the perceptron training. As obvious from Equation B.16 for every training sample the learning rule increases the weight w_i if $t_i - o_i > 0$ and decreases it if $t_i - o_i < 0$.

It remains important to node that the above defined learning rule is only converging to a stable solution if the applied training classes of the classification task are linearly separable. In that case a weight vector w_i can be found that classifies all training samples correctly. However, in the case of non linear separable classes the converging of the learning rule can not be guaranteed. One can enhance the delta rule in a way that is converging even if the training samples of the classification tasks are not linearly separable. The objective of learning is therefore to find the global minimum of all possible errors yielding to the minimum classification error. This approach is referred to as gradient descent learning and will be presented in the following.

Let X be a set of training samples, the idea of gradient descent learning is based on the results of the perceptrons activation function f_{act}. An error function $E(W)$ is introduced which calculates the training samples x_i sum of the squared differences between the perceptrons expected outputs t_i and its actual outputs o_i [55]:

$$E_X(W) = \frac{1}{2}\sum_{i=1}^{n}(t_i - o_i)^2. \tag{B.17}$$

The perceptron can then be trained by minimizing the error function $E_X(W)$ with respect to the weight vector W. A plot of the error function results in a multidimensional parabolic error surface in the "weight space" as shown in Figure B.3. Finding the global minimum of the error surface can be achieved by starting from an arbitrary point on the surface. Determined by the perceptrons initial weights and the initial inputs the objective is to move forward on the error surface towards a global minimum in a step-by-step fashion. The partial derivative of the error function $E_X(W)$ as shown in Equation B.17 with respect to a particular weight w_i is calculated by:

$$\begin{aligned}
\nabla Err_p(W) &= \frac{\partial Err_p(W)}{\partial w_i}, \\
&= \frac{\partial Err_p(W)}{\partial o_i}\frac{\partial o_i}{w_i}, \\
&= -(t_i - o_i)\frac{\partial o_i}{\partial net_i}\frac{\partial net_i}{\partial w_i}, \\
&= -(t_i - o_i)f'_{act}(net)x_i.
\end{aligned} \tag{B.18}$$

The gradient of $\nabla E_X(W)$ results in a vector yielding in the direction of the highest increase of classification error of the error surface. Consequently, the gradients inverse given by $-\nabla E_p(W)$ yields in the the direction of a local classification minima. The learning rule of

Equation B.16 can than be reformulated to find a minimum classification error according to Equation B.19:

$$\begin{aligned} w_{i,p+1} &= w_{i,p} + \Delta w_{i,p}, \\ &= w_{i,p} - \eta \sum_{i=1}^{n} (t_i - o_i) f'_{act}(net_i) x_i. \end{aligned} \tag{B.19}$$

However, it has to be mentioned in the case of nonlinear perceptrons the error surface exhibits a global minimum but perhaps also multiple local minima. To diminish the probability of converging to a local minima the so called *incremental delta* or *incremental gradient descent rule* can be applied. Using the incremental delta rule the weights of the perceptron are updated recursively after each applied training sample yielding to:

$$\Delta w_i = -\eta (t_i - o_i) f'_{act}(net_i) x_i. \tag{B.20}$$

Subsequently, the learning rule is less susceptible to converge against local minimas.

A parameter whose selection is in general very decisive for the perceptrons learning process is the learning rate η. The learning rate determines the size of the delta rules incremental optimization steps on the error surface and therefore the accuracy and speed of the learning procedure. Like in the majority of optimization problems choosing an appropriate learning rate is a tradeoff between the optimal result and time. While choosing a small η becomes computational expensive especially at narrow regions of the error surface, choosing in contrast a high learning the optimal solution could be jumped over. According to Kriesel [110] learning rates should be chosen from the range:

$$0.01 \leq \eta \leq 0.9. \tag{B.21}$$

However more sophisticted methods by adapting a momentum rate or stochastic approximation have been published by Orr and Leen [111] and Kushner and Yin [112].

B.3.1.1 Multilayer Perceptrons

Perceptrons are capable to discriminate linear separable data by dividing the input space using a linear hyperplane. However, even in simple classification tasks, like the XOR problem were the input data is not linear separable, the perceptron fails to learn a generalized decision boundary [52]. The appliance of single perceptrons to real world classification tasks is therefore limited.

A classifier capable to learn even more complex decision boundaries can be designed by arranging several perceptrons in multiple layers. This idea is implemented in so called *Multilayer Perceptrons (MLP)*. As shown in Figure B.3 multilayer perceptrons are referred

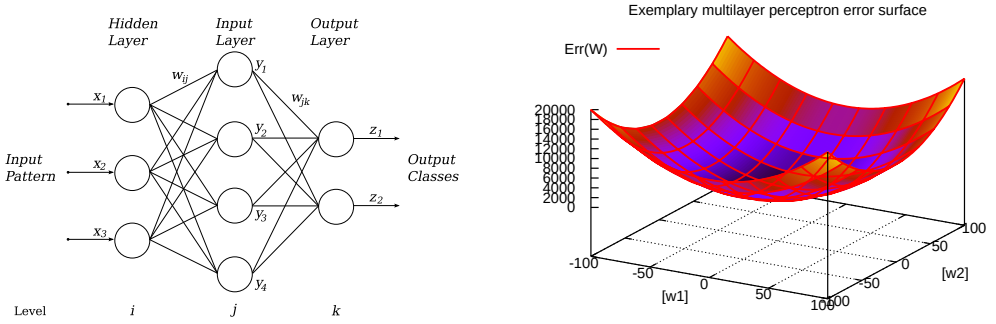


Figure B.3: Multilayer perceptron classification: (r) fully connected multilayer percepetron comprised of three distinct layers i, j, and k, (r) exemplary error surface of a multilayer perceptron having two trainable weights w_1 and w_2.

to as fully connected neural networks in which the calculated output of the actual layer is used as input of the subsequent layer. Multilayer perceptrons are characterized by two important properties. First, every perceptron in one layer is connected to all perceptrons of the successive layer. Secondly, within a multilayer perceptron each layer exhibits only directed connections to the neurons of the next following layer. Multilayer perceptrons can be seen as a special type of artificial neural networks referred to as feedforward networks that are forming a directed acyclic graph.

Mathematical a multilayer perceptron can be defined as a sorted triple (P, ν, W), where P defines the set of perceptron p_i and ν the sorted set $\{\nu_{i,j} | i, j \in \mathbb{N}\}$ of interconnections between two perceptrons p_i and p_j. As before W defines the set of weights corresponding to the interconnections ν. The weight of a single interconnection $\nu_{i,j}$ will be denoted by $w_{i,j}$ in the following. Figure B.3 depicts a simple multilayer perceptron consisting of an input and output layer and a third layer between the input and output neurons. In general these middle layers connecting input and output layers are referred to as *hidden layers* and can be of arbitrary number and size.

Given a randomly selected input vector x_i the output of the multilayer perceptron is calculated by forwarding x_i from the input layer i, through the hidden layer j to the output layer k in a feedforward manner. In every layer the input vector is transformed into a new vector by applying a two staged procedure. In the first stage, the linear combination of the input vector and the corresponding interconnection weights $w_{i,j}$ are calculated by the the propagation function f_{prop} within each perceptron of the layer as shown in Equation B.11. In the second stage, the calculated results are applied to the the perceptrons activation functions f_{act} leading to the layers output or activation vector.

Considering the multilayer perceptron shown in Figure B.3, let $x_i, x = 1, ..., n$ be an arbitrary input vector of the multilayer perceptron and $z_k, k = 1, ..., m$ its calculated output vector obtained from the output layer k. The activation within a single output layers perceptron z_k can than be calculated recursively solving the following equation:

$$
\begin{aligned}
z_k &: \mathbb{R}^n \rightarrow \mathbb{R}^m \\
z_k(x_i) &= f_{act,k}(y_j^{prop}), \\
&= f_{act,k}(f_{prop,k}(y_j w_{j,k})), \\
&= f_{act,k}(f_{prop,k}(f_{act,j}(x_i^{prop}) w_{j,k})), \\
&= f_{act,k}(f_{prop,k}(f_{act,j}(f_{prop,j}(x_i w_{i,j})) w_{j,k}),
\end{aligned}
\tag{B.22}
$$

where n defines the dimension of the input layer and m the output layer dimension and x_i^{prop} denotes a propagation functions result of x_i.

Backpropagation of Error

To train a given multilayer perceptron and adjust the weights $w_{i,j}$ of the net the so called backpropagation of error learning rule can be applied. Backpropagation of error previously published by Werbos [113] is based on the generalized delta rule, that calculates the squared error measure as shown in Equation B.17. Therefore, this approach can be seen as an enhancement of the gradient descent procedure to multilayer perceptrons.

Since a multilayer perceptron output nodes error can be easily obtained using the delta rule in contrast the error occurring at the hidden nodes is difficult to detected. For this purpose, Werbos expanded the delta rule from one trainable weight layer to several ones by recursively backpropagating the error. It can be shown selecting an arbitrary neuron y_j within the MLP performing the same derivation as for the delta rule the change within a particular weight $w_{i,j}^l$ corresponding to y_j can be determined by:

$$
\begin{aligned}
\Delta w_{i,j} &= \eta x_i \delta_j, \text{ where} \\
\delta_j &= \begin{cases} f'_{act}(net_j)(t_j - y_j), & \text{if } h \text{ is outer neuron} \\ f'_{act}(net_j) \sum_{k=1}^{n} (\delta_k w_{j,k}), & \text{if } h \text{ is inner neuron} \end{cases}
\end{aligned}
\tag{B.23}
$$

In the case of an outer neuron, the delta rule becomes the original delta rule already introduced in Equation B.20. While in the case of y_j being an inner or hidden neuron, the change in weight of y_j is depending on the weighted sum of the changes $\sum_k (\delta_k w_{j,k})$ of all subsequent neurons z_k of Y_j. As already mentioned η defines the learning rate and t_j the actual output of the perceptron.

B.4 Support Vector Machines

Support vector classification has become a prominent machine learning approach which can be found in a variety of application areas nowadays. Based decisively on the work of Vapnik in 1979 [114, 115] the underlying methodology of support vector machines (SVM) is

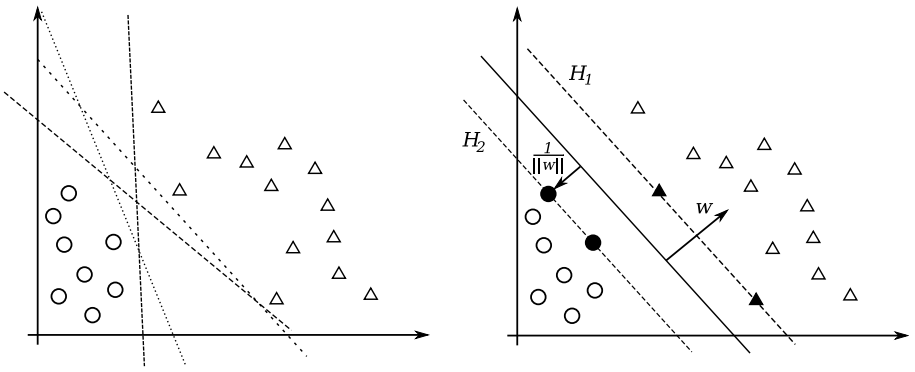


Figure B.4: Linear support vector classification: (l) possible choices of separating hyperplanes, (r) optimal hyperplane derived by margin optimization.

the construction of a separating hyperplane in a multidimensional feature space to separate between distinct sets of data. In the following an outline of the basis ideas behind support vector machines will be presented. For introductory purposes the explanations are based on the usage of SVM's for binary classification tasks. A more comprehensive and detailed overview can be obtained from the book of Schölkopf and Smola [116] or the introductory tutorial provided by Burges [58].

B.4.1 Linear Classification

Let $x_1, x_2, ...x_n \in X$ be the set of training samples representing the training data and $c_1, c_2, ...c_n \in \{-1, 1\}$ be their respective class labels. Supposing the training samples are linearly separable, the objective of SVM classication is to find a hyperplane separating both classes that can be expressed by:

$$\langle w, x \rangle + b = 0, \quad x \in X, b \in \mathbb{R}, \tag{B.24}$$

were w is the normal vector perpendicular to the hyperplane, $\langle w, x \rangle$ is the inner product of w and x_i. Normalizing w the inner product $\langle w, x_i \rangle$ depicts the length of x_i along the direction of w and b corresponds to the distance between the hyperplane and the origin.

The values of w and b can be scaled in a way that the hyperplane can be transformed into its canonical form as given by:

$$|\langle w, x_i \rangle + b| \geq 1, \tag{B.25}$$

implicating, that the distance between the hyperplane and the nearest training samples is $\frac{1}{||w||}$. The derived hyperplane can than be utilized for the prediction of new data points which are classified according to the following rules:

$$\langle w, x_i \rangle + b \;\; \geq \;\; +1 \quad for \; y_i = +1, \tag{B.26}$$

$$\langle w, x_i \rangle + b \;\; \leq \;\; -1 \quad for \; y_i = -1, \tag{B.27}$$

for all training samples $x_i, i = 1, ..., N$.

The classification decision function can than be expressed in the form:

$$g(x_i) = sgn(\langle w, x_i \rangle + b). \tag{B.28}$$

However, as illustrated in Figure B.4 there are various choices of hyperplanes that can be used to separate the training samples. A margin is introduced which is defined as the distance from the hyperplane to the closest sample vectors on either side. The objective during the SVM training procedure is to find the optimal hyperplane derived by the maximization of the margin between the training samples of both classes.

Considering the training samples for which the equalities in Equation B.26 and Equation B.27 holds. These samples lie on the hyperplanes $H_1 : \langle w, x_i \rangle + b = 1$ and $H_2 : \langle w, x_i \rangle + b = -1$ parallel to the optimal searched for hyperplane as shown in figure B.4. In the linear separable case the parameters w and b of the optimal hyperplane can than be obtained by maximizing the distance between the convex hulls created by the training samples of the distinct classes which is of size $\frac{2}{||w||}$.

The maximization of $\frac{2}{||w||}$ can be transfered into minimizing the formula $\frac{1}{2}||w||^2$. As a result, the optimization problem is converted into a constrained function that can be used to determine the separating optimal separating hyperplane:

$$\min_w \frac{1}{2}||w||^2 \quad \text{subject to} \quad |\langle w, x_i \rangle + b| \geq 1, \tag{B.29}$$

for all training samples $i = 1, ..., n$. The training samples lying on one of the hyperplanes H_1 and H_2 and determine the optimal hyperplane are referred to as *support vectors*. They are all equally close to the optimal hyperplane and it can be shown that they are enough to compute the separating hyperplane [117].

To solve the optimization problem formulated in Equation B.29 in even a high dimensional space lagrange multipliers are used. This is done for two reasons. First, the constraints $|\langle w, x \rangle + b| \geq 1$ can be replaced by the constraints on the Lagrange multipliers themselves. And second using this reformulation of the problem the training samples will only appear in the form of dot products between the samples. This is a very important property which will allow the generalization of the problem to the nonlinear case which will be shown in the subsequent section.

The given primal optimization problem in Equation B.29 is reformulated into the following Lagrange dual optimization problem:

$$\arg\max_{\alpha \geq 0} \ (\arg\min_{w,b} \ L(w, b, \alpha)), \tag{B.30}$$

where

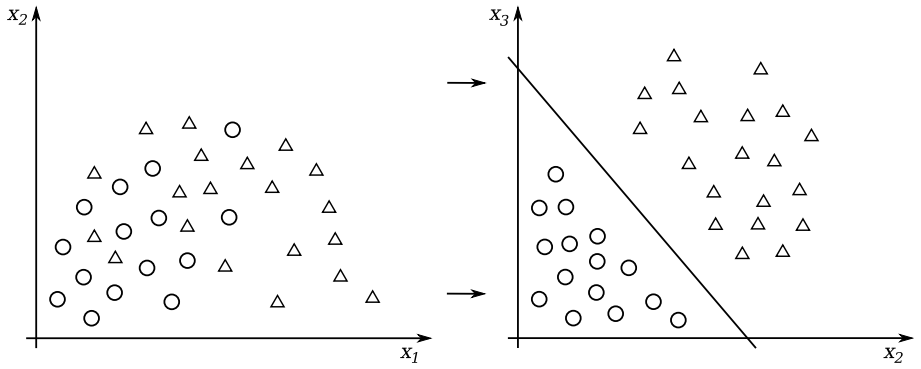


Figure B.5: Non-linear support vector classification: (l) non-linear separable feature data in the origin x_1x_2 dimension, (r) linear separable feature data mapped to the x_2x_3 dimension.

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i (c_i(\langle w, x_i \rangle + b) - 1). \tag{B.31}$$

It can be shown, that in the case of $\sum_{i=1}^{n} \alpha_i c_i = 0$ Equation B.31 is a convex quadratic programming problem, since the objective function is itself convex and the points satisfying the constraints also form a convex set. Using this observation and considering that $\sum_{i=1}^{N} \alpha_i y_i \neq 0$ resulting in $-\infty$ is not a maximum the dual problem statement can be simplified to:

$$\begin{aligned} &\arg\max_{\alpha \geq 0} (\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle), \\ &\text{subject to} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, \end{aligned} \tag{B.32}$$

for all training samples $i = 1, ..., N$. Solving this optimization task one will reveal a Lagrange multiplier α for every training point. The points for which $\alpha_i > 0$ are the searched for support vectors which lie on one of the hyperplanes H_1, H_2.

B.4.2 Non-linear Classification

In the preceding section the basic ideas of support vector classification for linear separable data were introduced. However, in real world classification scenarios the training samples might often not be linear separable. To obtain more complex decision surfaces, the feature space of the training vectors is mapped into higher dimensional space before undertaking the classification.

This can be achieved using a mapping ϕ to some other possibly infinite dimensional Euclidean space H equipped with a dot product as given by:

$$\begin{aligned} \phi : \mathbb{R}^{d_1} &\rightarrow \mathbb{R}^{d_2} \\ x &\mapsto \phi(x), \end{aligned} \tag{B.33}$$

were $d_1 < d_2$ as illustrated in Figure B.5. Consequently, the dual Lagrange problem statement derived from B.32 can be reformulated to :

$$\begin{aligned} \arg\max_{\alpha \geq 0} (\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle), \\ \text{subject to} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0, \end{aligned} \tag{B.34}$$

for all training vectors $i = 1, ..., n$. Solving this dual problem requires the calculation of the inner product $\langle \phi(x_i), \phi(x_j) \rangle$ in the higher dimensional space. This calculation can become very computational expensive or even unsolvable. As as shown by Schölkopf and Smola [116] the so called *kernel trick* can be applied in which the dot product is replaced by a specific positive definite kernel function in the way that:

$$\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j). \tag{B.35}$$

Using this replacement technique the inner product of the Lagrange optimization problem in B.32 can be calculated in the original feature space instead of the high dimensional space and therefore the computational complexity can be reduced strongly.

The theorem formulated by Mercer [118] reveals if k is a positive definite kernel function then there exists a Hilbert Space in which that kernel function is a inner product in the Hilbert Space. The kernel function k can then be utilized to learn a soft margin in the original space. However, it has to be noted that the coherence of a kernel function k and its mapping ϕ to higher dimensional space is not unique. Several kernel functions have been discovered so far. Within Table B.4.2 an overview of the most common kernel functions used in support vector classification is presented.

Table B.1: Common kernel functions used in support vector classification as introduced by Hsu [119].

Name	Function	
Linear	$\langle x_i, x_j \rangle$	
Polynomial	$(\gamma \langle x_i, x_j \rangle + r)^{dim}$	$\gamma, r \in \mathbb{R}, dim \in \mathbb{N},$
Sigmoid	$\tanh(\gamma \langle x_i, x_j \rangle + r)$	$\gamma, r \in \mathbb{R}$
Radial Basis	$e^{-\gamma \|x_i - x_j\|^2}$	$\gamma \in \mathbb{R}$

B.4.3 Non-Separable Classification

In practice, a separating hyperplane not always exists, this will be evidenced by a growing arbitrarily large Lagrangian function. Even if a separating hyperplane exists it might not always be the most feasible solution. This is for example the case if the training data is containing outlying samples and as a result becomes non-separable. Therefore, the aim is to tolerate a certain amount of training errors.

Cortes and Vapnik [120] introduced positive slack variables $\xi_i \geq 0,\ i = 1, ..., n$, which allow violations of the constraints in given by the Equations B.32 and B.34. The relaxed constraints can then be reformulated as:

$$|\langle w, x_i \rangle + b| \geq 1 - \xi_i, \tag{B.36}$$

for all $i = 1, ..., N$. It can be seen choosing ξ_i large enough, the constraints can always be fulfilled. Therefore, a penalty factor ρ of tolerated training vectors was implemented in the objective function. Subsequently, for a chosen parameter $\rho > 0$ the optimization problem becomes:

$$\min_w \frac{1}{2}||w||^2 + \rho \sum_{i=1}^{N} \xi_i \quad \text{subject to} \quad |\langle w, x_i \rangle + b| \geq 1,\ \xi_i \geq 0, \tag{B.37}$$

where the non-zero slack variables ξ_i correspond to margin errors. The chosen parameter ρ can be interpreted as the tradeoff between having a large margin and a few classification errors. The modified Lagrange problem statement given by Equation B.34 can than be enhanced using Equation B.37 to solve also non-separable training samples.

C. Experimental Results

In this Appendix the classification results obtained by global and local feature evaluation as described in Chapter 7 are presented.

C.1 Global Feature Evaluation Results

Table C.1: Global feature evaluation accuracy results obtained by C4.5 Decision Tree classification performing 10-fold stratified sampled cross validation.

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	68.37	**76.82**	73.36	**77.18**	91.82
Wavelet MRA	73.81	75.18	76.00	76.27	77.00
Gaussian Filter	72.27	73.18	69.27	74.27	94.63
Mean Filter	68.91	70.00	74.91	75.91	94.55
Median Filter	74.09	75.45	75.82	76.00	93.55
Gradient Prewitt	70.27	72.45	**77.00**	74.45	94.45
Gradient Sobel	67.82	74.73	71.91	76.73	96.36
All Spatial	**74.64**	74.45	74.09	73.09	**97.27**

Table C.2: Global feature evaluation accuracy results obtained by Multilayer Perceptron classification performing 10-fold stratified sampled cross validation.

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	68.27	74.45	74.09	**90.09**	96.36
Wavelet MRA	75.37	72.18	74.18	76.64	84.36
Gaussian Filter	71.55	76.45	73.91	75.18	95.45
Mean Filter	76.36	77.27	**79.64**	76.73	93.64
Median Filter	71.55	72.55	79.45	78.64	93.55
Gradient Prewitt	77.82	**77.82**	76.09	83.18	96.75
Gradient Sobel	73.36	74.64	76.82	82.18	**98.18**
All Spatial	**78.91**	76.45	75.55	87.82	95.45

Table C.3: Global feature evaluation accuracy results obtained by Support Vector Machine classification performing 10-fold stratified sampled cross validation.

Feature	100dpi	200dpi	300dpi	400dpi	800dpi
DCT Coefficents	72.64	80.95	**85.85**	**92.92**	**99.08**
Wavelet MRA	75.03	75.00	75.96	84.91	87.17
Gaussian Filter	80.55	**81.30**	79.62	80.58	97.24
Mean Filter	80.33	80.37	78.70	80.56	92.66
Median Filter	72.55	79.43	76.85	82.40	93.51
Gradient Prewitt	75.47	80.00	81.13	88.23	97.17
Gradient Sobel	80.56	79.44	82.40	87.04	98.07
All Spatial	**80.89**	80.19	80.37	91.57	94.50

C.2 Local Feature Evaluation Results

Table C.4: Local feature evaluation accuracy results obtained by C4.5 Decision Tree classification on the basis of all document classes within the ground truth database.

	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	58.49	26.86	66.25	34.53	63.21	29.71	65.35	30.62	79.02	28.36
Cooc Gauss	64.16	25.89	63.89	25.32	66.31	23.91	63.86	23.99	91.29	13.99
Cooc Lbp	57.99	23.46	63.54	27.83	65.63	25.00	75.73	20.97	91.08	19.01
Corr	43.50	24.36	41.87	22.90	43.28	37.56	46.39	38.45	85.48	25.69
Gray Dist	53.81	30.95	55.21	29.07	50.75	19.90	54.11	18.25	73.13	27.44
Edge Dist	55.99	22.46	56.59	23.36	62.96	17.75	59.83	17.58	91.15	9.52
Edge Rough	47.78	32.50	44.10	20.35	45.21	20.74	46.13	19.45	87.95	19.31
All	64.11	18.22	64.52	24.40	69.96	21.34	71.98	19.86	97.33	7.19

Table C.5: Local feature evaluation accuracy results obtained by Multilayer Perceptron classification on the basis of all document classes within the ground truth database.

	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	49.99	39.75	53.24	27.76	66.78	34.18	61.99	29.48	73.86	29.49
Cooc Gauss	58.92	36.54	61.88	28.23	69.12	26.28	65.74	25.53	90.51	14.16
Cooc Lbp	54.13	37.02	63.38	29.90	66.63	26.53	73.83	24.77	90.29	19.25
Corr	42.86	41.48	47.73	45.55	42.01	41.62	49.15	25.24	84.67	26.21
Gray Dist	43.13	41.07	48.02	47.93	49.67	31.08	47.22	22.76	66.12	31.44
Edge Dist	49.85	25.39	45.03	23.60	53.81	28.33	56.59	14.57	80.58	12.90
Edge Rough	42.88	41.41	36.01	35.98	42.64	40.69	45.65	20.43	87.47	19.42
All	69.65	27.89	71.45	29.29	72.99	26.24	77.55	22.11	98.03	4.48

Table C.6: Local feature evaluation accuracy results obtained by C4.5 Decision Tree classification on the basis of all laser printed and photocopied documents within the ground truth database.

	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	51.13	13.88	68.97	20.85	63.60	15.43	78.15	17.47	96.70	2.24
Cooc Gauss	72.07	19.84	74.64	20.65	78.43	16.16	74.73	22.29	97.96	3.49
Cooc Lbp	64.32	15.68	72.03	20.17	78.00	14.03	79.83	16.61	98.27	3.27
Corr	50.72	35.21	47.86	26.50	52.02	39.23	55.72	41.27	97.04	4.51
Gray Dist	60.15	13.99	65.76	21.01	65.22	25.11	60.22	9.64	84.43	6.30
Edge Dist	60.86	17.26	60.45	24.49	66.36	22.70	68.03	10.58	97.68	2.77
Edge Rough	52.59	7.42	52.20	7.15	52.72	9.96	53.40	11.29	97.60	2.70
All	73.69	16.56	77.27	19.63	78.88	16.54	78.49	15.61	99.75	0.51

Table C.7: Local feature evaluation accuracy results obtained by Multilayer Perceptron classification on the basis of all laser printed and photocopied documents within the ground truth database.

	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	49.32	47.73	65.47	16.07	67.39	19.31	76.74	17.74	94.70	5.94
Cooc Gauss	71.40	18.67	76.47	22.14	79.01	20.73	74.40	27.14	98.41	3.16
Cooc Lbp	64.19	20.88	70.51	15.77	78.08	19.48	79.24	19.61	98.46	1.59
Corr	50.01	49.99	42.86	41.48	50.34	46.78	58.54	28.62	96.32	3.84
Gray Dist	50.13	49.89	57.16	49.45	50.11	49.84	52.40	40.22	77.99	15.28
Edge Dist	57.02	11.03	55.50	17.71	57.61	21.07	60.23	12.31	96.81	3.31
Edge Rough	52.93	22.74	42.58	34.48	50.12	49.61	53.28	10.25	96.87	4.27
All	74.56	16.08	78.72	24.11	80.31	20.79	80.89	18.24	99.97	0.07

Table C.8: Local feature evaluation accuracy results obtained by C4.5 Decision Tree classification on the basis of all laser and inkjet printed documents within the ground truth database.

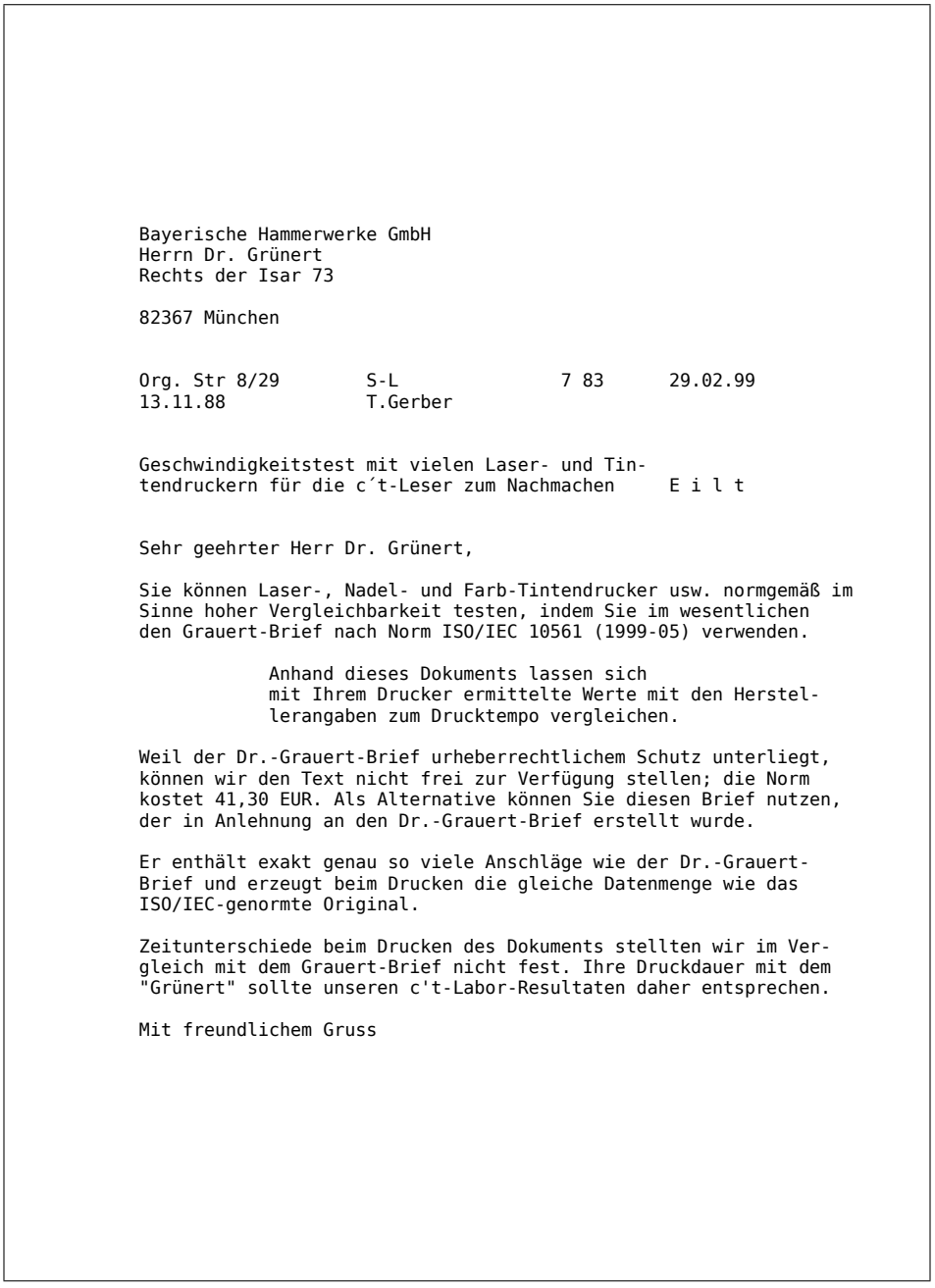
	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	72.30	38.50	73.30	35.81	72.84	40.43	75.47	30.28	77.38	36.63
Cooc Gauss	79.41	31.27	84.93	20.18	84.29	25.26	80.92	31.36	88.35	17.62
Cooc Lbp	81.54	31.39	88.31	22.63	80.21	29.16	83.01	27.77	91.30	14.03
Corr	72.00	43.30	73.04	36.73	72.42	37.61	75.15	37.70	80.14	32.53
Gray Dist	76.15	37.40	76.45	24.18	77.66	28.30	80.14	27.37	79.17	35.19
Edge Dist	74.69	35.79	87.37	18.29	89.24	16.85	83.44	21.76	96.52	2.31
Edge Rough	74.59	42.38	72.31	43.64	75.00	43.30	74.37	41.84	90.56	9.13
All	81.84	22.34	86.87	17.56	85.94	16.74	85.06	16.41	92.96	20.77

Table C.9: Local feature evaluation accuracy results obtained by Multilayer Perceptron classification on the basis of all laser and inkjet printed documents within the ground truth database.

	100dpi		200dpi		300dpi		400dpi		800dpi	
Feature	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Area Diff	77.27	36.73	87.89	24.75	75.00	43.30	74.93	43.26	72.98	44.30
Cooc Gauss	80.58	33.62	87.00	21.74	86.10	19.06	83.40	27.43	86.85	14.99
Cooc Lbp	81.40	34.03	84.95	23.47	80.30	27.66	89.90	23.75	91.72	16.57
Corr	75.00	43.30	71.24	41.85	72.30	38.87	74.86	35.69	79.05	34.71
Gray Dist	75.15	36.74	77.51	25.89	77.65	28.64	80.86	26.57	77.83	30.35
Edge Dist	75.06	41.36	88.07	19.21	86.52	16.27	80.72	21.46	91.43	6.51
Edge Rough	79.41	31.27	84.93	20.18	84.29	25.26	80.92	31.35	88.35	17.62
All	83.21	31.23	89.36	20.27	85.35	20.46	83.17	26.53	97.90	3.93

D. Template Document

In this Appendix the “Grünert” letter as used in ground truth document database creation is illustrated.



Bayerische Hammerwerke GmbH
Herrn Dr. Grünert
Rechts der Isar 73

82367 München

Org. Str 8/29 S-L 7 83 29.02.99
13.11.88 T.Gerber

Geschwindigkeitstest mit vielen Laser- und Tin-
tendruckern für die c´t-Leser zum Nachmachen E i l t

Sehr geehrter Herr Dr. Grünert,

Sie können Laser-, Nadel- und Farb-Tintendrucker usw. normgemäß im
Sinne hoher Vergleichbarkeit testen, indem Sie im wesentlichen
den Grauert-Brief nach Norm ISO/IEC 10561 (1999-05) verwenden.

Anhand dieses Dokuments lassen sich
mit Ihrem Drucker ermittelte Werte mit den Herstel-
lerangaben zum Drucktempo vergleichen.

Weil der Dr.-Grauert-Brief urheberrechtlichem Schutz unterliegt,
können wir den Text nicht frei zur Verfügung stellen; die Norm
kostet 41,30 EUR. Als Alternative können Sie diesen Brief nutzen,
der in Anlehnung an den Dr.-Grauert-Brief erstellt wurde.

Er enthält exakt genau so viele Anschläge wie der Dr.-Grauert-
Brief und erzeugt beim Drucken die gleiche Datenmenge wie das
ISO/IEC-genormte Original.

Zeitunterschiede beim Drucken des Dokuments stellten wir im Ver-
gleich mit dem Grauert-Brief nicht fest. Ihre Druckdauer mit dem
"Grünert" sollte unseren c't-Labor-Resultaten daher entsprechen.

Mit freundlichem Gruss

Figure D.1: ”Grünert“ letter used in feature evaluation and selection.

Bibliography

[1] J. S. Kelly and B. S. Lindblom, *Scientfic Examination of Questioned Documents*, 2nd ed. CRC Press, 2006.

[2] J. L. C. Chim, C.-K. Li, N. L. Poon, and S.-C. Leung, "Examination of counterfeit banknotes printed by all-in-one color inkjet printers," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 7, no. 2, pp. 69–75, 2004.

[3] C. K. Li and S. C. Leung, "The identification of color photocopiers: A case study," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 1, no. 1, pp. 8–11, 1998.

[4] J. D. Makris, S. A. Krezias, and V. T. Athanasopoulou, "Examination of newspapers," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 9, no. 2, pp. 71–75, 2006.

[5] J. L. Parker, "An instance of inkjet printer identification," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 5, no. 1, pp. 5–10, 2002.

[6] *Statement on Auditing Standard No. 99: Consideration of Fraud in Financial Statments*, American Institute of Certified Public Accountants (AICPA), October 2002.

[7] H. Cendrowski, J. P. Martin, and L. W. Petro, *The Handbook of Fraud Deterrence*. John Wiley and Sons, 2007.

[8] J. van Beusekom, F. Shafait, and T. M. Breuel, "Document signatures using intrinsic features for counterfeit detections," in *Proceedings of the 2nd Int. Workshop in Computational Forensics*, 2008.

[9] P. J. Smith, P. O'Doherty, C. Luna, and S. McCarthy, "Commercial anticounterfeit products using machine vision," in *Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE*, vol. 5310, 2004, pp. 237–243.

[10] A. Mikkilineni, G. Ali, P.-J. Chiang, G.-C. Chiu, J. Allebach, and E. Delp, "Printer forensics using svm techniques," in *Proceedings of the IS&T's NIP21: International Conference on Digital Printing Technologies*, vol. 21, Baltimore, MD, 2005, pp. 223–226.

[11] J. Mena, *Investigative Data Mining for Security and Criminal Detection.* Butterworth-Heinemann, 2003.

[12] K. Franke and S. N. Shrihari, "Computational forensics: Towards hybrid-intelligent crime investigation," in *Third International Symposium on Information Assurance and Security*, 2007, pp. 383–386.

[13] D. Ellen, *The Scientific Examination of Documents*, 2nd ed. Taylor and Francis, 1997.

[14] C. H. Lampert, L. Mei, and T. M. Breuel, "Printing technique classification for document counterfeit detection," in *Computational Intelligence and Security (CIS) 2006, Ghuangzhou, China*, 2006.

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

[16] J. Nickell, *Detecting Forgery: Forensic Investigations of Documents*, 1st ed. University Press of Kentucky, 2005.

[17] H. Dasari and C. Bhagvati, "Identification of printing process using hsv colour space." in *ACCV (2)*, ser. Lecture Notes in Computer Science, P. J. Narayanan, S. K. Nayar, and H.-Y. Shum, Eds., vol. 3852. Springer, 2006, pp. 692–701.

[18] T. W. Vastrick, *Forensic Document Examination Techniques.* The IIA Research Foundation, 2004.

[19] G. M. LaPorte, "Modern approaches to the forensic analysis of inkjet printing - physical and chemical examinations," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 7, no. 1, pp. 22–36, 2004.

[20] G. Herbertson, *Document Examination on the Computer - A Guide for Forensic Document Examiners.* WideLine Publishing Berkeley, California, 2002.

[21] A. Mikkilineni, G. Ali, P.-J. Chiang, G.-C. Chiu, J. Allebach, and E. Delp, "Printer identification based on graylevel co-occurance features for security and forensic applications," in *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents VII*, vol. 5681, San Jose, CA, 2005, pp. 430–440.

[22] B. Zhu, J. Wu, and M. S. Kankanhalli, "Print signatures for document authentication," in *CCS '03: Proceedings of the 10th ACM conference on Computer and communications security.* New York, NY, USA: ACM, 2003, pp. 145–154.

[23] T. F. Lunt, M. K. Franklin, D. L. Hecht, T. A. Berson, M. J. Stefik, D. Dean, A. G. Bell, T. M. Breuel, T. A. Cass, D. N. Curry, D. H. Greene, and R. T. Krivacic,

"Systems and methods for forgery detection and deterrence of printed documents," United States Patent, No. 6,790,259 B1, 2005.

[24] R. James, M. Long, and D. Newcomb, "Designing security holograms," in *Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE*, vol. 5310, 2004, pp. 264–274.

[25] L.-A. Peterson, "Surface treated security paper and method and device for producing surface treated security paper," United States Patent, No. 6,174,586, 2001.

[26] A. Mikkilineni, G. Ali, P.-J. Chiang, G.-C. Chiu, J. Allebach, and E. Delp, "Signature-embedding in printed documents for security and forensic applications," in *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents VI*, vol. 5306, San Jose, CA, 2004, pp. 455–466.

[27] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, August 2005.

[28] J. Oliver and J. Chen, "Use of signature analysis to discrimnate digital printing technologies," in *Proceedings of the IS&T's NIP18: International Conference on Digital Printing Technologies*, 2002, pp. 218–222.

[29] C. Schulze, M. Schreyer, A. Stahl, and T. M. Breuel, "Evaluation of graylevel features for printing technique classification in high-throughput document management systems," *Computational Forensics: Proceedings of the Second International Workshop, Springer Verlag*, vol. 5158, 2008.

[30] A. Mikkilineni, G. Ali, P.-J. Chiang, G.-C. Chiu, J. Allebach, and E. Delp, "Printer identification based on texture features," in *Proceedings of the IS&T's NIP20: International Conference on Digital Printing Technologies*, vol. 20, Salt Lake City, UT, 2004, pp. 306–311.

[31] J. Tchan, "The development of an image analysis system that can detect fraudulent alterations made to printed images," in *Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE, Volume 5310, pp. 151-159 (2004).*, ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, R. F. van Renesse, Ed., vol. 5310, jun 2004, pp. 151–159.

[32] Y. Akao, K. Kobayashi, S. Sugawara, and Y. Seki, "Discrimnation of inkjet printed counterfeits by spur marks and feature extraction by spatial frequency analysis," in *Optical Security and Counterfeit Deterrence Techniques V, Proceedings of the SPIE, Volume 5310, pp. 151-159 (2004).*, ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, R. F. van Renesse, Ed., vol. 5310, jun 2004, pp. 129–137.

[33] Y. Akao, K. Kobayashi, and Y. Seki, "Examination of spur marks found on inkjet-printed documents," *Journal of Forensic Science*, vol. 50, no. 4, pp. 915–923, July 2005.

[34] G. Ali, P.-J. Chiang, A. Mikkilineni, G.-C. Chiu, E. Delp, and J. 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*, vol. 21, Baltimore, MD, 2005, pp. 235–238.

[35] G. Gupta, R. Sultania, S. Mondal, S. K. Saha, and B. Chanda, "A structured approach to detect scanner-printer used in generating fake documents," in *Springer Lecture Notes of Computer Science*, 2007, pp. 250–253.

[36] J. S. Tweedy, "Class chracteristics of counterfeit protection system codes of color laser copiers," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 4, no. 2, pp. 53–66, 2001.

[37] C.-K. Li, W.-C. Chan, Y.-S. Cheng, and S.-C. Leung, "The differentation of color laser printers," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 7, no. 2, pp. 105–109, 2004.

[38] H. Gou, A. Swaminathan, and M. Wu, "Robust scanner identification based on noise features," in *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents IX*, 2007.

[39] N. Khanna, A. K. Mikkilineni, G. T.-C. Chiu, J. P. Allenbach, and E. J. Delp, "Scanner identification using sensor pattern noise," in *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents IX*, 2007.

[40] H. Kipphan, *Handbook of Print Media*, 1st ed. Springer, 2001.

[41] P. Doherty, "Classification of ink jet printers and inks," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 1, no. 2, 1998.

[42] R. A. Horton, "Identifiabilty of the flatbed scanner and its products (graphic files and printed results)," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 3, no. 1, pp. 41–46, 2000.

[43] A. S. Osborn and A. D. Osborn, "Questioned documents," *Journal of the American Society of Questioned Document Examiners (ASQDE)*, vol. 5, no. 1, pp. 39–44, 2002.

[44] H. S. Biard, "The state of the art in document image degradation modeling," in *Proceedings of the 4th IAPR Workshop on Document Analysis Systems*, 2000, pp. 1–13.

[45] E. H. B. Smith, *McGraw-Hill 2005 Yearbook of Science and Technology.* McGraw-Hill, 2005, ch. Document Scanning.

[46] J. Borch and R. G. Svendsen, "Paper material considerations for system printers," *IBM Journal of Research and Development*, vol. 28, no. 3, pp. 285–291, 1984.

[47] J. L. Crawford, C. D. Elzinga, and R. Yudico, "Print quality measurements for high-speed electrographic printers," *IBM Journal of Research and Development*, vol. 28, no. 3, pp. 276–284, 1984.

[48] E. H. B. Smith and X. Qui, "Statistical image differences, degradation features and character distance metrics," *International Journal on Document Analysis and Recognition*, vol. 6, no. 3, pp. 146–153, 2004.

[49] G. C. Holst, *CCD Arrays, Cameras and Displays.* JCD Publishing and SPIE Optical Engineering Press, 1996.

[50] A. E. Dirik, H. T. Sencar, and N. Memon, "Source camera identification based on sensor dust characteristics," in *Proceedings of the IEEE SAFE: Workshop on Signal Processing Applications for Public Security and Forensics*, 2007.

[51] S. Watanabe, *Pattern recognition: human and mechanical.* John Wiley and Sons, 1985.

[52] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley and Sons, 2000.

[53] S. J. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach*, 2nd ed. Prentice Hall International, 2003.

[54] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, Berlin, 2006.

[55] T. Mitchell, *Machine Learning.* McGraw-Hill, 1997.

[56] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, no. 31, pp. 249–268, 2007.

[57] A. K. Jain and R. C. Dubes, *Algorithms for clustering data.* Prentice Hall International, 1988.

[58] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[59] T. Bayes, "An essay towards solving a problem in the doctrine of changes," *Philosophical Transactions of the Royal Society of London*, no. 53, pp. 370–418, 1763.

[60] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Prentice Hall International, 2007.

[61] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, p. 146 – 165, Jan. 2004.

[62] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, no. 9, pp. 62–66, 1979.

[63] W. Burger and M. J. Burge, *Digitale Bildverarbeitung: Eine Einführung mit Java und ImageJ*, 2nd ed. Springer, Berlin, 2006.

[64] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, 1st ed. MIT Press, 1990.

[65] P. E. Ross, "Flash of genius," *Forbes*, pp. 98–104, November 1998.

[66] J. H. Bohórquez, B. P. Canfield, K. J. Courian, F. Drogo, C. A. E. Hall, C. L. Holstun, A. R. Scandalis, and M. E. Shepard, "Laser-comparable inkjet printing," *Hewlett Packard Journal*, February 1994.

[67] J. Lukas, J. Friedrich, and M. Goljan, "Detecting digital image forgeries using sensor pattern noise," in *Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents VIII*, 2006.

[68] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[69] M. Frigo and S. G. Johnson, "The design and implementation of fftw3," *Proceedings of the IEEE*, vol. 93, no. 2, pp. 216–231, 2005.

[70] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[71] P. Goupillaud, A. Grossman, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, pp. 85–105, 1984.

[72] D. W. Kammler, *A First Course in Fourier Analysis.* Prentice Hall International, 2000.

[73] T. Haenselmann, "Signalanalyseverfahren zur segmentierung von multimediadaten," Ph.D. dissertation, Universität Mannheim, 2004.

[74] I. Daubechies, *Ten lectures on wavelets*, 1st ed. SIAM: Society for Industrial and Applied Mathematics, June 1992.

[75] E. Brannock and M. Weeks, "Edge detection using wavelets," *Proceedings of the 44th annual ACM southeast regional conference*, pp. 649–654, 2006.

[76] s. Mallat, "A compact multresolution representation: The wavelet model," in *Proceedings of the IEEE Computer Society Workshop on Computer Vision*, 1987, pp. 2–7.

[77] B. B. Hubbard, *The world according to wavelets.* Birkhäuser Verlag, Berlin, 1997.

[78] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[79] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classifcation," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, pp. 610–621, 1973.

[80] T. Mäenpää and M. Pietikäinen, "Texture analysis with local binary patterns," in *Handbook of Pattern Recognition and Computer Vision*, 3rd ed., C. Chen and L. W. (eds), Eds. World Scientific Publishing Company, River Edge, NJ, USA, 2005, pp. 197–216, (invited chapter).

[81] "University washington document image database 1," http://www.science.uva.nl/research/dlia/datasets/uwash1.html, August 2008.

[82] "University washington document image database 2," http://documents.cfar.umd.edu/resources/database/UWII.html, August 2008.

[83] "University washington document image database 3," http://documents.cfar.umd.edu/resources/database/3UWCdRom.html, August 2008.

[84] "Medical article record system (mars) document image database," http://marg.nlm.nih.gov/index2.asp, August 2008.

[85] "Mediateam oulu document image database," http://www.mediateam.oulu.fi/downloads/MTDB, August 2008.

[86] D. H. Wolpert and W. G. Macready, "No free lunch theorems for search," Santa Fe Institute, Tech. Rep., 1995.

[87] ——, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–81, 1997.

[88] D. H. Wolpert, "The supervised learning no-free-lunch theorems," in *In Proc. 6th Online World Conference on Soft Computing in Industrial Applications.* Springer-Verlag, 2001, pp. 25–42.

[89] J. R. Quinlan, *C4.5: programs for machine learning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[90] D. Rumelhart, G. Hinton, and R. Williams, "Learning Internal Representations by Error Propagation in DE Rumelhart, JL McClelland (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations," *Foundations MIT-Press*, pp. 318–362, 1986.

[91] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

[92] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[93] S. N. Shrihari, C. Huang, H. Srinivasan, and V. Shah, "Biometric and forensic aspects of digital document processing," in *Digital Document Processing.* Springer Heidelberg, 2006.

[94] S. N. Shrihari, C. Huang, and H. Srinivasan, "Search engine for handwritten documents," *Document Recognition and Retrieval XII, SPIE*, pp. 66–75, 2005.

[95] K. Franke, L. Schomaker, C. Veenhuis, C. Taubenheim, I. Guyon, L. Vuupijl, M. van Erp, and G. Zwarts, "Wanda: A generic framework applied in forensic handwriting analysis and writer identification," in *Proceedings of the 9th IWFHR, Tokyo, Japan.* IEEE Computer Society, 2004.

[96] M. Kam, J. Wetstein, and R. Conn, "Proficency of professional document examiners in writer identification," *Journal of Forensic Science*, no. 39, pp. 5–14, 1994.

[97] M. Kam, G. Fielding, and R. Conn, "Writer identification by professional document examiners," *Journal of Forensic Science*, no. 42, pp. 778–786, 1997.

[98] ——, "Effects of monetary incentives on performance of non-professionals in document-examination proficency tests," *Journal of Forensic Science*, no. 43, pp. 1000–1005, 1998.

[99] M. Kam, K. Gummadidala, G. Fielding, and R. Conn, "Signature authenticitation by forensic document exminers," *Journal of Forensic Science*, no. 46, pp. 884–888, 2001.

[100] E. Alpayadin, *Introduction to Machine Learning.* MIT Press, Cambridge, 2004.

[101] B. D. Ville, *Decision Trees for Business Intelligence and Data Mining.* SAS Publishing, 2007.

[102] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[103] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees.* Chapman and Hall / CRC Press, 1984.

[104] C. E. Shannon, "A mathematical theory of commnication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[105] J. Quinlan, "Simplifying decision trees," *International Journal on Man-Machine Studies*, vol. 27, pp. 221–234, 1987.

[106] F. Esposito, D. Malerba, and G. Semeraro, "A comperative analysis of methods for pruning decision trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 476–491, 1997.

[107] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, 2000.

[108] F. Rosenblatt, "The perception: a probabilistic model for information storage and organization in the brain," *Psychological Review*, no. 65, pp. 386–408, 1958.

[109] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, no. 7, pp. 115–113, 1943.

[110] D. Kriesel, "A brief introduction on neural networks," World Wide Web, 2008. [Online]. Available: http://www.dkriesel.com/downloads/neuronalenetze-en-delta-dkrieselcom.pdf

[111] G. B. Orr and T. K. Leen, *Advances in Neural Information Processing.* MIT Press, Cambridge, 1997, no. 9, ch. Using carvature information for fast stochastic search, pp. 606–612.

[112] H. J. Kushner and G. Yin, *Stochastic Approximation Algortihms and Applications.* Springer-Verlag, New York, 1997.

[113] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard University, 1974.

[114] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data [in Russian].* Nauka, Moscow, (English translation: Springer Verlag, New York, 1982), 1979.

[115] ——, *The nature of statistical learning theory.* Springer-Verlag New York, NY, USA, 1995.

[116] B. Schoelkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press, Cambridge, 2001.

[117] M. Cord and P. Cunningham, *Machine Learning Techniques for Multimedia.* Springer-Verlag Berlin Heidelberg, 2008.

[118] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Proceedings of the Royal Society of London*, vol. 83, no. 559, pp. 69–70, 1909.

[119] C. Hsu, C. Chang, and C. Lin, "A practical guide to support vector classification," Department of Computer Science, National Taiwan University, Tech. Rep., 2003.

[120] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.