Document Forgery Detection using Printer Source Identification

Research Thesis

Thesis Submitted By:
Maryam Bibi

Supervised By:
Dr. Imran Ahmed Siddiqi

A dissertation submitted to the Department of Computer Science, Bahria University, Islamabad as a partial fulfillment of the requirements for the award of the degree of Masters in Computer Science
Session (2017-2019)

Bahria University
Discovering Knowledge

MS-13

Thesis Completion Certificate

Scholar's Name: MARYAM BIBI Registration No. 31890

Programme of Study: MS(CS)

Thesis Title: Document Forgery Detection Using Printer Source Identification.

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for Evaluation. I have also conducted plagiarism test of this thesis using HEC prescribed software and found similarity index at 9% that is within the permissible limit set by the HEC for the MS/MPhil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/MPhil thesis.

Principal Supervisor's Signature:

Date: 12. 2. 19 **Name:** Dr Imran Siddiqi

Bahria University
Discovering Knowledge

MS-14A

Author's Declaration

I, MARYAM BIBI hereby state that my ~~PhD~~ MS-thesis titled "Document Forgery Detection Using Printer Source Identification"

is my own work and has not been submitted previously by me for taking any degree from this university

BAHRIA UNIVERSITY ISLAMABAD CAMPUS

or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduate the university has the right to withdraw/cancel my ~~PhD~~ MS degree.

Name of scholar: MARYAM BIBI
Date: 12-Feb-2019

Bahria University
Discovering Knowledge

MS-14B

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled "Document Forgery Detection Using Printer Source Identification" is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Bahria University towards plagiarism. Therefore I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of ~~PhD~~ MS degree, the university reserves the right to withdraw / revoke my ~~PhD~~ MS degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of students are placed who submitted plagiarized thesis.

Student / Author's Sign: ____________

Name of the Student: MARYAM BIBI

Abstract

In recent years, identifying source printers has gained immense popularity for forgery detection. Paper employed for agreements and contracts have greater significance and to detect forgery among these papers validates the source more appropriately. With the advent of time, the document forgery is increasing and identifying source printer of document will result in validating a proper source as well as finding that document is forged or tampered. The proposed study aims to identify source printer of printed scanned document images. We employed a dataset for this study that contains 20 printers from which 13 are laser printers and 7 are ink-jet printers having 1200 document images. The documents also contains graphics, tables and text with different font style and sizes. We investigate features through hand-engineered techniques as well as machine-learning techniques. we also investigate the approaches related to text-dependent and text-independent mode. Development of such forgery detection systems are likely to facilitate the forensics community in analysis of printed scanned documents.

Acknowledgments

All praise is to Almighty Allah who bestowed upon me a minute portion of His boundless knowledge by virtue of which I was able to accomplish this challenging task. I am greatly indebted to my supervisor 'Dr. Imran Siddiqi'. It was through his continued support and guidance that I was able to finish this daunting task in time. I am honored to have had the opportunity to work with him. I would like to thank Dr. Faisal Shifaiat for sharing the dataset that is employed in this study. I extend my thanks to Sir Rashid Karim and Ms. Shehla Saif for providing with valuable feedback and support on my thesis work. A special thanks to my friends Romana Talat, Anmol Hamid and Fatima Akhlaq for helping me deal with the difficulties and providing emotional support and personal care.

Finally, I owe my special thanks to my family for putting up with all my demands and giving me the time and space required to successfully complete this work. I thank them for their patience, their love and their understanding and pray that my Allah protect them, and bless them with his greatest blessings. Ameen.

Maryam Bibi
Bahria University Islamabad, Pakistan

Feb 2019

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
U-LBP	Uniform Local Binary Pattern
LBP	Local Binary Pattern
OCR	Object Character Recognition
CNN	Convolution Neural Network
ReLU	Rectified Linear Unit
FC	Fully Connected
AUC	Area Under the Curve
GLCM	Gray Level Co-occurrence Matrix
GLCM_MD	Gray Level Co-occurrence Matrix_Multi-directional
GLCM_MDMS	Gray Level Co-occurrence Matrix_Multi-directional and Multi-scale
PTLS	Page Text Line Slope
PTLI	Page Text Line Interval
CMYK	Cyan, Magenta, Yellow, and blacK Color Space
DWT	Discrete Wavelet Transform
DFT	Discrete Fourier Transform
GPU	Graphical Processing Unit

Chapter 1

Introduction

1.1 Introduction

Paper has always remained the most widely employed mode to store and convey information in organizations, legal affairs, banking transactions and many more. Paper documents are specifically important when it comes to agreements and contracts. With the advancements in technology, it is now fairly straight forward to modify printed documents for malicious purposes or change the content of a document that is being digitally transferred. Printed documents can be forged and may be used for illegal purposes such as creating fake documents or currency, altering the contents of official or testimonial documents etc. Hence, coming to conclusion about the authenticity of a document has remained an attractive research area in forensic sciences. Traditional approaches for detection of forged documents are generally based on laboratory/chemical analysis that can damage the investigated document either partially or entirely.

With the recent advancements in different areas of image analysis and machine learning, examination of documents by forensic experts is being facilitated by computerized systems. Document forensics technology, focused on identifying the source of a document or on detecting forgery, has developed rapidly in recent years. Document forensics currently faces many challenges that limit its development. The techniques are currently limited to text documents with black text on white background. Various methods in digital image forensics are reviewed by many researchers. Recent research in document forensics has focused on forgery detection using source-printer identification [40, 1].

In the present research, we aim to develop techniques which are able to distinguish the documents produced by different printers. Identifying the source printer of a given document can be useful in detecting forgery or identifying tampered documents. In this research, we analyze the characteristic of different laser and ink-jet printers to determine the key signatures produced by the different 'physical and technical processes' involved in several types of printing. Based on the analysis of these signatures, we aim to detect the source-printer for query document.

Among different modalities to detect document forgery, identification of source printer is known to be an effective technique. Forgery or tampering can be identified if the questioned document has been printed using a different printer than the expected one. A number of traditional feature extraction techniques have been investigated for printer identification. These features mainly target the imperfections in printing of characters to characterize the source. In the recent years, conventional hand crafted features are being replaced with machine learned features extracted using deep learning techniques. Such automatic feature extractors are known to outperform traditional techniques on a number of classification problems like Skin Cancer Classification in [23, 11], Classification of Graphomotor Impressions [34],Scene classification of multisource remote sensing data [56] and Classification of Handwritten documents [10, 5].

The proposed approach is tailored to identify source printer of textual documents. Our proposed research aims to investigate features through both hand-crafted and machine-learned features. For Identifying source printer, the only available data is in form of scanned versions of textual documents more specifically, the training set consists of scanned textual documents of suspect printer or the printer that is targeted. Our experimental study is based on a database consisting of 1200 documents from 20 different printers. Out of these, 13 are laser printers and 7 are inkjet printers. The images are scanned in grayscale and contain images, tables and text (text-independent mode). The available training documents are printed with different fonts also, the different font-sizes of the document. we are not bounding our method neither to work with a single font and font-size, nor to work with a fixed character. Forgery of tampering can be identified if a questioned document has been printed using a different printer than the expected one.

1.2 Motivation and Problem Description

In today's era, the use of different types of printers is within the reach of everyone, hence printers can be used for malicious purposes. Documents can be forged and tampered by anyone therefore having ways of detecting who printed such documents is important to identify suspects in such cases. Moreover, being able to identify which printer printed the document is also a way to verify the authenticity of the document.

To address this issue, in literature, many traditional feature extraction techniques have been applied for printer recognition; however, such existing researches have been applied on different dataset either private or public that are available for this problem. Some reported techniques used color-documents, text-documents and some used datasets with majority of the printers being of the same brand but most of the techniques were text-dependent. Text-dependent techniques work with characters and same textual content is used at the time of training and testing. In this thesis, we have presented a solution based on work on text-independent approach. In text-independent solution, we identify the source printer that extracts the information from the textual content that is different in both training and testing. Furthermore, we also identify the best hand-crafted and machine-learned features for the problem of source printer identification.

1.3 Research Contribution

The main contributions of this thesis are:

1. A novel technique is presented for characterizing the source printer from scanned images of printed documents.

2. Performance of proposed techniques is investigated in text-dependent and text-independent methods. The analysis is performed by applying hand-crafted and machined learned features. For machine learned features, we have evaluated our experiments on different pre-trained Convolution neural networks. Our analysis is based on three level i.e. page level, patch level and character level.

3. For increasing accuracy of our system, our analysis is based on the discriminating power of different characters in identifying printer source. We implement deep learning to individually all characters (a-z) and picked out the best performance characters. Afterwards, we combined best performance characters and evaluated the results.

4. We analyzed the characteristic of different laser and ink-jet printers to determine the key signatures produced by the different physical and technical processes involved in several types of printing.

5. Performance exceeding the current benchmarks is reported on publicly available dataset.

1.4 Thesis Organization

This thesis consists of a total of 5 chapters. Chapter 2 covers the background and literature work of source-printer-identification, text-dependent approaches, data extraction and printer techniques that are used for color and text documents. Chapter 3 discusses the proposed methodology and frame work of system and the detailed analysis of algorithms that are used in proposed methodology. Chapter 4 discusses the experimental setup with results and discussions based on our extracted results. Chapter 5 concludes the research work with final considerations and highlight future directions.

Chapter 2

Literature Review

Printed documents are an integral part of almost every organization. A wide variety of printer types and models are available from numerous vendors. In the recent years, laser printer has become the most widely employed printing technique due to its speed and reduced costs. A laser printer employs a dry painting process. In this process, a black sooty powder and a paper is used for printing. A charged drum is involved in the process and when it revolves, a beam of laser is reflected by a mirror and the laser prints the letters and images as a pattern. After reflection of the laser, the positively charged toner is attracted as the paper rolls under the drum. The fusing process then fuses the images and letters on the paper permanently. The process of printing a document through a laser printer is illustrated in Figure 2.1. The intrinsic characteristics generated on paper can be seen as some imperfections in the manufactured parts. These imperfections maybe caused by the electromagnetic charges, different drum size or its revolving speed, the placement of parts etc. and can be exploited to identify the printer.

In the literature on identifying source printer, banding has been the most discussed intrinsic characteristic. Banding refers to the light and dark lines in a perpendicular direction to where the paper is moved inside the printer [12]. In most cases, different models of printers have unique banding frequencies and can be characterized by these bands. Identifying such banding artifacts has been the common focus of researchers in the literature. As a function of type of documents or images, techniques for identifying the printer are broadly categorized into two subsection, colored-documents and textual-documents (binary or gray scale) as discussed in Section2.1 and 2.2.

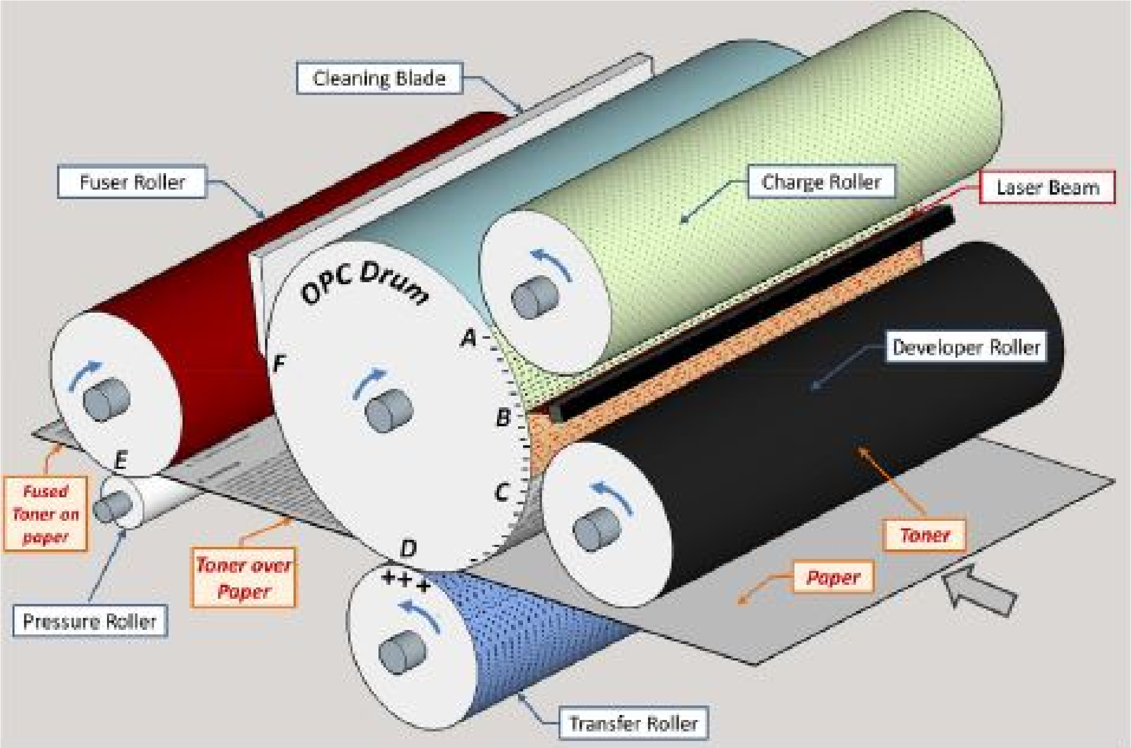


Figure 2.1: 'Laser printer printing process: (a): Charging (b): Exposure (c): Development (d): Transfer (e): Fusing (f): Cleaning'.[13]

2.1 Colored Documents

Colored printed documents typically comprise of images (in addition to textual content). Intrinsic signatures in printing process like commotion (noise) and geometric distortions or, insights inferred from the transformed scanned pictures are typically exploited to characterize the printer.

- **Geometric-distortion Analysis:** Among notable contributions exploiting geometrical distortions, Mao et al. [3] proposed a technique which relies on generating the geometric-distortion signature of electro-photographic printers. The signature of questioned document is compared with those in the reference base to identify the printer. They consider clustering printed images that are originated from the same printer model and type with different gray levels from each printer. Authors carried out experimental study of the proposed technique on a dataset of six printers and reported promising identification rates.

  Wu et al. [57] proposed a method for detection of half-tone dots arrangements on paper to examine images (for forgery) and identify the source printer. Forgery detection and printer identification performance were tested separately. The position of halftone dot is estimated by constructing a Gaussian Model and then calculate the correlation coefficient between that model and halftone dot. They established a printer model that is based on hexagon structure and it contains 6 radiuses and 6 angles. This printer model is thus used to identify printer and detect forgery. The performance is evaluated on dataset that contains 5 printers of which 4 printers have HP label and 1 is of Canon Printer. Each printer prints 10 images and then these images are scanned by 'Epson Perfection 1200' at resolution of 600dpi. To identify printer, Euclidean distance is used to calculate the correlation between printer known and unknown model. For Forgery detection they used K-mean clustering that distinguish the forged part from the original scanned image. The technique realized an overall accuracy of 87.92% for printer identification and forgery detection.

- **Noise Analysis:** Choi et al. [4] studied printed colored images that are in RGB scale and then converted into CYMK scale. These images are then decomposed into 4 bands by using 2-D discrete wavelet transform (DWT) out of these 4 bands statistical features are only computed from high frequency band. Authors extracted 39 features from each image. 'Support Vector Machine' (SVM) were used as a classifier and the technique was evaluated on a database of images printed using four different colored printers with 99 pages per printer. Authors classify the brand of color-laser printer, color toner, and model of color laser printer and achieved accuracy 97.89%, 92.28%, and 80.24%, respectively.

Hae & Jung [28] incorporated 60 noise features based on 'Wiener filters' and 'gray level co-occurrence matrices' extracted in the CMY color space. The images are in RGB space and they were converted into CMY color space. Using Wiener filter removes unnecessary noise from the images and authors found that noise removed version is much smoother than original images. Noise feature were extracted by calculating the difference between CMY-image and 'wiener filtered' image. To extract resultant feature of size 60, GLCM was used to compute 5 statistical features, 3 color channels at 4 directions. The statistical features were also computed by using weiner filter. These extracted features are then fed to SVM classifier. To support multi-class classification problem, authors used radial basis function as their kernel function. The employed techniques, used 7 color printers of different models and total of 2597 images used in experiment. To evaluate the performance of their proposed methodology, two tests were applied i.e. 'brand identification test' and 'model identification test' and concluded that they achieved overall best performance in comparison with previous studies discussed in literature.

In another study, Wang et al. [49] explored the relationship between scanned colored images and colored laser printers. Authors used scanned color images for feature extraction process and then train SVM classifier on those features. Statistical analysis based features are extracted using discrete wavelet transform. Authors applied several statistical techniques to extract features from DWT sub-band. Total 45 statistical features are extracted such as skewness, kurtosis, covariance, and standard deviation. Best features were selected on the basis of accuracy rate. Total features and best selected feature both were divided into training and testing part. Authors performed comparative analysis on 10 models of color laser printer of 6 different brand and carried out classification using SVM realizing an accuracy of 92.4%. Van et al. [55] worked on detecting the color laser printer using machine identification codes (CPS). CPS dots are also called yellow dots that are printed on document. Manual binarization method was also used to extract cps dot because other methods have some anomalies. The following binarization method was:

$$h(x,y) = \begin{cases} 0 & \text{if } min(h_R(x,y), h_G(x,y)) - h_B(x,y) < T_1 \\ & \text{and } h_R(x,y)\, T_2 \\ & \text{and } h_G(x,y)\, T_2 \\ 255 & else \end{cases}$$

After extracting these yellow dots, author used separating distance method for extracting horizontal and vertical pattern. For CPS comparison an accuracy of 91.3% is achieved.

- **Textural & Transformed-Image Statistical Analysis:** Lee et al. [39] investigated the distinguishing properties of printers based on halftone textures. Authors employed textual regions in the image and computed the CMYK histograms to characterize the printer. For calculating the histogram, transforming RGB to the CMYK color space that helps to enhance the halftone texture. Each channel of 'CMYK' is organized with set of dots or lines. 'Hough transform' is used to extract lines, which are described by angle and distance. The binarization was performed at each channel to apply transformation and then merging all channels of CMYK domain to get the histogram. Separately, authors extracted the reference pattern using histogram of RGB image and then compute correlation between reference pattern and histogram. Experiment are performed on 9,000 images of 9-printers.

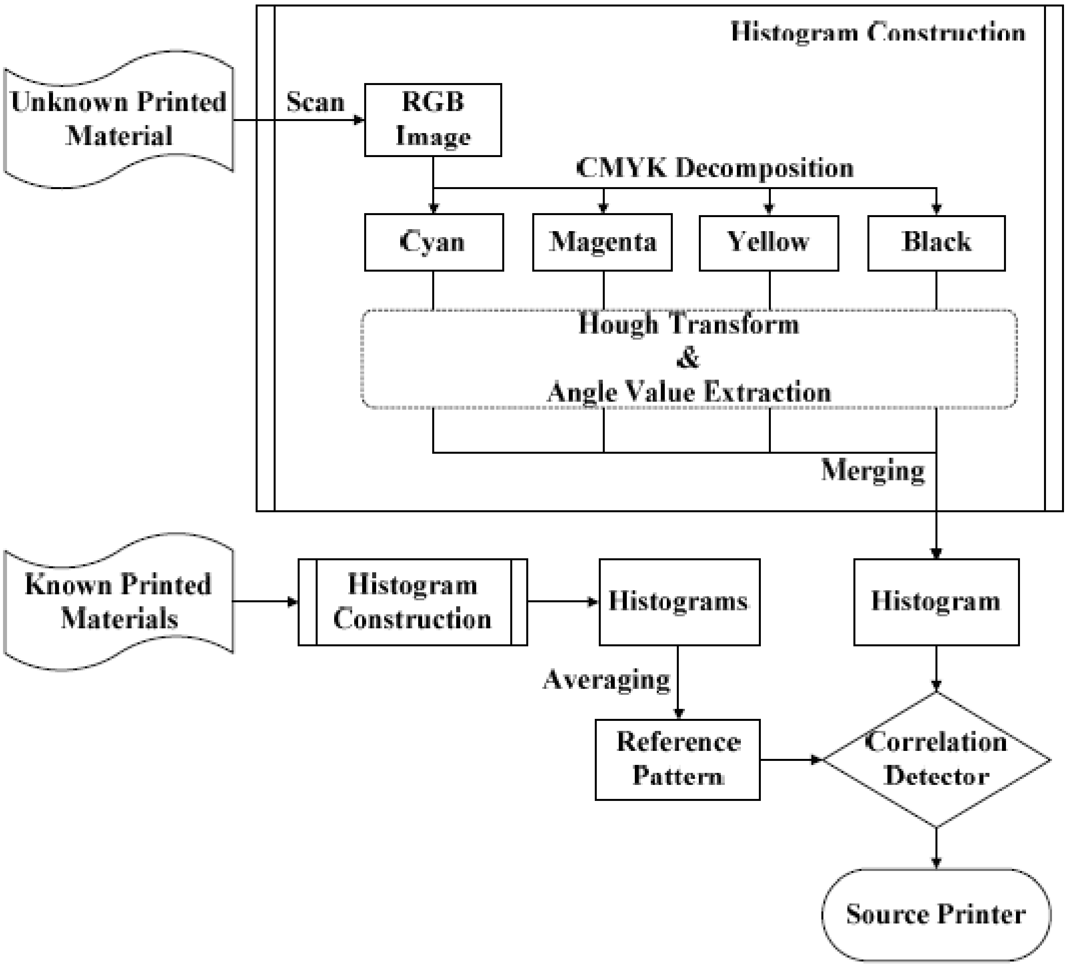


Figure 2.2: Methodology of Source printer Identification [39]

The study concluded that most of the classification errors resulted due to documents printed from different models of the same brand.

In another similar study [25], authors extracted the pattern of variation of illumination in printed images. halftone texture features were extracted using discrete Fourier Transform (DFT). CMYK color space is used to extract features such as the printing angle features, the printing resolution features, and statistical features. Features capturing the illumination variation were fed to an SVM to learn to discriminate between different printers. The technique reported better classification rates as compared to those reported in [39]. Cruz et al. [7] proposed classification-based-approach for document forgery detection. Authors used 'Uniform Local Binary Patterns' (LBP) to capture discriminant texture-feature and descriptor for contextual information. Different pre-processing methods were applied on the images for removal of noise for textural analysis. The patches of images were extracted around each connected component to cover the discontinuities and then compute 'LBP descriptors' on each patch. 'LBP' is also used to extract contextual-information from neighbor patches to extract the descriptor. All patches are classified as a forge and non-forge classes. Suitable results are reported using SVM classifier on 7 color laser printers.

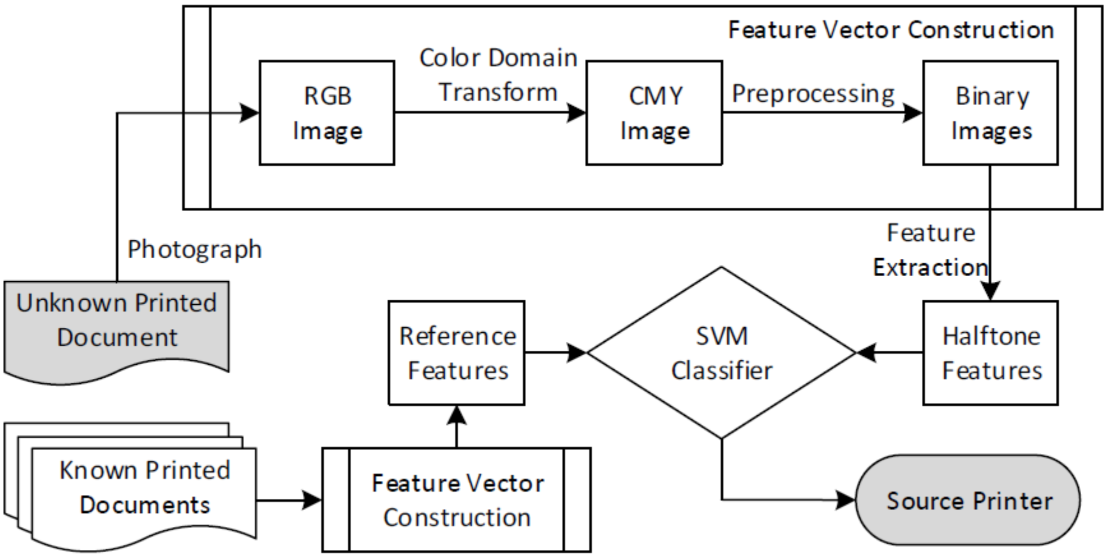


Figure 2.3: Proposed architecture of identification of color laser printer [25]

A summary of printer identification techniques from colored documents is presented in Table 2.1.

Table 2.1: Overview of Printer Identification Techniques on Colored Documents

Type	Study	Techniques	Dataset	Year	Accuracy
Geometric Distortions	Bulan et al. [3]	Geometric distortion signature, Correlation of images and printer signatures	Corpus of EP printers	2009	-
	Shang et al. [57]	Half-tone dots arrangements, Euclidean distance, k-means	2 Brands (HP, Canon), 5 Models	2015	87.92%
Noise Analysis	Choi et al. [4]	Noise features, DWT, SVM	9 models of 4 brands, Xerox, Konica, HP, Canon (99 images)	2009	97.89%, 97.3%, 92.28%, and 80.24%
	Lee et at. [28]	Wiener filters, Noise features, GLCM features	9 models of 4 brands, Xerox, Konica, HP, Canon (99 images)	2010	98.9%, 98.9% 98.4%, and 96.5%
	Tsai et al.[49]	DWT with SVM (45 features)	10 models of colored laser printers from 6 brands	2011	92.4%
	Van et al. [55]	Machine learn dot(CSP Dots)	10 models of colored laser printers from 6 brands	2013	91.3%, 93%
Transformed Texture Images	Ryu et al. [39]	CMYK histograms	9 printers of 4 brands ('HP', 'Canon', 'Konica', 'Xerox')	2010	High errors on printers from same brand
	Kim et al.[25]	Half-tone features with SVM		2014	
	Cruz et al. [7]	Uniform Local Binary Patterns (LBP), descriptor for contextual information and SVM classifier	7 laser printer	2017	suitable accuracy achieved

2.2 Textual Documents

Most of the existing work on identification of printers from textual documents relies on noise or geometric-distortion or textual features extracted from the printed text. Notable techniques are discussed in the following while an overview of the presented techniques is summarized in Table 2.2.

- **Geometric-distortion and noise Analysis:** Gupta et al. [21] explored the distortions produced during printing process to identify the source and reported high accuracy (99%) on a small dataset of 13 printers. Ferreira et al. [12] presented the first solution based on deep learning to identify source printer. Authors used characters to extract the features using multiple representations of characters. Character are extracted using template matching technique. A separate CNN was trained on each representation of characters to extract the three feature vectors (raw-image, median-residual and average-residual filtered image) which were concatenated together. Scores of different characters were combined using majority voting to arrive at final decision. Authors demonstrated the superiority of machine learned features over traditional textural features in characterizing the printer in a text-dependent mode. Experiments are performed on different characters and discuss the strength of characters. Highest accuracy achieved with character 'e' that is 98.3%.

  Hao et al. [18] proposed features based on geometric distortions to characterize a printer. Some pre-processing methods are used to extract the distortion and noise. The images are saved in gray scale after scanning the document page and apply thresholding technique to get a binarize image. Text lines in printed-document-page are not parallel because of distortion. Hence, authors extract features including geometric distortion measure of PTLS ('Page Text Line Slope') and PTLI ('Sequence and Page Text Line Interval') that describe the distortion horizontally and vertically. Authors proposed matching-euclidean distance for measuring the similarity of features which have different length. Reference document is not needed for this technique and performs well when document has partial text. Authors reported high accuracies in terms of printer identification rates on 10 printers from 8 models of 3 brands.

  Shafait & Elkasrawi [9] investigated forgery detection using the analysis of low-high resolution of scanned documents. Author used text-independent text with extracting the noise from text-line and classify the source printer. They extracted features from text lines so, text-line were segmented with the help of Tesseract. The clean images are extracted by applying the Otsu thresholding method, the thresholding is used as a mask to binarized the image. They extracted noise-image by subtracted the clean-image and original-image. Authors exploited the distinct noise-image produced by different printers to characterize the printer model. Different features are extracted from noise image such as mean, standard deviation, skewness, total 15 features are used to trained SVM and experiments carried out on a database of 20 printers reported an accuracy of 76.7%.

Shang et al. [40] used scanned document for identifying the source printer. Characters of document are segmented by using the thresholding and extracted the information of text-region, background-region, and edge-region separately. Text and edge-region information is used to extract the noise-energy, contour roughness, and average-gradient from scanned-image. 'Average-gradient' is used to distinguish inkjet and laser printers. This method also detects tampered-documents produced by a mixture of sources. Authors used SVM for classification of characters. They used majority voting technique to classify the scanned-image based on the individual character and achieved 90% accuracy. Experimental dataset contains ten-laser-printers, six-inkjet-printers, and nine-electrostatic-copiers, each printer contain 10 pages consisting of English words and letters. The documents were scanned with most commonly used range of resolutions ranging from '300 to 1,200 dpi'.

Gebhardt et al. [16] presented a system that used unsupervised anomaly detection to detect documents and the difference in edge-roughness technique to distinguish laser-printed-pages from inkjet-printed-pages. Authors extracted connected components in pre-processing step using image binarization with the Otsu thresholding method. Edge detection and value extraction process repeated for each extracted object. For further processing, OCR is used to extract characters from document images. No prior training is used for this method and evaluation is carried out using outlier rank score on a dataset (7 inkjet and 13 laser printer and document domain invoices, scientific papers, contracts) that contains 1200 documents. They achieved best outlier rank score with comparison to the 'state of the art' method.

Bertrand et al. [2] present forgery detection method at character level using intrinsic method that is based on shape of character and irregularity in document. They prefer two techniques copy-paste and imitation forgery to create a fraudster document. They describe the technique for forged character based on character outlier, similarity and dissimilarity. For Detecting the forged character, firstly they extracted character using OCR and detected the imitate or copy-paste character using shape comparison and distance measure. Different imprecision occurs at character level in case of imitation or copy-paste forgery. Firstly, they detected similar character and find out the forged ones. For detection of forged character, they used imprecision in documents such as orientation, size, alignment of lines. Authors combine these measures as vector for each character and apply mahalanobis distance for calculating dissimilarity. Threshold is applied on distance results to classify the character as a genuine or forged. Author used comparative analysis method and reported 77% recall and 82% precision on custom dataset that contains synthetic images of fraudulent documents.

- **Textural Analysis:** Joshi and khanna [22] employed texture-based features with a single classifier for printer identification on printed letters. Characters were extracted using connected-component labelling followed by morphological-operations and it substituted the need of an OCR. Each printed letter is subdivided into two regions i.e. flat region and edge region. Before extracting features each image is preprocessed that involves cropping of image margins for removing the dominating effect of noise presents at the border of the page. Textural features were extracted from characters based on local ternary patterns, combination with Gabor filter using gabor filter bank having 3 scales and 2 orientations of 0°and 90°. The filter size was fixed at 10x10 before training and the extracted feature vectors are pooled into a one average vector and resulted vector is used for training the SVM classifier. Technique was evaluated on two datasets, one is publicly available English documents having documents from 10 printers in single font and second is German language dataset with 400 dpi pre-processed scanned documents. Authors prepared a dataset that contains 720 documents from 18 printers at two scanned resolutions i.e. 600dpi and 300dpi. A comparison analysis is performed with the 'state-of-the-art' methods. The highest mean accuracy reported 97.68% using 5x2 validation set. They also reported results on printed pages that were printed with Arial, Cambria, Times New Roman, Comic Sans fonts at two different scanning resolutions.

In another work, Tsai et al. [47] used scanned document with setting of 8-bit and scanning resolution of 300 dpi. Authors used one specific character ('シ') for feature extraction because this character contains moderate information. They captured the textural information in a document using a set of features based on DWT, GLCMs, Wiener-filters and Gabor-filters. They collected different features of size 34 to 209 and applied feature-selection technique to reduce the dimensionality of feature-vector. Classification was carried out using SVM. Different combinations of features were investigated to identify printer source from a set of printed Japanese characters.

In a later study [54], they used documents which contain text and images that are scanned with ‘8 bits/pixel and 300 dpi’ resolution. Authors implemented statistical features such as ‘Spatial filters, Wiener filter, Gabor filter, Haralick, fractal filters, Gray Level Co-occurrence Matrix (GLCM), and Discrete Wavelet Transform (DWT)’ for printer identification. They extracted the effective features by applying five feature selection methods and achieved average accuracy of 97.68% with SVM. Experiments are performed on 10 sets of images randomly generated by 12 printers in which 500 and 300 images are used for training and testing respectively. They used different dimension of feature for experiments and selected 165 as effective feature dimension after analysis.

In another study, Tsai et al. [52] used texture feature, spatial features and fractal features for identifying the printer based on microscopic images. Different characters from English ‘e’,Urdu ع , German ‘シ’, and Chinese‘永’ are used to analyze the identification rate of source printer. Authors explored both textual and image analysis techniques and also used different microscopes such as ‘Olympus CX 41’, ‘BX 51 M’ and USB microscope. They achieved average accuracy of 95.29% on BX 51 M microscope and overall 99.89% accuracy was achieved on character ‘e’ and ‘永’. Likewise, Ferreira et al. [13] proposes a new solution that is not based on simple textural measures rather they investigated the ‘multi-directional glcm’ (GLCM_MD) and ‘multi-directional and multi-scale’ matrices using Glcm (GLCM_MDMS). Their proposed methodology is based on three possible solutions that aimed in identifying printer source and explores intrinsic signatures. Firstly, authors used two descriptors that are implemented on text letters. Secondly, convolution Texture Gradient Filter (CTGF) is used to filter parts(inner and outer) of letter and figures that are being printed, also extracts the low-gradient feature that are created at the time of making to create visual effects not identified by human-eye. Afterwards, they recognize the document source from where it is printed and identified its availability, unavailability and other problems even if any part of document is presented. If the document is available fusion strategies are implemented. The proposed solution is implemented on Wikipedia dataset having 1184 images in TIFF format and contains different sizes of letters, fonts and images. They used SVM classifier and reported an overall accuracy of 99.4%.

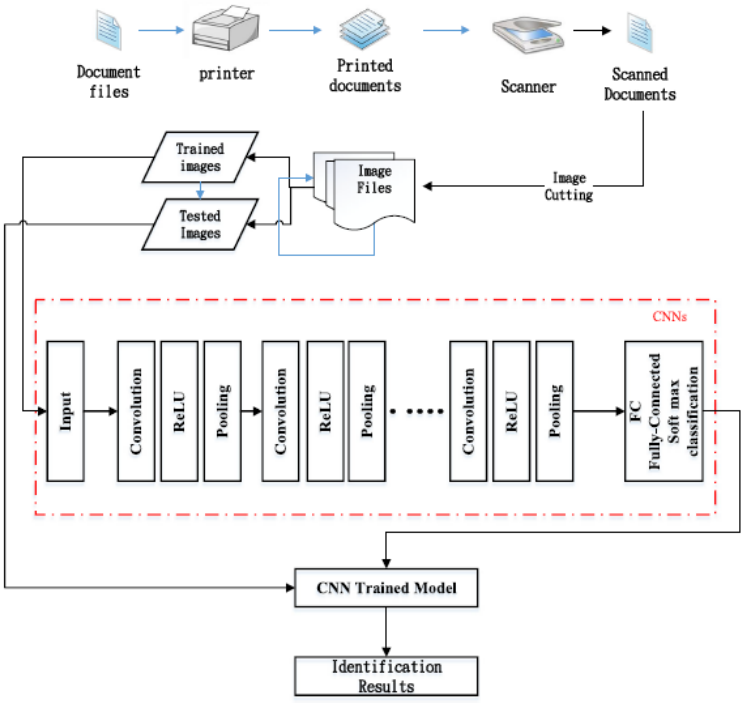


Figure 2.4: Convolution Neural Network Architecture for printer identification based on text and image [53]

In 2018, Tsai et al. [53] proposed a deep learning-based solution and compare results with hand-crafted features. They used hand-crafted feature as discussed in [54] with additionally SFTA feature and after that used decision-fusion model to extract best feature for classification. They used 4-layer architecture of CNN for printer identification. The proposed deep learning architecture is illustrated in Figure 2.4. Authors applied these approaches on scanned images and compare text document and images results. They used two types of architecture, using 7 Layers of CNN, they achieved 98.41% on text and 99.93% on natural-image using scanned documents and using 13-layer architecture, they achieved 97.37% on text and 97.7% on image using microscopic images.

The above discussion is summarized in Table 2.2.

Table 2.2: Overview of Printer Identification Techniques on Text Documents

Type	Study	Techniques	Char/Text level	Dataset	Year	Accuracy
Distortions & Noise Analysis	Jain et al. [21]	Distortions in printed characters with SVM classifier	Text-line	25-page dataset	2017	99%
	Ferreira et al. [12]	Multiple representation of characters, CNN, Early fusion and late fusion	Char-level	10-printer dataset	2017	-

	Hao et al. [18]	Page Text Line Slope, Sequence and Page Text Line Interval, Page geometric distortions in horizontal and vertical directions	Text-line	30-page dataset	2015	94%
	Elkasrawi et al. [9]	Noise produced by printers with SVM	Text-line	20-printer dataset	2013	76.7% for hole doc, 93.57% on inkjet-printer
	Gebhardt et al. [16]	unsupervised anomaly detection using edge roughness technique, Connected component and OCR for character extraction	Char-level	20-printer dataset, 7 inkjets and 13 lasers	2014	Best outlier rank score achieved
	Shang et al. [40]	AWGN energy, Impulsive noise energy, Contour roughness, and Average gradient with SVM classifier	Char-level	Laser ink jet copier printer	2014	90%
	Bertrand et al [2]	Mahalanobis distance for calculating dissimilarity, intrinsic method that is based on shape of character and irregularity	Char-level	Synthetic images Dataset	2013	77% recall and 82% precision
Texture Analysis	Joshi et al. [22]	Variants of LBPs with Gabor-filters	Char-level	Public dataset & 18-printer dataset	2017	97% and 99%
	Tsai et al. [53]	Convolution neural network, Spatial-filters, Wiener-filter, Gabor-filter, Haralick, fractal-filters, Gray-Level-Co-occurrence Matrix (GLCM), and Discrete-Wavelet-Transform (DWT) and SVM classifier	Char-level	12-printer Dataset	2018	text 98.72%, image 99.95% using SVM and text 97.7%, image 99.95% using CNN (10 layer)
	Tsai et al. [51]	LBP, GLCM, DWT, Wiener filters, Gabor filters, Harlick and SFTA features with SVM	Char-level	Multiple datasets	2017	99%
	Tsai et al. [47]	DWT, GLCM, Wiener filters and Gabor filters with SVM	Char level	Japanese dataset	2015	94%
	Ferreira et al. [13]	Low-level gradient textures, multi-scale and multi directional texture features, GLCM; SVM with majority voting	Char-level	Public-dataset	2015	98% on fragments, 97% on characters and 92% on documents
	Tsai et al. [54]	Spatial-filters, Wiener-filter, Gabor-filter, Haralick, fractal-filters, Gray-Level-Co-occurrence-Matrix (GLCM), and Discrete-Wavelet-Transform (DWT) and SVM classifier	Char-level	12-printer Dataset	2017	97.68% on text, 99.67% on images

	Tsai et al. [52]	SFTA feature, Spatial-filters, Wiener-filter, Gabor-filter, Haralick, fractal-filters, Gray-Level-Co-occurrence-Matrix (GLCM), and Discrete-Wavelet-Transform (DWT) and SVM classifier	Char-level	12-printer Dataset	2018	99.89% accuracy on character 'e' and '永'
	Tsai et al. [50]	CNN with different layers, SFTA feature, Spatial-filters, Wiener-filter, Gabor-filter, Haralick, fractal-filters, Gray-Level-Co-occurrence-Matrix (GLCM), and Discrete-Wavelet-Transform (DWT) and SVM classifier	Char-level	12-printer Dataset	2019	Text 99.96%, Image 99.98% using hand-crafted and Text 99.93% Image 97.7% using CNN

2.3 Summary

It can be noted from literature that there are several text-dependent solutions, some authors used one or two characters, and some tried combinations of frequent characters (i.e. character 'a' , 'e') in [47, 12, 18]). In literature, researcher working on characters and text-line extraction, from the analysis of these work, we conclude that characters perform well on printer identification problem with comparison of text-line because in this problem we required fine information of printing that are best extracted from character level in text-dependent mode. Basically, text-dependent approaches contain same textual content in training and testing while text-independent approaches contain different textual content in training and testing. A limited literature is available on text-independent approach and the authors reported state-of-the-art results.

Chapter 3

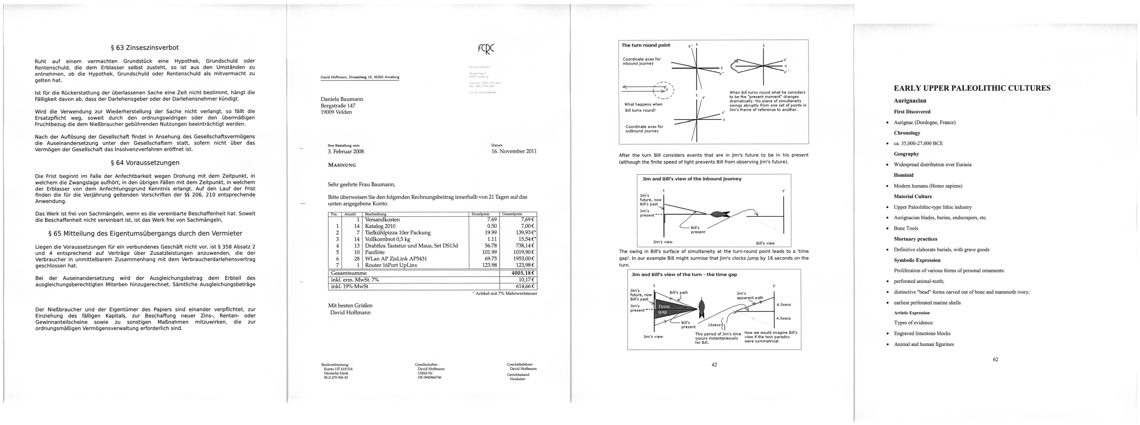
Methodology

In this chapter, we discussed about our proposed approach in which we discuss text-dependent and independent approaches that we used. We have chosen to use a hand-crafted as well as machine-learned features for source printer identification problem. A set of contextual information is extracted and is used to train the classification model. The classification module relies on characterizing the printers using convolution neural networks and texture features. A number of pre-trained models are used in our study as feature extractors.

Since we target on classification and not on the detection of textual content. Our proposed methodology is based on three levels i.e page-level, patch-level and character-level. At patch-level, we constructed patches of fixed size from full page-image (Figure 3.2) and at character level, we extract characters using OCR (Figure 3.3). The extracted patches and characters are then used to extract features which are used for system training and evaluation. An overview of the steps involved is presented in Figure 3.11 and 3.10. In the following, we first present the details of the datasets followed by the data preparation, feature extraction and classification.

3.1 Dataset

For evaluation of printer identification techniques, a number of datasets have been developed an overview of these datasets is discussed in Table 3.1. Few of these are private while others have been made publicly available. In our experimental study, we have employed the database presented in [9]. The dataset comprises 1200 documents from 20



§ 63 Zinseszinsverbot

Ruht auf einem vermachten Grundstück eine Hypothek, Grundschuld oder Rentenschuld, die dem Erblasser selbst zusteht, so ist aus den Umständen zu entnehmen, ob die Hypothek, Grundschuld oder Rentenschuld als mitvermacht zu gelten hat.

Ist für die Rückerstattung der überlassenen Sache eine Zeit nicht bestimmt, hängt die Fälligkeit davon ab, dass der Darlehensgeber oder der Darlehensnehmer kündigt.

Wird die Verwendung zur Wiederherstellung der Sache nicht verlangt, so fällt die Ersatzpflicht weg, soweit durch den ordnungswidrigen oder den übermäßigen Fruchtbezug die dem Nießbraucher gebührenden Nutzungen beeinträchtigt werden.

Nach der Auflösung der Gesellschaft findet in Ansehung des Gesellschaftsvermögens die Auseinandersetzung unter den Gesellschaftern statt, sofern nicht über das Vermögen der Gesellschaft das Insolvenzverfahren eröffnet ist.

§ 64 Voraussetzungen

Die Frist beginnt im Falle der Anfechtbarkeit wegen Drohung mit dem Zeitpunkt, in welchem die Zwangslage aufhört, in den übrigen Fällen mit dem Zeitpunkt, in welchem der Erblasser von dem Anfechtungsgrund Kenntnis erlangt. Auf den Lauf der Frist finden die für die Verjährung geltenden Vorschriften der §§ 206, 210 entsprechende Anwendung.

Das Werk ist frei von Sachmängeln, wenn es die vereinbarte Beschaffenheit hat. Soweit die Beschaffenheit nicht vereinbart ist, ist das Werk frei von Sachmängeln,

§ 65 Mitteilung des Eigentumsübergangs durch den Vermieter

Liegen die Voraussetzungen für ein verbundenes Geschäft nicht vor, ist § 358 Absatz 2 und 4 entsprechend auf Verträge über Zusatzleistungen anzuwenden, die der Verbraucher in unmittelbarem Zusammenhang mit dem Verbraucherdarlehensvertrag geschlossen hat.

Bei der Auseinandersetzung wird der Ausgleichungsbetrag dem Erbteil des ausgleichungsberechtigten Miterben hinzugerechnet. Sämtliche Ausgleichungsbeträge

Der Nießbraucher und der Eigentümer des Papiers sind einander verpflichtet, zur Einziehung des fälligen Kapitals, zur Beschaffung neuer Zins-, Renten- oder Gewinnanteilscheine sowie zu sonstigen Maßnahmen mitzuwirken, die zur ordnungsmäßigen Vermögensverwaltung erforderlich sind.

FCRC

David Hoffmann

Drosselweg 15
45302 Arneburg

Telephon: [illegible]
Fax: [illegible]

Ust-Id: [illegible]

David Hoffmann, Drosselweg 15, 45302 Arneburg

Daniela Baumann
Bergstraße 147
19009 Velden

Ihre Bestellung vom
3. Februar 2008

Datum
16. November 2011

MAHNUNG

Sehr geehrte Frau Baumann,

Bitte überweisen Sie den folgenden Rechnungsbeitrag innerhalb von 21 Tagen auf das unten angegebene Konto.

Pos.	Anzahl	Beschreibung	Einzelpreis	Gesamtpreis
	1	Versandkosten	7.69	7,69 €
1	14	Katalog 2010	0.50	7,00 €
2	7	Tiefkühlpizza 10er Packung	19.99	139,93 €*
3	14	Vollkornbrot 0,5 kg	1.11	15,54 €*
4	13	Drahtlos Tastatur und Maus, Set DS13d	56.78	738,14 €
5	10	Panflöte	101.99	1019,90 €
6	28	WLan AP ZisLink AP5431	69.75	1953,00 €
7	1	Router 16Port UpLinx	123.98	123,98 €
Gesamtsumme				**4005,18 €**
inkl. erm. MwSt. 7%				10,17 €
inkl. 19% MwSt				614,66 €

* Artikel mit 7% Mehrwertsteuer

Mit besten Grüßen
David Hoffmann

Bankverbindung:
Konto 137 618 016
Deutsche Dank
BLZ 470 906 43

Gesellschafter:
David Hoffmann
UStId-Nr
DE 0945866746

Geschäftsführer:
David Hoffmann
Gerichtsstand:
Neukalen

The turn round point
Coordinate axes for inbound journey
t''
t
x
x''
t
x'
x
x''
What happens when Bill turns round?
t
t'
x'
x
Coordinate axes for outbound journey
When Bill turns round what he considers to be the "present moment" changes dramatically. His plane of simultaneity swings abruptly from one set of points in Jim's frame of reference to another.

After the turn Bill considers events that are in Jim's future to be in his present (although the finite speed of light prevents Bill from observing Jim's future).

Jim and Bill's view of the inbound journey
t
t'
Jim's future, now Bill's past
Jim's present
Bill's present
Jim's view
Bill's view

The swing in Bill's surface of simultaneity at the turn-round point leads to a 'time gap'. In our example Bill might surmise that Jim's clocks jump by 16 seconds on the turn.

Jim and Bill's view of the turn - the time gap
t
t'
Jim's future, now Bill's past
Bill's path
Jim's apparent path
Jim's present
Time gap
4.5secs
4.5secs
Bill's present
16secs
Jim's view
This period of Jim's time occurs instantaneously for Bill.
How we would imagine Bill's view if the twin paradox were symmetrical.

42

EARLY UPPER PALEOLITHIC CULTURES

Aurignacian

First Discovered

- Aurignac (Dordogne, France)

Chronology

- ca. 35,000-27,000 BCE

Geography

- Widespread distribution over Eurasia

Hominid

- Modern humans (Homo sapiens)

Material Culture

- Upper Paleolithic-type lithic industry
- Aurignacian blades, burins, endscrapers, etc.
- Bone Tools

Mortuary practices

- Definitive elaborate burials, with grave goods

Symbolic Expression

Proliferation of various forms of personal ornaments:

- perforated animal-teeth;
- distinctive "bead" forms carved out of bone and mammoth ivory;
- earliest perforated marine shells

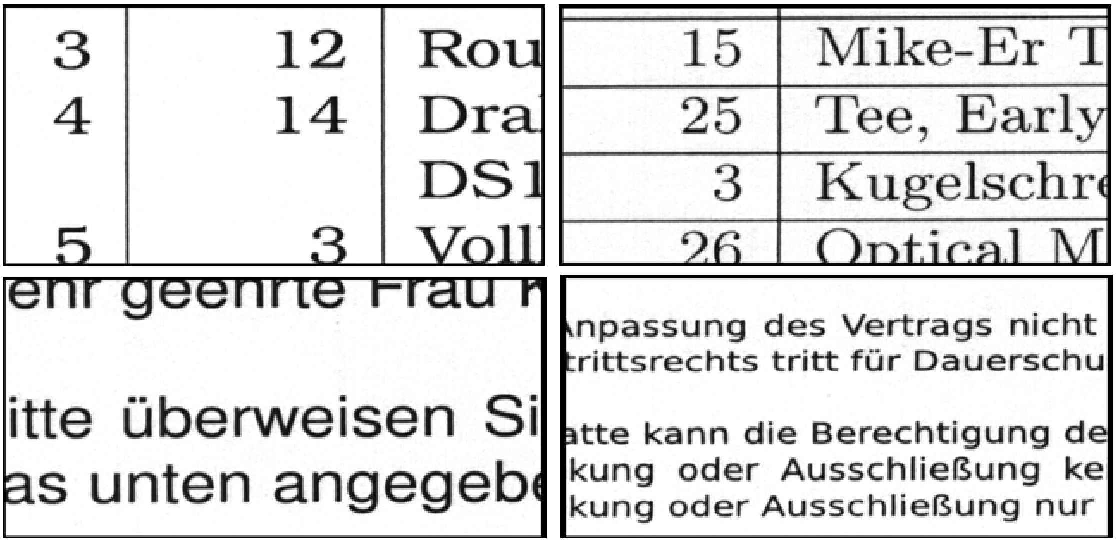
Artistic Expression

Types of evidence:

- Engraved limestone blocks
- Animal and human figurines

62

Figure 3.1: Scanned Document of Original Dataset



3	12	Rou
4	14	Dra
		DS1
5	3	Voll

15	Mike-Er T
25	Tee, Early
3	Kugelschr
26	Optical M

ehr geehrte Frau B

itte überweisen Si
as unten angegebe

npassung des Vertrags nicht
trittsrechts tritt für Dauerschu

atte kann die Berechtigung de
kung oder Ausschließung ke
kung oder Ausschließung nur

Figure 3.2: Patches Datasets Extracted from Original Scanned Documents

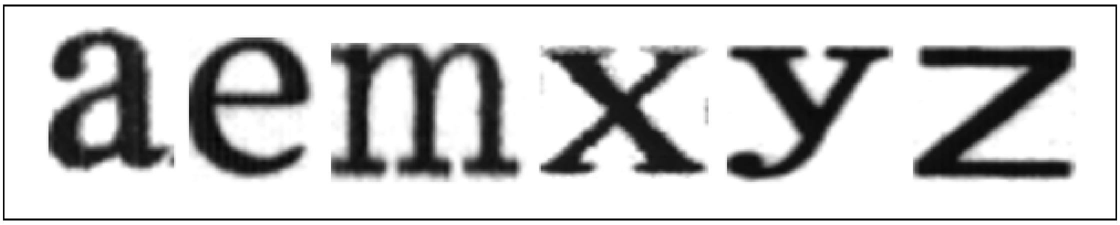


Figure 3.3: Different Character Datasets Extracted from Original Scanned Documents

Table 3.1: An overview of printer identification datasets

Dataset	No. of printers	Language	No. of Docs	Type	Content
Laser inkjet copier dataset [40]	10 lasers printers, 6 inkjets & 9 copiers	English	10 pages per device	Private	Different characters in the pages
Multiple languages [51]	12 lasers printers (Most were HP)	English, Arabic, Japanese & Chinese	500 microscopic images per device for training and 300 for test	Private	Analysis of different letters
Japanese dataset [47]	12 lasers printers (Most printers are HP)	Japanese	500 microscopic images per device for training and 300 for test	Private	Analysis of different Japanese letters
Public dataset [12]	10 lasers printers	English	Around 120 Wikipedia docs per printer	Public	They used letter 'e' , fragments from documents & whole doc
German dataset [9]	Laser & inkjets printer	German	13 laser and 7 inkjet printers with 400 pages	Private	German Business letters

different printers. Out of these, 13 are laser printers and 7 are inkjet printers. The images are scanned in grayscale and contain images, tables and text (text-independent mode). The employed dataset images are shown in Figure 3.1 and detail of each printer in illustrated in Table 3.2.

3.2 Data Preparation

Our Dataset contain full pages that are not directly used for training process, as we employed different pre-trained models and these models have pre-define input image size that why, we construct patches and characters. Features are extracted from patches and characters for training. The following sections present the detail of extraction.

3.2.1 Patches Extraction

Identifying information explicitly is a difficult task in printed documents. To highlight the information more prominently we construct patches of size 300x512 from scanned printed document images [29]. These images are illustrated in Figure 3.4. Some patches contain portions of text, lines, numbers, tables and some contain data at the border

Table 3.2: Dataset Used for Experimental Evaluation

ID	Brand	Model	Documents
01	Ink-jet	Officejet-5610	60
02	Laser	Samsung-CLP-500	60
03	Laser	Ricoh-Aficio-MPC2550	60
04	Laser	HP-LaserJet-4050	59
05	Laser	OKI-C5600	60
06	Laser	HP-LaserJet-2200dtn	60
08	Laser	Ricoh-Afico-Mp6001	60
11	Ink-jet	Epson-Stylus-Dx-7400	59
13	Ink-jet	unknown	59
19	Laser	HP-Color-LaserJet-4650dn	60
20	Laser	Nashuatec-DSC-38-Aficio	60
21	Laser	Canon-LBP7750-cdb	60
22	Ink-jet	Canon-MX850	60
23	Ink-jet	Canon-MP630	60
24	Laser	Canon-iR-C2620	60
26	Ink-jet	Canon-MP64D	60
31	Laser	Hp-Laserjet-4350-o.4250	60
32	Ink-jet	unknown	59
49	Laser	Hp-Laserjet-5	60
50	Laser	Epson-Aculaser-C1100	60

Table 3.3: Multiple combination of character dataset

Dataset	Character	Image	Total Samples	Each printer sample	Size	Dimension
D_{ae}	a and e		166,672	2845	106MB	different dimensions
D_{de}	d and e		188,192	2886	121MB	
D_{eu}	e and u		171,772	2851	107.3MB	
D_{abe}	a, b, and e		229,024	3041	147.2MB	
D_{ade}	a, d, and e		243,598	3150	146.2MB	
D_{aeu}	a, e, and u		211,199	2911	132.5MB	

Table 3.4: Multiple character dataset Used for Experimental Evaluation

Dataset	Character	Image	Total Samples	Each printer sample	Size	Dimension
D_a	char-a	a	39,427	994	25.20MB	All images are in different dimensions
D_b	char-b	b	62,352	2605	41.40MB	
D_c	char-c	c	14,887	827	7.90MB	
D_d	char-d	d	60,947	2985	39.40MB	
D_e	char-e	e	127,245	1851	81.60MB	
D_f	char-f	f	60,032	1056	38.65MB	
D_g	char-g	g	40,910	940	26.60MB	
D_h	char-h	h	60,932	1103	39.30MB	
D_i	char-i	i	18,661	765	13.10MB	
D_j	char-j	j	15,675	750	12.00MB	
D_k	char-k	k	42,614	1971	27.20MB	
D_l	char-l	l	70,900	2506	43.60MB	
D_m	char-m	m	62,645	2270	40.30MB	
D_n	char-n	n	55,845	2089	68.10MB	
D_o	char-o	o	13,909	584	7.11MB	
D_p	char-p	p	14,280	604	8.10MB	
D_q	char-q	q	2,581	50	1.95MB	
D_r	char-r	r	89,401	3065	39.10MB	
D_s	char-s	s	35,879	1506	45.80MB	
D_t	char-t	t	13,270	545	57.90MB	
D_u	char-u	u	44,527	1754	25.70MB	
D_v	char-v	v	20,436	854	15.50MB	
D_w	char-w	w	42,907	1806	31.40MB	
D_x	char-x	x	7,310	225	4.93MB	
D_y	char-y	y	9,499	298	6.31MB	
D_z	char-z	z	10,825	303	6.59MB	
D_{patch}	patches	 adkosten ihlpizza 10er Pack ng 2010 fos Tastatur und 1	28,855	1866	1.23GB	300 X512
D_{full}	full-page		1200	60	2.21GB	3312 X4677

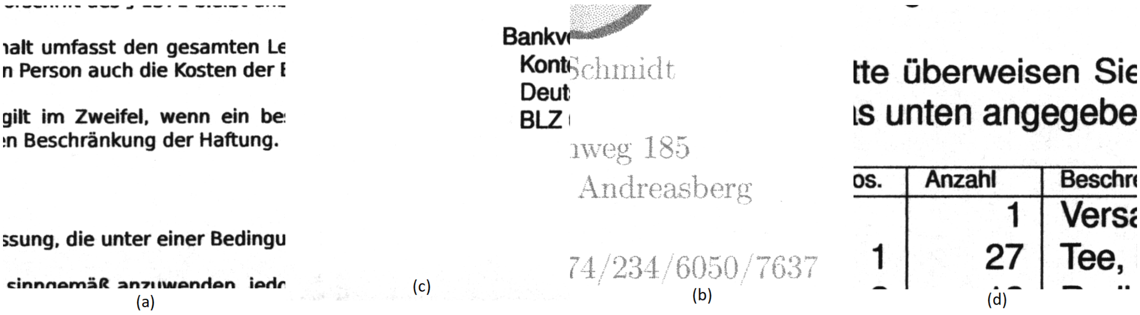

ıalt umfasst den gesamten Le
n Person auch die Kosten der I
gilt im Zweifel, wenn ein be
:n Beschränkung der Haftung.
ssung, die unter einer Bedingu
sinngemäß anzuwenden, jedc
(a)
(c)
Bankv
Kont Schmidt
Deut
BLZ
ıweg 185
Andreasberg
74/234/6050/7637
(b)
tte überweisen Sie
ıs unten angegebe
os.
Anzahl
Beschr
1
Versa
1
27
Tee,
(d)

Figure 3.4: Patches Dataset Extracted from Original Scanned Documents (a) Text (b) Text on boarder (c) Light text and logo (d) Table and text

of patches. These images contain enough information to characterize the document source printer. we analyze these patches by two different techniques hand-crafted feature extraction technique and machine-learned features. we have taken input of full page image and constructed patches of size 300x512 in a way that it divides the page equally in horizontal direction. The illustration is presented in Figure 3.5. The highlighted part shows that the extracted patches are in horizontal direction.

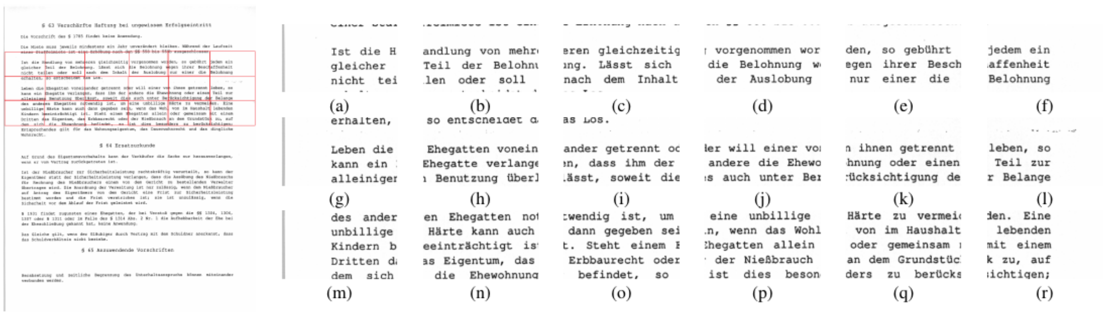

Ist die H
gleicher
nicht tei
(a)
andlung von mehr
Teil der Belohnu
len oder soll
(b)
eren gleichzeitig
ıng. Lässt sich
nach dem Inhalt
(c)
vorgenommen wor
die Belohnung w
der Auslobung
(d)
den, so gebührt
egen ihrer Besch
nur einer die
(e)
jedem ein
affenheit
Belohnung
(f)
Leben die
kann ein
alleiniger
(g)
Ehegatten vonein
Ehegatte verlange
Benutzung überl
(h)
ander getrennt oc
ın, dass ihm der
Lässt, soweit die
(i)
ler will einer vo
andere die Ehewo
s auch unter Be
(j)
ı ihnen getrennt
hnung oder einen
ücksichtigung de
(k)
leben, so
Teil zur
r Belange
(l)
des ander
unbillige
Kindern b
Dritten d
dem sich
(m)
en Ehegatten nol
Härte kann auch
eeinträchtigt is
as Eigentum, das
die Ehewohnung
(n)
wendig ist, um
dann gegeben sei
t. Steht einem E
Erbbaurecht oder
befindet, so
(o)
eine unbillige
n, wenn das Wohl
hegatten allein
der Nießbrauch
ist dies beson
(p)
Härte zu vermei
von im Haushalt
oder gemeinsam
an dem Grundstüc
ders zu berücks
(q)
ien. Eine
lebenden
mit einem
k zu, auf
ichtigen;
(r)

Figure 3.5: Patches extracted from full page image

3.2.2 Character Extraction

In addition to printed text, the documents that we consider, contain plenty of equations, graphs, drawings, logos and tables etc. Choosing an appropriate input data is very important to solve a printer identification problem. Selected input should contain enough information to characterize the different printers. Since, we characterize the document by printed text and not by other objects (e.g. graphics).

By motivated from state-of-the-art methods in document analysis, using characters analysis yield into promising results, that's why we also decided to validate our proposed approach through character analysis [13]. The printed document allows an automatic

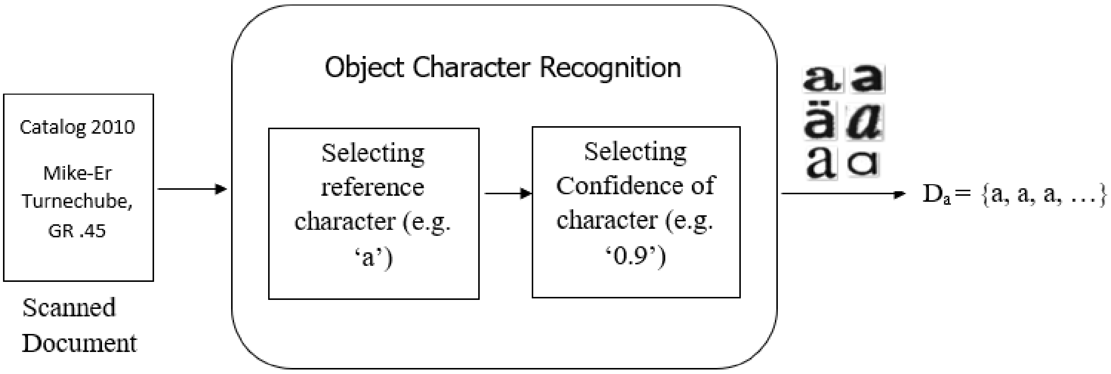

Object Character Recognition
Catalog 2010
Mike-Er
Turnechube,
GR .45
Scanned
Document
Selecting reference character (e.g. 'a')
Selecting Confidence of character (e.g. '0.9')
D_a= {a, a, a, ...}

Figure 3.6: Character-extraction pipeline. Extract character of scanned document by using OCR. The set D is containing character 'e'.

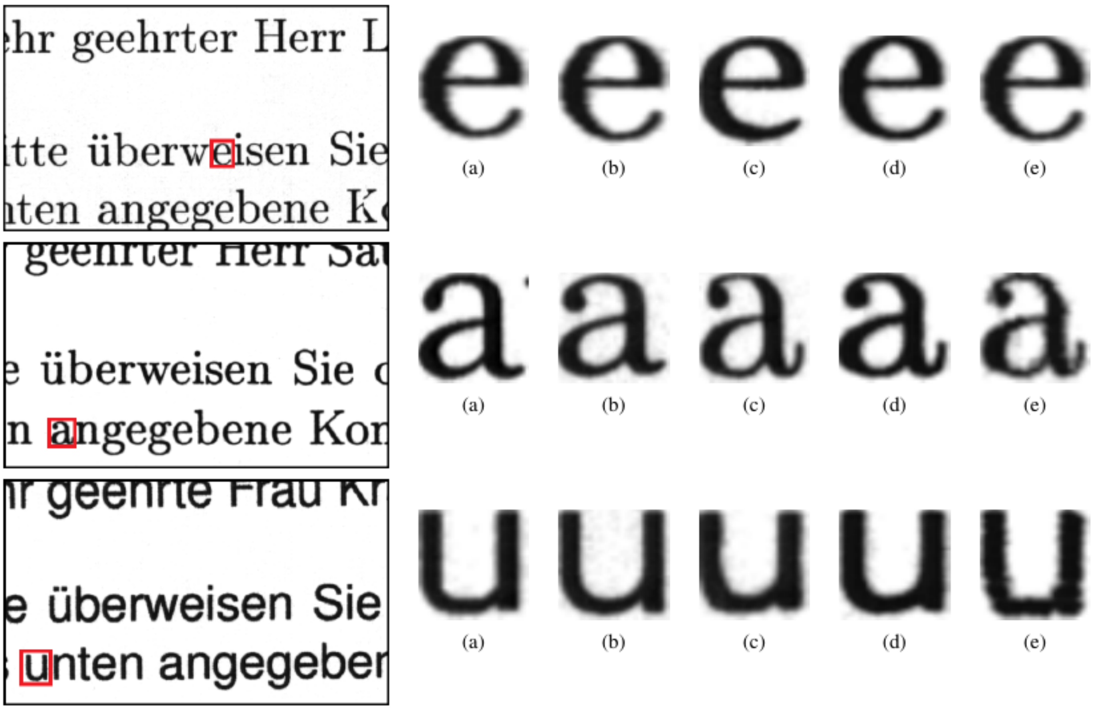


Figure 3.7: 'Image patch samples in 40 × 40 pixel size. The printer brand and models are (a) Ricoh Aficio MPC2550, (b) HP LaserJet 4050, (c) HP LaserJet 2200dtn, (d) Ricoh Afico Mp6001, (e) Epson Stylus Dx 7400'

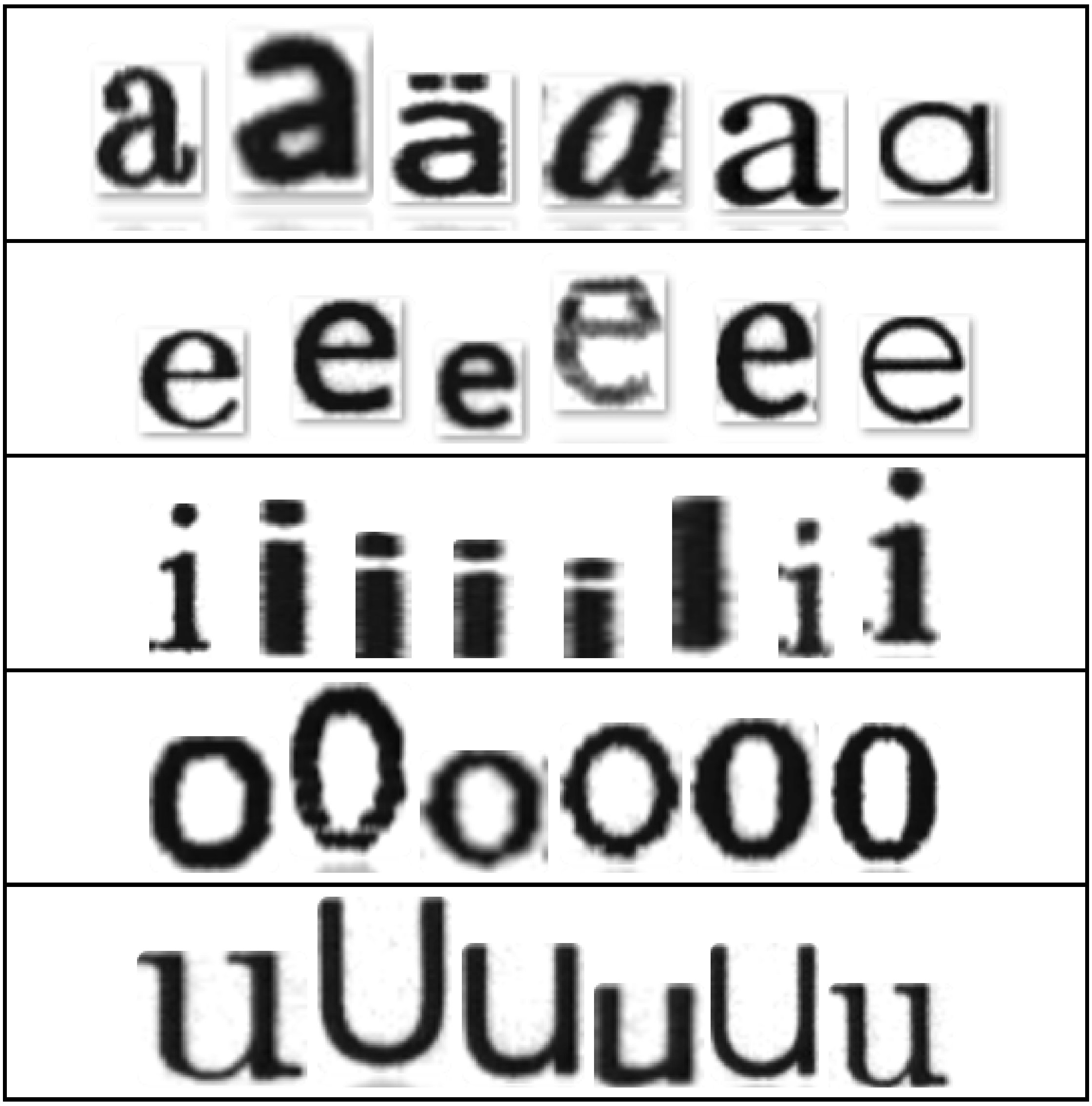


Figure 3.8: Multiple Input character representation with different font and sizes

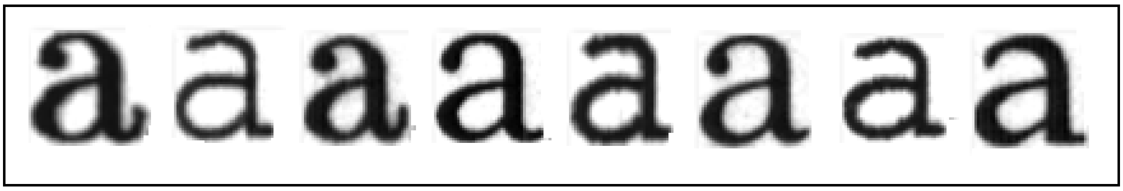


Figure 3.9: Character ’a’ printed by different printers

segmentation of text (e.g. characters) areas. We therefore carry out a segmentation of text by cropping a part of image containing characters using Object Character Recognition (OCR)[43]. OCR is the main extractor for this, we choose specific character and as the result of this process, we get characters in different font and different sizes. Character-extraction process is shown in Figure 3.6. OCR takes reference character and confidence value as input and extract characters accordingly. Most of extracted characters are same as reference letter characteristics but vary in different styles and sizes that are illustrated in Figure 3.8. Segmented characters printed by different printers as discussed in Figure 3.7.

To validate the performance of our proposed method, we also extracted all characters(a-z) and create separate dataset for each of them. The details of each character dataset is summarized in Table 3.4. To evaluate the text-dependent and text-independent approach we combine different characters.The different combinations of character datasets are D_{ae}, D_{de}, D_{eu}, D_{abe}, D_{ade} and D_{aeu}. The details of these combination datasets are reported in Table 3.3

In our study, two approaches have been employed for document forgery detection using source printer identification i.e. text-dependent and independent approach. The detailed analysis is discussed in next sections.

3.3 Text-dependent Approach

In Text-dependent mode, the entities which are compared need to have same textual content for e.g. features extracted from particular character like ’a’ are compared with the feature extracted from the same character. Text-dependent approach is illustrated in Figure 3.10. Features extraction is done at character-level in text-dependent approach, it contains same textual content in training and testing (e.g. character ’e’ is present in both training and testing).

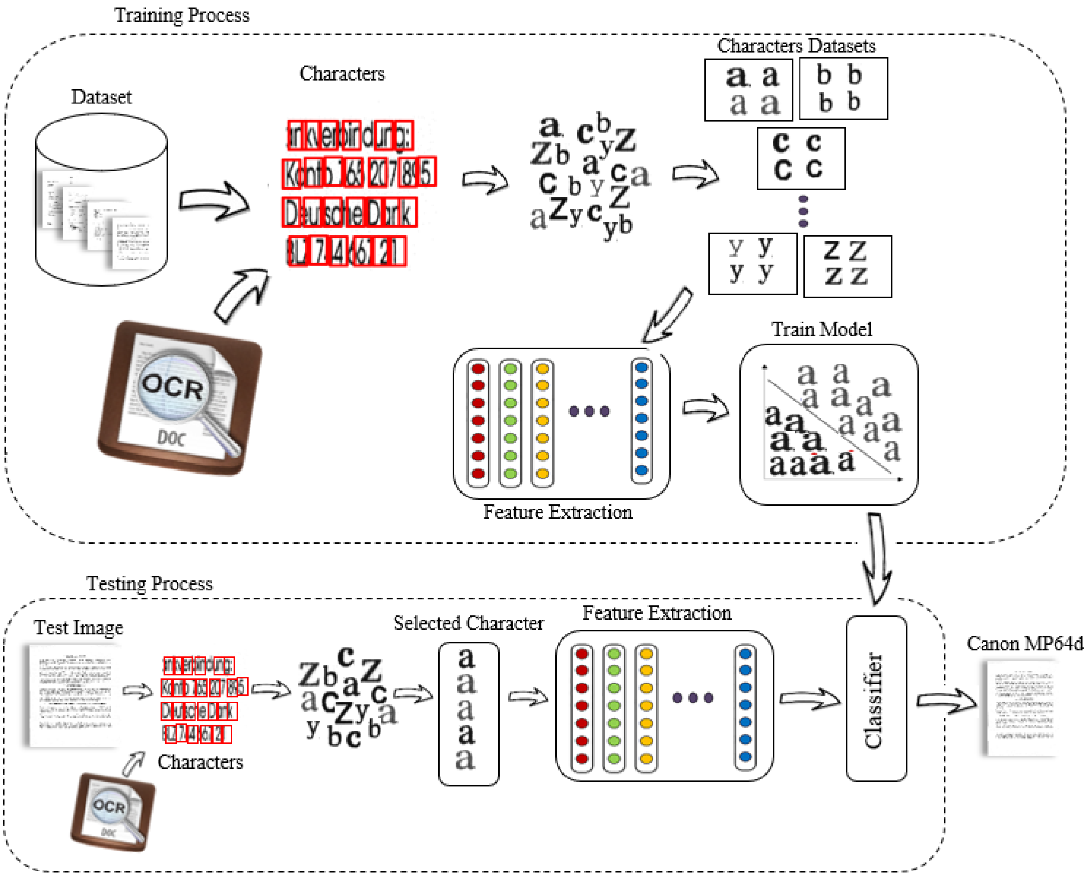


Figure 3.10: Flow of Identification of Source Printer Using Deep-learning Feature

3.4 Text-Independent Approach

For text-dependent approach the same textual content is necessary in training and testing and this is a hard constraint which is difficult to meet in practical situation Therefore, from the view point in real-world applications text-independent approaches are more practical where the two images too be compared can contains different textual content. Text-independent approach is illustrated in Figure 3.11. It presents that patch level is text-independent approach because the training and testing contains different textual content and features are extracted through both hand-crafted and machine-learned feature techniques. The classification is based on these feature extraction techniques and predict the source printer label after applying majority voting.

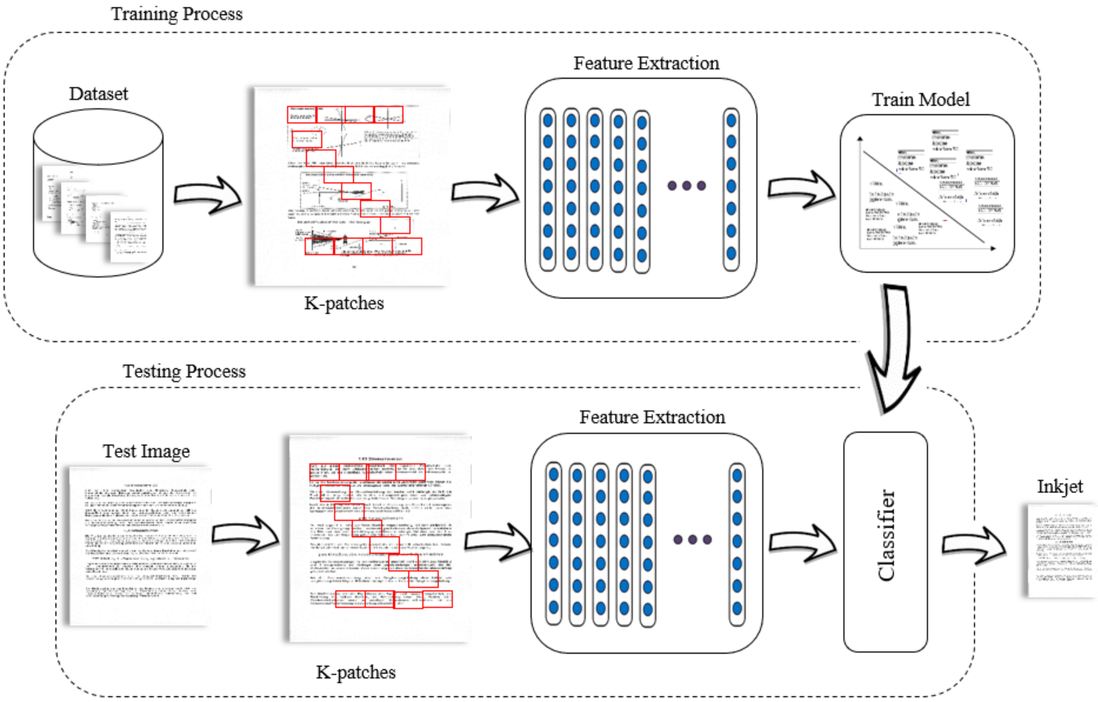


Figure 3.11: Flow of Identification of Source Printer Using Texture Feature

3.5 Feature Extraction and Classification

In this study, we performed classification by using hand-crafted feature techniques as well as machine learned feature techniques. For this the framework involves extracting features from each individual (character/patch/page) and train a classifier to classify printer of the scanned document. Sec 3.5.1 and Sec 3.5.2 summarizes some of the hand-crafted techniques and convolution neural network known as CNN (machine learned technique) as well as the layers which we used in our experiments. In our proposed approach, we have used pre-trained models of deep learning feeding them the training images to extract n-dimensional feature vectors and the same process is then repeated for testing images. More specifically, we used pre-trained network as a feature extractor. The network itself learns to discriminate the printers by learning the appropriate characteristics from the input images. We trained the network, by using stochastic gradient descent having 0.001 as a learning rate and a fixed batch size of 32 images. In the following, we first present the details of the hand-crafted techniques followed by the machine-learned techniques.

3.5.1 Feature Extraction Using Handcrafted Techniques

It is obvious from literature survey, that mostly authors employed texture- based techniques for the detection of forgery using source printer identification [48, 33]. In a number of studies [30, 14, 24] Local Binary Patterns (LBPs) [20] and its variants have been employed to extract features globally for characterizing the source printer. In this study, we propose to investigate some of the popular texture features. Dimensionality of texture features are illustrated in Table 3.5. We briefly explain these techniques in the next paragraph.

Table 3.5: Summary of Features Employed

Feature	Description	Dimensionality
f1	GLCM Features	16
f2	HOG Features	216/756
f3	U-LBP Features	59
f4	LBP Features	256

- **Local Binary Pattern (uniform and non uniform)**

  Local Binary Pattern (LBPs) is considered as the most employed texture feature. It extracts various texture representation from the given input image. The intensity value of each pixel is computed by comparing it with the intensity values of its neighborhood. Fixing the neighborhood size, two values (0 and 1) are assigned to each pixel. The intensities smaller to reference pixel assigned “0” value otherwise it is assigned “1” value. A resultant binary string is generated followed by an LBP code for the referenced pixel. The normalized histogram is computed using these codes and then used as a texture descriptor.

  Uniform LBP [36, 37] is the common variant of LBP. Uniform LBP distinguished between all the uniform and non-uniform patterns. The uniforms patterns are

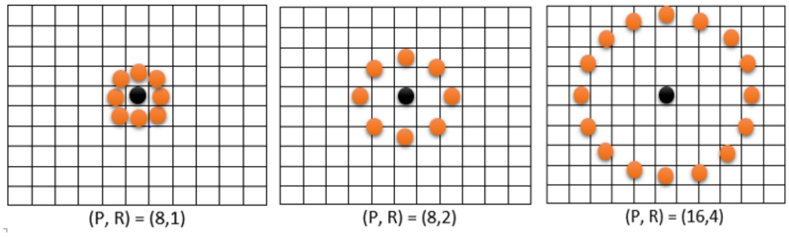


Figure 3.12: Different LBPs Operator ('P' is neighboring pixel and 'R' is the radius)

computed separately, a fixed threshold is used and the transitions between 0 and 1 which are less than this threshold termed as uniform patterns and thus, other patterns termed as non-uniform patterns. In our study, the feature vector dimensionality is 256 by computing the histogram of gray scale image for LBP. Inspired from [31], for uniform LBP, the size of feature vector is 59 that is computed, using cell size 3x3 with pixel value is equal to 8 and radius is set to value 1.

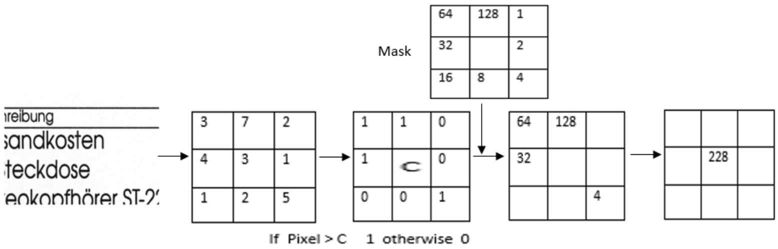


Figure 3.13: Example of Local Binary Pattern

- **Gray-level Co-occurrence Matrix**

  Gray-level Co-occurrence Matrices (GLCMs) is used for a statistical textural analysis. GLCMs has been employed in numerous studies. Given an input image, glcm computes the co-occurrence frequency of neighboring pixels in four directions (i.e. 0°, 45°, 90°and 135°) with a one pixel displacement. The matrix size is dependent on the intensity levels existing in an image i.e. 0 and 1. 'Contrast', 'Entropy', 'Homogeneity' and 'Correlation' are the known statistics computed by glcm. These statistics then employed as a feature later. For each input image a 16-dimensional feature vector is then generated. A summary of GLCM based features implemented in our experiments is provided in a Table 3.6.

Table 3.6: Summary of GLCM based features (‘P’ represents the matrix) [15]

SNo.	Feature	Computational Details
1.	Contrast	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$
2.	Correlation	$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{(\sqrt{(\sigma_i^2)(\sigma_j^2)})} \right]$
3.	Homogenity	$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$
4.	Entropy	$\sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j})$

- **Histogram of Oriented Gradients**

  Classifying source printer based on scanned documents is a difficult task, because documents contain different type of data in the form of tables and graphics. To accomplish this, a feature descriptor named Histogram of oriented gradients (HOG) is used. HOG is also known as a dense feature extraction method. It means it extract features from all the locations (area of interest) existing in an image. HOG captures the object structures from the gradient information presented in an image. It takes both formats of image i.e. RGB and gray scale. Higher magnitude values are selected from each plane separately [8]. In our study, we used 64x64 cell size for hog feature extractor. different cell sizes are illustrated in Figure 3.14.

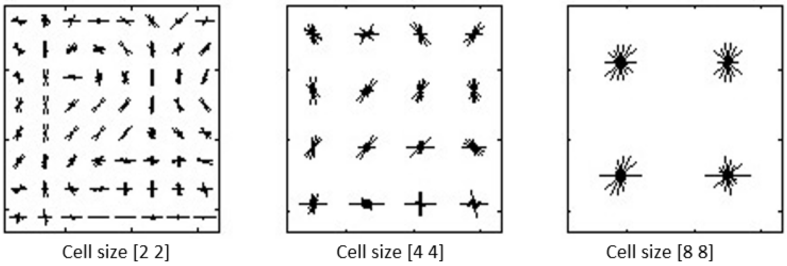


Figure 3.14: Operators based on different cell sizes

We can also discriminate the features of one printer from another using different textural- features. In our study, we have employed GLCM, LBP, Uniform LBP and HOG so we performed a analysis based on these texture-features. The images from same printer learns similar feature values while the image from different printer may have discriminating feature values. The difference between two printers is illustrated in Figure 3.15, 3.17, 3.16 and 3.18. The features learned from two different printers are both visualized at character and patch. This also discriminates the performance at character and patch level. We also plot the feature vector on vowel characters to illustrate the effect of different characters using texture features. The representation is shown in Figure 3.19.

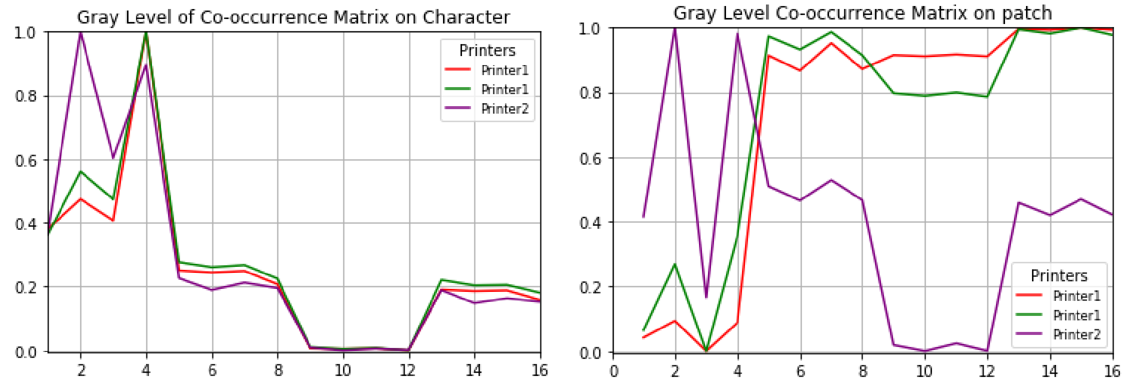


Figure 3.15: Representation of Gray Level Co-occurrence Matrix on character and patch

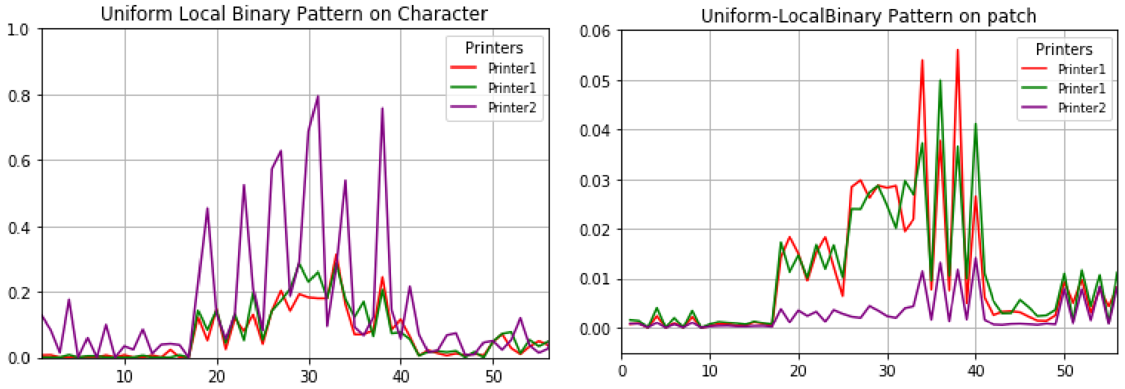


Figure 3.16: Representation of Uniform Local binary pattern feature on character and patch

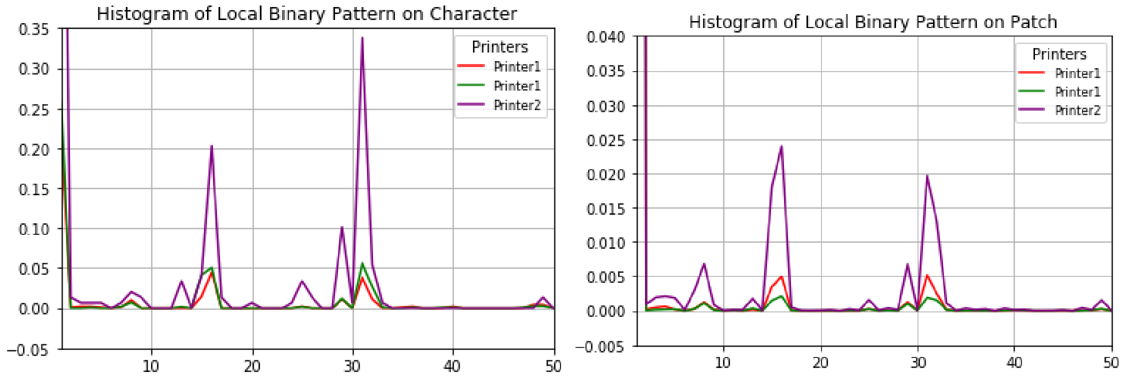


Figure 3.17: Representation of Local binary pattern feature on character and patch

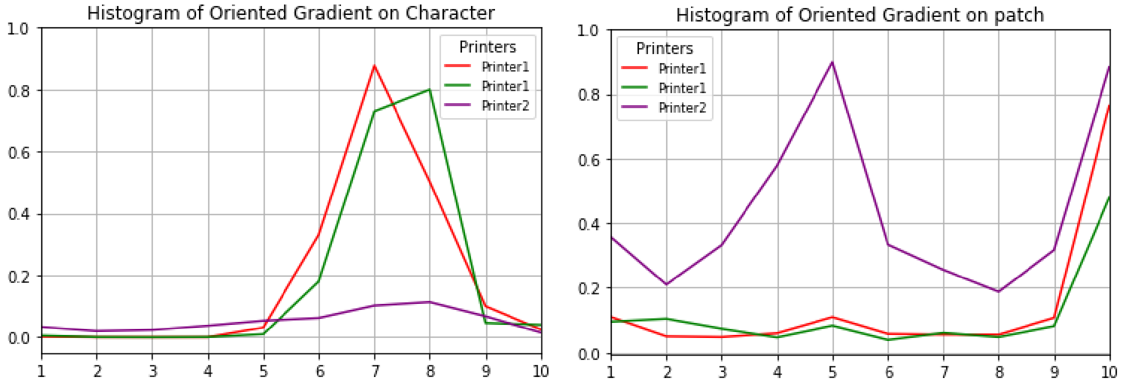


Figure 3.18: Representation of Histogram of Oriented Gradients on character and patch

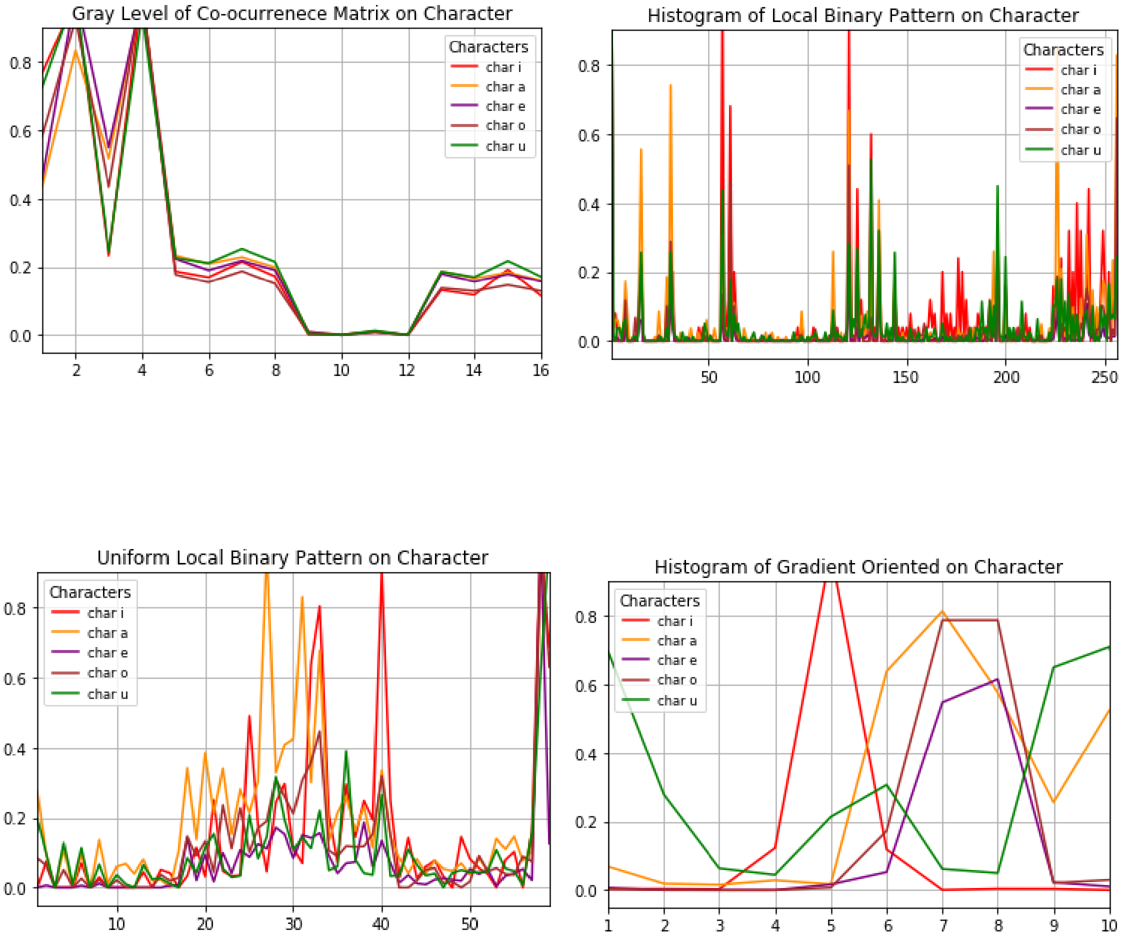


Figure 3.19: Representation of Texture Feature on different characters

3.5.2 Feature Extraction Using Machine-learned

In the recent years, hand-crafted techniques have been replaced by Convolution Neural Networks (CNNs) also known as machine-learned techniques. We also employed CNN to detect forgery using source printer attribution. In the following sections, an overview of CNN is presented followed by the discussion about different models and their architectures.

3.5.3 Convolution Neural Networks

For the first time [27] in 1990, Convolution Neural Networks were presented. The researchers did not pay much attention at that time because CNN needs large dataset as well as high performance computing devices. As time passes, the availability of high-end devices get increased like graphical processing units (GPUs) and large datasets like Imagenet [38]. In most of the classification problems, CNN outperformed the conventional techniques. CNN include different layers, i.e. convolution, pooling, ReLU and fully connected. Detail of the layers are discussed below:

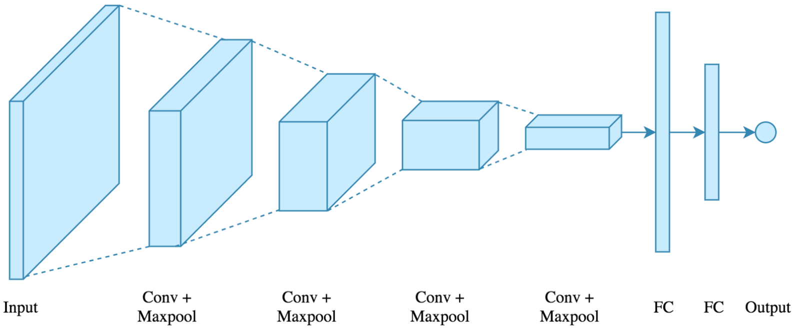


Figure 3.20: Architecture of Convolution Neural Network

- **Convolution Layer**
  Convolutional layers extracts feature from the given input samples. Each convolution layer set the filters for extracting features. As the layer architecture increases the complexity is also increased for e.g. The starting convolution layers learn low level (edges, regular, different orientations and curves) features as it goes to deeper level, the CNN started to learn domain specific features. CNN also have some hyper-parameters. The basic ones are depth, zero padding and stride.

- **Pooling Layer**
  Pooling layer, it is also known as Down Sampling layer. To overcome, the problem of over-fitting this layer is used, and it also reduce number of parameters. It works at each depth of layer. There are three different types of pooling; Max polling, Min

poling and average polling. Mostly used operation on every filter is max polling. The dimensionality of output is also reduced with keeping the important information, where each filter consists of different feature.

- **Relu Layer**
  Rectified Linear Unit (ReLU) layer is used after convolution layer. In deep learning models this the most used activation function. The function uses f(x) = max(0, x) for all input values it returns 0 for negative values and the positive values returns back. As some linearity commonly occurs from conv layer this layer is used to perform some non-linear operations means the images have different borders, colors, intensities (non-linear features). Using this layer, the training becomes fast and without disturbing the accuracy of model computation efficiency is also increased. It also overcomes the issue of gradient problem that effects the network to be trained slowly in the lowest layers because the gradient is decreased to certain level through those layers. Without affecting any previous layer this layer increases non-linearity to the model.

- **Fully Connected Layer**
  The last layer of architecture is Fully connected layer (FC). This layer serves the purpose of classifier. The features we get through from both conv and polling layers are then passed to fc layer these features are work as an input. The working of fully connected layer is like basic simple neural network. The neurons at this layer are fully connected to previous layer neurons. The final layer consists neuron equal to number of classes used to classify the problem

3.5.4 Pre-trained Neural Networks

CNNs have widely been adopted as the most useful feature extractor for images but they have the same issues as all deep learning, i.e. a large amount of labelled training data, time and memory resource for the required heavy computations. An effective method to resolve these issues is to make use of pre-trained models instead of training a network from scratch [17]. The pre-trained model can be modified to be used in the new domain by making use of any one of the following strategies:

- **Fine-tuning:**
  Fine-tuning [35, 41], is done by precisely adjusting the parameters and learning the weights by using back-propagation that improves the performance of pre-trained models on the problem under study. It is also possible to froze early layers of networks as it learns generic features.

- **Feature Extraction:**
  Convolution neural network can be used as a powerful feature extractor. In feature extraction process, fully connected layers of CNN's are replaced by the new classifier and extracted features from pre-trained model are then fed to a classifier for classification.

3.5.5 Different Architecture of Pre-trained Networks

Extraction of features from each class and to train a classifier is the basic framework for a classification problem. In different researches, standard classifiers have been mostly employed for classification task. In recent years, for feature extraction task the manual techniques have been replaced by machine-learned techniques. Convolution layers are made up of small sized kernels. These kernels help to extract high-level features which are then fed to FC layers for classifying data properly. CNN performed training by using stochastic gradient descent and back-propagation techniques. The mis-classification error drives the weight updates of both convolutional and fully-connected layers. The basic layers of a CNN are input-layer, convolution-layer, pooling-layer, rectified linear unit layer, fully-connected-layer, and soft-max-layer. We also employ different CNN architectures in our study that are discussed as follows:

- **Alex-Net**
  AlexNet [26] is considered one of the revolutionary network which revived the deep networks. Alexnet reported the lowest error on image-net dataset in 2012 ILSVRC challenge. The architecture consists of 5 conv and 3 FC layers with 3 pooling layers. The input layer has an image size of 227x227. Pre-trained models are able to learn edges, lines in the early layers of architecture as it goes into deeper layer of architecture it started learning the shapes which contains the information that is used to classify the image properly. We used Alexnet's first and last convolution layer to visualize the features at both patch and character level. The illustration of features on convolution layer is showing in 3.21 and 3.22.

- **GoogleNet**
  'Carvana Image Masking Challenge' was held by ILSVRC in 2014. Googlenet/Inceptionv1 [45] architecture is emerged as a winner of this challenge. The architecture contains 57 convolution and RELU layers with 1 FC layer. 4 million parameters have been learned by this architecture in comparison with Alexnet that learns nearly 60 million parameters.

- **VGG-Net**
  The Visual Geometry Group at Oxford developed VGG-Net [42] nets. They increased the depth of already employed architectures. They released two architectures i.e. VGG-16 and VGG-19 where 16 and 19 refer to the number of (convolutional+ fully connected) layers. It takes RGB image as an input size of 224x224.

- **Resnet101/Resnet50**
  Residual Neural Network also known as ResNet [19] is based on the concept of 'Skip Connections'. The multiple parallel residual modules are presented in this architecture. Every residual module can perform some functions on input or can skip it. In the end, like a GoogleNet, to complete the network all the modules are placed one over another. The architecture contains 177 layers in ResNet50 and 347 layers in Resnet101. The main advantage of ResNet is to add extra residual layers and then trained them accordingly.

- **Inception-v3**
  In 2012 ILSVRC challenge Inception-v3 [46] architecture was presented. The network architecture takes an input image of size 299x299. It has 103 convolution, 94 RELU, and 1 fully connected (FC) layer. In Image-Net classification challenge inception-v3 achieved error of 3.58%. The network is based on the concept of Inception module. Inception module performs convolution on input with different filter sizes and the outputs are concatenated and sent to next module. Number of inception modules are stacked up to achieve better accuracy. It has several versions that involves iterative improvement w.r.t previous version.

- **Inceptionvresnet2**
  In 2015 ILSVRC challenge Inceptionvresnet2 [44] architecture was presented. The network architecture takes an input image of size 299x299. It has 205 convolution, 204 RELU, and 1 fully connected (FC) layer. In ImageNet classification challenge inceptionresnetv2 achieved error of 3.08%. The architecture is a combination of inception network with residual connections.

A summary of different deep learning architectures that are employed in our study are illustrated in Table 3.7.

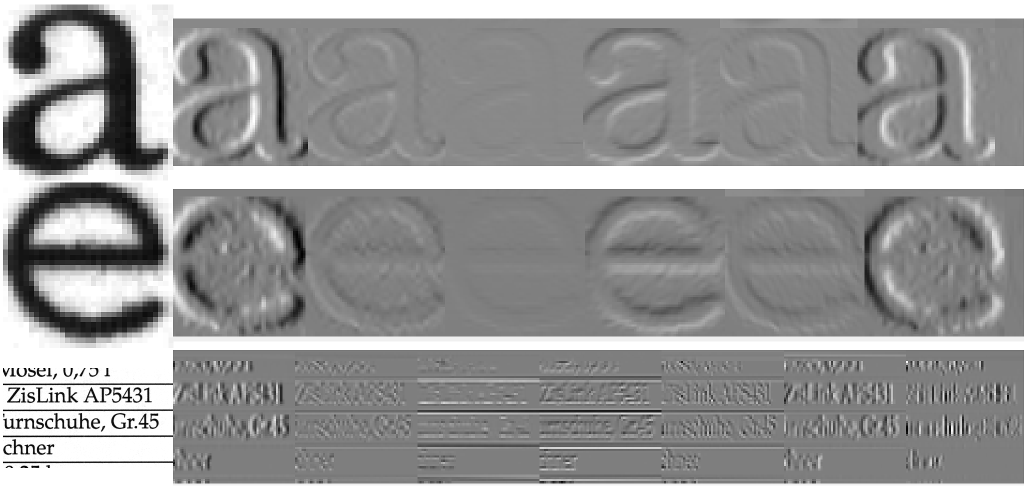


Figure 3.21: Visualization of features from First Convolution Layer of Alexnet at characters ’a’, ’e’ and a patch

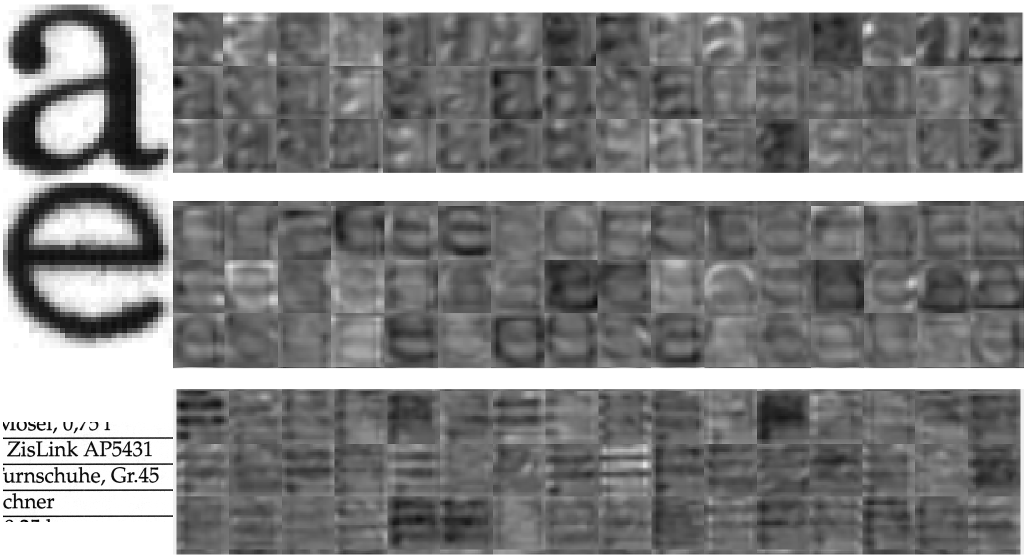


Figure 3.22: Visualization of features from last Convolution layer of Alexnet at characters ’a’, ’e’and a patch

Table 3.7: Machine-learn networks

Networks	Input size	Conv layers	Relu Layer	FC layers	layers
Alex-net [26]	227x227x3	5	7	3	25
VGG-16 [42]	224x224x3	13	15	3	47
VGG-19 [42]	224x224x3	16	18	3	47
Resnet-50 [19]	224x224x3	54	49	1	177
Resnet-101 [19]	224x224x3	104	101	1	347
Google-net [45]	224x224x3	57	57	1	144
Inceptionv3 [46]	299x299x3	103	94	1	316
inceptionresnetv2 [44]	299x299x3	205	204	1	164

3.5.6 Classification

More specifically, we used pre-trained network as a feature extractor that are presented above, we used them as feature extractors using characters and patch as already discussed. The network itself learns to discriminate the printers by learning the appropriate characteristics from the input images. We trained the network, by using stochastic gradient descent having 0.001 as a learning rate and a fixed batch size of 32 images. Classification is carried out using classifier When the models are used as feature extractors and we employed SVM-classifier with Radial Basis Function (RBF) kernel.

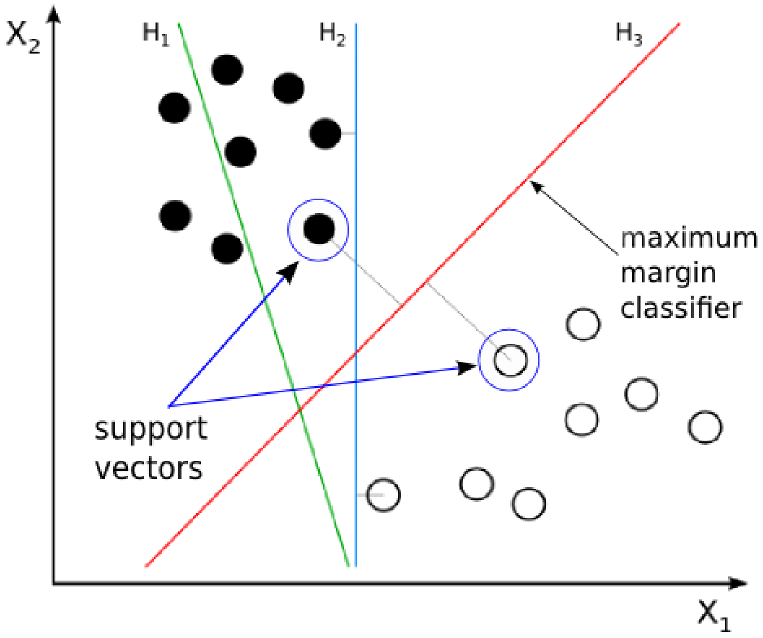


Figure 3.23: Illustration of Support Vectors

3.5.6.1 Support Vector Machine

'Support vector machine' (SVM) is a machine learning algorithm, It was proposed by Cortes and Vapnik in 1995 [6]. Here, each element is plotted on an n-dimensional space which represents the number of features. Classification is then performed by finding a hyper plane that best separates the two classes. SVM performs well when there is a large amount of training data, there are not many missing values, when the data is non-linear and when the classes are balanced. SVM uses different kernel functions and different kernels perform better on different types of problems. we employed Radial basis function (RBF) kernel for classification.

3.6 Summary

This chapter introduced the dataset used in our study, the data preparation step that used to extract different dataset at two levels; patch and character, the detail discussion on texture features and pre-trained models. we used total 28 dataset for experiments, one is original-dataset, one is patches-dataset and 26 character datasets. Detail of all dataset that we employed in our study are presented in Table 3.3 and 3.4. Furthermore, explain the pre-trained models that we employed for transfer learning. Different models are considered for using as feature extraction followed by classification. In the next chapter we present the details of the experiments and the realized results.

Chapter 4

Experiments and Results

In this chapter we discuss the results obtained from various experiments done on hand-crafted features and machine-learned features. The results are categorized into ’Text dependent solutions’, that mostly comprise of character-level datasets and ’Text independent solutions’, which contains page-level, patch-level and character-level datasets. Both the solutions are evaluated with hand-crafted and machine learned feature representations. Comparison of the techniques is done with the proposed model that is build using Deep Convolution Neural Network.

The result comparison is also done on various representations of input data as well as different models and feature representations. For this purpose, we used as input individual characters and combination of some characters. The feature representations used in this study are discussed in detail in chapter 3, and include Glcm-based method from Mikkilineni et al. [32], which describes the signatures present in the banding with 22 statistics calculated per matrix. We refer to this approach in the experiments Glcm (Using 16 statistical feature) method. Next we use ‘local binary patterns’ (LBP) [36] and ‘histogram of oriented gradients’ (HOG) [8]. We employed various combinations of Glcm, Hog, U-LBP and LBP to assess their performance. All deep learning experiments were performed using the methodology presented in Chapter 3 using the datasets (in sec 3.1).

4.1 Performance Measure

Performance of proposed methodology in Section 3.3 is validated by computing the accuracy of classifier on each dataset. Accuracies are calculated on the test set, whereas the split between test and train data is fixed to a 30 to 70 ratio respectively. The sensitivity of a classifier is measured using the ROC curve. The ROC curve calculates the ratio of true positive to true negative rate to find how effective the classifier is. We used confusion matrix to measure the direction and effect of deviations of our classifier as well.

4.2 Text-dependent Experiments

Text dependent approaches depend on using images of fixed size where each image is centered on a character. In our study, selected images contain plain text, text from graphs and text from tables. This content has different variations of font style and size and is extracted using an OCR.

For text-dependent approach, the extracted features from characters are compared with the features of same character in training and testing. At a character-level, we extracted characters from dataset that are discussed in Sec 3.1. Since, vowels are frequently used characters, we initially applied different texture features on only vowels characters-datasets. Texture feature results on character dataset are presented in Table 4.1 and 4.2. An analysis of Table 4.1 suggests that any one feature technique does not consistently perform better than others on all types of characters. Therefore, we moved on to using combination of features as well, the results of which are depicted in Table 4.2. Performance of combination of HOG and LBP is better than other texture features. The detail of this techniques is discussed in section 3.3.

Next, we applied deep learning architectures and found that deep learning results are better than texture results by a significant margin. For Deep learning, individual characters are used to train a CNN models. Results of six different models are reported in Table 4.3. We used vowels characters dataset to report the performance of different pre-trained models. Out of all employed models, Alexnet outperformed other models. For reporting the performance of characters, we performed our analysis using vowels and from results, it can be noted that character ’e’ achieved best accuracy from Alexnet model and that is 94.36%. Other employed models such as; Resnet50, VGG19, and Resnet101 achieved 93.3%, 93,8%, and 92,37% accuracy respectively. Reported results show that char ‘i’ achieved lowest accuracy in comparison to other vowels characters.

Table 4.1: Proposed approach Results Using Texture Features I

Input Data	LBP	HOG	U-LBP	GLCM
char-a	54.2	54.55	53.44	39.45
char-e	50.38	56.21	43.56	41.33
char-i	50.92	65.55	43.48	50.39
char-o	43.56	10.91	40.71	34.26
char-u	64.71	73.72	58.76	54.99

Table 4.2: Proposed approach Results Using Texture Features II

Input Data	GLCM+LBP	HOG+LBP	LBP+U-LBP	LBP+U-LBP +HOG	LBP+U-LBP +GLCM
char-a	55.83	**66.96**	56.33	58.01	57.51
char-e	52.37	**66.29**	54.91	56.16	55.31
char-i	52.87	**70.94**	54.31	60.86	55.79
char-o	44.71	**55.57**	43.46	47.52	44.88
char-u	66.25	**80.16**	67.44	69.37	68.37

Table 4.3: Proposed approach Results Using Deep Learning Architectures

Input Data	Resnet50	VGG19	Resnet101	InceptionResnetv2	Alexnet	Googlenet
char-a	88.90	92.12	87.32	84.37	83.50	83.32
char-e	93.30	93.80	92.37	90.49	**94.36**	87.62
char-i	73.66	76.88	71.18	68.34	81.68	60.04
char-o	76.70	79.22	74.78	72.61	82.65	70.39
char-u	87.30	89.62	86.58	84.30	90.00	79.11

As the performance of Alexnet was significantly better than other deep models and hand-crafted representations on character level datasets, we use Alexnet model for further analysis, starting off by training all characters (a-z all datasets) individually. The trained Alexnet's results are displayed in Figure 4.1 that shows results on all character. The highest accuracy is achieved on character 'e' which is 94.36% and lowest on character 'x' is 76.3% because character 'e' frequently appear in documents and 'x' is not frequent. After conducting experiments on single character, we used multiple character datasets and performed classification by concatenating character features. Another analysis is performed by using different printer datasets such as laser, inkjet and a full (both inkjet

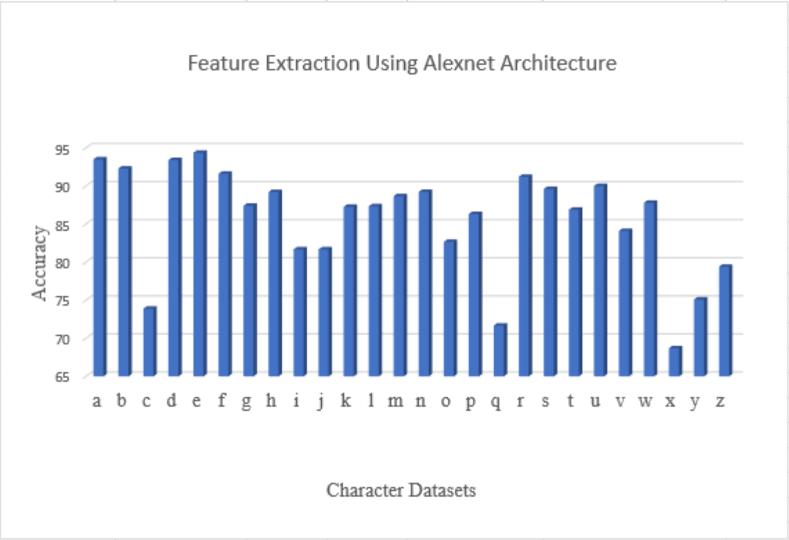


Figure 4.1: Performance of Individual Characters (a-z) Using Feature Extraction

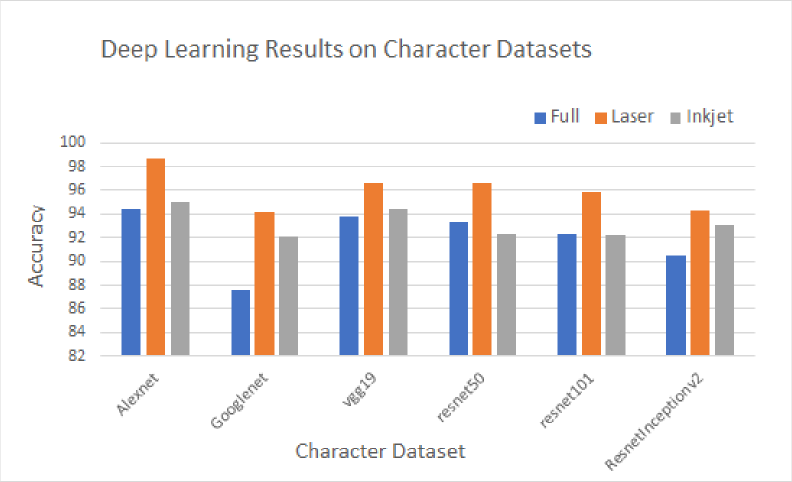


Figure 4.2: Results of Deep learning Different Models on Characters ‘e’ Datasets

Table 4.4: Concatenated Feature Vector Results Using Alexnet

Dataset	Character	Feature Vector	Accuracy
D_{ae}	a and e	8192	96.81
D_{de}	d and e	8192	95.21
D_{eu}	e and u	8192	94.62
D_{abe}	a, b, and e	12288	95.16
D_{ade}	a, d, and e	12288	**98.06**
D_{aeu}	a, e, and u	12288	97.62

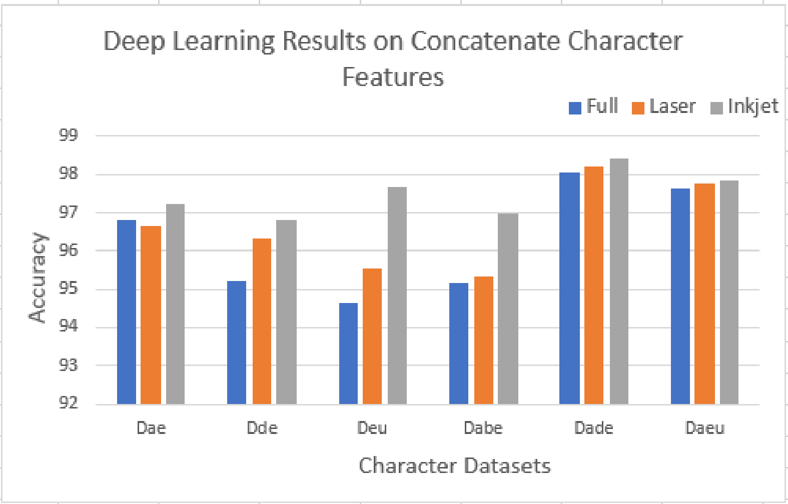


Figure 4.3: Results of Concatenation of Character feature on Different Datasets

and laser printer) datasets to perform classification by concatenating different character features. It can be observed from Figure 4.3 that laser and inkjet printer achieves better classification accuracy than combining both printer datasets. Results of this experiments is shown in Table 4.4

4.3 Text-independent Experiments

Using text-independent approach we performed the analysis at page, patch and character level. For text-independent approach, the extracted patches are used to learn features that are able to identify printer label. The patch data used in training is compared with a different patch in testing. From experiments, we also tried to extract features from characters that are compared with the features of different characters in training and testing. This make the technique independent at character level. The analysis at both levels is discussed in next sections.

4.3.1 Page and Patch-Level Results

The page level results are computed by extracting features from full page image and patch level results are computed by extracting patches from full page image. We extracted patches of size 300x512 that divides the A4 document evenly. The patch size covers the full A4 page of scanned document. Both handcrafted texture-features and machine-learned features are used to perform classification. The employed texture techniques include: LPB, Uniform LBP, HOG and GLCM. Just as in Text dependent analysis, comparative analysis for text independent techniques is also performed by extracting features by employing them individually as well as their combinations. SVM is used for classification purpose.

Table 4.5: Texture Feature Results on patch-level using Text-independent

Feature	Dimensionality	Page-level	Patch-level
U-LBP Features	59	74	67
HOG Features	756/216	75	30
LBP Features	256	84	78
GLCM Features	16	39	26
GLCM_LBP	262	50	45
HOG_LBP	1012/472	72	57
U-LBP_LBP	315	86.96	83.19
LBP_ U-LBP_HOG	1071/531	85.06	82.06
LBP_ U-LBP_GLCM	331	**87.55**	**84.44**

Table 4.6: Deep Learning Results on patch-level using Text-independent

Model	Feature Layer	Acc
Alexnet	fc-6	45.64
VGG16	fc-6	45.01
VGG19	fc-6	52.56
Googlenet	loss3-classifier	55.08
Inceotionv3	predictions	58.25
Resnet101	fc1000	60.78
Resnet50	fc1000	63.89

We also investigated the performance using deep learning models. The classification rates against different deep learning models are presented in Table 4.6 which shows that highest accuracy achieved is 63.89% using CNN as a feature extractor. We can conclude that deep learning performance is not efficient on patch-level datasets in comparison to texture techniques. For page-level results using deep learning models are not meaningful because input of CNN models are of fixed size. Table 4.5 presented the results of texture features where we can see that individual feature performance was satisfactory but combination of texture features outperforms the results of individual texture feature results. Best results are achieved using different combination of texture features with accuracy **84.44%**, **87.0%**,and **91.0%** on patch-level and **87.55%**,**83.03%**,and **77.95%** achieved by using LBP _ U-LBP _ GLCM on full-dataset, laser-dataset and ink-jet-dataset respectively. The results are displayed in Figure 4.5).

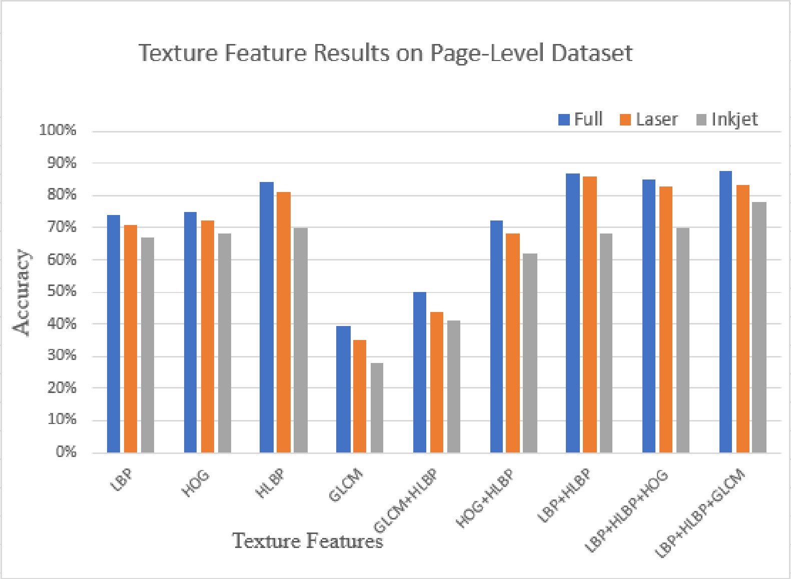


Figure 4.4: Texture Feature Results on page level Dataset

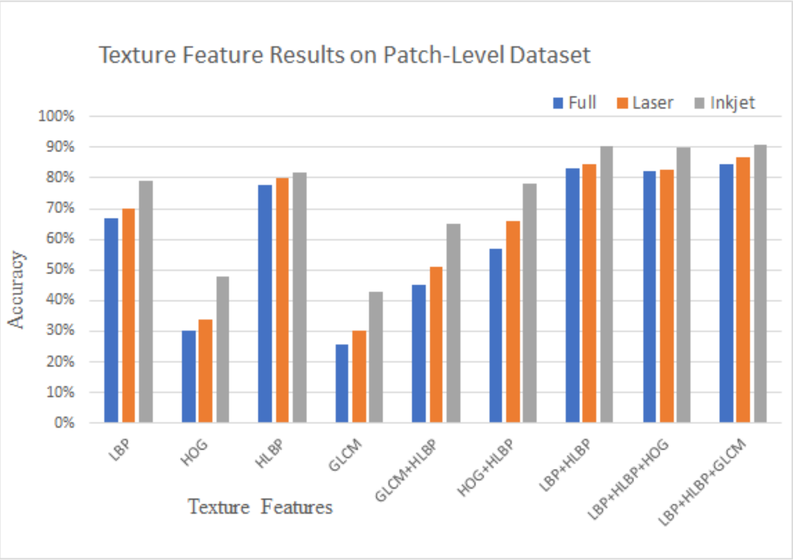


Figure 4.5: Texture Feature Results on patch level Dataset

Table 4.7: Results on Multiple combination of character Using Alexnet

Dataset	Character	Feature Vector	Accuracy
D_{a-z}	**a to z**	**4096**	**88.5**
D_{ae}	a and e	4096	**95.81**
D_{de}	d and e	4096	93.12
D_{eu}	e and u	4096	92.33
D_{abe}	a, b, and e	4096	93.71
D_{ade}	a, d, and e	4096	**95.75**
D_{aeu}	a, e, and u	4096	80.01

4.3.2 Character-Level Results

The text independence was tested at character level as well. For achieving independence, we create variations in character datasets such that the characters on which we trained our model were different from the ones on which we tested it for accuracy evaluation. For this purpose, we made different dataset that includes all (a-z) characters and different combination of datasets. The characters ‘a’, ‘b’, ‘d’, ‘e’, ‘u’ is used in combination to make different datasets. Initially, we computed the accuracy by combining all the characters in-spite of the fact that we got different characters on training as well as on testing.

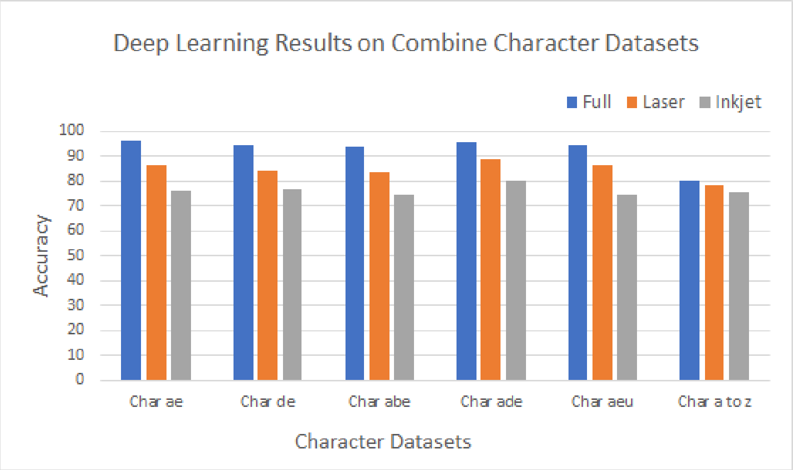


Figure 4.6: Results of Deep learning on Combination of Different Character Datasets

Furthermore, we computed accuracy on different combination of highest accuracy characters, i.e. ‘a’, ‘b’, ‘d’, ‘e’, ‘u’ (in Figure 4.1). Table 4.7 shows all the result that were extracted from the combinational dataset. We got **95.8%** and **95.7%** accuracy on the combination of character ‘a’, ‘e’ and ‘a’, ‘d’, ‘e’ dataset respectively. We also achieved **88.8%** accuracy when we combine all characters in one dataset for the experiment. The drop in accuracy can be attributed to the variation in shape of each character, particular since it was made sure that feature vector for each test was kept constant in size: 4096. The results indicate that as texture doesn’t perform better on characters, we only computed results by using deep learning models. Performing an analysis after using full, laser and inkjet printer datasets, the best accuracy achieved is 95.81% by using a combination of character ‘a’, ‘e’. Detailed results are displayed in Figure 4.6.

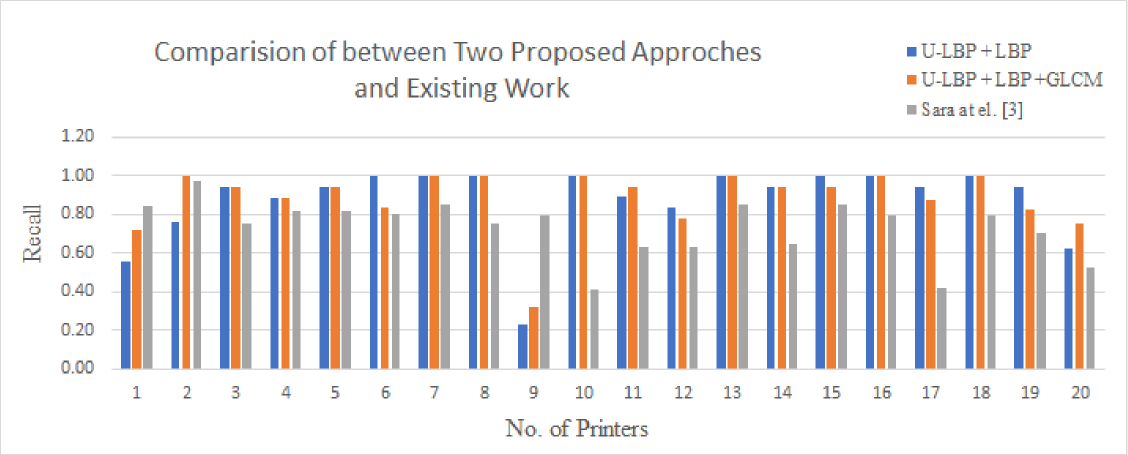


Figure 4.7: Comparison of existing and proposed work with Recall (20-printers)

4.4 Comparison with Existing Technique

We also performed a comparative analysis of proposed techniques with existing techniques. For evaluation we employed two metrics precision and recall, graphical representations of which are given in Figure 4.8 and 4.7. All 20 printers were used for analysis. For recall at printer 1 and 9 existing techniques proposed by Sarah et al. [9] outperformed our proposed techniques, while for the rest of the methods our proposed model performed better. Numerous printers achieved the recall value of 1.00 including printers labelled as: 7,8,10,13,15,16 and 18.

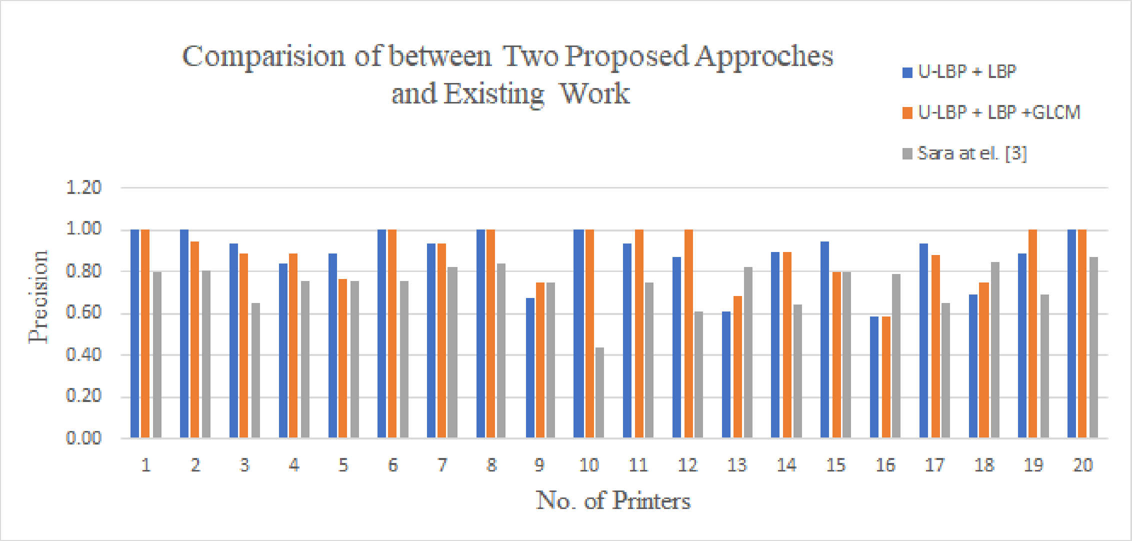


Figure 4.8: Comparison of existing and proposed work with Precision (20-printers)

Table 4.8: Comparison with Text-Independent Existing Techniques

Method	Extraction-level	Dimension	Dataset	Acc
Elkasrawi et al. [9]	Noise	15	German Dataset [9]	76.75
LBP_U-LBP_GLCM	Patch-level	331	German Dataset [9]	84.44
LBP_U-LBP	Patch-level	315	German Dataset [9]	83.19
LBP_U-LBP_HOG	Patch-level	531	German Dataset [9]	82.06
LBP_U-LBP_GLCM	Page-level	331	German Dataset [9]	87.55
LBP_U-LBP	Page-level	315	German Dataset [9]	86.95
LBP_U-LBP_HOG	Page-level	1071	German Dataset [9]	85.06
Alexnet	Char-level	4096	German Dataset [9]	95.81

Table 4.9: Comparison with Text-dependent Existing Techniques

Method	Feature-level	character	Dataset	Acc
Ferreira et al. [12]	Char-level	char e	Letter/Character [13]	98.30
Alexnet	Char-level	char a,d, and e	German Dataset [9]	98.06
Alexnet	Char-level	char a,e, and u	German Dataset [9]	97.62
Alexnet	Char-level	char a, and e	German Dataset [9]	96.81
Alexnet	Char-level	char d, and e	German Dataset [9]	95.21

Existing work performance is increased at printer labelled 16 and 18. It can also be noted from these images that combining two features somehow gives competitive results with existing results but combining more texture features give best results. Table 4.8 shows the comparison of our proposed technique with the existing technique. Our proposed technique achieved results comparable to existing techniques in text-dependent mode, but in text-independent mode our model outperforms the existing techniques, as seen from results reported in Table 4.9.

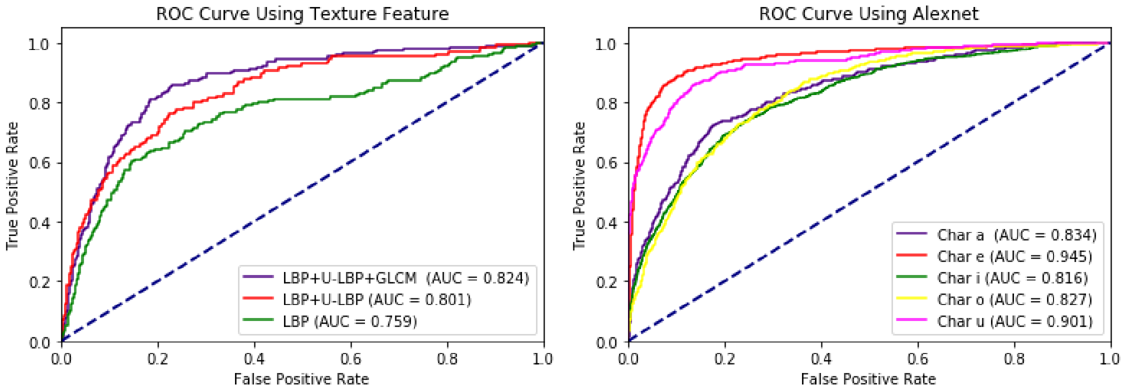


Figure 4.9: ROC Curve of Texture Feature and Deep Learning (Alexnet)

4.5 Analysis and Discussion

To detect forgery from printed scanned documents, we performed experiments based on text-dependent and text-independent approach. For text-dependent the results are computed at character level whereas, for text-independent the results are computed at page, patch and character level. In text-dependent approach we achieved best accuracy 98% at character level and for text-independent we achieved 84.44% on patch level and 95.81% on character level.

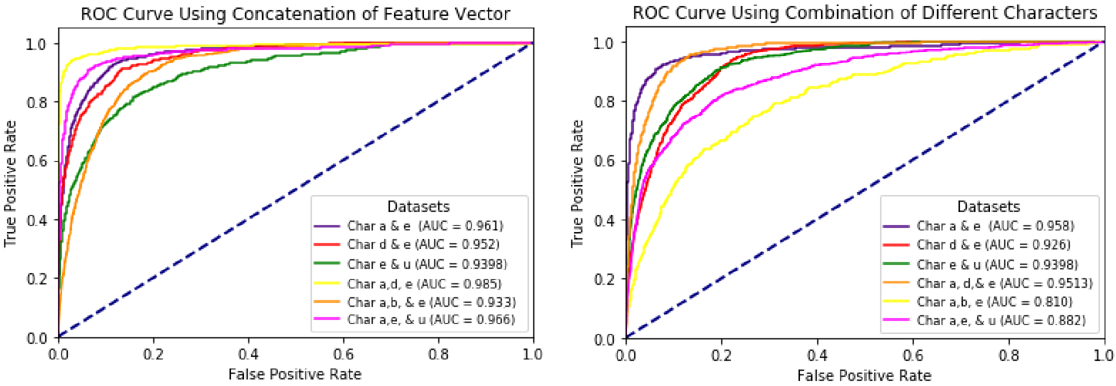


Figure 4.10: ROC Curve on Character datasets Using Text-dependent and Independent mode

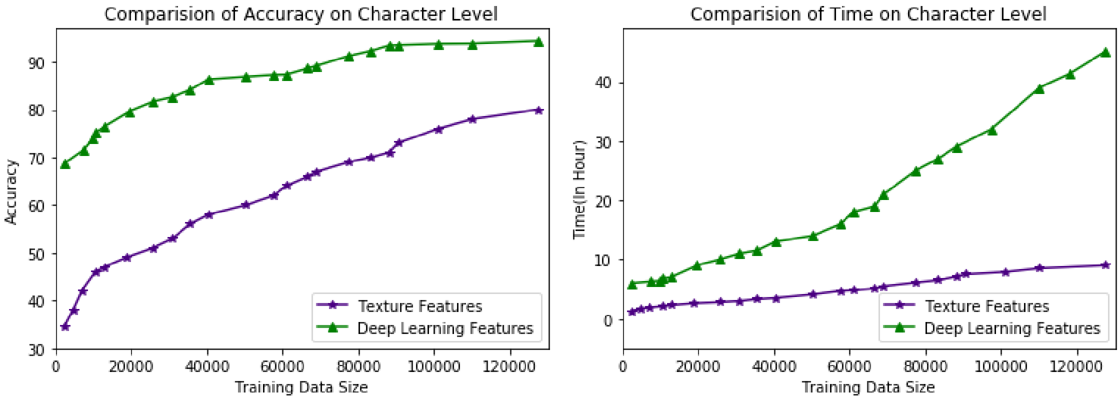


Figure 4.11: Comparison of Variation of Data with Time and Accuracy

The analysis results are presented by using ROC curve. The ROC curve is computed from using the dataset of characters in which features are concatenated (text-dependent approach) and also using combined characters (text-independent approach) datasets. The curves are also constructed to discriminate the features learned from using conventional techniques and also machine-learned techniques. From our results we achieved best results from using Alexnet model so we reported that results in curve. ROC curves at this analysis are illustrated in Figure 4.9 and Figure 4.10.

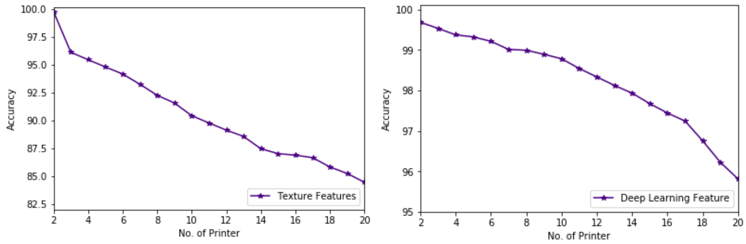


Figure 4.12: Comparison of Independent Approach with Texture and Deep Learning Techniques

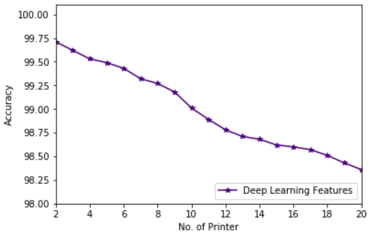


Figure 4.13: Dependent Approach with Deep Learning Techniques

Another analysis is performed between the accurately identified printer using both texture and deep-learning techniques. It is noted from Figure 4.11, that deep learning achieves high accuracy values in comparison with texture values and as deep learning needs large datasets the accuracy increased to a maximum value. Increasing training data size also effects the performance of system w.r.t time. As the data set increases the time complexity for predicting the label of source printer is also increased. The resulting values can be visualized from Figure 4.11.

We also perform a comparison analysis between text-dependent and text- independent techniques that shows that increasing number of printer will result in decreasing of accuracy. From Figure 4.12 and 4.13, it can be observed that when the number of printers are less the high accuracy values are achieved and as the printer started increasing the accuracy dramatically decreases to lower values.

Confusion Matrix is also generated on both highest achieving accuracies proposed in our study. The highlighted diagonal in matrix shows the correctly classified source printers. We are able to classify source printer using both text- dependent and independent approach more accurately. The matrix is represented in Table 4.10 and Table 4.11. Our results outperform the discussed techniques in literature.

Table 4.10: Confusion Matrix of Best Proposed Approach of Text-Independent (Alexnet results on Combine Dataset of Character a d and e)

char a d and e		Predicted No. of printer																			
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Actual No. of printer	01	668				1		9	3	3				3	1		8		27		
	02		717				1			1							2				
	03	1	2	679	3	7	6		1	1	0	2	11	4	1	2			1		2
	04			1	695	13	5	1	1		1		2		1					3	
	05		2	2	2	704			4		1	2		1		1			1	1	2
	06	1		1	1	3	700	2	1	1		2	2	2	3			1	1	1	1
	07	1			1		1	712	5									2			
	08	4						1	704	3					1		5	2	3		
	09	3					2	1	12	680				4			8		12		1
	10	1			2	1	12	2	2	1	691	5	4				1				1
	11	2		6	4	2	3		4	1	8	684	4			1				2	2
	12	3		4	1	1	3	2			6		684	3	6	1					3
	13	2				4	3		4	10	1		1	648	43		5	1			1
	14		1	3		2	2			2	2	1	3	16	684		4			2	1
	15	1	2			9			2		1			1	2	701	2		1		1
	16	5	3						2	5			1	1			694	2	9		1
	17		1					8	4	1		2	1					705	1		
	18	28	1					1	9	4				1			26		653		
	19				6	4					1	1	1	2		1			1	705	1
	20		1	1						2	1		2		1		1	3			711

Table 4.11: Confusion Matrix of Best Proposed Approach of Text-dependent (Alexnet results on Concatenate Feature Vector of Character a d and e)

char a d and e		Predicted No. of printer																			
		01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20
Actual No. of printer	01	703				1				3					4				12		
	02	2	717						3	4											
	03	3		706					2				6						5		
	04			1	713						3				6						
	05			6		707			4					1		6					
	06				6		700			7		6						4			
	07							723													
	08								723												
	09			9	6		2	2		680				16			8				
	10				4			5			703						10				1
	11					4			4	9		687	10			5				4	
	12												723								
	13					8	5				1		1	701				7			
	14		1								9				713						
	15					12				3				3		701					4
	16		3						4								704		12		
	17		1			5				4								712	1		
	18									13				1					707		
	19								5			2		3					2	711	
	20						4						3		3						713

Chapter 5

Conclusions & Perspectives

5.1 Conclusion & Perspectives

In this thesis, we proposed a comparison-based analysis between hand-crafted and machine-learned features for document forgery detection using Printer source identification. It would be a powerful tool to detect crimes involving documents. We employed both hand-engineered and machine learned features to classify these printed documents accurately. Experiments are carried out in text dependent and text-independent modes. Features are then extracted at page level, patch level and character level. We also break down our dataset into patches and character level both for laser and ink-jet printers.

We first employed hand-engineered features that are Local Binary Patterns (uniform and non-uniform), Gray level Co-occurrence Matrix and HOG descriptor separately to our dataset and then by combining them we compute results. We also employed different pre-trained model as feature extractors but deep learning results are not satisfactory at patch-level in comparison with textural feature results. It can be seen from the above analysis that combination of different texture features leads to more accurate results than applying them individually. We achieved 84% accuracy on patch level and 87% on page level using combination of texture features (text-independent).

At character-level, we tested both approaches text-dependent and independent. We employed different pre-trained convolution neural networks as feature extractor. Deep learning techniques performed best on character level datasets. Using text-dependent technique, we conducted experiments and gathered results at all individual 26 characters and then picked the best performance. By concatenating different characters in text-dependent approach, the best accuracy achieved is 98% on characters ('a', 'd' and 'e') with

Alexnet. Using text-independent technique, we combined best performance characters and applied both hand-crafted and machine-learned feature extraction techniques. We achieved accuracy of 95.81% on a and e, 95.75% on a, d and e, and 88% on a-z character dataset.

Future extensions of the study include comparison of the performance of text-dependent and text-independent feature extraction. An analytical study on discriminating power of different characters (in a text-dependent framework) would also be an interesting direction. Furthermore, identifying the most suitable scale of observation (character, word, patch or document) for this type of problem also requires further investigation.

Bibliography

[1] Amr Gamal Hamed Ahmed and Faisal Shafait. Forgery detection based on intrinsic document contents. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 252–256. IEEE, 2014. Cited on p. 1.

[2] Romain Bertrand, Petra Gomez-Krämer, Oriol Ramos Terrades, Patrick Franco, and Jean-Marc Ogier. A system based on intrinsic features for fraudulent document detection. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 106–110. IEEE, 2013. Cited on pp. 13 and 17.

[3] Orhan Bulan, Junwen Mao, and Gaurav Sharma. Geometric distortion signatures for printer identification. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1401–1404. IEEE, 2009. Cited on pp. 6 and 11.

[4] Jung-Ho Choi, Dong-Hyuck Im, Hae-Yeoun Lee, Jun-Taek Oh, Jin-Ho Ryu, and Heung-Kyu Lee. Color laser printer identification by analyzing statistical features on discrete wavelet transform. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1505–1508. IEEE, 2009. Cited on pp. 7 and 11.

[5] Dan Cireşan and Ueli Meier. Multi-column deep neural networks for offline handwritten chinese character classification. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–6. IEEE, 2015. Cited on p. 2.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. Cited on p. 41.

[7] Francisco Cruz, Nicolas Sidere, Mickaël Coustaty, Vincent Poulain D'Andecy, and Jean-Marc Ogier. Local binary patterns for document forgery detection. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 1223–1228. IEEE, 2017. Cited on pp. 10 and 11.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. Cited on pp. 32 and 42.

[9] Sara Elkasrawi and Faisal Shafait. Printer identification using supervised learning for document forgery detection. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 146–150. IEEE, 2014. Cited on pp. 12, 17, 19, 21, 50, and 51.

[10] Mohamed Elleuch and Monji Kherallah. An improved arabic handwritten recognition system using deep support vector machines. In *Computer Vision: Concepts, Methodologies, Tools, and Applications*, pages 656–678. IGI Global, 2018. Cited on p. 2.

[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017. Cited on p. 2.

[12] Anselmo Ferreira, Luca Bondi, Luca Baroffio, Paolo Bestagini, Jiwu Huang, Jefersson A dos Santos, Stefano Tubaro, and Anderson Rocha. Data-driven feature characterization techniques for laser printer attribution. *IEEE Transactions on Information Forensics and Security*, 12(8):1860–1873, 2017. Cited on pp. 5, 12, 16, 18, 21, and 51.

[13] Anselmo Ferreira, Luiz C Navarro, Giuliano Pinheiro, Jefersson A dos Santos, and Anderson Rocha. Laser printer attribution: Exploring new features and beyond. *Forensic science international*, 247:105–125, 2015. Cited on pp. v, 6, 15, 17, 24, and 51.

[14] Yin-rong Fu and Sheng-yun Yang. Ccs-ltp for printer identification based on texture analysis. *International Journal of Digital Content Technology and its Applications*, 6(13), 2012. Cited on p. 30.

[15] Cheng-Cheng Gao and Xiao-Wei Hui. Glcm-based texture feature extraction. *Computer Systems & Applications*, 6:048, 2010. Cited on pp. vii and 32.

[16] Johann Gebhardt, Markus Goldstein, Faisal Shafait, and Andreas Dengel. Document authentication using printing technique features and unsupervised anomaly detection. In *ICDAR*, pages 479–483, 2013. Cited on pp. 13 and 17.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. Cited on p. 36.

[18] Jianyuan Hao, Xiangwei Kong, and Shize Shang. Printer identification using page geometric distortion on text lines. In *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*, pages 856–860. IEEE, 2015. Cited on pp. 12, 17, and 18.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. Cited on pp. 38 and 40.

[20] Marko Heikkilä, Matti Pietikäinen, and Cordelia Schmid. Description of interest regions with local binary patterns. *Pattern recognition*, 42(3):425–436, 2009. Cited on p. 30.

[21] Hardik Jain, Gaurav Gupta, Sharad Joshi, and Nitin Khanna. Passive classification of source printer using text-line-level geometric distortion signatures from scanned images of printed documents. *arXiv preprint arXiv:1706.06651*, 2017. Cited on pp. 12 and 16.

[22] Sharad Joshi and Nitin Khanna. Single classifier-based passive system for source printer classification using local texture features. *IEEE Transactions on Information Forensics and Security*, 2017. Cited on pp. 14 and 17.

[23] Jeremy Kawahara, Aicha BenTaieb, and Ghassan Hamarneh. Deep features to classify skin lesions. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1397–1400. IEEE, 2016. Cited on p. 2.

[24] Shihab Hamad Khaleefah and Mohammad Faidzul Nasrudin. Identification of printing paper based on texture using gabor filters and local binary patterns. *Journal of Theoretical & Applied Information Technology*, 86(2), 2016. Cited on p. 30.

[25] Do-Guk Kim and Heung-Kyu Lee. Color laser printer identification using photographed halftone images. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 795–799. IEEE, 2014. Cited on pp. v, 10, and 11.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. Cited on pp. 37 and 40.

[27] Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990. Cited on p. 35.

[28] Hae-Yeoun Lee and Jung-Ho Choi. Identifying color laser printer using noisy feature and support vector machine. In *Ubiquitous Information Technologies and Applications (CUTE), 2010 Proceedings of the 5th International Conference on*, pages 1–6. IEEE, 2010. Cited on pp. 7 and 11.

[29] Qingyong Li, Zhen Zhang, Weitao Lu, Jun Yang, Ying Ma, and Wen Yao. From pixels to patches: a cloud classification method based on a bag of micro-structures. *Atmospheric Measurement Techniques*, 9(2):753–764, 2016. Cited on p. 21.

[30] Li Liu, Jie Chen, Paul Fieguth, Guoying Zhao, Rama Chellappa, and Matti Pietikainen. A survey of recent advances in texture representation. *arXiv preprint arXiv:1801.10324*, 2018. Cited on p. 30.

[31] Topi Mäenpää and Matti Pietikäinen. Texture analysis with local binary patterns. In *Handbook of pattern recognition and computer vision*, pages 197–216. World Scientific, 2005. Cited on p. 31.

[32] Aravind K Mikkilineni, Pei-Ju Chiang, Gazi N Ali, George T-C Chiu, Jan P Allebach, and Edward J Delp. Printer identification based on texture features. In *NIP & Digital Fabrication Conference*, volume 2004, pages 306–311. Society for Imaging Science and Technology, 2004. Cited on p. 42.

[33] Aravind K Mikkilineni, Pei-Ju Chiang, Gazi N Ali, George TC Chiu, Jan P Allebach, and Edward J Delp. Printer identification based on graylevel co-occurrence features for security and forensic applications. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, pages 430–441. International Society for Optics and Photonics, 2005. Cited on p. 30.

[34] Haris Bin Nazar, Momina Moetesum, Shoaib Ehsan, Imran Siddiqi, Khurram Khurshid, Nicole Vincent, and Klaus D McDonald-Maier. Classification of graphomotor impressions using convolutional neural networks: An application to automated neuro-psychological screening tests. In *Document Analysis and Recognition (ICDAR),*

2017 14th IAPR International Conference on, volume 1, pages 432–437. IEEE, 2017. Cited on p. 2.

[35] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449. ACM, 2015. Cited on p. 36.

[36] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *European Conference on Computer Vision*, pages 404–420. Springer, 2000. Cited on pp. 30 and 42.

[37] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. Cited on p. 30.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. Cited on p. 35.

[39] Seung-Jin Ryu, Hae-Yeoun Lee, Dong-Hyuck Im, Jung-Ho Choi, and Heung-Kyu Lee. Electrophotographic printer identification by halftone texture analysis. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 1846–1849. IEEE, 2010. Cited on pp. v, 8, 9, 10, and 11.

[40] Shize Shang, Nasir Memon, and Xiangwei Kong. Detecting documents forged by printing and copying. *EURASIP Journal on Advances in Signal Processing*, 2014(1):140, 2014. Cited on pp. 1, 13, 17, and 21.

[41] Arjun Sharma et al. Adapting off-the-shelf cnns for word spotting & recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 986–990. IEEE, 2015. Cited on p. 36.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. Cited on pp. 38 and 40.

[43] Raymond W Smith. Hybrid page layout analysis via tab-stop detection. In *2009 10th International Conference on Document Analysis and Recognition*, pages 241–245. IEEE, 2009. Cited on p. 27.

[44] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017. Cited on pp. 38 and 40.

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. Cited on pp. 37 and 40.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. Cited on pp. 38 and 40.

[47] Min-Jen Tsai, Chien-Lun Hsu, Jin-Sheng Yin, and Imam Yuadi. Japanese character based printed source identification. In *Circuits and Systems (ISCAS), 2015 IEEE International Symposium on*, pages 2800–2803. IEEE, 2015. Cited on pp. 14, 17, 18, and 21.

[48] Min-Jen Tsai and Jung Liu. Digital forensics for printed source identification. In *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*, pages 2347–2350. IEEE, 2013. Cited on p. 30.

[49] Min-Jen Tsai, Jung Liu, Chen-Sheng Wang, and Ching-Hua Chuang. Source color laser printer identification using discrete wavelet transform and feature selection algorithms. In *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, pages 2633–2636. IEEE, 2011. Cited on pp. 8 and 11.

[50] Min-Jen Tsai, Yu-Han Tao, and Imam Yuadi. Deep learning for printed document source identification. *Signal Processing: Image Communication*, 70:184–198, 2019. Cited on p. 18.

[51] Min-Jen Tsai and Imam Yuadi. Digital forensics of microscopic images for printed source identification. *Multimedia Tools and Applications*, pages 1–30, 2017. Cited on pp. 17 and 21.

[52] Min-Jen Tsai and Imam Yuadi. Digital forensics of microscopic images for printed source identification. *Multimedia Tools and Applications*, 77(7):8729–8758, 2018. Cited on pp. 15 and 18.

[53] Min-Jen Tsai, Imam Yuadi, and Yu-Han Tao. Decision-theoretic model to identify printed sources. *Multimedia Tools and Applications*, pages 1–45, 2018. Cited on pp. v, 16, and 17.

[54] Tao Tsai, Yuadi and Yin. Source identification for printed documents. In *3rd IEEE International Conference on Collaboration and Internet Computing (CIC)*, pages 54–58, 2017. Cited on pp. 15, 16, and 17.

[55] Joost van Beusekom, Faisal Shafait, and Thomas M Breuel. Automatic authentication of color laser print-outs using machine identification codes. *Pattern Analysis and Applications*, 16(4):663–678, 2013. Cited on pp. 8 and 11.

[56] Zhen Wan, Ronghua Yang, Yangsheng You, Zhilin Cao, and Xinan Fang. Scene classification of multisource remote sensing data with two-stream densely connected convolutional neural network. In *Image and Signal Processing for Remote Sensing XXIV*, volume 10789, page 107890S. International Society for Optics and Photonics, 2018. Cited on p. 2.

[57] Han Wu, Xiangwei Kong, and Shize Shang. A printer forensics method using halftone dot arrangement model. In *Signal and Information Processing (ChinaSIP), 2015 IEEE China Summit and International Conference on*, pages 861–865. IEEE, 2015. Cited on pp. 6 and 11.