Chapter 1

INTRODUCTION

In this Chapter, background and importance of the proposed problem, a detailed literature survey of the state of the art methods, the motivation for the writer and script identification, objectives and applications of the research, mathematical tools and techniques, and classifiers used in the research work are explained clearly.

1.1 Background

Twenty first century became the information era due to the changes occurred in hardware and software technology. The automation of the manual process in almost all fields is the need of the hour as the customer/user wants the desired information on finger tips. Hence there is a need to design and develop the algorithms to process the desired task automatically without any human interference. Document image processing in general, writer and script identification in particular is the challenging field of research.

Handwriting is regarded as a sign of the identification of writer and each writer is characterised by his own handwriting style, repetition of certain details of text and unconscious practices. A long history before, people have realized the importance of identifying the true writer of an unknown handwritten document. In fact, writer identification based on the handwriting (there are many forms of handwriting, such as signatures, letters, notes, etc.) has a large number of applications: E.g. the confirmation of the document authenticity in the financial sphere, determining the authenticity of a document (e.g. :- a will), in banks (signature verification), to solve the expert problems in criminology, also performed on legal papers by way of a signature, etc. The documents which do not contain signature also need to be identified using only handwritten text. As a result, writer identification enjoys a huge interest from both industry and academia with multiple applications such as document examination, palaeography, security, financial activity, forensic and user
access control.

Traditionally, the ways of personal identifications are identification cards (ID cards) and passwords, while they cannot provide us an unique, secured and consistent personal identification. For example, passwords and ID cards can be shared by others and therefore are not unique. Furthermore, there may be chance of forgetting ID cards to carry along with us or forget the passwords and thus they are not consistent. So there is a need for better solutions to personal identification. Six most commonly used biometrics for personal identification are fingerprint, iris, hand geometry, face recognition, speaker identification, signature verification to overcome the problem of loss, sharing and forgery (Nalini K et al. [75]). Writer identification is another biometric modality and a solution to personal identification problem. Writer identification is to determine the writer (person) from his/her handwritings (including signatures, letters, notes, etc.), and is such a technique that satisfies four requirements of personal identification: accessible, cheap, reliable and acceptable. In spite of the existence and development of other techniques of personal identification based on DNA, iris, fingerprint, etc the writer identification still remains an attractive application due to the seemingly uniqueness of physiological and behavioural characteristics of each individual and existence of a certain degree of stability in every individuals handwriting.

In this computer era, most of the manual works are carried out by computer automatically and a great deal of time, efforts and money are saved. Automatic writer identification is a complex problem concerned with many science disciplines, including computer vision, pattern recognition, machine intelligence etc and has many challenging problems which are still not well solved. Naturally, the automatic writer identification comes into scientists’ views and is receiving growing interests from both academia and industry. And more and more researchers and firms put money and energies on it. Generally speaking, writer identification is a typical pattern recognition problem, a pattern is a sample handwritten document or handwritten text and a class represents one writer. **Writer identification is a process of one-to-many matching the query handwritten image with a set of N writers** (N
documents) already stored in the database. It is a modality based on the handwriting behaviour which proceeds by matching unknown handwritten text against a database of the samples of texts with known authorship. Writer verification is one-to-one matching process i.e., query document is matched with another document of unknown authorship and coming to conclusion that whether they are written by the same writer or not.

Writer identification can be classified in several ways, however the most straightforward one is to classify it into on-line and off-line writer identifications. The former assumes that a transducer device is connected to the computer, which can convert writing movement into a sequence of signals and then send these signals to the connected computer. The most common form of the transducer is a tablet digitizer, which consists of a plastic or electronic pen and a pressure or electrostatic-sensitive writing surface on which the user writes down his/her handwritings with the electronic pen. Since dynamic information of the writing process captured by the transducer device contains many useful writing features (shape, speed, pressure) in on-line writer identification, compared to the off-line writer identification, it is easier to achieve a high identification accuracy. On the other hand, off-line writer identification deals with handwritings scanned into a computer file in two-dimensional image representation. Despite continuous effort, off-line writer identification still remains as a challenging issue. In fact, off-line systems rely on more sophisticated architectures to accomplish the same identification task, and their identification results are still lower than those of on-line identification systems under same testing conditions. Unfortunately, on-line systems are inapplicable in many cases. For example, on-line systems cannot help us to find out the writer of an existing handwriting document. Therefore, designing and developing effective techniques for off-line writer identification is the need of the present scenario.

Further offline writer identification problem may be text-dependent or text-independent or script dependent or script independent. Text dependent method is to check a person's handwriting with two or more text materials with the same contents. It is very relevant with the text contents and we can utilize font, word
position, aspect ratio, distribution of strokes orientation, strokes arrangement as its characteristics. Some text dependent methods have been widely used, such as orthogonal transform, histogram method, standard template transform, higher-order moments, orientation index histogram and strokes matching, etc where as in text independent writer identification problem, the above features are not suitable. In text independent approach the writers may write any text as a line, paragraph or document without any restriction. Hence it is much more difficult problem.

Further the writer identification problem attempted in the literature can be classified as global and local approaches. The global approaches are based on the overall look and feel of the writing, on the other hand, local approaches identify the writer, based on localized features of writing, which are inherent in a way writer specifically writes the characters using allograph level, graphemes level features, etc.

With the aid of modern technology it is possible to produce, process, store, and transmit document images efficiently. Processing of not only printed documents but also the handwritten document has become an inherent part of office automation process. In multi-script and multi-lingual problem where documents of many scripts exist script recognition research is still an active research area across recognition of such documents. As the world is moving towards the paperless office, large quantities of printed or handwritten documents are digitized, stored as images in the databases for future access. Many organizations currently use and depend on handwritten document image database and hence the document images have retained their importance as information source even today. Searching for relevant handwritten documents from large complex document image repositories is a crucial problem in document image analysis and retrieval. Handwritten based document image retrieval is a process of retrieval of the handwritten documents that are similar to query image and has many applications in indexing and document image retrieval from archival of digital library of handwritten documents, resolving the disputes of the authorship of a document etc.
1.2 Document Image Analysis

Document Image Analysis involves investigation of document image data for a specific application. *Image analysis process requires the use of tools such as image segmentation, image transforms, feature extraction and pattern classification.* Since information contained in text images is basically bi-level in nature, we convert a gray scale document image into binary image through binarization process. Image segmentation is often one of the first steps in finding higher level objects from the raw image data, it involves partitioning of an image into several constituent components. Feature extraction is the process of acquiring higher level image information such as shape or color information, and may require the use of image transforms to find spatial frequency information. Image processing involves the manipulation of images to extract information to emphasize or de-emphasize certain aspects of the information contained in the image, to perform statistical analysis. This higher level information is taken up by pattern classification and objects within the image are identified.

1.3 Review of Literature

In this section review of literature on writer Identification and script identification are presented. A number of approaches to writer identification problems have been proposed in the literature. A survey on writer identification problems is discussed in Sreeraj et al. in [99].

**Offline Text Independent Writer Identification:**

A number of approaches to text independent writer identification also exist in the literature. Hertel et al. [22] used a subset of IAM database containing 250 pages written by 50 writers in English with 5 pages of text each. One page comprised of about 8 lines of text on the average amounting to the total data set of 2185 text lines. A set of connected component features, enclosed regions, lower and upper contours, fractal and basic features were extracted from a text line to identify the author using KNN classifier and achieved 90% for 50 writers and they have also shown that the combination of all lines increased the identification rate to 100%. Lambert
Schomaker et al. [59] used features of the contours of fragmented connected components in mixed-style handwritten samples of limited size. The writer was considered to be characterized by a stochastic pattern generator, producing a family of character fragments (fraglets). Using a codebook of such fraglets from an independent training set, the probability distribution of fraglet contours was computed for an independent test set. Each of the 150 paragraphs of the 150 writers was divided into a top half (set A) and a bottom half (set B). Writer descriptors \( p(\text{FCO3}) \) were computed for set A and B. Using Kohonen self-organized map (KSOM) of fragmented connected component contours of free style writing achieved a performance of 70\% which might be attributed to both the smaller number of characters and its variable text content. Imran Siddiqi et al. [51] extracted a set of features from contours of handwritten images at different observation levels. At the global level, they extracted the histogram of the chain codes, the first and second order differential chain codes and the histogram of the curvature indices at each point of the contour of handwriting. At the local level, the handwritten text was divided into a large number of small adaptive windows and within each window the contribution of each of the eight directions (and their differentials) was counted in the corresponding histograms. Two writings were then compared by computing the distances between their respective histograms. Using IAM database of 250 writers with 2 samples per writer, they achieved an identification rate of 86\% using chi-square distance with minimum distance (kNN with k=1) classifier.

Bulacu Marius et al. [69] has evaluated the performance of edge-based directional probability distributions as features in writer identification in comparison to a number of non-angular features. They have noticed that the joint probability distribution of the angle combination of two "hinged" edge fragments outperforms all other individual features such as Edge-direction distribution, Run-length distributions, Autocorrelation and Entropy. They [64] also demonstrated that fusing multiple features (directional, grapheme, run-length PDFs) yield very high writer identification and verification performance. Their fusion method based on simple distance averaging diminishes the risk of a biased solution. Bensefia Ameur [18] extracted graphemes as features for describing the individual properties of
handwriting and tested on dataset consisting of 88 different writers at PSI lab and further they analysed concatenation of graphemes as features. Hong Ding et al. [45] extracted the Local Contour Distribution Feature (LCDF) from the fragments which are parts of the contour in sliding windows. In order to reduce the impact of stroke weight, the fragments which do not directly connect the center point were ignored in the feature abstraction procedure. The edge point distributions of the fragments were counted and normalized into LCDFs and found the better performances for writer identification problem. Naeem Akther et al. [74] considered handwriting as n-branching ink fragments originating from a central point. Branchlets of 1-degree, 2-degree, 3-degree, and 4-degree were identified, probability distribution functions of these branching structure's orientation were used as features to overcome allographic constraints. They also characterized these branchlets as script independent unit of one's writing, since every script hold two attributes, one it's unique character set, second the sub patterns oriented in different directions. It considered linear and curvilinear patterns, which were not true about edge-hinge features. Time to generate a codebook including four types of branchlets was 17 hours for 250 writers.

Chetan Ramaiah et al. [26] proposed a method for accent detection (writer's native script on his/her writing style in another script) in handwriting based on writing styles. They have experimented with contour direction distribution, fractals, structure and concavity features at the line level, found that combination of these features also increases the accuracy of the system.

Vivian Blankers et al. [112] proposed a method for writer identification based on sub-allographic structural properties: shape of loop and lead-in strokes of characters used in handwriting. The Loop features such as length, area, width/height ratio, Relative Height, direction, average direction, average standard deviation, curvature of loop and Lead-in features such as length, direction, average direction, average standard deviation and Curvature of lead-in were extracted. They have tested their dataset on equal letters and similar letters using kNN classifier and achieved 98% writer identification accuracy for 41 writers. Romero et al. [21] proposed a writer identification based on handwritten text analysis for samples of 30 writers. The
features used are long of words, quantity of pixels in black, estimation of wide one of letters, height of the medium body of writing, heights of ascending and descending, height relation between of the ascending and medium body, height relation between descending and medium body, height relation between descending and ascending, height relation between medium body and the wide of writing. And achieved success rate of 94.66% by applying neural network classifier.

Utpal Garain et al. [109] devised a two dimensional autoregressive (AR) model to characterize writers and presented a framework to deal with writers writing in more than one script, that required no script specific knowledge. They tested on datasets of two different scripts, namely RIMES dataset containing 382 French writers and ISI dataset consisting of samples from 40 Bengali writers and found encouraging results.

Wong Yee Leng et al. [115] reviewed Chinese handwriting based writer identification and observed that for Chinese character set which contained more number of straight lines based features and has shown lesser identification accuracy compared to other script writers, since straight lines have less discriminating power than the curved lines.

Cheng-Lin et al. [25] used Moment-Based Feature Method to identify Chinese writers that normalized individual geometric moments of character images. The extracted features were invariant under translation, scaling, and stroke-width.

Yiu-ming Cheung et al. [117] extracted 2-D Gabor transformation as the global feature and Local Binary Pattern (LBP) as the local feature for Chinese handwriting of 10 writers and shown that the combination of global and local feature outperforms the utilization of each single one in writer identification.

Cong Shen et al. [27] proposed writer identification using two-dimensional Gabor Wavelet to extract global features from preprocessed, normalized, automatically zoomed and segmented image. Writer identification accuracy of 97.6% was achieved using k-NN classifier for 110 specimens of 50 writers. Said et al. [42] proposed writer identification methods for non uniformly skewed handwriting images (128x128 block size images) considering each writers’ handwriting as a
different texture. Multi-channel Gabor filtering technique is applied to handwritten samples for \( \Theta = 0^0, 45^0, 90^0, 135^0 \) and \( f = 4, 8, 16, 32 \) and extracted totally 32 features and obtained writer identification accuracy of 96.0\% using Weighted Euclidean Distance (WED) and 77\% using KNN classifiers for 15 training and 10 testing images. Gray Scale Co-occurrence Matrix based features are extracted from handwritten samples for distances \( d = 1, 2, 3, 4, 5 \) and \( \Theta = 0^0, 45^0, 90^0 \) and \( 135^0 \) (3 x 5 x 4 = 60 features). Obtained high accuracy of 88.0\% for 15 training and 10 testing images using kNN classifier. Shahabi et al. [34] extracted a set of features based on Gabor and co-occurrence matrix of document images. Initially they preprocessed the handwritten document images by smoothing and padding process. Their experimental results demonstrated that Gabor-energy and Fourier transform of Gabor output could outperform others methods, the comparative analysis of results shown the better performance in frequency domain for higher frequencies (details) and in spatial domain for small distances and hence obtained a higher identification rate by using features which contain details of texture. Behzad Helli et al. [17] proposed a method to identify writer of Persian (Farsi) handwritten documents based on extended Gabor filter, the features were based on the percentage of the different angles one uses in his/her word lines and the percentage of the amount and style of the curves that one uses in his/her writing style and achieved 100\% accuracy for 20 writers for same text written by the authors and 82\% accuracy for 70 writers on arbitrary texts.

Zhenyu He et al. [121] considers the global styles of different people’s handwritings are obviously distinctive and hence used the histogram of the wavelet coefficients of preprocessed handwriting images characterized by the Generalized Gaussian Model (GGD) in wavelet domain. They [120] also proposed a method based on hidden Markov tree (HMT) model in wavelet domain for writer identification of 1000 Chinese handwriting documents of 500 persons. They showed that the method when compared with two-dimensional Gabor model achieves better identification results as well as reduces the elapsed time on computation.

Marius Bulacu et al. [70] developed an automatic writer identification and verification method that used probability distribution functions (PDFs) extracted
from the handwriting images to characterize writer individuality that were designed to be independent of the textual content of the handwritten samples. The methods operated at two levels of analysis: the texture level and the character-shape (allograph) level. At the texture level, they used contour-direction PDF, contour-hinge PDF, direction co-occurrence PDFs, and other texture level Features: Run-length PDFs, autocorrelation that encode orientation and curvature information. At the allograph level, they characterized the writer by a stochastic pattern generator of ink-blob shapes, or graphemes and computed the PDF of these simple shapes in a given handwriting sample using a common shape codebook obtained by grapheme clustering. They showed that combining multiple features (directional, grapheme, and run-length PDFs) can yield enhanced writer identification and verification performance and their methods are applicable to free-style handwriting (both cursive and isolated). Schomaker et al. [60] proposed a Sparse-Parametric writer Identification method using heterogeneous feature groups and by evaluating edge-based directional probability distributions as features, and they found the features edge-hinge distribution, horizontal co-occurrence edge angles could outperform the non-angular features like run-length distributions, autocorrelation, ink density at stroke endings. They also found that only by the combinations of these features reliable results can be obtained.

Mohamed Nidhal et al.[71] computed six feature vectors from the minimum-perimeter polygon (MPP) contours of Arabic words for off-line text-independent writer identification. They extracted the feature vectors in the form of probability distribution functions (PDFs) based on the length, direction, angle and curvature measurements of edge, evaluated using 82 writers from the IFN/ENIT database and obtained identification accuracy of Top1 as 90.2% and Top10 as 97.5%.

Sami Gazzah et al. [90] proposed a writer identification approach for Arabic handwriting based on the combination of global and structural features. They extracted structural features such as line height, spaces between sub-words, inclination of ascenders, dots boldness and shape, global features such as Daubechies wavelet transforms, entropy of the word image. Genetic algorithm was
employed for optimal features selection, modular MLP classifier for classification and achieved 94.73% of average accuracy.

Xin Li et al. [116] writer identification of Chinese Handwriting using Grid Microstructure (histogram based) feature. The characteristics of a local grid around every pixel was described by the positions of edge pixel pairs and considered the pdf distributions of every pixel pairs as a feature. For HIT-MW Chinese handwriting database involving 240 writers, achieved top-1 identification accuracy of 95% and Top-20 accuracy of 99.6% by using improved weighted Chi-square distance metric.

Seropian et al. [2] proposed a fractal construction of a reference base for writer identification with accuracy of 85%. Soheila et al. [96] proposed a three step approach for writer identification: the first step comprised of grapheme feature to select first candidates, the second one comprised of area features and fuzzy approach to restrict candidate domain; and finally third step used gradient features to finalize selection. They applied fuzzy approach for the classification and clustering of handwritten texts and obtained the accuracy of 90% for 50 writers.

Andreas Schlapbach et al. [5] proposed HMM based recognizers for handwritten text lines, used 2,200 text lines from 50 writers and obtained 94.47% of accuracy in writer identification. Asim Imdad et al. [1] proposed a writer identification method based on Steered Hermite features to characterize oriented features, curves and segments of handwritings with image size is of 256X256 pixels. For 30 writers of IAM database, they achieved 83% accuracy using SVM classifier.

Imran Siddiqi et al. [52] have developed a two step approach, first one dividing the handwritings into distinct classes according to writing style by applying the Gabor filter to handwritten document images and second one by dividing the handwritten text into large number of small sub-images to extract local features to form clusters and thereby identifying the writer of the unknown document by searching within a specific writer class. Obtained 92% writer identification accuracy using the Bayesian classifier for 100 writers with two documents per writer.
Imran Siddiqi et al. [54] proposed a method involving codebook generation and contour, polygon based feature extraction to characterize contour-based orientation and curvature of redundant writing patterns. By combining codebook and contour features they achieved identification rates of 91%, 84% and 88% on the 650 writers of IAM, 375 writers of RIMES and the combined data sets, respectively.

A Grid Micro-structure Feature (GMSF) based text-independent writer identification method was proposed by Lu Xu et al. [63], using new weighted Chi-square distance metric, the recognition accuracy of 92% (testing category pen-width is 0.3mm) and 88% (testing category pen-width is 1.0mm) was achieved.

Bertolini et al. [19] used two texture descriptors (Local Binary Patterns and Local Phase Quantization) on two different datasets, the Brazilian Forensic Letters database and the IAM database. Their approach based on LPQ features was able to achieve accuracies of 96.7% and 99.2% on the BFL and IAM and databases respectively and showed the use of both LBP and LPQ as interesting alternatives for describing this kind of texture.

Zhenyu et al. [119] proposed Contourlet-based GGD method. They compared to 2-D Gabor filter and wavelet based methods and found Contourlet-based method achieves a higher accuracy, because of capacity of Contourlet transform to capture comparatively richer directional information, which is important in representing style of a handwriting.

Rajiv Jain et al. [82] presented a method for performing offline writer identification by using K-adjacent segment (KAS) features in a bag-of-features framework to model a user’s handwriting. They achieved a top 1 recognition rate of 93% on the 650 writers of IAM English handwriting dataset and demonstrated that identification performance improves as the number of training samples increase, and additionally, that the performance of the KAS features extend to Arabic handwriting found in the 325 writers of MADCAT dataset.

Imran Siddiqi et al. [53] proposed an approach for writer identification and handwriting classification independent of writing instrument (pen width) and texts
written with different inks. The divided the handwritten word into connected components and thinned each component repeatedly to a point where there is no change in the component area. The component area at each iteration was plotted against number of iterations and a blob fraction was computed which was found to be quite stable for a writer. Three samples of handwritten text with three different instruments were considered for evaluation.

Yoon Sungsoo et al. [118] described that handwriting originates from a particular copybook style such as Palmer or Zaner-Bloser that one learns in childhood. They proposed a system to identify a questioned document’s handwriting style, since the questioned document has an important investigative and forensic role in many types of crime and therefore the origin or nationality of an unknown writer of a questioned document was found. They collected 33 Roman alphabet copybook styles from 18 countries. Each character in a questioned document was segmented and matched against all of the 33 handwriting copybook styles and observed that the more characters present in the questioned document, the higher the accuracy observed.

Most of these methods need finding of contours, the creation of codebook of graphemes, fragments or allographs etc.

**Online Writer Identification:**
Guo Xian Tan et al. [41] exploited the alphabet knowledge to identify the writer and improved the performance from 66% to 87% by considering alphabet knowledge and observed that such alphabet knowledge can only be found if the prototype templates are built at the character level.

Anoop Namboodiri et al. [7] proposed a text independent writer identification system for online documents. They used sub character level features for writer identification and demonstrated that even with one line of data one can get a high confidence about the identity of the writer. To improve the accuracy and robustness of the system, they suggested the use of other high level primitives based on character, word, line and paragraph and shown that different primitives like shape, size and other higher level features in combination for the scripts like Chinese and
Roman and observed the use of spline for more robust representation of shape primitives.

**Script Recognition:** Many authors have contributed for script identification and brief review is as follows:

For handwritten script identification, various types of features can be extracted such as horizontal and vertical histogram, curvature information and local extreme of curvature, topological features such as loops group of white pixels surrounded by black ones, dots a cluster of say 1-3 pixels and junction as a point with more than 2 neighbours all in thinned black and white images, parameters of polynomial, contour information where contour is the outside boundary of a pattern, PCA as a way of identifying and expressing pattern in data [73].

A survey of script identification methods reported by Abirami et al. [87]. In the survey for texture (global approach) measurements are classified as Grey level Co-Occurrence Matrices (GLCM), Multichannel Gabor filters, Rotation invariant texture features, Wavelet based coefficients, Multichannel Log Gabor filters and Multiclass Gabor filters. Methods of local approaches are classified as upward concavities, Token based approach for block level and Stroke based, Structural based and water reservoir methods for line based script identification and morphological reconstruction for word based identification to provide some background information about the past researches on both global based approach as well as local based approach for script identification in document images. Both the approaches can perform Script/Language identification in document images at document, line and word level. But to be specific, Global based approaches gives an excellent result only if good quality document image prevails whereas local approaches can be adopted in low quality document images also.

By observing the text as a distinct visual texture, Andrew Busch et al. [6] investigated the use of texture as a tool for determining the script of a document image. They extracted texture features like Gray-level Co-occurrence Matrix Features, Gabor energy features, Wavelet energy features, Wavelet log mean deviation features, Wavelet co-occurrence signatures, Wavelet scale co-occurrence
signatures and observed that by using such features, it is not necessary to extract individual script components, and are found ideal for degraded and noisy documents or situations where such segmentation is not possible. When classifying scripts containing a single font, they observed the texture features outperformed the other script recognition techniques, and the lowest overall classification error rate with the wavelet log co-occurrence features.

Chanda et al. [103] used directional chain-code based histogram based features along with a Gaussian Kernel-based Support Vector Machine for Chinese, Japanese, Korean and Roman scripts. Obtained script recognition accuracy of 98.39% across character level and 99.85% at block level by considering 30 documents of each script, dividing equally into training and testing set.

Lijun Zhou et al. [61] proposed an approach for Bangla/English script identification with applications to the destination address block of Bangladesh envelope images. The approach was based upon the analysis of connected component profiles extracted from the destination address block images. During the preprocessing phase prior to feature analysis, they deleted special connected components that are either too small or too large and achieved accuracy of about 95% is for handwritten envelope images.

Sangame et al. [92] proposed an approach based on structural outline of characters of two handwritten scripts Kannada and English at block level and word level by the process of hole filling of image and finding number of on pixels and obtained script recognition accuracy of 95%.

Joshi et al. [37] have extracted features consistent with human perception and used hierarchical classification scheme for printed script identification from documents of 10 Indian scripts (Devanagari, Bangla, Tamil, Gurumukhi, Kannada, Malayalam, Roman, Oriya, Gujarati, Urdu) using a global approach. They have extracted local energy based features using multi-channel log-Gabor filter for printed script classification at block level. With KNN and Parzen window classifier achieved the recognition accuracy of 97.11%. 
Gopakumar et al. [81] proposed a zone based structural feature extraction algorithm for recognition of South-Indian printed scripts. They segmented the document image into lines and divided each line into is three zones and in each zone number of horizontal, vertical, right diagonal, left diagonal lines, normalized length of horizontal, vertical, right diagonal, left diagonal lines, and area of the line image were calculated. Using optimal feature selection approaches, obtained 100% by wrapper subset selection approach, 99.13% by filter approach.

1.4 Objectives of the Research Work

To address the problem of writer and script identification which are challenging problems since their application in forensic and digital libraries. In a multi-script and multi-lingual country like India, a writer knows more than one script/language, so he may write a document in any of the script/language known to him. But the algorithm developed for single script writer identification may not work for multi-script problem. Identification of writers who write in more than one script/language is a difficult and challenging problem of research, since the handwriting of a writer differ from script to script. Hence, there is need to develop a method which identifies the writer irrespective of the script of the handwritten document. The scripts considered in the proposed work are Roman (English), Devanagari (Hindi) and Kannada. The motivation behind considering Roman, Devanagari and Kannada Script is that, normally the persons belonging to Karnataka state will know English as an international language, Hindi as a national language and Kannada as a regional language. From literature survey it is clear that there is no unified approach for identifying the writer and script of the handwritten documents. Hence an attempt is made in this direction, since writer identification for single script handwritten documents are available in the literature.

Following are the objectives of the research:

- Text dependent writer identification
- Text independent writer identification
- Identify the writer and then script of the handwritten document
• Identify the script and then search for author of the handwritten document
• Document Image Retrieval based on Writer
• Script Identification

Writer and Script identification is one of the important pre-processing problem for automatic processing of multilingual document images, optical character recognition and writer identification. The problem of writer identification can be addressed as Text dependent and Text Independent Writer Identification. The problem of script identification may be addressed by considering bi-scripts, tri-scripts and multi-scripts documents. In multi-script and multilingual scenario, a writer knows to write in more than one scripts. There it is necessary to identify the writer of the handwritten document as well as the script underlying writer's document.

With the development in e-technology and widespread use of Internet in the last few years, the possibility of automatic archival of documents without any manual aid has become a challenging problem. In this view, the task of writer identification involves not only the identification of author of a piece of handwriting but also optionally in retrieval of handwritten samples from a database using an handwritten query as an input.

Applications of Proposed System:
• It can be adopted by forensic department for criminal investigation
• To check the authenticity of writer in bank cheque processing
• Indexing and Retrieval of handwritten documents written by the query writer.
• Automatic segregation / categorization of multilingual documents.
• Countering repudiations of handwritten mail
• Verification of claimed security

This thesis, addresses some of the issues related to the offline text independent writer and script identification in Roman-Devanagari-Kannada scripts written by the same writer. Feature extraction methods based on global and local approaches are
proposed. For subsequent classification purpose nearest neighbor (NN), kNN and linear discriminant analysis (LDA) classifiers are used.

1.5 Mathematical Tools and Techniques

The features used in the proposed methods are either textural or structural or combination of both.

Texture Features: Features used for pattern classification are extracted from the whole block of text (texture level) to be identified. The document image is simply considered as an image or a texture i.e., overall look and feel of the writing from a writer. Ex: Application of DCT, DWT, RT, GLCM etc are texture features.

Structural (Shape) Features: In this case the extracted features describe the shape/Structural properties of handwriting. Pattern classification based on localized features of writing which are inherent in the way writers specifically write characters and joins between characters. Eg: Average Height, Average Width, Average Slope, Average Stroke length and Average Stroke density etc are structural features.

The mathematical tools and techniques used for feature extraction in the thesis are given below:

1.5.1 Discrete Wavelet Transform (DWT)

Wavelet Transform is a family of transforms that satisfy specific conditions. Wavelet transform contains not just frequency information, but spatial information as well. Discrete wavelet transform (DWT) performs sub-band decomposition on an image in terms of spatial and frequency components and analysis of image from coarse to finer level. The different wavelet transform functions filter out different range of frequencies (i.e. sub bands). Thus wavelet can be used as a powerful tool, which decomposes the image into low frequency and high frequency sub band images. In the proposed method only three sub band images of single level decomposition namely Approximation (A) and Horizontal (H), Vertical (V) of DWT with Coiflet-5 family are considered.
1.5.2 Discrete Curvelet Transform (DCvT)

A discontinuity point affects all the Fourier coefficients in the domain. Hence the FT does not handle point discontinuities well. Wavelets are very effective in representing objects with isolated point singularities, but failed to represent line singularities. Using wavelets, it affects only a limited number of coefficients and the Wavelet Transform (WT) handles point discontinuities well. Discontinuities across a simple curve affect all the wavelets coefficients on the curve. Hence the WT doesn’t handle curves or line discontinuities well. Ridgelet transforms deal effectively with line singularities in 2-dimensional. Handwritten documents often contain curves rather than straight lines, so Curvelet transform is designed to handle it using only small number of coefficients. It allows to represent edges and other singularities along lines in a more efficient way than the other transforms. Hence the DCvT handles curve discontinuities well.

Candes and Donoho developed [31] a new multiscale transform called Curvelet Transform which was designed to represent edges and other singularities along curves much more efficiently than traditional transforms, i.e., concentrates the information content using fewer coefficients for a given accuracy of reconstruction. Curvelet Transform is developed by Candès and Donoho (1999) and is a new multi-scale representation most suitable for objects with curves.

Having an object in the domain $[0,1] \times [0,1]$, the approximation rates are given below using certain system of functions.

- Using the Fourier Transform:
  \[
  \| f - \hat{f}_m \|^2_2 = O\left( m^{-\frac{1}{2}} \right)
  \]

- Using the Wavelet Transform:
  \[
  \| f - \tilde{f}_m \|^2_2 = O\left( m^{-1} \right)
  \]

- Using the Curvelet Transform:
  \[
  \| f - \tilde{f}_m \|^2_2 = O\left( m^{-2} \log^3 m \right) \approx O\left( m^{-2} \right)
  \]
The curvelet decomposition [31] can be equivalently stated in the following form.

**i. Subband Decomposition**

We define a bank of subband filters $P_0, (\Delta, s \geq 0)$. The object $f$ is filtered into subbands:

$$f \mapsto (P_0 f, \Delta_2 f, \ldots).$$

(1.1)

The different subbands $\Delta_i f$ contain details about $2^{-2i}$ wide.

**ii. Smooth Partitioning**

We define a collection of smooth windows $w_Q(x_1, x_2)$ localized around dyadic squares

$$Q = \left[ k_1 / 2^s, (k_1 + 1) / 2^s \right] \times \left[ k_2 / 2^s, (k_2 + 1) / 2^s \right]$$

(1.2)

with $k_1$ and $k_2$ varying but $s$ fixed, produces a smooth dissection of the function into 'squares'.

Each subband is smoothly windowed into ‘squares' of an appropriate scale:

$$\Delta_s f \mapsto (w_Q \Delta s f)_{Q \in Q_s}$$

(1.3)

**iii. Renormalization**

The partitioning introduces redundancy, as a pixel belongs to 4 neighboring blocks. So, each resulting square is renormalized to unit scale.

$$g_Q = 2^{-s} (T_Q)^{-1} (w_Q \Delta s f), \quad Q \in Q_s$$

(1.4)

**iv. Ridgelet Analysis.**

Each 'square' is analyzed in the orthonormal Ridgelet system. This is a system of basis elements $\rho_{\lambda}$ making an orthobasis for $L^1(R^2)$:

$$\alpha_\mu = \langle g_Q, \rho_\lambda \rangle, \quad \mu = (Q, \lambda)$$

(1.5)
Chapter 1: Introduction

The curvelets scheme can be used to represent the curve distribution as a superposition of functions of various lengths and widths obeying the scaling law \[31\] width \(\approx\) length\(^2\). More details on Curvelet Transform are given in papers \[31, 67, 62\].

The image block size of \(m \times m\) was chosen since the digital Curvelet requires a \(2^n\) square image; this is discussed more detail in \[31\]. The texture features used in the algorithm are derived from the Discrete Curvelet Transform (DCvT) and is a discretization of their continuous Curvelet Transform \[67\], which uses a “wrapping” algorithm. The transform consists of four steps: application of a 2-dimensional fast Fourier transform of the image, formation of a product of scale and angle windows, wrapping this product around the origin, and application of a 2-dimensional inverse fast Fourier transform. The approximate scales and orientations are supported by a generic ‘wedge’.

The DCvT can be calculated to various resolutions or scales and angles. Two parameters are involved in the digital implementation of the Curvelet Transform: number of resolutions and number of angles at the coarsest level. Several features can be calculated on the curvelet coefficients. The most common statistics calculated on wavelets are mean and standard deviation. A discrete Curvelet frequency tiling domain for 16 angular orientations and 5 levels of scale and a wedge sample is shaded as shown in Figure 1.1.

![Figure 1.1: Discrete Curvelet frequency tiling domain for 16 angular orientations and 5 levels, a wedge sample is shaded.](image)
1.5.3 Discrete Cosine Transform (DCT)

The DCT expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies that are necessary to preserve the most important features. The DCT concentrates energy in its coefficients. With an input image, \( I(x, y) \), the two dimensional DCT coefficients for the transformed output image \( C(p, q) \) are computed according to equation 1.6 shown below. In the equation, \( I \), is the input image having \( M \)-by-\( N \) pixels, \( I(x, y) \) is the intensity of the pixel in row \( x \) and column \( y \) of the image and \( C(p, q) \) is DCT coefficient in row \( p \) and column \( q \) of the DCT matrix.

\[
C(p, q) = \alpha(p)\alpha(q)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x, y)\cos\left(\frac{\pi(2x+1)p}{2M}\right)\cos\left(\frac{\pi(2y+1)q}{2N}\right)
\]

(1.6)

for \( p, q = 0, 1, 2, \ldots, N-1 \), \( \alpha(p) \) and \( \alpha(q) \) are defined as

\[
\alpha(p) = \frac{1}{\sqrt{M}} \text{ for } p = 0 \text{ and } \frac{2}{\sqrt{M}} \text{ for } p \neq 0
\]

(1.7)

\[
\alpha(q) = \frac{1}{\sqrt{N}} \text{ for } q = 0 \text{ and } \frac{2}{\sqrt{N}} \text{ for } q \neq 0
\]

(1.8)

1.5.4 Radon Transform (RT)

Radon transform (RT) computes the projection of input word image \( I(x, y) \) as a set of line integrals from multiple sources along parallel paths in the specified direction. The sum of energy of the pixels in each direction is line integral and is given by \( R(\theta, \rho) : \)

\[
R(\theta, \rho) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y) \delta(x \cos \theta + y \sin \theta - \rho) \, dx \, dy
\]

(1.9)

where \( \rho = x \cos \theta + y \sin \theta \) and \( \delta \) is the Dirac delta function with \( 0 \leq \theta < \pi \) and \( \delta \in [-\infty, \infty] \).
1.5.5 Gray Level Co-occurrence Matrix

Texture is related to properties such as smoothness, coarseness, roughness, and regular patterns. The features based on first order histogram probability are mean, standard deviation, skew, energy, and entropy. One of the approach of measuring texture is to use second-order histogram of the gray levels based on a joint probability distribution model. The second-order histogram provides statistics based on pairs of pixels and their corresponding gray levels. The second-order histogram methods are also referred to as gray level co-occurrence matrix or gray level dependency matrix methods. Spatial gray level co-occurrence estimates image properties related to second-order statistics. These features are based on two parameters: distance and angle. The distance is the pixel distance between the pairs of pixels that are used for the second order statistics, and the angle refers to the angle between pixel pairs. Typically, four angles are used corresponding to vertical, horizontal, and two diagonal directions. The pixel distance chosen depends on the resolution of the image and the coarseness of the texture of interest, although it is typical to use 1 or 2.

One of the defining qualities of texture is the spatial distribution of gray levels [43]. The use of statistical features is therefore one of the early methods proposed in the machine vision literature. \( I \) is a NXN image with G gray levels \( \{I(x,y), 0 \leq x \leq N-1, 0 \leq y \leq N-1\} \). Center pixel \((x,y)\) and neighbouring 8 pixels in image \( I \) which are at a distance 1 from the center pixel at eight different angles is shown in Figure 1.2.

![Figure 1.2: Relationship between neighboring pixels](image)

A statistical method that considers the spatial relationships of pixels is the Gray-Level Co-occurrence Matrices (GLCM) of the image, also known as the gray-level
spatial dependence matrix. Five distances $d = 1, 2, 3, 4, 5$ and four directions $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ are used to create twenty GLCMs. For each GLCM matrix common statistical properties can be extracted. Few of the common statistical methods applied to co-occurrence probabilities are given in equations 1.10, 1.11, 1.12 and 1.13, where $c(x_i, y_j)$ is the $(i,j)^{th}$ entry in the GLCM and is probability that a pixel with value $x_i$ will be found adjacent to a pixel with value $y_j$ at a distance $d$ and an angle $\theta$ or $c(x_i, y_j)$ is the $(i,j)^{th}$ element in the co-occurrence matrix normalized by dividing the number of pixel pairs in the matrix, for a given distance and angle [43].

When GLCM method is applied on the input binary image, number of gray levels are only two either 0 or 1 and hence the GLCM of input binary image is a 2X2 matrix for a particular distance and direction, since the number of gray levels are two, $i$ and $j$ vary from 1 to 2, and $x_i$ and $y_j$ varies from 0 to 1 for binary image, where the black pixels having value 0’s correspond to object and white pixels having value 1’s correspond to background.

The co-occurrence matrix reveals certain properties about spatial distribution of the gray levels in the texture image. Haralick [43] has proposed a number of useful texture features that can be computed from the co-occurrence matrix and those are energy, contrast, homogeneity, correlation and entropy etc. Few of them are given below:

i. **Energy**

$$\text{Energy} = \sum_i \sum_j c(x_i, y_j)^2$$

(1.10)

where $c(x_i, y_j)$ is the probability that pixel with value $x_i$ is neighbor to pixel with value $y_j$ at a distance $d$ an angle $\theta$. This statistic is also called Uniformity or Angular second moment. It measures the textural uniformity that is pixel pair repetitions. It detects disorders in textures. Energy reaches a maximum value equal to one. High energy values occur when the gray level distribution has a constant or periodic form. Energy has a normalized range. The GLCM of less homogeneous image will have large number of small entries.
ii. Contrast

\[
Contrast = \sum_i \sum_j (x_i - y_j)^2 c(x_i, y_j)
\]  

(1.11)

It is the variation between the highest and the lowest values of a contiguous set of pixels. It measures the amount of local variations present in the image. A low contrast image presents GLCM concentration term around the principal diagonal and features low spatial frequencies.

iii. Homogeneity

\[
Homogeneity = \sum_i \sum_j \frac{1}{1+(x_i - y_j)^2} \cdot c(x_i, y_j)
\]  

(1.12)

This statistic is also called as Inverse Difference Moment. It measures image homogeneity as it assumes larger values for smaller gray tone differences in pair of elements. It is more sensitive to the presence of near diagonal elements in the GLCM. It has maximum value when all elements in the image are same. GLCM contrast and homogeneity are strongly, and inversely correlated in terms of equivalent distribution in the pixel pairs population. It means homogeneity decreases if contrast increases while energy is kept constant.

iv. Correlation

\[
Correlation = \sum_i \sum_j \frac{(x_i - \mu_x)(y_j - \mu_y) c(x_i, y_j)}{\sigma_x \sigma_y}
\]  

(1.13)

where \( \mu_x, \mu_y, \sigma_x \) and \( \sigma_y \) are the means and standard deviations of \( c_x \) and \( c_y \).

\[
\mu_x = \sum_i \sum_j x_i c(x_i, y_j), \quad \mu_y = \sum_i \sum_j y_j c(x_i, y_j),
\]

\[
c(x_i) = \sum_j c(x_i, y_j), \quad c(y_j) = \sum_i c(x_i, y_j),
\]

\[
\sigma_x = \sqrt{\sum_i \sum_j (x_i - \mu_x)^2 c(x_i, y_j)} \quad \text{and} \quad \sigma_y = \sqrt{\sum_i \sum_j (y_j - \mu_y)^2 c(x_i, y_j)}
\]

It is a measure that a pixel is correlated to its neighbor over the whole image. The correlation feature is a measure of gray tone linear dependencies in the image.
1.6 Classifiers:
When an unknown sample is submitted to a recognition system, it decides to which class the sample belongs. To perform this, the given system is trained using set of samples whose classes are known (training samples), then unknown test sample's class is decided. In the recognition or classification phase the features extracted from test sample are compared with the training samples whose classes are already known. A system that performs this type of classification process is called classifier. There are two types of classification tasks generally, one supervised classification in which pattern is identified as a member of predefined class and another unsupervised classification in which pattern is assigned to a hitherto unknown class. The type of classifier used for classification affects the recognition process a lot. Higher the discrimination power of the classifier, the better the performance. Some supervised classifiers are used in this thesis and are given below:

1.6.1 Nearest Neighbor (NN)
Nearest neighbor is one of the basic classifier. Basically NN stores the training data \( X \). Then finds the minimum distance \( d \) between training sample \( X \) and testing sample \( Y \) using different distance functions or similarity measures such as Euclidean distance, city block distance, cosine, correlation etc. Some of them are given below:

i. Euclidian Distance
\[
d(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}
\]  
(1.14)
where \( n \) is feature vector size.

ii. City Bock Distance
\[
d(X,Y) = \sum_{i=1}^{n} |X_i - Y_i|
\]  
(1.15)
where \( n \) is feature vector size.
NN classifier assigns the test sample to the training class to which it is nearest in distance or closest in similarity.
1.6.2 K-Nearest Neighbor

KNN is one of the Statistical pattern recognition approach. KNN uses the same functions as used in nearest-neighbor (NN) classifier, but by examining the labels on the k nearest training samples and taking a majority voting to decide the class of the test sample. It is good to choose k value as an odd number to avoid tie votes. In particular NN classifier is a special case of KNN classifier.

1.6.3 Linear Discriminant Analysis

Linear Discriminant Analysis is one of the most commonly used classification technique. It preserves class discriminating information to the higher extent by reducing dimensionality of feature space. It also optimizes separability between the classes by maximizing the ratio of between-class variance to the within class variance. In this thesis, LDA is employed on a dataset $X=[x_1, ..., x_i]$ of dimension $N \times n$ ($N=K\times M$) and the sample $x_i$ belongs to one of the class $C_i$, where $i = 1$ to $K$, $K$ is number of classes. Further, the dimension of $x_i$ is $m \times p$, where $m = 1$ to $M$, $M$ is number of samples of each class and $p = 1$ to $n$, $n$ is total number of features of a sample. Then the classification function is defined as

$$ f(X) = Z^T X $$

where $Z$ is the linear projection, and which maximizes between-class scatter

$$ S_{\text{between}} = \sum_{i=1}^{K} m_i (\mu_i - \mu)(\mu_i - \mu)^T $$

whereas it minimizes the within-class scatter

$$ S_{\text{within}} = S_1 + S_2 + ... + S_K = \sum_{i=1}^{K} \sum_{x \in C_i} (X - \mu)(X - \mu)^T $$

where $\mu_i$ is the mean over class $C_i$, $\mu$ is the mean over all samples, and $m_i$ is the number of samples in class $C_i$. The classification of a new sample $X$ of class label $\omega \in C_i$, is done based on the nearest neighbor classification rule. For this purpose, the Euclidean distance $d$ of $f(X)$ and the centers $V_i = Z^T \mu_i$ in LDA space are compared

$$ \omega = \text{argmin}_{1 \leq i \leq K} d(f(X), V_i) $$
1.6.4 Support Vector Machine

The SVM is a hyper plane classifier with the aim of maximizing a geometrical margin of hyperplane. Generally, an SVM classifier is a two class (binary) linear classifier in kernel induced feature space. Subset of training samples closest to margin that determines optimal hyperplane are called support vectors. It involves mapping input vectors $X$ into a high dimensional feature space $Z$ through nonlinear transformation. Linear decision function is used for separable training points.

A decision function of an SVM has the form

$$f(x) = w.x + b$$

(1.20)

where $w$ is the weight vector and $b$ is a bias (-$b$ is also called threshold). Classification is given by $\text{sgn}(f(x_i))$ which correspond to +1 or -1 for $x_i \in \mathbb{R}^n$. Thus for all points $x_i$, $i=1,2,\ldots,m$ with class labels $y_i \in \{+1, -1\}$. An SVM finds the hyperplane with maximum margin from the training set. Multiclass classification is accomplished by combining multiple binary SVM classifiers.

To deal with non-separable points, the margin constraints are relaxed. Different kernel functions are used which use different methods of support vector machines. Nonlinear kernel functions are:

1. Sigmoid Kernel:

$$k(x, x_i) = \tanh(\kappa(x.x_i) + c)$$

(1.21)

2. Polynomial Kernel:

$$k(x, x_i) = (\kappa(x.x_i) + 1)^p$$

(1.22)

polynomial classifier with order $p$.

3. RBF Kernel:

$$k(x, x_i) = \exp\left(-\frac{|x-x_i|^2}{2\sigma^2}\right)$$

(1.23)

where $\kappa$, $\sigma$, $c$ are hyperparameters which have to be tuned according to dataset. Depending on the level of nonseparability of dataset, the kernel function is chosen. In this thesis the SVM with Radial Basis Function (RBF) kernel is used.
1.7 Organization of the Thesis

The thesis is organized into seven chapters. The offline handwritten samples are considered for the proposed research work. In Chapter 1, a brief description of background of document image processing, writer identification, script recognition, document image retrieval based on writer, a detailed literature survey of the state of the art methods, the motivation and objectives of the research problem, related mathematical tools and techniques, and classifiers used for research work in this thesis are discussed.

The Chapter 2 describes the problem of text dependent writer identification addressed at word level based on combination of global and local features. The directional multi-resolution spatial features of Radon Transform and Discrete Cosine Transform and structural features such as aspect ratio and on-pixel ratio of word images are extracted, and the classified using NN (Nearest Neighbor) classifier with Euclidian distance.

The Chapter 3 presents two methods to address the problem of text independent writer identification at block (whole Page, half page and quarter page) level. Firstly, method based on Directional Stroke Based Features and secondly, method based on Gray Level Co-occurrence Matrix Based Features, the created dataset of Roman, Devanagari and Kannada documents along with Standard IAM dataset of handwritten English documents of 100 writers is used for evaluation of the two proposed techniques and classified using NN, kNN and LDA classifiers.

The Chapter 4 provides two approaches to address the problem of writer and script identification at block (whole Page, half page and quarter page) level for the writers who know to write in more than one script. The two approaches are: 1) Writer and Script identification: writer of the input handwritten document is identified first then the corresponding script. 2) Script and Writer identification: script in which the input handwritten document is written is identified first then the underlying writer. To perform this, two feature extraction techniques are employed. One, Directional Stroke Based Features and another, Gray Level Co-occurrence Matrix Based Features. The created dataset of Roman, Devanagari and Kannada documents
written by same 100 writers is used for evaluation of the two proposed techniques. The accuracy of the proposed techniques has been estimated by NN, kNN and LDA classifiers for writer and script identification.

In Chapter 5, document image retrieval based on writer using directional multi-resolution spatial features of Discrete Wavelet Transform and correlation of GLCM of input images is described. NN classifier with Euclidean and City block distance measures are employed to retrieve the documents.

In Chapter 6, Script recognition at handwritten text block level based on two potential feature extraction techniques are described. Firstly, Script Recognition using Discrete Curvelet Transform Features, Secondly, Script Recognition using GLCM and DWT Features. The accuracy of the proposed techniques has been estimated by using NN, LDA and SVM classifiers for six Indic scripts viz. Roman, Devanagari, Kannada, Telugu, Tamil and Malayalam.

Finally, the conclusion and future scope of the present research work are discussed in Chapter 7.

The outcome of research work is expected to be useful for writer identification, writer based document image retrieval and script identification.