CHAPTER 1
INTRODUCTION

Artificial Intelligence (AI) techniques are applied to optical character recognition (OCR), handwriting recognition, pattern recognition and speech recognition domains. Many research works are being carried out through last two and half decades on different aspects of OCR system. Research work in document processing and OCR is being seen and felt every day. The research initially started with printed text and later extended to handwritten texts also.

1.1 DOCUMENT IMAGE PROCESSING

Document image processing (DIP) is the process of converting the scanned images of machine printed or handwritten letters, symbols and numerals into a format such as American Standard Code for information Interchange (ASCII), which could be processed by a computer. Documents include machine printed documents (like letters, reports and books), offline hand written documents (like personal letters, addresses on envelopes and notes in the margins of documents) and online handwritten documents (written on Personal Digital Assistants (PDA) or tablet Personal Computers (PC)).

Documents are converted to computer readable form through DIP, which encompasses the process of OCR. A well-known DIP product is the OCR software that recognises characters from a scanned document. OCR makes it possible for a user to edit or search the contents of a document. An automated OCR system could reduce the time needed to convert a document to computer readable form to 25 percent of the time a human needs to enter the same data manually (Elisa Barney Smith).
Document image analysis (DIA) performs overall interpretation of the document images. DIA aims to develop algorithms and processes through which computers could automatically read and develop some basic understanding of documents. Elliman and Lancaster (1990) identified two main processing units in the OCR system – a character separator and a character classifier. Character separation, also known as segmentation, could work in fixed (constrained) spacing mode, where the character size is known in advance or variable (arbitrary) spacing mode, where no prior information on character size could be assumed.

Kurt Alfred Kluever (2008) opined that OCR is the process of translating images of hand-written, typewritten or printed text into a format understood by machines for the purpose of editing, indexing, searching and compression. An OCR system takes a document image as input and generates a character set in editable form as output.
Majority of errors in the results of OCR systems are due to image degradation of the scanned characters. Main causes for receiving images, that are not recognised could be bad quality of printing, use of proportional fonts with serifs and inappropriate scanner settings. Noise is introduced during image transmission or by medium of transfer also. Bokser (1992) presented an OCR engine, which was omni font and reasonably robust on individual degraded characters.

Developing well performing offline OCR systems has been an elusive goal for researchers. This is true for recognition of unconstrained texts, where the vocabulary might be large. Contemporary OCR systems cope with different fonts. But if the character structure changes or the character images become either touching or broken, the recognition rate falls considerably and turns that as major challenge for the OCR systems.

1.2 HANDWRITTEN CHARACTER RECOGNITION

Since the beginning, paper prevailed as a medium for writing. The convenience of paper, its pervasiveness for communication and the quantity of information already on paper emerged for quick and accurate methods to automatically read the information and adapt it into electronic form (Albadr and Haralick 1995).

Handwritten character recognition (HCR) is the technique by which a computer system could recognize characters and other symbols written in natural handwriting. Early HCR techniques were based on template matching, simple line and geometric features, stroke detection and their derivatives extraction. It could be divided into online recognition and offline recognition.

Online recognition is the process of recognizing the characters as they are being written. Examples of systems that employ online recognition include Apple Newton and Palm Pilot. Online recognition is the interpretation of the handwriting real time, as it is created using electronic pen. In online recognition, the computer could trace the process of writing.
Hence the strength and the order of each segment when it is written could be recorded for recognition. Larger amount of input data available makes online recognition easier than offline recognition.

Offline recognition is recognising the handwritten text, which was written already. That is, the input to the recognition engine is a binary or grayscale image, containing the handwritten text. Offline recognition consists of four processes - acquisition, segmentation, recognition and post-processing. Due to lack of stroke and timing information, offline recognition is a much harder problem than online recognition. It is less accurate than online recognition. Because, only spatial information is available for offline recognition, while both spatial and temporal information are available for online recognition.

Most of the methods designed for HCR mainly address the issues of global skew, but not robust for local skew with concern in skew angle detection. Some methods designed for specific applications like character types with special features are computationally expensive. Although lot of efforts had been paid to develop methods to automatically convert paper documents into electronic form, many documents that are easy for humans to read still have only 92 percent recognition accuracy (Elisa Barney Smith). Handwritten documents are difficult to be processed because they lack a specific structure.

Character identification is a pattern recognition application with the aim of simulating human reading capabilities on machine printed and handwritten cursive texts. Systems in the market are now available for interpreting a wide range of writing styles and character sets including Chinese, Japanese, Korean, Cyrillic and Arabic. Character recognition is a primary and essential component in document processing (Bunke et al 1994, Bunke and Wang 1997, Lu 1995, Casey and Eric Lecolinet, 1996). The task of the character recognition is to recognize omni font characters which might be separated, touching, overlapping or broken.

Low accuracy rates are arrived due to poor quality in printing, scanning, photocopying and FAXing the documents. Problems are severe with illegible writing and similar shaped
characters like U and V. A general solution to this problem involves sophisticated recognition algorithms as well as syntax and semantics to resolve ambiguities.

In DIP, segmentation is the process that locates and extracts the entities to be recognized. Segmentation procedures are more heuristic than recognition procedures; and proper segmentation requires a prior knowledge of the patterns that form meaningful units. Segmentation of unconstrained handwritten text is a complicated and diverse problem by the nature of handwriting. Handwriting could vary depending on the user skill, disposition and cultural background. There is no universally accepted solution in automatic handwritten document recognition systems (Gomathi alias Rohini 2012).

1.3 RECENT TRENDS AND APPLICATIONS

As exhaustive research and development on HCR went by and with several conferences and workshops, modern techniques advanced rapidly. The subject of HCR gained considerable momentum and grew swiftly. People could write the way they normally did; characters need not have to be written like specified models. As of now, many new algorithms and techniques in preprocessing, segmentation and classification have been developed.

In many industries, there is significant demand for automatic processing of different types of handwritten materials. As research on HCR and OCR got advanced and since a lot of data were written by hand and they had to be fed into the computer for processing, demand on handwriting identification also increased.

Even though human handwriting processing has gone through substantial development during the past decades, relatively well performing systems are available only for vertical applications, where the utilized vocabulary is very narrow and well defined such as
- addresses on envelopes
- bank check automation
- medical prescriptions
- handwritten invoices and tax forms
- sorting large document datasets
- FAX messages
- signature verification
- PDA or tablet PC technology
- conversion of books to digital libraries
- handwritten historical manuscripts
- old degraded documents

Currently organizations are interested in implementing digital mailrooms to improve the efficiency of paper-intensive workflows and reduce the burden of information processing. Multi-script document recognition, historical document transcription, word spotting, video document analysis and recognition are other challenging application areas. HCR systems also find applications in newly emerging areas like development of electronic libraries, multimedia database and systems which require handwriting data entry.

1.4 PREPROCESSING

The unconstrained handwritten text image obtained as scanner output may have some impurities, due to quality difference in paper, ink, color, pen and scanner. Those impurities are removed in the preprocessing stage. Preprocessing plays an important role in HCR. The scanned image undergoes a series of operations like filtering, binarisation, thinning and smoothing to produce a cleaned image for segmentation and recognition. Poor preprocessing increases the complexity of techniques used in further stages.

Skewness is formed in images, when the document pages are not fed properly into the scanner. A small skew angle in a document image would result in the failure of segmentation, since the distance between the characters get reduced. Most of the OCRs and other document retrieval systems are sensitive to skewness in document images. It is important to detect and correct the skewness. Skew in scanning process, skewed handwriting and skewed original document are some of the reasons for skewness in document images. The noises and deviation
in the document resolution or types are main challenges in skew detection and correction (Lee et al 1994).

Noise may occur in foreground or background of an image due to paper quality, typing machine or scanning process. Uneven contrast, show through effects, interfering strokes, background spots, aging of documents, humidity absorbed by paper and uneven backgrounds are examples of background noise in the scanned document images. Such degradations could destroy the blank spaces between lines and words. Noise reduces the accuracy of subsequent tasks of OCR. There are many challenges addressed in degraded handwritten documents image processing.

Document image binarisation converts a gray-scale document image into a binary document image. It is carried out as a preprocessing task in DIP, which aims to segment the foreground text from the document background. Binarisation marks the foreground pixels with intensity ON and background pixels with intensity OFF. Data reduction reduces the quantity of data of the original image to fewer bits, to be retained in the preprocessed image to effectively utilize the resources like memory and time. It is achieved through binarisation. This process works on the basis of threshold value.

1.5 SEGMENTATION

Text segmentation is the process through which the text component of an image is isolated from the background. The document image is first preprocessed and passed through the process of text segmentation and character recognition. In the preprocessing step, skew correction, background subtraction and binarisation techniques remove majority of noise from the input image. The cleaned image obtained after preprocessing is the input for segmentation, which decomposes the image into subcomponents like lines, words and characters based on various segmentation approaches.

For proper reconstruction of the editable text lines from the recognized characters, the line is first segmented. From the segmented line, words are segmented and from words,
characters are segmented. Sridhar and Kimura (1995) described that recognition of words in a document follows a hierarchical scheme as given below:

1. removing tilt (skew) of the document
2. extracting lines from the document
3. removing skew from each line
4. extracting words from each line
5. extracting characters from each word
6. concatenating characters
7. recognizing characters to recognize each word

Lot of methods had been proposed for machine printed document segmentation (Kim et al 2002). But they could not be directly applied to handwritten documents because of the spatial variability in handwriting. The impressive results of OCR techniques on printed scripts are not usable for handwritten documents. If handwritten documents are less structured than the printed ones, the segmentation process could be benefited by the use of prior knowledge about the structure of the document and contextual information (Stéphane Nicolas et al 2006). The segmentation process could handle perfectly fitting documents with varying skew angle, overlapping words and documents with frequently appearing accents.

Higher the segmentation accuracy, greater the character recognition. A few existing methods dedicated to handwritten documents focus on a particular type of documents or a particular task of segmentation. Human writing style varies from person to person and even for the same person, it varies depending on the mood, speed, environment etc.

Due to aging factor and the quality of ink and paper, the documents are generally degraded. Many manuscripts are written on both the sides of the paper producing back-to-front ink transposition effect. Some documents have large black borders, adhesive marks or damaged papers. Absence of knowledge about the content of the document poses a lot of difficulties for segmentation. These factors make the segmentation process a challenging one.
There are several methods of segmentation such as white space and pitch, projection profile analysis and connected component (CC) analysis. The most common approach is projection profile analysis, since it is simple and fast. Cursive handwritten texts are also segmented using more advanced methods like Hidden Markov Model (HMM), Artificial Neural Networks (ANN), linear programming, genetic programming and contextual methods. To achieve greater accuracy on complex problem domains, segmentation and recognition could not be treated independently.

1.5.1 Line Segmentation

In text line segmentation, the document image is divided into different regions each one representing a text line. Once the lines of handwritten text have been separated, they are subjected to word segmentation, character segmentation and other indexing steps for recognition or retrieval operations. Text line segmentation provides essential information for further document processing steps like baseline detection, skew identification and correction, text feature extraction, word segmentation and character segmentation.

When dealing with handwritten text, line segmentation has to deal with some obstacles like skewed lines (variability in skew angle between different text lines, in the same text line and in the skew directions), curvilinear lines, fluctuating lines, lines close to each other, touching and overlapping components, overlapping words, touching adjacent text lines, accents from other languages like French and Greek, words or letters between lines, irregularity in geometrical properties like line width, height and leftmost most position in the line and distance between words and lines.

The existing methods of text line segmentation could be grouped into two classes: top-down approach and bottom-up approach. Top-down methods start from the whole image and iteratively subdivide into smaller blocks to isolate the components. Bottom-up methods group small units of image (pixels, connected components, etc.) into text lines and then text regions.
Most of the existing methods have limitations when applied to unconstrained handwritten documents; because they more or less assume horizontal, straight, parallel and untouched text lines. Methods based on CCs are faster, but suffer from touching or close proximity of components. A few methods like projection profile, Hough transform and smearing failed to segment the text lines properly in case of very closely spaced lines. Some additional techniques were required as post processing steps to isolate touching text lines.

The overall performance of a HCR system strongly relies on the results of text line detection process. If text line detection does not give good results, the accuracy of further segmentation and recognition procedures would be affected. Thus, there is a need for an optimal text line detection stage.

1.5.2 Word Segmentation

Words segmentation approaches for handwritten text lines are based on various heuristics and assumptions. Main assumptions that most word segmentation approaches adopt are:

i) the document is already segmented into text lines

ii) each CC belongs to only one word and

iii) gaps between words are greater than gaps between characters of the same words.

Efficiency of DIA is affected by the precision of word segmentation process. Word segmentation of unconstrained handwritten text lines is an important stage, due to high variability and uncertainty of human writing style. Machine-printed text has inter-word gaps (gaps between words) larger than intra-word gaps (gaps within words). In handwritten text, various inter-word and intra-word spacing exist from different writing style. It is observed that in some cases of handwritten document images, the character gap was greater than the word gap. Most of the word segmentation techniques consider a spatial measure of the gap between successive CCs and define a threshold value to classify the inter-word and intra-word gaps (Seni and Cohen 1994).
Words extraction from handwritten text lines involves calculation of a threshold for gap classification. The text line is decomposed into series of single connected or horizontally overlapped CCs (Marti and Bunke 2001, Seni and Cohen 1994). Following decomposition, the threshold value decides the gap as inter-word gap or intra-word gap (Mahadevan and Nagabushnam 1995) and finally the words are extracted. The problem of word segmentation concerns in the way the distance between adjacent components is calculated and the approach used to classify the calculated distances as inter-word and intra-word gaps.

Separating handwritten text into words is challenging. Because handwritten text lacks the uniform spacing normally found in machine-printed text. Text characteristics could vary in font, size, shape, style, orientation, alignment, texture, color, contrast and background information. These variations turn the process of word detection complex and difficult (Kavalliieratou et al 2002).

1.5.3 Character Segmentation

In character segmentation phase, an image of sequence of characters is decomposed into sub images of individual symbol or character. A character is a pattern that resembles one of the symbols the system is designed to recognize (Casey and Erics 1996).

Segmentation of cursive handwritten words into isolated characters is a process, which decides whether the isolated components are characters, parts of characters or noises. Perfect character segmentation is required to reduce the error rate of recognition. Researchers have acknowledged that correct segmentation is the base of higher recognition accuracy. Tanzila Saba et al (2010) opined that character segmentation is an important task in the field of OCR.

Character segmentation is necessary to isolate each character for recognition. Therefore character segmentation should be closely coupled with character recognition. In 1996, Casey and Lecolinet surveyed and defined the following three categories of offline character segmentation methods:
1. Classical or Dissection Approach: It identifies the segments based on ‘character like’ properties. The word image is divided into segments, each representing a character. It precedes classification and recognition and has no feedback information.

2. Recognition based Approach: It searches the image for components that matches classes in its character. It splits words into segments and passes each segment to a classifier. If the classification is not satisfactory, again segmentation is called with feedback information.

The classical approach and recognition based approach together is called as analytical approach. The most accurate analytical systems belong to the recognition based segmentation group. Most of the algorithms constructed in recent years fall in this group.

3. Holistic Approach: It seeks to recognize words as a whole to avoid the need of segmenting into characters. It is based on the analysis of the order of ascender, descender, loops, vertical strokes and their count. It is heavily on dictionary searching that is costly and prone to mislead by spelling errors.

Segmentation algorithms for unconstrained handwritten words could be generalized into two categories (Liang et al 1994) - external segmentation, in which, character boundaries are determined prior to recognition and holistic segmentation in which, segmentation and recognition are carried out at the same time and the final character boundaries are determined dynamically by semantic analysis and classification performance.

Low quality of some scanners and printers and wide variety of fonts and poor thresholding process in binarisation cause problems in segmenting touching, overlapping, merged and broken characters in cursive handwriting. Touching characters make the segmentation process more crucial. Two touching character segmentation methods are explicit
character segmentation and implicit character segmentation. Segmentation of touching cursive characters is more complicated compared to linked or non-touching cursive characters.

The complexity of character segmentation stems from the wide variety text styles and image characteristics also. Moreover, when a document is composed of multiple languages, it is difficult to segment characters due to differences in each language. It is a challenge to develop a practical system with high recognition accuracy, independent of the quality of input documents and character fonts. (Tappert et al 1990). Due to non uniform spacing between adjacent words in handwritten document images, there are cases that parts of adjacent words are merged (under segmentation) and cases where parts of the same word are split into two or more words (over segmentation).

1.6 PROJECTION PROFILE

For a binary image S with height N and width M, the projection profile is the count of foreground pixels in each row or column. Projection profile can be a horizontal projection profile or vertical projection profile.

Horizontal projection profile is defined as the sum of black pixels perpendicular to the x axis. This is represented by vector $P_h$ of size N as Equation (1.1). Horizontal projection profile will have peaks at text line positions and troughs in between successive text lines.

$$P_h[j] = \sum_{i=1}^{N} S[i,j]$$ (1.1)

Vertical projection profile is defined as the sum of black pixels perpendicular to the y axis. This is represented by vector $P_v$ of size M as Equation (1.2). The height of $P_v$ depends on the font size. Vertical projection profile will have peaks at character positions and troughs in between successive characters.

$$P_v[i] = \sum_{j=1}^{M} S[i,j]$$ (1.2)
The horizontal or vertical projection profile is represented as a histogram of the number of black pixels along the horizontal or vertical axis. A histogram is the graphical representation, which plots the intensity distribution of an image.

![Graph](image)

**Figure 1.2 Projection Profile**  
Source: Naazia Makkar and Sukhjit Singh (2012)

The histogram for a very dark image will have majority of its data points on the left side and center of the graph and for a very bright image with few dark areas will have most of its data points on the right side and center of the graph.