1.1 Introduction to Segmentation

People have always tried to develop machines which could do the work of and as a human being. The reason is obvious, man has been very successful in using the machines which are developed to reduce the amount of physical labor needed to do many tasks. With the invention of computer, it became reality for the fact that machines could also reduce the amount of mental labor needed for many tasks. Over the past four to five decades, with the development of different computers ranging from one’s capable of becoming the world chess champion to one’s capable of understanding speech, it has come to seem as though there is no human mental faculty which is beyond the ability of machines. With the advancement of technology, the hardware rate heavily came down. Now huge amount of data can be stored in a small chip. With this revolution in hardware technology, computers are widely used to preserve documented data. Earlier transmission and storage of data was done by paper documents, but the advancement of software and hardware technology, paper documentation is replaced by computers. But still there are large numbers of documents which are on paper and are not preserved in computers. Still, there is a gap between paper documents and computer storage. To integrate these two mediums of information flow, the solution is to make computer capable to “read” paper documents. Arica and Yarman [2001] stated that machine simulation of human reading has been a very challenging research field since the advent of digital computers. This is one of the areas, which has been the subject of intensive research for the last three decades, yet it is still far from the final frontier. So, works are still going on in this direction.
Automatic processing means the work is processed automatically with negligible or less supervision of human. The objective of automatic document processing is to recognize text, graphics and pictures in digital images and extract the intended information, as would a human. Textual and graphical are two categories of document processing dealing, respectively, with the text and the graphics components of a document image. Document processing leads to the theory of optical character recognition (OCR). Chaudhuri and Pal [1998] concluded that the objective of optical character recognition (OCR) is automatic reading of optically sensed document text materials to translate human readable characters to machine readable codes. Lu [1995] explained that textual processing includes the steps as to determine the skew (any tilt at which the document may have been scanned), finding columns, paragraphs, text lines and words and then to perform optical character recognition. But before recognizing the character it is required to segment area which would have that character. So to have such area, a perfect segmentation of characters is required, before the recognition of individual characters. Therefore segmentation techniques are to apply to word images before actually putting those images to reorganization process.

1.2 Natural Language and its Processing

Language is a term most commonly used to refer to so called “natural languages” i.e. the forms of communication considered to humankind. In linguistics, the term is extended to refer the type of human thought process which creates and uses language. Essential to both meanings is the systematic creation and usage of systems of symbols, each referring to linguistic concepts with semantic or logical or otherwise expressive meanings. The most obvious demonstrations are spoken languages such as English or Chinese. However, there are also written languages. When discussed more technically as a general phenomenon then, “language” always implies a particular type of human thought which can be present
even when communication is not the result, and this way of thinking is also sometimes treated as indistinguishable from language itself. In Western philosophy for example, language has long been closely associated with reason, which is also a uniquely human way of using symbols. According to ancient Greek philosophical terms, the same word, logos, was used as a term for both language or speech and reason. Languages change with time and live, die, and move from place to place. Any language that opposes to change is categorized as a dead language. Conversely, any language that is in a continuous state of change is known as a living language or modern language.

1.2.1 Natural Language Processing

Natural language processing is a field of science and linguistics concerned with the interaction between computers and human languages. Natural language generation systems convert information from computer databases into readable human language. The term “natural” language refers to the languages that people speak, like English and Japanese and Hindi, as opposed to artificial languages like programming languages or logic. “Natural Language processing” (NLP), programs deal with natural language in some way or another. The study of human languages developed the concept of communicating with non human devices. The goal of NLP is to design and build software that will analyze, understand and generate languages that humans use naturally. Natural language understanding systems convert samples of human language into more formal representations such as parse trees or first order logic that are easier for computer programs to manipulate. In theory, natural language processing is a very attractive method of human – computer interaction. Early systems working in restricted “blocks worlds” with restricted vocabularies, worked extremely well, leading researchers to excessive optimism, which was soon lost when the systems were extended to more realistic situations with real world ambiguity and complexity. If one goes to history of NLP, then
one would find a paper titled "Computing Machinery and Intelligence" and this was written by Alan Turing published in 1950 as Turing [1950]. Nowadays, the contents of this paper are known as the “Turing test” as a criterion of intelligence. This criterion depends on the ability of a computer program to impersonate a human in a real time written conversation with a human judge. In 1954, sixty Russian sentences were fully translated into English and that experiment was known as the “Georgetown experiment”. There was a slow progress in this field till 1980, when the first statistical machine translation system was developed. In late 1980, machine learning algorithms for language processing were introduced. This revolutionized the concept of NLP. This increased the computational power. IBM research groups were working in the field of machine translation and these type of systems were able to translate some documents successfully. Nowadays, the stress is on less supervised or unsupervised learning algorithms. But, this type of unsupervised learning is much more difficult to achieve.

NLP has significant overlap with the field of computational linguistics, and is often considered a sub field of artificial intelligence. There are many applications of Natural Language processing developed over the years. The main are text based applications, which involves applications such as searching for a certain topic or a keyword in a large document, translating one language to another or summarizing text for different purposes. Therefore, natural language processing is the branch of computer science which deals to convert documents written in one of the natural languages to machine understandable format. For this type of conversion, one or combinations of computer languages are used. Here question arises that whether computer languages differ than natural language. Yes, these differ. The term natural language is used to distinguish human languages (such as English, Spanish, and Swedish etc.) from formal or computer languages (such as C++, Java, and VB etc.).
Here is some interesting comparison between natural language and computer language. As natural language referred to as human language, where as computer language is a language acceptable to a computer system. In natural language, each word has a definite meaning and can be looked up in a dictionary. In the similar manner, all computer languages have a vocabulary of their own. Each word of that vocabulary has a definite unambiguous meaning, which can be looked up in the manual meant for that language. The main difference between a natural language and computer language is that natural language has large vocabulary but most computer languages use a very limited or restricted vocabulary. This is because a programming language, by its nature and purpose, does not need to say too much. Every problem to be solved by computer has to be broken down into discrete (simple and separate), logical steps, which basically comprise of four fundamental operations like input and output operations, movement of information within the CPU and memory, and logical or comparison operations. Each natural language has a systematic method of using the words and symbols of that language, which is defined by the grammar rules of the language. Similarly the words and symbols of a computer language must also be used as per set rules, which are known as the syntax rules of the language. People can use poor and incorrect vocabulary and grammar, and still make them understood. However in the case of computer language, one must stick to the exact syntax rules of the language, if one wants to be understood correctly by the computer. Yet, no computer is capable of correcting and deducing meaning from incorrect instruction. So, computer languages are smaller and simpler than natural languages, but they have to be used with great precision, only then correct results can be expected.

1.2.2 Automatic Document Processing

Philip and Samuel [2009] asserted that an automatic document processing system is one of the most fascinating and challenging areas of pattern recognition with a wide range
of practical applications. In this modern digital day and age, it has become obligatory to have all the available information in a digital form recognized by machine. The main reasons to convert paper into electronic form are:

- Electronic document is more economical and useful. No need to waste a lot of time and money because everything is done by machine with less human efforts.
- Unlimited copies can be made at any time.
- Small space is required for electronic formatted documents. So, redundancy decreases.
- Information on the documents can be shared by everyone, if permitted to do so. It is not limited to one person.
- Electronic formatted documents can be secured in a better way.
- Documents can be edited as per the need and the requirement of any particular organization or individual. For example, there is a record of teacher on a document and want to include record of new teacher inducted in the organization, then that document can be edited to include the same.
- Documents in the electronic form can be converted to other useful formats with fewer efforts.
- Whole file cabinets of paper can be stored on a few dozens CD ROMs or a single hard drive. Less space is required.
- The documents acquired by automatic processing can be published on the web too for distribution to masses.

Document data can be scanned through image scanner but there are some limitations in using these type of scanned documents directly. Since the input document is stored as an image instead of text, therefore it is not possible to do any word processing of the document because the computer cannot interpret the image of stored documents as letters.
or words, numbers and special characters. On the other hand, the storage required for storing the document as an image is much more than that required for storing the same document as a text. For example, a page of printed text, having 3020 characters, can be stored as 3020 bytes by using the ASCII representation. A bitmap image representation of the same document will require 10 to 15 times more storage, depending on the resolution of grid points. Therefore, it is always felt that the scanner must have the capability to convert the scanned image to digital format. That capability is known as optical character recognition (OCR). The OCR technology is used to overcome these limitations. The scanner is equipped with character recognition software, called OCR software, which convert bitmap images of characters to equivalent ASCII codes. The scanner first creates the bitmap image of the document and then the OCR software translates the array of grid points into ASCII text, which the computer can interpret as letters, numbers and special characters.

1.2.3 Character Recognition

Arica and Yarman [2001] stated that character recognition is the term, which covers all types of machine recognition of characters in various application domains. The intensive research effort on the field of character recognition was not only because of its challenge on simulation of human reading, but also, because it provides efficient applications such as the automatic processing of bulk amount of papers, transferring data into machines and web interface to paper documents. A character recognition system can be either “online” or “offline. According to the mode of data acquisition, character recognition methodologies are categorized into two systems as Online Character Recognition Systems and Offline Character Recognition Systems:

1.2.3.1 Online Character Recognition

Online character recognition is the process of recognizing handwriting, recorded with
a digitizer, as a time sequence of pen coordinates. It captures the temporal and dynamic information of the pen trajectory. Applications of online character recognition systems include small handheld devices, which call for a pen only. Computer interfaces complex multimedia systems, which use multiple input modalities including scanned documents, keyboard and electronic pen. These systems are useful in social environments where speech does not provide enough privacy. Pen based computers, educational software for teaching handwriting and signature verifiers are the examples of popular tools utilizing the online character recognition techniques.

1.2.3.2 Offline Character Recognition

Offline character recognition is the process of converting the image of written document into bit pattern by an optically digitizing device such as camera or optical scanner. The recognition is done on this bit pattern data for machine printed or handwritten text. Applications of offline recognition are large scale data processing such as postal address reading, cheque sorting, and office automation for text entry, automatic inspection and identification. Offline character recognition is a very important tool for creation of the electronic libraries. Also, the wide spread use of web necessitates the utilization of offline recognition systems for content based internet access to paper documents. According to the text type, Handwritten and/or Machine Printed Character Recognition Systems are two main areas of interest in character recognition field:

Machine printed text includes the materials such as books, newspapers, magazines, documents, and various writing units in the video or still image. Machine printed characters are mostly uniform in height, width, and pitch assuming the same font and size. These problems for fixed font, multi font and omni font character recognition are relatively well understood and solved with little constraint.
Handwritten text can be further divided into two categories: cursive and hand printed script. Recognition of handwritten characters is a much more difficult problem. Characters are non uniform and can vary greatly in size and style. Even characters written by the same person can vary considerably. The location of characters is not predictable, nor the spacing between them. In an unconstrained system, characters may be written anywhere on the page and may be overlapped or disjoint. A typical recognition system will require some sort of constraints, or added information, about the data being processed.

1.3 Optical Character Recognition (OCR)

OCR is the process of converting scanned images of machine printed or handwritten text into a computer processable format. This is a branch of computer science that helps in reading text from paper and translating the images into a form that the computer can manipulate. It involves computer software designed to translate images of typewritten text into machine printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. Chaudhuri and Pal [1998] said that research in OCR is popular for its various application potentials in banks, post offices and defense organizations. An OCR system enables you to take a book or a magazine article, feed it directly into an electronic computer file, and then edit the file using a word processor. If one scans a text document, one might want to use optical character recognition (OCR) software to translate image into text that can be edited. When a scanner first creates an image from page, image is stored in computer’s memory as a bitmap. A bitmap is a grid of dots; one or more bits represent each dot. The job of OCR software is to translate that array of dots into text that computer can interpret as letters and numbers.

1.3.1 Historical Background of OCR

The first conceptual idea of OCR is due to Tauschek in 1929 and Paul W. Handle in
Tauschek obtained a patent on OCR in Germany in 1929, followed by Handle who obtained a U.S. patent on OCR in U.S.A in 1933. Tauschek was also granted a U.S patent on his method in 1935. His Machine was a mechanical device that used templates. Chaudhuri and Pal [1998] quoted that the OCR technology took a major turn in the middle of 1950s with the development of digital computer and improved scanning devices. For the first time OCR was realized as a data processing approach, with particular applications to the business world. The first commercial OCR system was installed at the Reader's Digest in 1955, which, many years later, was donated by Reader Digest to the Smithsonian, where it was put on display. The United States Postal Service has been using OCR machines to sort mail since 1965 based on technology devised primarily by the prolific inventor Jacob Rabinow. In 1974, Ray Kurzweil started the company Kurzweil Computer Products, Inc. and led to development of the first omni font optical character recognition, a computer program capable of recognizing text printed in any normal font. He decided that the best application of this technology would be to create a reading machine for the blind, which would help blind people to understand written text by having a computer read it out to them. Kurzweil Computer Products started selling a commercial version of the optical character recognition computer program in 1978. After two years, Kurzweil sold his company to Xerox, which was having an interest in further commercializing paper to computer text conversion.

Earlier OCR systems, worked by matching the scanned images against stored bitmaps based on specific fonts. The hit or miss results of such pattern recognition systems helped to establish OCR’s reputation for inaccuracy. Today, OCR software can recognize a wide variety of fonts, but handwriting and script font that mimic handwriting are still a problematic area for researchers. Developers are taking different approaches to improve script and handwriting recognition. Advances are being made to recognize characters,
based on the context of the word in which the software will use knowledge of the parts of speech and grammar to recognize individual characters. Typical accuracy rate exceed 99%, although certain applications demanding even higher accuracy require human review for errors.

1.3.2 Stages in OCR

A hierarchical approach for most of the systems would be from pixel to text, as follows:

\[ \text{Pixel} \xrightarrow{} \text{Feature} \xrightarrow{} \text{Character} \xrightarrow{} \text{Sub word} \xrightarrow{} \text{Word} \xrightarrow{} \text{Meaningful text} \]

Lehal and Singh [1999] stated that the above hierarchical tasks are grouped in the stages of the character recognition for image acquisition, pre-processing, segmentation, feature extraction, recognition, and then finally post processing. As shown in figure 1.1, the process of optical character recognition of any script can be broadly broken down into the different stages.

![Figure 1.1: Steps in Word Interpretation](image)

i. \textit{Image acquisition}: This involves a step to scan a document and storing it as an image. The resolution (number of dots per inch, dpi) determines the rate of acquisition process. When the paper document is being scanned, it translates paper document into an electronic format which can be stored in computer. The
input documents can be typed text, pictures, graphics, or even handwritten material.

ii. *Pre-processing:* It is a process of representing the scanned image for further processing. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that is easy for the character recognition systems to operate accurately. It reduces noise and distortion, removes skewness and performs skeletonizing of the image thereby simplifying the processing of the rest of the stages.

iii. *Segmentation:* The pre processing stage yields a ‘clean’ document in the sense that sufficient amount of shape information, high compression and low noise on normalized image is obtained. The next stage is to segment the document into its sub components. It separates the different logical parts, like text from graphics, line of a paragraph, and characters of a word. Segmentation is an important stage, because the extent to which one can reach in separation of words, lines or characters directly affects the recognition rate of the script. There are mainly two types of segmentation:

a) *External segmentation* decomposes the page layout into its logical units. It is the most crucial part of the document analysis, as it is an essential step prior to the off line character recognition. It gives the extraction of various writing units, such as paragraphs, sentences or words.

b) *Internal segmentation* is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. It gives the
extraction of letters, specially, in cursively written words.

iv. **Feature Extraction:** Character recognition involves analyzing segmented part of the image and comparing its features against a set of rules stored on OCR engine that distinguishes each character to identify a character. The selection of a stable and representative set of features is the main part of pattern recognition system design. It is the most consequential point in the designing issues involved in building an OCR system.

v. **Classification:** It is the main decision making stage of an OCR system and to use the features extracted in the previous stage to identify the text segment according to preset rules.

vi. **Post-processing:** It is the final stage, improves recognition by refining the decisions taken by the previous stages and recognizes words using context. It is ultimately responsible for outputting the best solution and is often implemented as a set of techniques that rely on character frequencies, lexicons, and other context information.

1.3.3 Issues Affecting OCR

Tanner [2004] explained that there are a number of key issues to consider when observing a printed resource and assessing whether it will produce the text resource accuracy desired through OCR technologies. Some of the main factors are listed below:

1.3.3.1 Method Used for Scanning

The number one factor that can improve OCR accuracy is the method used for scanning or resolution at which the document is scanned. 300 dpi is the recommended best scanning resolution for OCR accuracy. Higher resolutions do not necessarily result in
better accuracy and can slow down OCR processing time. Resolutions below 300 dpi may affect the quality and accuracy of OCR results. Remember that all OCR engines will struggle to recognize anything well if the resolution is below 300 dpi and that this is the absolute minimum baseline for scanning.

1.3.3.2 Type of Paper Used

There are other factors related to the type of paper used that will affect the standard of the scanned image and these should be accounted for as well. The paper on which text appears is critical to the scanned image standard. If the OCR engine cannot discriminate between the character and the paper background noise then it will be more likely to misrepresent the character.

1.3.3.3 Nature of Printing

The nature of the printed text in the original may make a significant difference to OCR accuracy. Obviously if the text is printed poorly or if it was typed and characters are broken, faded or have indistinct edges then this will affect the ability of an OCR engine to recognize patterns and differentiate between similar shaped characters. So the clarity of the printing is a factor to consider. Some fonts may also have improved print clarity over others. Similarly, character sizes of below 6 points in the original will limit the accuracy likely to be achieved.

1.3.3.4 Formatting Complications

Variations in font size and type face may result in misunderstanding the characters. Broken character and touching character stemming from excess ink or paper degradations may not be recognized. Discolored and Older documents must be scanned in RGB mode to capture all the image data, and to maximize OCR accuracy.

These are the few reasons which indicates that why OCR vendors never claim their
software to be 100% accurate. OCR has been the subject of a large body of research because there are numerous commercial applications for this technology. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. OCR can be used for

- Speeding up the data entry. For input task of many documents, OCR is the most cost effective and speedily available method.
- To reduce data entry errors.
- To reduce the storage space required by paper documents. Each year, the electronic storage technologies will free storage space required to store documents. If these technologies are not used then large space (in the form of cabinets and boxes) is required to keep the documents.

Philip and Samuel [2009] concluded that some of the most significant applications of OCR include the following potential area where the OCR system can be useful:

- Reading aid for the blind
- Automatic text entry into the computer for desktop publication
- Library cataloging
- Ledgering
- Automatic reading for sorting of postal mail
- Bank cheques and other documents
- Document data compression: from document image to ASCII format
- Language processing

The application of handwritten recognition system can be used in a number of different areas. Few of them are discussed below:

*To Read Handwritten addresses*: After scanning the image of address part of the postal
envelop, next task is to assign a mail piece image to a delivery address, by interpreting handwritten addresses. The address, for the purpose of physical mail delivery, is made up of the organization name or personal name, primary number which could be a street number or a post office box number, secondary number such as an apartment or suite number, then followed by street, post office, city, state, country and pin code. This type of system can save the efforts of keying the address to system.

*Reading Bank cheque:* The system can be used for reading and interpreting the bank cheques. This will reduces the efforts needed to enter the data written on the cheque. But bank cheque recognition presents several research challenges in the area of document analysis and recognition. The reason behind this challenge is because of the colored backgrounds and has complex patterns to crack.

*To Read Filled form:* The filled forms are generally collected by the different organization, to collect inputs data from different users. Sometimes the number of these forms is so large that just to key in data is a time consuming. So the handwritten recognition system can be used to interpret the field data of the data forms.

### 1.4 Segmentation

Segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Text segmentation is a process in which the text image is segregated into units of patterns that seem to form characters. For textual processing, following general procedure is used.

Casey and Lecolinet [1996] suggested that to segment a document, start from a given point in the scanned image of document. Segment an area to find or extract the next character. Then extract distinguishing attributes of character. These attributes should
match a member of a given symbol set, so next step is to find that symbol.

The procedural sequence, mentioned above, is repeated while all additional character images are not found. Here one simple question should be raised to a researcher, who is working on segmentation, that what does make a character? The answer is difficult. The possible answer to this query can be that a character is a pattern that resembles one of the symbols to which the system is designed to recognize. But to determine such a resemblance, the pattern must be segmented from the document image. Therefore each stage depends on the other. Segmentation process can broadly put in three categories namely line segmentation, word segmentation and character segmentation. These are discussed in the following section.

1.4.1 Line Segmentation

Line segmentation is the process in which the lines are extracted from scanned image. Only lines are differentiated. Horizontal projection of a document image is most commonly employed to extract the lines from the document. If the lines are well separated, and are not tilted, the horizontal projection will have separate peaks and valleys, which serve as the separators of the text lines. These valleys are easily detected and used to determine the location of boundaries between lines. Line segmentation is shown in figure 1.2.

![Figure 1.2: Line segmentation](image)
1.4.2 Word Segmentation

Word segmentation is the process in which words are differentiated or extracted from the segmented lines. As this is a known fact that two different words are separated by a gap, this concept is used for word segmentation. Word segmentation is the problem of dividing a string of written language into its component words. Word splitting is the process of parsing concatenated text (i.e. text that contains no spaces or other word separators) to infer where word breaks exist. A vertical projection profile gives the column sums. One can separate lines by looking for minima in horizontal projection profile of the page and then separate words by looking at minima in vertical projection profile of a single line. Valleys, in the vertical projection of a line image, can be used in the extraction of words in a line, as well as extracting individual characters from the word. Word segmentation is shown in figure 1.3.

![Figure 1.3: Word segmentation](image)

1.4.3 Character Segmentation

Character segmentation is the process in which from the segmented word, characters are extracted. Character segmentation is a crucial step of OCR systems as it extracts meaningful regions for analysis. This step attempts to decompose the image into classifiable units called character, as shown in figure 1.4. A poor segmentation process
produces misrecognition or rejection. Segmentation process is carried out after the pre
processing of the image.

![Character segmentation](image)

**Figure 1.4: Character segmentation**

Casey and Lecolinet [1996] stated that character segmentation is an operation that
seeks to decompose an image of a sequence of characters into sub images of individual
symbols. Character segmentation is a necessary pre requisite step for character recognition
in almost all OCR systems. Character segmentation has been a well investigated field over
the last decade and its main aim was to provide individual character to optical character
recognition algorithms. It is important to note that this step (segmentation) only produces a
sequence of fragments, while the segmentation of characters is confirmed at the
classification stage.

All recognition algorithms depend on the segmentation algorithm to break up the
image into individual characters. It is the process of dividing written text into meaningful
units, such as words, sentences. The term applies both to mental processes used by humans
when reading text, and to artificial processes implemented in computers.

Lu [1995] said that character segmentation is a technique, which partitions images of
lines or words into individual characters. Character segmentation is an operation that seeks
to decompose an image of a sequence of character into sub images of individual symbols.
So it can be said that it is one of the decision processes in optical character recognition
system. Its decision, that a pattern isolated from the image is that of character (or other
identifiable unit), can be right or wrong. Character segmentation is all too often ignored in the research community, yet broken and touching characters are responsible for the majority of errors in automatic reading of both machine and hand printed text. Character segmentation is fundamental to character recognition approaches, which rely on isolated characters. It is a critical step because incorrectly segmented characters are not likely to be correctly recognized.

Segmentation is the process of separating out the individual character, which make up a word. Although it is a straightforward task when dealing with typewritten or well written numbers, it can be quite difficult when the words are handwritten. The main reason is that handwritten characters often overlap and, in some cases, may be disjointed. Also, the wide variations in handwriting styles make it very difficult to make generalizations for making segmentation heuristics. The problem arises if any of the characters in words are connected or disjointed, then any of the connected components could represent a single character, more than one character, or part of a character. According to a survey of vast literature done by Casey and Lecolinet [1996], there are three pure strategies for segmentation, plus numerous hybrid approaches that are weighted combinations of these three. These elementary strategies are discussed in detail in the following section.

1.5 Segmentation Strategies

Different methods used can be classified based on the type of text and strategy being followed like the Classical Approach in which segmentations are identified based on character like properties. This process of cutting up the image into meaningful components is also called dissection. Recognition Based Segmentation, in which the system searches the image for components that match classes in alphabet. Holistic Methods, in which the system seeks to recognize words as a whole, thus avoiding the need to segment into
characters. There are many strategies for segmentation, which are combinations of one or more of above three pure ones. Davessar et. al. [2003] stated that hybrid methods can be represented as weighted combinations of these lying at points in the intervening space. Figure 1.5 is used to show the three dimensional space representing the segmentation strategies.

![Figure 1.5: Three dimensional space representing the strategies of segmentation](image)

### 1.5.1 Classical Approach for Segmentation

In this approach, the image is usually decomposed into sequence of sub images using general feature like approximating character size, pitch and white space. This is also called as dissection method. Dissection means the decomposition of the image into a sequence of sub images using general features, but the image is not divided into sub images independent of content. Dissection is an intelligent process that analyses the image, however classification into symbol is not involved at this point. Casey and Lecolinet [1996] stated that methods used for this approach are projection analysis and using white space and pitch.

*Using white space and pitch:* In machine printed, vertical white space often serves to separate successive characters. These types of applications are used for involving limited font sets, each character occupies a block of fixed width. The pitch, or number of
characters per unit of horizontal distance, provides a basis for estimating segmentation points. The sequence of segmentation points obtained for a given line of print should be approximately equally spaced at the distance corresponding to the pitch. This provides a global basis for segmentation, since separation points are not independent. But it's not the case with handwritten document, where the boxes, for handwritten symbols in which to print individual symbols, can be provided. Especially, where OCR is concerned, an additional space between characters is given because it is easy to segment characters separated by white space. Pitch or number of characters per unit of horizontal distance, provides the basis for estimating segmentation points, so applying this rule permits correct segmentation in a case where several character along the line are merged or broken.

*Projection analysis:* The vertical projection of printed line consists of simple running count of black pixels in each column. It can serve for detection of white space between successive letters. Moreover, it can indicate locations of vertical strokes in machine prints or region of multiple lines in handprint. Thus, the vertical histogram is the basis of segmentation process, but it fails to make clear distinction between merged characters. A vertical projection is less satisfactory for the slanted characters commonly occurring in handprint. For a script, which requires two dimensional processing, horizontal projection can be used to find the cut points.

1.5.2 Recognition Based Segmentation

This is the segmentation in which the system searches the image for components that match classes in its alphabet. In this, no feature based dissection algorithm is employed. Rather, the image is divided systematically into many overlapping pieces without regard to content. As per Casey and Lecolinet [1996], these are classified as part of an attempt to find a coherent segmentation/recognition result. Systems using such principle perform
“recognition based segmentation.

Methods considered here also segment words into individual units (letters). Letter segmentation is a by product of letter recognition, which may itself be driven by contextual analysis. The main interest of this category of methods is that they bypass the segmentation problem. No complex “dissection” algorithm has to be built and recognition errors are basically due to failures in classification. Recognition based segmentation consists of the following two steps:

I. Generate segmentation hypothesis.

II. Choice of the best hypothesis i.e. verification step.

There were many algorithms given for this segmentation method. One algorithm was reported, that is, recursive algorithm for machine printed characters. This algorithm systematically tests all combinations of admissible separation boundaries until it either exhausts the set of cut points, or else finds an acceptable match which is one in which every segmented pattern matches a library prototype within a pre specified distance tolerance. Here, various features and their positions of occurrence are recorded for an image. Each feature contributes an amount of evidence for the existence of one or more characters at the position of occurrence. The positions are quantized into bins such that the evidence for each character indicated in a bin can be summed to give a score for classification. These scores are subjected to contextual processing using a pre defined lexicon in order to recognize words. The method is being applied to text printed in a known proportional font.

1.5.3 Holistic Approach

It recognizes the entire word as a unit. So there is no requirement for the character segmentation. Casey and Lecolinet [1996] concluded that the main flaw is that, it requires
a pre defined lexicon because this is not dealing with characters but with words. For this approach, training session is a must to modify the lexicon. Therefore, this kind of method is usually suitable for applications having static lexicon where there is a limited domain of the OCR that is cheque recognition, passport reader, paper checking etc. There are two steps for this approach:

I. First is, feature extraction.

II. Second is, comparison of the word with the words in the lexicon.

In actual practice, combination of the above three approaches is used to segment the document which can be a machine printed or hand written. Now various hybrid techniques are available to segment a character. One of these techniques is “Hit and Deflect” algorithm which attempts to find an optimal path for cutting a connected component by literally hitting and deflecting their way through the connected component. Similarly, another algorithm is “Drop Fall” algorithm which attempts to build a segmentation path by mimicking an object falling or rolling in between the two characters which make up a connected component. These types of algorithms can be tried to segment handwritten Gurmukhi script.

1.6 Terminology Used

**Optical Character Recognition:** Optical character recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format. It involves computer software designed to translate images of typewritten text into machine printed editable text, or to translate pictures of characters into a standard encoding scheme representing them in ASCII or Unicode. If a text document is scanned, then one might want to use optical character recognition (OCR) software to translate image into text that can be edited.
**Image Acquisition:** This involves scanning a document and storing it as an image. The resolution (number of dots per inch, dpi) determines the rate of process.

**Pre Processing:** Process of representing the scanned image for further processing. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre processing aims to produce data that are easy for the character segmentation systems to operate accurately.

**Segmentation:** A clean image means sufficient amount of shape information, high compression and low noise on normalized image is obtained after the scanned document is passed through the pre processing phase. Segmentation is the next stage to pre processing stage. The output of Segmentation phase is breakage of scanned document into its sub components. Segmentation separates the different logical parts, like text from graphical image, then paragraph from text, followed by extraction of line from a paragraph, then finally words and characters from lines. Segmentation is an important stage, because the extent one can reach in separation of words, lines or characters directly affects the recognition rate of the script. If the segmentation is not correct then character recognition can not be correct.

**Water Reservoir:** As per the water reservoir principle, if water is poured from a side of a component, the cavity regions of the component where water will be stored are considered as reservoirs of the component.

**Classification:** It is the main decision making stage of a character recognition system and uses the features extracted in the previous stage to identify the segmented part as per the preset rules.
**Thresholding:** This is a process to differentiate the foreground from the background. For that purpose, get the histogram of gray scale values of a document image. These would consist of two peaks. A high peak would be corresponding to the white background and a smaller peak corresponding to the foreground. So, to determine the threshold gray scale value is to determine optimal value in the valley between the two peaks. Using this value, the gray scale value is assigned to white if it is above the optimal value and if below then it is assigned to black.