CHAPTER 1

INTRODUCTION

1.1 INFORMATION RETRIEVAL

Information Retrieval (IR) has become an important area of research in the past decades due to the rapid growth of information in diverse areas. Technically, information retrieval can be defined as a system for representing, indexing (organizing), searching (retrieving) and recollecting (delivering) documents.

An information retrieval system begins when a user enters a query into the system. Queries are formal statements, such as search strings, which characterize user’s information need. Either a single document or several documents may match the query, perhaps with different degrees of relevancy. The IR system ranks the matching documents based on the relevancy and presents to the user.

Due to the advent of computer technology, digital documents have become more popular for storage and transmission, when compared to the traditional paper documents. However, billions of volumes of traditional newspapers, books, magazines, etc. in the paper format need to be transferred to the digital domain for various purposes. As the digital storage is becoming cheaper, paper documents are easily archived with the help of the digitization equipment like scanners into digital images or document images. In general, document image refers to an image generated from a scanner, a fax machine
or an image obtained through the conversion of an electronic document into an image format (e.g. TIFF or JPEG).

Document images are gaining popularity and importance in the recent decades, since a lot of efforts has been made to build digital libraries, which digitize high-volume archives of paper documents (patents, historical and business documents) to provide the public with a free and easy on-line access. As a matter of fact, many organizations currently practice business on document image archives. However, such archives are often poorly indexed which makes them unfit for IR purposes. In addition, text retrieval techniques cannot be applied directly over imaged documents. As a result, the user experiences a lot of difficulty in retrieving relevant information from imaged data and there arose a need for efficient document image retrieval tools. Therefore, the study of information retrieval in document images has become a compelling area of research.

1.2 DOCUMENT IMAGE RETRIEVAL (DIR)

Recently, many efforts have been devoted to the digitization of paper collections in order to archive them as document images. However, the presence of large digital archives could be fruitful only when the information residing in the document images is accessible. As a result, a document image retrieval system is becoming more important, since, patents, historical and business documents are converted into a scanned image format.

To perform information retrieval on document images, a proper mechanism is needed to characterize the document content in a meaningful manner. This technique is known as Document Image Analysis (DIA). In general, the DIA refers to the algorithms and techniques applied over the document images to fetch computer readable description from pixel data.
The objective of DIA is to recognize the text and graphical components in document images, and to extract the underlying information for any interpretation. Two categories of DIA exist: they can be defined (Kasthuri et al 2002) as (1) Graphical processing and (2) Textual processing. Graphical processing deals with the non-textual line and symbol components contained in the diagrams, pictures, company logos etc. Textual processing deals with the text components of a document image, such as determining the skew (any tilt at which the document may have been scanned into the computer), determination and location of each region or blocks in a document image, (Page segmentation or Page layout analysis), Separation of the Text from Images, Script/ Language Identification and finally understanding the text represented in the document image (possibly its attributes).

In this thesis, we emphasize Textual processing in Document Image analysis. In Textual processing, understanding the text (which enables retrieval) of document images cannot be done directly. The reason behind this is due to the existence of lot of structures in the document. Text understanding techniques can be applied to the image only after understanding the structures and relationships involved in the document and its underlying script/language.

The process of determining the document structure involves the understanding of physical appearance of the document and the relationships between the entities that make it up. To analyze the structure, preprocessing, skew detection and page layout analysis of document images are necessary.

In the preprocessing phase, some low level processing (pixel level processing) shall be performed over the document images to improve the quality of the image. In skew detection, the tilt occurring in the orientation of the text lines need to be identified and corrected. Page layout analysis (or Page Segmentation) determines the formatting of the text and understands the meaning associated with the blocks in which the text is located (Kasthuri et al 2002).
Page layout analysis / Segmentation techniques could be twofold: (1) Structural layout analysis (also called physical and geometric layout analysis) and (2) Functional layout analysis. A physical layout analysis is performed to divide the document image into homogeneous zones, each consisting of one physical layout structure such as text, graphics and picture. To be specific, these layout structures can be extracted by isolating it from the layout gaps or columnar width in the document image. Further, in textual blocks, words, text lines and paragraphs can be extracted. The functional layout analysis (Dengel et al 1992) (also called syntactic and logical layout analysis) uses domain-dependent information, to perform labeling of the structural blocks such as headings, titles of the page, content of the page, footer etc. After page segmentation, the text present in the text areas could be understood after script recognition.

Most of the text understanding systems have been developed with the assumption that the script or language of the input image is known. In multi-script or multi-lingual environment, DIA invokes manual intervention to select an appropriate text understanding system. Therefore, Script Identification (Script Recognition) is a key factor in the automatic processing of document images before choosing an appropriate text understanding system. Script Recognition recognizes a particular script among multi-scripts automatically whereas language identification is needed to discriminate the languages of the same script. Script/ Language can be recognized in three ways. They are:

- Document level
  - Input could be a monoscript/ monolingual document image.

- Paragraph/ Line level
Input could be a multi-script/multi-lingual image.

- Word level
  - Input image contains multi-script in a single line of text.

Once the layout and script information has been identified, document image understanding techniques could be applied to retrieve information. Document Image Understanding (or interpretation) is the formal representation of the abstract relationships indicated by the two dimensional arrangement of symbols (Nagy 2000). Researches have addressed two different techniques (Doermann 1998) to identify or understand the text from document images, namely, Optical Character Recognition (OCR) and Keyword Spotting technique.

Optical Character Recognition (OCR) lies at the core of the discipline of pattern recognition. The objective is to develop computer algorithms to identify the characters of the alphabet. Optical Character Recognition deals with the machine recognition of characters present in an input image obtained using scanning operation (Doermann 1998). It refers to the process by which scanned images are electronically processed and converted into an editable text (ASCII representation). Standard text retrieval techniques could be applied over the recognized text to retrieve information from the document images.

In contrast, Keyword Spotting approaches understand the text at the word level of document images. These approaches understands the image properties of text at word level and converts it into an intermediary representation instead of converting the entire document image into ASCII representations character by character. Since, information retrieval is concerned with the keyword, (i.e.) obtaining a query word from the user and retrieving the documents relevant to the user query, the Keyword Spotting
technique has gained its popularity. By applying Keyword Spotting technique, relevant document images could be fetched by matching the word image representations of documents directly with the user query word.

1.3 PROBLEMS ENCOUNTERED IN DIR

The DIR process relies heavily on various factors of Document Image Analysis. Therefore, the major challenges of DIA involved in Document image retrieval are Page Segmentation, Text/Image separation, Script Recognition and Text Image Understanding. The challenges encountered in the above said factors of DIA have been explained in the following sections.

1.3.1 Physical Layout Analysis

Page segmentation is a crucial step in many applications related to document images such as text extraction using OCR and layout-based document retrieval (Cattoni et al 1998). In page segmentation, content blocks may be extracted based on the columnar width in document images and the layout analysis heavily depends on the page segmentation algorithms. In general, the page segmentation algorithms can be classified into three types: top-down, bottom-up and integration methods.

The Top-down method uses X-Y projection, (Cesarini et al 2002, Marinai et al 2004) which is a process of dividing the whole page into smaller sub- components from larger components. For example, the page may be split into one or more column blocks of text, then each column split into paragraph blocks, then each paragraph split into text lines, etc. (Marinai et al 2005b). However, this approach is successful only for Manhattan layouts (layouts separable by horizontal and vertical segmentation) and fails to analyze the complex and non-Manhattan layouts. The reason is that in a non Manhattan
layout, text and graphics gets intermixed and are placed irregularly in the image. Bottom-up approaches are incremental merging methods, guided at first by sparse “local” evidence, and using the connected component analysis, merges the evidence from lowest level of detail and then rises, merging words into lines, lines into columns, paragraphs and physical blocks etc., until the page is completely assembled covering all kinds of layouts.

Examples of this approach include Run Length Smearing Algorithm (Wong et al 1982), connected component analysis, Docstrum (Gormann 1993) and Voronoi (Kise et al 1998). None of the above reported methods are generic, parameter free and threshold free. In addition, a detailed prior layout model is required for analysis and a proliferation of the task specific rule is essential for success. However, this can be achieved through hybrid methods.

Hybrid approaches, combines the robustness of top-down methods with the speed of bottom-up methods (Bruel 2003), but this depends heavily on heuristics. Though this is threshold free and eliminates the requirement of a prior model, heuristics are involved and the impact of heuristics gets reflected in segmentation, resulting to an over-segmentation or under-segmentation of layouts. However, none of these systems have performed page segmentation by eliminating heuristics (which reduces recall), task specific rules and prior knowledge. In addition, the elimination of the connected component analysis (this requires more computations or more pixel visits) has not yet been considered for page segmentation, and hence, the need for a Hybrid approach to address all these issues arises.

1.3.2 Script Recognition

Script Recognition approaches can be broadly classified into two categories, namely, local and global approaches. The local approaches
Pal and Chaudhury 2001, Pal et al 2003) analyze a list of connected components (Line, word, char) in the document images, to identify the script (or class of script). In contrast, global approaches (Joshi 2006) employ an analysis of regions (block of text) comprising at least two lines (or words) without finer segmentation. In general, global approaches work well based on texture measurement, but this relies heavily on a uniform block of text (Busch et al 2005), and extensive preprocessing (to make the text block uniform) is required to measure the texture. Even though local approaches rely on the accuracy of character segmentation or connected component analysis, it could work well on the documents irrespective of their quality or uniformity in the block of text.

In the literature, many works have been reported for script recognition at the document, line and word levels, using local approaches. In this context, researchers have made a number of attempts to discriminate the Han and Latin script (Spitz 1997, Lu and Tan 2008) at the document level and exploited many Indian scripts at line level and word level (Pal and Chaudhury 1997, Pal and Chaudhury 1999, Padma and Nagabhushan 2003, Dhandra et al 2007). However, all the techniques reported in the literature are script dependent. Since this research is intended to develop an information retrieval system for Tamil document images, Script Recognition, to discriminate the Tamil from English scripts in bilingual document images is becoming important.

In this connection, few local approaches are reported in the literature, such as spatial spread analysis (Dhanya et al 2002), Aspect Ratio (Tan et al 1999), Structural features (Pal and Chaudhury 1999), and Water Reservoirs (Pal et al 2003). However, all the above mentioned techniques produce a low discrimination rate due to its incapability in exploration of the scripts. Global approaches (Pati et al 2004, Pati and Ramakrishnan 2006)
reported in the literature discriminates well but the performance degrades in noisy images. Therefore, a Script Recognizer which could produce more discrimination in English and Tamil scripts across font variations and quality of images could well suit the needs of the users.

1.3.3 Text Image Understanding

In recent years, Information Retrieval from Document images has been successfully performed using Keyword Spotting technique. Numerous keyword spotting techniques and IR systems have been reported to retrieve the information from Roman and Chinese document images (Lu and Tan 2002b, Lu and Tan 2004a) and some of the Indian languages such as Hindi (Chaudhury et al 2003, Harit et al 2005a) and Telugu (Jawahar et al 2004a, Balasubramanian et al 2006). However, feature extraction techniques discussed in the literature are specific and language dependent and cannot be applied to the Tamil since the shapes of the Tamil characters are complicated and varied. In addition, Tamil text recognition systems could not be utilized for information retrieval since it suffers inherent weaknesses, namely (1) Failure in the discrimination of a set of characters that closely resemble with others in the character set, (2) Restrictions in font faces and sizes (Seethalakshmi et al 2005) and (3) Post processing to correct the errors occurred during the recognition.

Hence, an effective Keyword Spotting technique is required for the Tamil language to understand the text image at the word level and to represent it as a string without explicit conversion.

To initiate query processing for document image retrieval, strings representing the word images in the documents should match the textual query words supplied by the user. If the query word supplied by the user is in the ASCII format and the word images are represented in an intermediary
form, a suitable mapping scheme is required for matching the relevant documents. Therefore, a suitable retrieval framework with a mapping (Transliteration) scheme which matches the query word with intermediate form could promote the information retrieval effectively.

Many systems that were proposed till date attempt to address a specific factor involved in the Document Image Analysis. Information retrieval system attempted in the literatures has concentrated more on text image understanding rather than other aspects. Apart from text image understanding, factors such as layout analysis and script recognition would require greater attention especially for the document images which arise from Patents, newspapers and magazines, where a large number of complicated layouts and scripts need to be processed. Few systems exist in the literature that incorporates the entire spectrum of factors that affect the document image retrieval process.

As a result, Document Image retrieval can be done through two processes: offline processing and online processing. During offline processing, document images in the corpus could be preprocessed, layouts could be analyzed, text blocks, paragraphs, lines and words can be segmented further, the script can be identified either at the line or word level, and later, the text could be understood from the word images and get represented (vector representation) properly. During online processing, when a query word has been supplied by the user, the relevant document images could be retrieved by matching the representations obtained in the offline process. An information retrieval system which could integrate all these factors and work as stated above would be quite ideal to fulfill the needs of the user.
1.4 MOTIVATIONS OF THIS RESEARCH

This research has been performed, since guided by the motivation of Information Retrieval (IR) from Tamil document images due to the explosion of information and increasing number of scanned documents in Digital Libraries, Patent Development and Business Domain applications in Tamil language and lacking of information retrieval concepts for Tamil document images (Regional language).

Therefore, all the factors discussed so far, give a complete scenario of what is required to improve Information retrieval from Tamil document images and this requirement, in fact, makes information retrieval a compelling area for research. In the process of building such an effective information retrieval system for Tamil document images, the various factors which motivated this research are as follows:

- A Hybrid layout analysis technique to segment the pages thereby accommodating homogeneous and heterogeneous images with an improved recall rate (less over segmentation and under-segmentation), and reduced computational complexity and threshold (less number of pixel visits over the image) than those of the existing techniques.

- Isolation of the text and images from the identified physical layouts.

- A Script Recognizer which recognizes the script of the word images in bilingual documents (Tamil and English) across font variations and qualities with a better discrimination rate in order to route the word images automatically to the text understanding engine.
• A Tamil Word Image Understanding (Keyword Spotting) technique to understand the text at the word level across font variations and touching characters.

• A Tamil Keyword Spotting technique which eliminates post-processing and spell check.

• The development of a Tamil information retrieval framework (GUI and a mapping scheme) to match the query word given by the user with the intermediary representation of word images in documents.

• The integration of the entire spectrum of factors to provide a complete Information retrieval system for Tamil document images.

• Evolving evaluation methodologies for the information retrieval system.

1.5 CONTRIBUTIONS

The major contributions of our research have been listed below:

• A Hybrid approach for the physical layout analysis of document images
  o As part of this idea, a **Rectangular White Space analysis algorithm (RWSA)** is proposed in this thesis, to analyze and identify heterogeneous (Manhattan and non-Manhattan) layouts of document images in an effective way using the white spaces as delimiters.

• Isolation of the text from image blocks using Statistical properties
As part of this idea, two Statistical properties called as Black Run length and Transition rate have been considered in this thesis to discriminate the text blocks in images. Since these properties obtain high regularities over the text blocks, these parameters are considered for discrimination.

- **Word level Script Identifier for Tamil word images using a local approach**
  - This idea gives birth to the contribution of the SFBSR, the Spatial Features based Script Recognizer, which identifies the script of the word images (in bilingual documents consisting of Tamil and English words) using the initial character of the word image across its variations in font faces and sizes. Here, the density and black runs of the initial character have been analyzed for script recognition.

- **Understanding of Keyword images from Tamil documents**
  - This idea leads to a feature string generation technique called as LR-TB-FS (Left to Right- Top to Bottom- Feature String). The LR- TB- FS extracts the features of the word images based on their basic characteristics and represents them as feature string. In order to generate a feature string at the word level, characters which touch with one another have also been dealt in the LR-TB-FS technique through the Touching character analysis.

- **Tamil Information Retrieval framework**
  - The overall framework (overall idea) to initiate online processing leads to the generation of two different
techniques namely, the **Feature String Table** and the **Deviation string algorithm**.

- Based on the feature extraction LR-TB-FS described above, a **Feature String table** has been devised in this thesis to represent every character in a shape-based Tamil character set using a standard primitive string. During online processing, the feature string for the query word can be synthesized from the primitive string of the feature string table.

- A **Deviation String algorithm** has been proposed in this thesis to adapt the font variations for a query word.

- Integration factors to build a complete Information Retrieval system
  - This idea gives birth to the development of a **Keyword-based complete information retrieval system for Tamil document images** by integrating various factors such as the Physical Layout analysis, Script Recognition, Word Image Understanding and Retrieval framework. Using this system, document images pass through various levels to represent the word images and retrieval is initiated by matching the feature strings with the query word supplied by the user.

1.6 **ORGANIZATION OF THESIS**

The thesis is organized as follows: Chapter 2 gives an overview of the literature associated with the Layout analysis, Script recognition and Keyword Spotting works and their implications.
Chapter 3 gives a detailed view of the proposed keyword based information retrieval system and described the various processes involved in the system. This chapter on a whole provides an overall view of our work.

Chapter 4 deals with the hybrid approach of the physical layout analysis in document images, together with the process of isolating text blocks from the images after layout identification. Chapter 5 discusses about the recognition of the script from bilingual document images at the word level and its comparison with various other techniques.

Chapter 6 discusses the word image understanding scheme proposed for Tamil word images and a retrieval framework which enables the user to obtain relevant documents when a query word has been supplied. In addition, this chapter also discusses the integration of the various factors dealt within Chapters 4, 5 and 6 and illustrates a complete keyword based information retrieval system by encompassing the various factors. This chapter also includes an account of the performance of the information retrieval system after integration.

Chapter 7 concludes the various contributions of our work and the possible enhancements that may be carried out in future.