CHAPTER 3

SYSTEM OVERVIEW

3.1 SYSTEM STRUCTURE

Document image retrieval is becoming more important in today’s world, due to the advent of digital libraries, where huge archives are digitized as imaged documents rather than text. Users of digital libraries need to retrieve the document images for multiple purposes and need of the document image information retrieval gets increased. Even though, many researchers have turned towards the development of retrieval mechanism from Roman and Indian languages, lack of retrieval techniques exist for the Tamil imaged documents. In addition, number of researches has been attempted in text understanding techniques and IR techniques but none of the research has concentrated towards a complete information retrieval system encompassing various factors such as Page segmentation (layout identification), Script recognition, Text Understanding and Retrieval framework, which could promote Information Retrieval effectively.

Therefore, this research work proposes a keyword spotting based information retrieval system for Tamil document images encompassing various factors such as Layout analysis, Script recognition, and a Retrieval framework with a Tamil word image understanding scheme, which intends to fetch the documents specific to the user supplied query word. The basic block diagram of the system is shown in Figure 3.1.
Keyword spotting based Information Retrieval system comprises of two phases: Offline Processing and Online Processing. Need for Offline Processing arises to organize and annotate the textual blocks of document images in the corpus into a feature representation. This enables the document images which are relevant to the user query to get presented directly. Online Processing involves the graphical user interface and query-processing framework, which enables the user to obtain the desired documents. Offline Processing, which has been carried out prior to online processing, has been detailed in the next section.
3.2 OFFLINE PROCESSING

In Offline processing, document images are first preprocessed and the physical layout analysis is performed to identify the blocks. Then, text blocks are annotated by isolating them from image blocks and the segmentation process is carried over those textual blocks to identify the lines and words. Later, the script of the word image is determined as either English or Tamil from the (bilingual) word images. Subsequently, the Tamil word images are represented as a feature string, by making use of the Feature String Generation algorithm. With respect to each document image, a feature code file is constructed with complete information about the textual blocks and its location word images, containing in it and feature string of every word image. These feature code files are stored in a text database, and they are used instead of the original document images for query processing. Various modules that comprise the offline processing of the Tamil information retrieval system are discussed in the subsequent sections.

3.2.1 Document Image Corpus

In this thesis, an information retrieval system has been implemented using a Tamil document image corpus. Here, the corpus has been created with the images acquired from scanned hard copy of newspapers and magazines containing English and Tamil words (bilingual magazines) using flat bed scanner, greater than or equal to 200 dots per inch and images acquired from the Web. The corpus size is about 2000 document images, and each image represented as a two dimensional array holds a maximum size of 1000*600 pixels. Each image in the corpus undergoes the following processes defined so that the word images get represented as feature strings at the end.
3.2.2 Preprocessing

Archived document images often suffer from noise and various types of document degradation such as poor image resolution, impulse noise and physical degradation of documents (Lu and Tan 2008). Therefore, document images need to be preprocessed in order to extract the features of word images properly. Here, preprocessing involves noise removal and binarization, which enhances the input images to a great extent. Document images are first smoothed to suppress the noise by a simple median filter. Later, the filtered images are binarized (color images are converted into black and white images) using the global thresholding technique for further processing.

3.2.3 Layout Analysis

Document images are composed of a variety of physical entities or regions such as text blocks, lines, images, figures, captions etc, since they are generated by digitization using scanners from physical documents such as newspapers, magazines and brochures which contain complex layouts. As a result, understanding an arbitrary document image with a complex layout is an extremely difficult task and the process of a layout analysis is required to decompose a given document image into its component regions and understand their relationships.

This thesis envisages to design a Rectangular White Space Analysis (RWSA) technique for document layout analysis, since white spaces are a generic delimiter for layouts. Today publishers and printers in many languages use white spaces to separate blocks of text since they are constrained by common printing technology. Also this RWSA technique is not much restricted to thresholds and heuristics and adapts to various heterogeneous structures.
The RWSA technique consists of white space Section Finding, Section Merging, Cropping Extraneous section and Rectangular Formation phases. The Section Finding procedure analyzes and records the presence of sections containing white spaces (even smaller) in the horizontal directions. In Section Merging, consecutive white space sections within the line and between the horizontal lines have been merged to produce horizontal white space rectangles. On the other hand, Section merging takes place vertically to produce vertical white space rectangles. Non-intersecting edges of the rectangles and deviated sides of the rectangles are cropped. Later, Rectangular Formation goes through the intersecting horizontal and vertical white space rectangles to identify the layouts (required content blocks).

Once the physical regions have been identified, the next step is to isolate the textual regions from the non-textual regions in order to enable the text to get retrieved.

### 3.2.4 Text Image Analyzer

An automatic separation of the textual blocks from the non-textual blocks (graphics) is required to understand and represent the text given in the image. Textual blocks could be discriminated using either statistical or textual properties, since text blocks possess some regularity in their characteristics, such as frequency and orientation information as well as spatial cohesion (characters and words in the same line are of similar heights, spacing and orientation).

In this thesis, text and image blocks have been isolated using the statistical properties of a text obtained through heuristics. Two statistical properties such as Black Run length and Transition Rate, which span in horizontal direction, have been applied to identify the text. The Black Run
length observes the ratio of the black pixels to the total pixel area, whereas the Transition count records the black to white movement over a unit area. Mean Black Run length is produced by the average between the black run length measurement and transition count in a unit area and Mean Transition count is obtained by producing an average between the total transition count and pixels in the unit area. The above two properties produce uniformity in every unit area for the text blocks whereas a maximal variance has been produced for image blocks.

3.2.5 **Script Recognition**

Once the text blocks are isolated, they undergo the segmentation process to obtain the lines and word images. Since text understanding and information retrieval operates at the keyword level, a word spotting process is essential. In this thesis, Horizontal and Vertical projection profiles are employed to spot the word images.

A horizontal projection profile observes the pixel concentration of each row in the image and detects the presence of text lines between two consecutive peaks of the histogram. On the other hand, the Vertical projection profile detects the presence of word images by recording the peaks and valleys in a vertical direction.

Script/Language identification is essential in a multi-script environment, before understanding the text. In a document image of several script forms (Pal and Chaudhury 2001), it is necessary to separate the scripts before feeding them to their individual text understanding systems.

In this thesis, a Spatial Features Based Script Recognizer (SFBSR) has been introduced to identify the script of the word image (either English or
Tamil) using the first character of the word by analyzing its pixel density irrespective of font faces and sizes. This is a kind of character level script identification, which is less attempted in Indian scripts. Here, character images are segmented from the word images, and the spatial spread of the character over its total area extracted through zone segmentation and the extracted features transformed into digital values. The script is then determined according to the classification rules imported by the decision tree classifier using digital values. The results of the experiments and simulations with the proposed technique are promising for script identification.

3.2.6 Word Image Understanding

After script determination, Tamil word images pass through the word image understanding module, which extracts the features of the word images and generates the feature string for them. In this research, a new feature string generation algorithm called the Left-to-Right, Top-to-Bottom Feature String (LR-TB-FS) has been proposed, which frames a feature string as a result of left to the right and top to bottom scan of the word image. Line and Transition features are used here to represent the word images.

In the feature string generation process, Tamil word images are explicitly segmented into discrete entities from left to right using vertical projection profiles and each entity called as primitive, is represented using seven attributes such as Vertical Line (Vl), Horizontal Line (Hl), Vertical Transition rate (Vt), Three Horizontal Transition rates (Ht1, Ht2 and Ht3) and Outline attribute (Ol) after undergoing vertical and horizontal scanning. All entities (represented by the seven attributes) are concatenated together to represent the word image. The above said process can also be applied to generate the feature string for English word images in bilingual document images.
3.2.7 Representation of Document Images

A text file has been created for every document image in the corpus. Once the feature strings are generated for word images, they are written in their corresponding text files in a line. In other words, each line of the text file contains the nature of the script (as either English or Tamil) and feature string representation for every word present in the document image. In addition to this, the size of the word image has also been represented to determine its importance. Therefore, the storage contains the representation files for every document image in the corpus as part of the offline processing. Later, it is possible to search in these text files, to fetch the relevant documents.

3.3 ONLINE PROCESSING

Online processing, accepts the user query word to retrieve the relevant results. In this research work, Online processing consists of three processes, namely, Query word Rendering, Mapping process which generates the feature vectors and the Display of Images.

3.3.1 Query Word Rendering

A Graphical User Interface allows the user to type the query word either in English or Tamil to be searched over the document images. Users can also have the option to search the query word for partial matching in documents instead of an exact match. Later, the query word rendered is converted into a feature vector using the Mapping (transliteration) process.
3.3.2 Mapping Process

In this thesis, feature string tables have been devised, which consists of a primitive string (string) for each character in the Tamil and English Character set. These primitive strings are generated based on the LR-TB-FS algorithm stated above. Once the query word has been obtained from the user, a feature vector is generated by synthesizing the primitive string of each character in the word from the feature string table. The feature vector contains both the information about the script of the word image and the primitive string (composed character by character) for the word image (synthesized from the feature string table). Once the feature vector gets composed, it is searched over the collection of text files by matching the feature vector with the feature strings of relevant script (contained in text files) to fetch the relevant results.

3.3.3 Display of Images

When a set of text documents matches the query word, their corresponding document images are retrieved and the resultant images are ranked before displaying the images. The results are ranked based on the font sizes and frequency of occurrences of the query word in the text files and presented to the user. In the case of a Two-word query representation, Boolean ranking has also been done (And / Or operation) to rank the query results. This helps the users to easily evaluate the relevance of the retrieved pages to their need. Various metrics such as Precision, Recall and MAP score have been computed to measure the performance of the information retrieval system.

Therefore, the major contributions of developing the Keyword spotting based information retrieval system are: Design of a RWSA technique
for efficient physical layout analysis, Statistical properties for Text/Image separation, Spatial features-based script recognizer (SFBSR) for Tamil-English script recognition, Feature string representation using LR-TB-FS technique for Tamil word images, Touching character analysis, Feature string table for unique characters in the Tamil character set, Deviation string algorithm to accommodate various font faces and sizes in query processing and integration factors to achieve information retrieval system.