CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

The advancement in the field of electronics in the past decades has contributed greatly to the management of Information system. Many software tools such as word processors, drawing tools, and computer aided design software, and compilers are extensively used in the generation, storage and retrieval of Information in a format that is understood by computers. Storing of information in computers required format resulted in easy editing of Information, its printing or its distribution across the globe electronically by world-wide networks. Additional tools are also available for simulation, pattern searching, and improving the visual quality of the books, pictures, or video images and other areas.

An automatic recognition tools are developed with an intention of converting the textual materials available on physical media into digital form. This will bring life to the texts on the physical media as these texts can be processed and analyzed using a digital computer. The solution to this problem lies in the intersection of the fields of pattern recognition, image and natural language processing. Different areas of a typical character recognition system are shown in the Figure 2.1. OCRs are available to recognize printed text documents with simple layouts that to with few languages such as English, Japanese. However for Kannada language is concerned, only limited success have been achieved for printed Kannada characters and not much progress is made for handwritten Kannada character recognition. In spite of the intensive effort in the past decades, the recognition of free style and cursive handwriting continues to be important research area.
Large amount of data or information is available on the paper. Bulk of data being produced every day is available on the paper and that needs to be processed for various reasons. It has been estimated that a huge cost is required to key-in all these data into computers and is a laborious process. If the process of data entry is automated, significant cost can be saved. Added to this, the percentage of data that is brought online can be drastically increased. In all of the applications, the major goal is to extract the information contained in the documents and to store them in a computer expected format. This requires scanning of these documents. The background of the image, which is the color of the paper, affects the proper extraction of the foreground characters. This leads to poor recognition. In India one can find lots of information related to history and culture inscribed on the palm leaves and stones, on copper sheet and other media available at that time. Deformation, aging, dirt, cracks and other environmental parameters causes deterioration of these information. These information need to be extracted and stored in the modern gadgets for further understanding and analysis. To retrieve the information from these faded out, degraded images, it is necessary that background of the image need to be eliminated to make foreground image more legible.

Many researchers have focused on this issue using various techniques. Thresholding technique is popularly used for this purpose. Background of the image can be eliminated using posterior global thresholding process [Santos, José Eduardo B]. The tool based on Otsu’s method can be used for identification and elimination of the background [Sahoo, P.K.,

![Figure 2.1: The different areas of character recognition system](image-url)
Soltani, S. and Wong, A.K.C]. One of the most common approaches to this problem is color or grayscale background subtraction. Typical problems with this technique include foreground objects with the same colors as the background (produce holes in the computed foreground). To overcome this problem background estimation and removal based on the joint use of range and color is proposed [G. Gordon, T. Darrell, M. Harville, J. Woodfill]. Another interesting proposal discusses historical document image analysis using background light intensity normalization [Z. Shi and V. Govindaraju]. Histogram and background normalization method can be used to remove the background of palm leaf manuscript [Zhixin Shi and Govindaraju]. As all the pixels in the images are considered for processing, these methods are computationally intensive. We have proposed graph pyramid for background elimination and image segmentation. This technique reduces the dimension of the image which in turn reduces computation time.

### 2.3 Challenges in character recognition

It is the desire of human beings to bring in automation to make computers to mimic humans and get relieved from the monotonic operations he is doing. In spite of lot of research activities in this area, we have not reached a stage where a printed page can be inserted into an OCR system and a coded file comparable to a keyed in version emerges. Unusual typefaces may baffle the system, but much problem is due to missegmented characters. Lines with dropcaps, large mathematical symbols, or superscripts may be missed altogether. Most OCR systems still cannot read degraded documents and handwritten characters/words.

We learn to read and write during our initial stage of education. With time we acquire necessary skills for very good reading and writing, including the ability to read most texts, whether texts are printed in different fonts and styles, or handwritten neatly or sloppily. Most of the time we come across the many documents with varieties of attributes, but still we are able to read them. Light prints or heavy prints; upside down prints; advertisements in fancy font styles; characters with flowery ornaments and missing parts; and even characters with funny decorations, stray marks, broken, or fragmented parts; misspelled words; and artistic and figurative designs. At times, the characters and words may appear rather distorted and yet, by experience and by context, most people can still figure them out.
To make machine to learn different attributes of the font is challenging. Many attempts have been made toward this in English, Japanese, and Chinese languages. Font recognition single Chinese character based on wavelet feature has been reported [Li Chen, Ding, Xioqing, 2004]. Font detection based on global texture analysis has been proposed [Yong, Zhu, Tan, Tieniu, Wang, Yunhuai, 2001]. The local features of the font such as serifness, boldness etc has been used for estimation of font attributes in an OCR systems. Font types are also identified by using page properties such as histogram of word length and stroke slopes [Shi and Paulidis]. In Kannada language few attempts has been made in this direction. Taking the advantage of representation of data in a database, an attempt is also been made for converting printed Kannada text image to machine editable format using database approach [B.M.Sagar, Shoba, Ramakanthkumar P] and this works only when the text is represented by a single font type. A simple and efficient Optical character recognition system for basic symbols in printed Kannada text [R. Sanjeev Kunte, Sudhaker Samuel R. D] does not work well for different font type. Recognition of handwritten Kannada numerals based on zone based angle system extraction [S.V. Rajashekararadhya, P Vanaja Ranjan] is able to achieve good results for few digits. Vowel recognition system for Kannada language using fuzzy statistical approach has been developed [P. Nagabhushan, S.A. Angadi, B.S. Anami] works only for Kannada vowels. We have attempted to identify the font type based on empirical mode decomposition, which works at paragraph level. The font type is assumed to be same throughout the paragraph.

2.4 Segmentation Process

Image segmentation sub divides an image into its constituent parts. The level of the details to which subdivision is carried out depends on the problem in hand. It has been an important and challenging issue in the field of computer vision over decades. Segmentation of nontrivial images is one of the most difficult tasks in the image processing. It plays a critical role for most image analysis tasks, such as object recognition, object-based image compression and content-based indexing. Lots of techniques and algorithms have been proposed, and most of them are dedicated to the development of fully automatic segmentation. However, it still lacks a robust and unified algorithm that can be applied to all kinds of images.
2.4.1 Text segmentation

Segmentation of text is a process by which the text is partitioned into its coherent parts. The text image contains a number of text lines. Each line again contains a number of words. Each word may contain a number of characters. The following segmentation scheme is proposed where lines are segmented then words and finally characters. These are then put together to the effect of recognition of individual characters. The individual characters in a word are isolated. Spacing between the characters can be used for segmentation.

2.4.1.1 Line Segmentation

Horizontal projection profile of the document image is used to detect the valleys between the text lines is used to extract the lines.

2.4.1.2 Word Segmentation

Vertical projection profile is used to extract the words.

2.4.1.3 Character Segmentation

Vertical projection profile is used to separate the first character of the word.

2.4.2 Image segmentation

2.4.2.1 Pixel-based methods

Pixel-based methods directly work in the feature domain without considering spatial information into consideration. Thresholding is the simplest technique that produces segments with similar intensities. It plays an important role in many forms of pattern recognition and classification. The thresholding requires that an object has uniform intensity and a background with a different intensity.

2.4.2.2 Finite mixture segmentation

Finite mixture segmentation technique assumes that the input image (or images) consists of a fixed number of distinct objects (or tissue types). The intensities of pixels in each class conform to a multivariate Gaussian distribution, and the entire histogram can seemingly be approximated by the summation of these Gaussian distributions.
2.4.2.3 K-means clustering segmentation

Another approach to the segmentation is to cluster image pixels based on their intensities into $k$ partitions, referred to as K-means clustering segmentation. It clusters the data by interactively computing a mean intensity for each class and segmenting the image by classifying each pixel in the class with the closest mean.

2.4.2.4 Edge-based segmentation

Edge-based segmentation method exploits the spatial information by detecting the edges in an image. Edge-detection and linking algorithms rely on an edge-detecting filters such as Prewitt, Sobel, and Laplacian, that is then thresholded to produce pixels that are flagged as being a part of an edge. These edges, once extracted, are connected by some method appropriate for the problem domain.

2.4.2.5 Active contour models

Active contour models are another technique for segmentation. The existing active contour models can be divided into parametric snakes and geometrical snakes. Based on the features being used in leading the evolution, active contour models can also be classified as edge-based and region-based models. Edge-based models are based on the notion of shrinking/expanding a curve until it reaches high-gradient areas.

2.4.2.6 Region based image segmentation

Regions in an image are a group of connected pixels with similar properties. In region approach each pixel is assigned to a particular object or region. It is based algorithms are more global characteristics of the image and relies on entire regions instead of individual pixels. Assumptions are typically used such as constant image intensity within each region. These methods can accurately segment images where these assumptions are warranted.
2.5 Classification

The classification is the process of identifying each character and assigning to it the correct character class. Different types of classifiers such as minimum distance classifiers, statistical classifiers and neural networks are used for this purpose. Each of these classification techniques are briefly described below.

2.5.1 Matching

Matching covers the groups of techniques based on similarity measures where the distance between the feature vectors, describing the extracted character and the description of each class is calculated. Different measures may be used, but the common is the Euclidean distance. This minimum distance classifier works well when the classes are well separated, that is when the distance between the means is large compared to the spread of each class. When the entire character is used as input to the classification, and no features are extracted (template-matching), a correlation approach is used. Here the distance between the character image and prototype images representing each character class is computed.

2.5.2 Optimum statistical classifiers

In statistical classification a probabilistic approach to recognition is applied. The idea is to use a classification scheme that is optimal in the sense that, on average, its use gives the lowest probability of making classification errors.

A classifier that minimizes the total average loss is called the Bayes’ classifier. Given an unknown symbol described by its feature vector, the probability that the symbol belongs to class $c$ is computed for all classes $c=1...N$. The symbol is then assigned the class which gives the maximum probability.
2.5.3 Neural Networks

Neural networks to recognize characters are composed of several layers of interconnected elements. A feature vector enters the network at the input layer. Each element of the layer computes a weighted sum of its input and transforms it into an output by a nonlinear function. During training the weights at each connection are adjusted until a desired output is obtained. A problem of neural networks in OCR may be their limited predictability and generality, while an advantage is their adaptive nature.

In our approach, we are using Empirical mode decomposition method to extract the feature of font type and Non negative matrix factorization to extract the feature of the characters. We used Euclidean distance and Earth movers distance to compare the distance between the features and classify them accordingly.

2.6 Objectives of the thesis

The Design of automatic recognition system is quite challenging. Even though the computers are presently available with large computation speed, large memory space it is difficult to make computers to mimic human beings. A lot of processing needs to be done before the compute is able to recognize the documents which is present in the digital form. Researchers are partially successful in this automation process. The documents are prepared with different fonts, styles, sizes. The information may be present on different type's media, the data present as a digital image may have different backgrounds which affects the extraction of the data of interest. The documents may come with varieties of background to make it more attractive which poses many problems. Segmentation of the image is another important issue that affects the selection of region of interest. The documents may be of the type printed or handwritten. Printed documents have specific formats and algorithm may be designed to recognize them. The recognition of handwritten documents is very difficult as each individual prefers his own style of writing and as such automation of such documents is difficult. Identification of the writer based on his handwriting is another challenging issue. Looking at these problems, it becomes clear that there is a much complexity is involved in the design of pattern recognition system. This brings the question: How to solve such problems and develop good automation system?
Attempts are made for the development of automatic character recognition system with the view that the documents are available with single font type. However, in practice, the documents are prepared with different fonts, may have different background, and may contain different languages. In such situations, the recognition rate comes down drastically. Documents need to be preprocessed before they are sent to the character recognition system. There is not much work reported for the recognition of handwritten documents. Considering all these facts into account, we have developed novel methods to address the problems that are encountered in pattern recognition systems. We are sure these new techniques will enhance the capabilities of Kannada OCR system.