CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Digitization of a country essentially requires equipping its people with the state of art technological advancements and thereby, providing them means of social mobilization and accessibility. To reach the lowest strata of society it is very essential that the channels of communication are of very native origin. In India where a large number of people reside in rural areas and have a little knowledge of English language it is imperative to facilitate localized channels of communication. In addition to this a very large number of languages face the severity of extinction primarily due to neglect of native textual resources and the vast wisdom hidden. In this era of English dominance native languages including Kannada face the problem of negligence towards it's ancient and wisdom oriented textual material, this makes it necessary for the development of digital based support system, which would make the rich indigenous resource of the language universally accessible and hence lead to its higher standing in the global linguistic arena. So the solution to all this is development of a text recognition system which helps individuals who have little knowledge about the language and hence requires assistance to read any information presented in this language. Such a system has wide applications such as form filling, word processing and text-to-speech conversion. Different techniques have been explored for character recognition and efficient practical applications exist in English language. However, such a system does not exist in other languages.

Majority of the research reported are for Indian languages and have dealt either with a subset of characters such as only the numerals or the base characters. Given to its vast character set and confusing curvatures it is very hard to create a system with easy and involves considerable complications.

1.1.1 PROBLEM STATEMENT AND MOTIVATION

In the present scenario when an enormous number of languages are on the verge of losing its valuable literary resources due to negligence by stake holders, Kannada is also facing a slight heat of the situation. the necessity of a reliable Kannada OCR system is clearly noticeable. With some research, it has been found that there only a few Kannada OCR systems available on the market. These systems, however, aren’t reliable and accurate enough to provide the best results. As per our research, using Convolutional Neural Networks, or CNN’s, provides the most accurate results. Unless the system is trained by appropriate methods, the run time will be much more than what it should be. With the use of CNN, the run time decreases rapidly and an agile system with good accuracy is functional.

The motivations to take up this project include the points mentioned below

1.Development of native dwellers

2.Effort towards preserving ancient Kannada literature

3.Creation of e-libraries with universal accessibility

4.Increase International linguistic competitiveness

5.Overcome language barriers and increase feeling of international fraternity

1.2 OPTICAL CHARACTER RECOGNITION

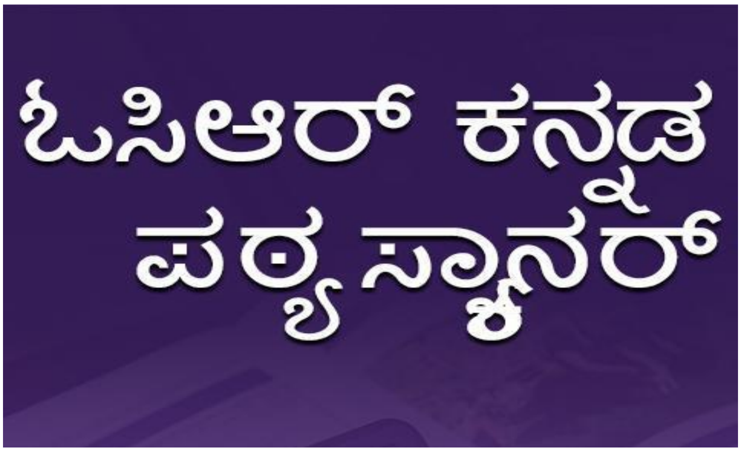

Figure 1.1: Optical Character Recognition

OCR is the mechanical or electronic conversion of images of typed, printed or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast) as shown in the figure 1.1 is the kannada translation of OCR.

It is a common method of digitizing printed text so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computational analysis, Data extraction, machine translation (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. With the ever-growing field of data mining, data extraction and hence Policy making, it is essential for an OCR in vernacular language.

1.3 APPLICATIONS OF PRINTED CHARACTER RECOGNITION

1.3.1 Digital Character Conversion

Some documents may be damaged therefore unable to recognize the characters in those documents. A digital character conversion system identifies characters easily and converts them into a people-readable format. Figure 1.2 shows that a printed text contains some letters that are hard to read. In such situations, a digital character conversion system can convert the text into a readable format. This application facilitates humanoid robots to read printed characters and words. It helps in data extraction algorithms as well.

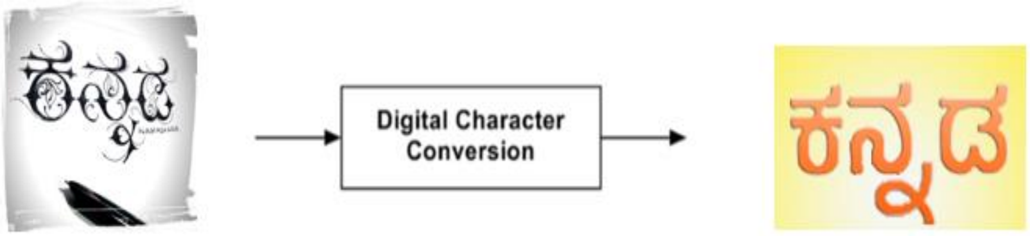

Figure 1.2: Block Diagram of Digital Character Conversion

1.3.2 Meaning Translation

With the help of printed recognition system, one language can be converted into another language. Many people wrote stories, documents, novels, research works in their own languages. This meaning translation system can be used to convert these images into another language. Figure 1.3 illustrates Kannada to English meaning Translation System. This offline printed character recognition system can be used to recognize characters from Kannada-based document images and translate them into English meaning. This can also be adapted for other indigenous languages as well and helps common people.

Figure 1.3: Block Diagram of Kannada To English Meaning Translation Application

1.3.3 Content Based Image Retrieval

Many researchers are developing various content-based image retrieval methods, such as those based on text and color. Figure 1.4 displays an example of content-based image retrieval based on text, in which the user enters a word in the search box, the search engine then retrieves the document images that contain the search word. Here the word image is searched for and highlighted in the document as shown.

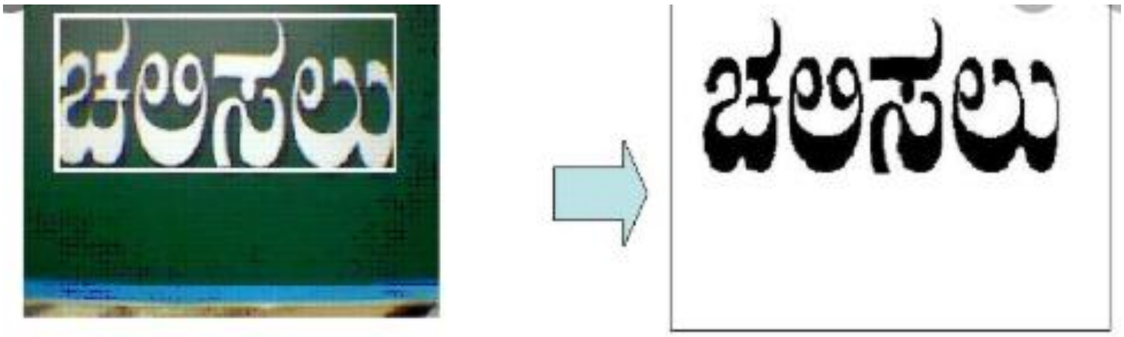

Figure 1.4: Content Based Image Retrieval

1.3.4 Signboard Translation

Signboard Translation is necessary to translate one language to another language. In many countries, public display boards are in their native languages. When people from other countries need to understand these display boards, they use this Signboard Translations system. In signboard translation, character-by-character translation is necessary in order for the original meaning to stay unchanged. In India some village's names may have other meaning, therefore, character-by-character translation is necessary. Figure 1.5 displays the signboard translation process.

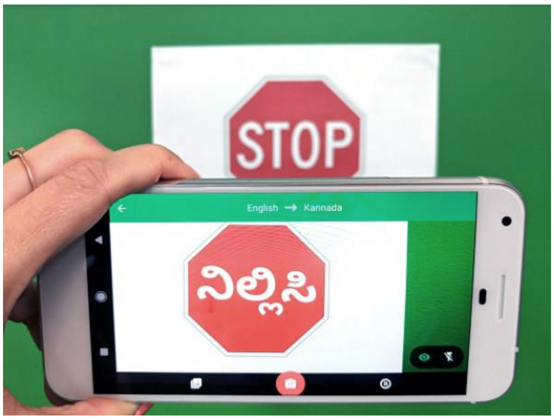

Figure 1.5: Signboard Translation System

1.3.5 Text-to-Speech Conversion

Converting the text of any printed document image into speech is called text-to speech conversion. This application is very useful for visually handicapped people The Text to Speech Conversion application is used to convert the image to audio form. The application takes help of offline printed characters recognition system to convert the text to speech format as shown in Figure 1.6.

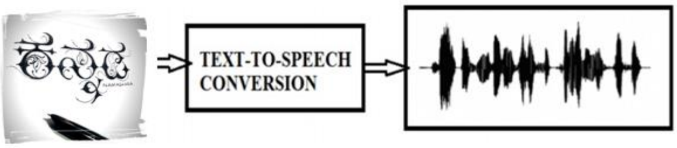

Figure 1.6: Text-to-Speech Conversion System

1.4 OBJECTIVES

Pre-processing of scanned text to remove noise, blur, and skew correction:

Pre-processing is required as noise, blur and skew are important aspects of old textual resources therefore it is essential to develop a means to make corrections required.

Segmentation of lines, words and characters:

Due to large number of character units there needs a mechanism to reduce it to smaller set of symbols and hence segmentation is required.

Train a model that uses Convolutional Neural Network:

For classification a CNN is thought of and is to be worked on to obtain good accuracy and result.

Converting to machine readable format (e book, pdf):

Once the segmentation and classification are done with satisfactory levels of accuracy, machine readable format implementation, that is creation of e-books and hence e-libraries could be created.