Chapter 1

Introduction

1.1 Background Study

This project aims to propose work towards automatic recognition, extraction and translation of Kannada language natural sign board characters. In particular, it focuses on recognizing every single character in textual images. Figure 1, 2 and 3 apotheosizes the reason for this to be a flinty job. Even after ignoring the difficulty of text partitioning momentarily, the sources which brings problems are as follows:

i. font style and font size
ii. graphics in background and foreground
iii. camera alignment yielding to geometric distortion
iv. illumination
v. resolution of image
vi. removal of unwanted objects
vii. identifying threshold value for different images and
viii. edge detection

The above listed features give rise to the problem of object recognition. Henceforth, this technique cannot be used. Moreover, within the realm of possibility of such systems, OCR systems have been designed and developed for very less languages. Recognition of all languages of Indic origin is beyond their capability.

Figure 1.1: Sample Source Image in Our Dataset

Figure 1.2: Examples of Different Classes of Texts in Different Dimensions

Figure 1.3: Sample images of Kannada Characters belonging to separate classes.

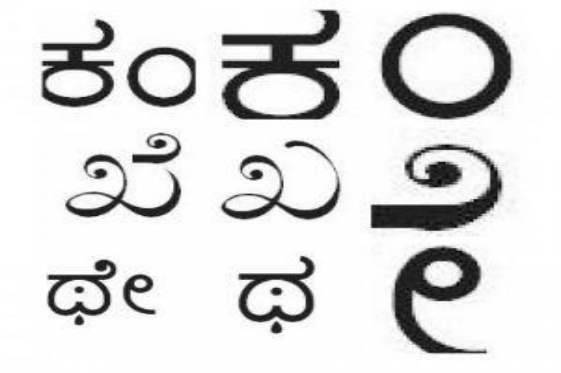

Figure 1.4: Characters followed by the attachments making it two different styles of As per Kannada Kagunitha, vowels on combining with consonants, it creates new character.

Various problems are to be solved so as to read Kannada texts from natural scene images in conjunction with finding text, identification of characters, partitioning of words. We try to highlight via this paper to recognize characters; documenting recognized characters into editable file format and lastly translating the phrase/ sentence recognized

to another language (say English here). A standard database comprising of images/ characters of Kannada language is introduced. To measure the practicability of presenting the work as Kannada Character recognition, we establish a paradigm to measure the effectiveness of different attribute based on a standard data set as described above. The result delineates that indeed the confinement of character acknowledgment assignment is a dull work. The number of classes for Kannada characters is almost 619 counting consonants, vowels, digits and other images utilized with exceptionally small inter-class disparity as finished by Figures 2 and 3. This issue is eminently exceptionally ghastly for Kannada dialect where two typical characters in the letter set can contradict fair by substitution of a single speck, hyphen, bar, comma, accentuation like structure as appeared in Figure 4. The distinguished characters will be diverted to an editable record arrange. These writings can be afterward utilized for different purposes as required. The recognized characters can be archived in Latin script.

1.2 Overview

Right from the dawn of human civilization, people have started migrating from places to places for various reasons. Our Country; India being a nation with diverse culture and languages, people on moving away from their native may find it very much difficult to understand different languages. Let us say a non-native speaker of Kannada language comes to the state of Karnataka may not be able to read the Kannada script. So, we propose a system that reads the text encountered in natural scenes with the aim to provide assistance to the non-native speakers of Kannada language.

The area of scene text recognition focuses on the problem of recognizing arbitrary text in images of natural scenes. Examples of scene text may include street signs, name of shops, grocery item labels, and name plates etc. With the increased use of smart phones and digital cameras, the ability to accurately recognize text in images is becoming increasingly useful and many people will be benefited from advances in this area.

1.3 Problem Statement

This project presents work towards reading Kannada text from natural scenes. After recognizing text, it will be converted into editable format of Kannada text. After recognizing the text, it will be converted into any other desirable language. For now we focus to convert the text into English language.

1.4 Introduction to OCR

Optical character acknowledgment (moreover optical character recognition, OCR) is the mechanical or electronic transformation of pictures of written, manually written or printed content into machine-encoded content, whether from a checked report, a photo of a report, a scene-photo (for illustration the content on signs and bulletins in a scene photo) or from subtitle content superimposed on an picture (for case from a TV broadcast).

In OCR processing, the scanned-in picture or bitmap is analyzed for light and dim regions in arrange to recognize each alphabetic letter or numeric digit. When a character is recognized, it is changed over into an ASCII code. Extraordinary circuit sheets and computer chips planned explicitly for OCR are utilized to speed up the acknowledgment process.

OCR is being utilized by libraries to digitize and protect their property. OCR is moreover utilized to prepare checks and credit card slips and sort the mail. Billions of magazines and letters are sorted each day by OCR machines, impressively speeding up mail conveyance. The point of this venture is to recognize Kannada Normal Sign Board Characters. This can be valuable for both local speakers and non-native speakers as well for different reasons.

1.4.1 Types of OCR

- Optical character recognition (OCR) – targets typhoid content, one glyph or character at a time.
- Optical word recognition – targets typhoid content, one word at a time (for dialects that utilize a space as a word divider). (Ordinarily fair called "OCR".)
- Intelligent character recognition (ICR) – moreover targets transcribed print script or cursive content one glyph or character at a time, ordinarily including machine learning.
- Intelligent word recognition (IWR) – too targets manually written print script or cursive content, one word at a time. This is particularly valuable for dialects where glyphs are not isolated in cursive script.

OCR is for the most part an "offline" handle, which examinations an inactive report. Penmanship development examination can be utilized as input to penmanship acknowledgment. Instep of simply utilizing the shapes of glyphs and words, this

procedure is able to capture movements, such as the arrangement in which fragments are drawn, the course, and the design of putting the write down and lifting it. This extra data can make the end-to-end handle more exact. This innovation is moreover known as "on-line character acknowledgment", "energetic character acknowledgment", "real-time character acknowledgment", and "cleverly character acknowledgment".

1.5 Motivation for Work

We have seen people migrating from place to place leaving their native place and going to different linguistic place. It often becomes difficult to read and understand another language. So, we came up with the idea to design a program to overcome the above problem. This can be beneficial not only to non-native speaker but also to native speaker as well.

1.6 Problem with Existing System

- OCR system is not present for all languages.
- No systems have been developed for Indic Languages. E.g. Kannada
- Identification and processing of character is difficult.
- Redirecting the character in the file is difficult.
- Translation is not available.

1.7 Proposed Work

Our objective is to develop a system that is able to perform various works which meets the requirement of user.

The Proposed system works in following manner.

- Segmentation - HOG
  i. Browse any scene text image.
  ii. Convert RGB to GrayScale.
  iii. Calculate threshold from GrayScale.
  iv. Calculate Canny Edge Detection.
  v. Removal of unwanted objects.
  vi. Feature extraction.

- Training a SVM model
- Testing a SVM model
- OCR

i. Browse text image.

ii. Convert the recognized character into Latin Script.

iii. Redirect into file.