Chapter 2

LITERATURE SURVEY

2.1 Existing System

Text extraction from the images has gained research interest. On this note OCR was invented. OCR is a kind of pattern recognition which contributes towards the recognition of text in the documents. Computer vision, artificial intelligence and pattern recognition contains OCR as part of research. When a document or paper is scanned by computer, it gets only image file. The text on the page is not understandable by the computer, so it is not possible to search or edit the page. OCR software can be used to produce editable file which is more flexible. The existing system of OCR works on grid infrastructure, i.e. works without a grid infrastructure. It deals with homogeneous character recognition of single language. Existing work focus on recognizing characters using bounding boxes. Three different classifiers are trained for this purpose i.e. k-nearest neighbor, random forest and neural networks. Further segmentation technique is used where text pixels are isolated from the background. Due to few drawbacks of OCR technique convco-HOG was proposed where it has more differentiating power by repeatedly examining all possible image pixels. This was developed to recognize characters in natural scenes. Since natural scene characters are type of hand written, their style varies recognizing each style is not easier task. Along with this there are few more features that affect the recognition process i.e. color texture of background and foreground, geometric distortion which is caused by camera position, illumination and resolution of the natural scene image and this difficulty can be observed in Figure 2.1. This is done by introducing a database of images containing English and Kannada text.

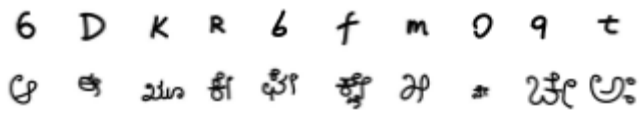

Figure 2.1- Handwritten Characters

2.2 Related Work

The work of Kannada characters acknowledgment in characteristic scenes is related to issues considered in camera arranged record examination. Larger part of the work in scene content acknowledgment is particularly based [10], [9], [4] and [3] on finding and adjusting the

content zones and taking after the OCR application strategies. [8] Such approaches are in this manner limited to environment where OCR works well. From this time forward such approaches are controlled to environment where OCR works appropriately. In expansion to amendment prepare, it does not straightforwardly relate to our work, as it points on discovery of printed characters. The edge detection is carried out by the technique described by J Canny, called Canny Edge Detection Technique [24] and also by an improved edge detection technique [23].

The technique for static recognition of hand written characters have been efficiently solved by intra-class variation due to non-identical styles of writing [15], [14]. Such scenes prototypically assume only a finite number of appearance classes, unable to resolve differences in foreground/background color and texture, especially the graphics present. This is achieved by identification and removal of extraneous graphics in a commercial OCR operation [25]. For occurrence, [16] we have utilized cognizance from NLP and display a Markov chain system for parsing pictures. [5] Presentation of composition machines for developing probabilistic progressive picture models. This makes a difference in obliging relevant connections. This approach permits re-usability of parts among different substances and non-Markovian disseminations. [16] Proposed a strategy that amalgamates picture highlights and dialect data a single demonstrate and coordinating disparity data between character pictures.

Acknowledgment of digits utilizing pipelines based on crude pictures classifications have been broadly utilized [12]. [21] By shape coordinating procedure, this is too done [1]. The classification is carried forward by HOG technique, known as Histogram Oriented Gradient. In this line by line detection of characters and words is done.

2.3 Data Sets

We aim to recognize Kannada characters from natural scene images. To do so, we design a database containing images of natural scene having Kannada characters. These images have been gathered from around the streets of Tumkur and Bangalore in Karnataka, India. The natural scene images comprises not only of street symbols but also of sign boards, hoardings, posters, pamphlets, banners, name plate, number plate etc. However, collection and annotation of huge sample of images is a costly as well as costly job. So, we acquired a database of characters generated by computer fonts of different size as shown in Figure 2.2.

Figure 2.2: The Standard Dataset containing all possible Kannada language Characters and Digits

English language has characters separately in two cases namely upper case and lower case, but in case of Kannada Language it's not the same. Kannada language alphabet does not have the system of upper-case and lower-case characters. It has 37 consonants and 16 vowels. By combining the vowels with consonants, it generates around 603 distinct classes. It has numerals from 0 to 9 which can further have combined to generate infinite number of terms. Digits can be identified separately.

2.4 Drawback of Existing System

The drawback in the early OCR system is that they only have the capability to convert and recognize the textual image of English language to any desired language. That is, the older system is unilingual.

2.5 Proposed System

The goal of this project is to develop methods for improving natural scene text recognition. Here we have focused on recognizing characters in situations that would traditionally not be handled well by OCR (Optical Character Recognition) techniques. We will have an annotated database of images containing English and Kannada characters. The problem is addressed in an object categorization framework based on a bag-of-visual-words representation. We will assess the performance of various features based on nearest neighbor and SVM (Support Vector Machine) classification.