ABSTRACT

‘Optical Character Recognition’, or OCR, is basically converting an image containing text to an editable text format. The image could either be a scanned document, or a simple newspaper cut-out. Using Supervised Learning in the form of Neural Networks will make the system produce the output required with a much larger accuracy.

For a country to be completely digitized it is very important that each and every section of society should be able to interact with technologies (banking sectors, health care industries) comfortably. For this people should know English. But India is a multilingual country consisting of 22 official languages. Only 10% of Indian population know English. So, it becomes very important to develop a system in which people should be able to interact with these technologies in their regional language. There is also scarce availability of Kannada e-books, this effort towards creation of an OCR would give an essential mean towards such undertakings. The scarcity of e-books specifically refers to older literary work.

The objective of the project is to develop an OCR system that can recognize characters of Kannada language and then convert them to a machine-readable format. Kannada script consists of 6,23,893 characters. The complexity is reduced by segmentation approach. The vattaksharas are extracted/differentiated from the words by using base-line technique. When the characters are recognized, they are compared with Unicode’s available on the system and then printed. In the above method, CNN plays a pivotal role in reading the character and comparing it with the Unicode look up table values to print the output. Dataset consisting of 340 classes of Main Aksharas and 32 classes of Vatt Aksharas are used to train the CNNs. An average of 25 images per class have been collected and this amounts to a total of around 9000 images. For determining the accuracy of the system, five paragraphs with distinct font have been used. The Prediction Accuracy obtained is 82% and the Segmentation Accuracy obtained is 96%.