Vilas M. Rokade, An optical character recognition system for Devanagari script, M. Tech. Thesis, Department of Electrical Engineering, Indian Institute of Technology Bombay, 1999.

Supervisor(s): Prof. R. K. Shevagoankar and Prof. P. C. Pandey

***Abstract* –** This project aims at developing an optical character recognition system for the Devanagari script. The approach is based on syntactic pattern recognition principles. Input to the system is to be unifont printed text. However it can contain different sized letters for the same font.

The process of recognition begins with preprocessing the scanned image. It consists of binarisation, tilt correction and thinning. The image obtained after preprocessing is segmented into lines, words, matras and characters (single and compound) in that order. The image of each character so obtained is partitioned into nine equal regions by placing a 3x3 grid over it. This takes care of scaling and aspect ratio.

Each section of the partitioned character is then searched for features. The features used in this system are 1) joint points, 2) crosses and 3) tails. The feature from each section is coded into a 12 bit binary vector, forming a 108 bit binary vector for the entire character. This vector is then compared with stored vectors obtained from a training set and assigned a code if a match is found. The matras are similarly analysed and coded. The grid for matras is kept 3x2 as they have aspect ratio of approx. 3:2, whereas most characters have aspect ratio of 1:1. The recognition accuracy of the system is 80%.