335 C

Optical Character Recognition (OCR) of Sinhala characters by feature analysis followed by matrix matching

In this study, a scanned image of a character written in black on white paper is converted to a monochrome bitmap file. The bitmap file is then read into a 2-D matrix in a popular mathematical application software.

The elements of the 2-D matrix were made either 0 (representing black) or 1 (representing white). We analyzed the matrix using many algorithms developed by us to find out many features such as

(i) height/ width ratio,
(ii) location of the centre of gravity of the convex hull of the character,
(iii) presence of long straight line sections such as what we have in "rayanna",
(iv) the number, sizes and the centres of gravity of closed regions of the character,
(v) whether the roof of the character is curved down at the left and right sides so that it does not "retain rainwater" .

After testing 100 samples of characters, the number of times a particular character associated with a certain feature was recorded in a database.

Results showed that by using this database, a scanned image of an unknown character could be identified with a good degree of probability. The identification can be made fool proof by using matrix matching of the scanned image with a template of the most probable character.