Abstract

Tesseract OCR Engine is one of the most efficient open source OCR engines currently available. Recently, Tesseract OCR 3.01 is capable of recognizing Hindi language but still it needs some enhancement to improve the performance. The Hindi language recognition accuracy is quite low even for the printed text, as the conjunct character combinations of Hindi Language are not easily separable due to partial overlapping. The proposed approach solves this problem, so that Devanagari conjunct characters can easily be segmented and recognized using Tesseract OCR Engine. This paper presents a complete methodology to improve The Hindi Language Recognition accuracy. This paper also presents comparison with other Devanagari OCR engines available on the basis of recognition accuracy, processing time, font variations and database size.
References

- Smith, R. “An Overview of the Tesseract OCR” in proc. ICDAR 2007, Curitiba, Paraná, Brazil.
- tesseract-ocr An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google. Available at: http://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3/ (last accessed 8 January, 2012)

Index Terms

Computer Science

Pattern Recognition
Keywords
Tesseract  Hindi  OCR  Shirorekha Chopping  Character Segmentation