Abstract

Many valuable paper documents are usually scanned and kept as images for backup. Extracting text from the images is quite helpful and thus a need for some tool for this extraction is always there. One of the important applications of this tool is its use in Braille Translation. Braille has been the primary writing and reading system used by the visually impaired since the 19th century. This application that extracts text from images and then converts it to Braille will prove to be quite useful for converting old valuable documents or books into Braille format. In this paper the complete methodology used for the extraction of texts from scanned images and for the translation of texts to Braille is presented. The scanned images are initially pre-processed and converted to grayscale and then passed through an adaptive threshold function for conversion to binary image. Then it is sent for Recognition using Google’s powerful Tesseract recognition engine which is considered to be the best Open Source OCR Engine currently available. The generated text is then post-processed using a spell checking API JOrtho for removing the errors in the previous step. The final corrected text is then translated to a six dot cell Braille format using a set of rules provided by www. iceb. org. The translation to Braille includes conversion of numbers, alphabets, symbols and compound letters. The translated text can then be saved for printing the document later or for sending it to a Refreshable Braille Display.
An Open Source Tesseract based Tool for Extracting Text from Images with Application in Braille Translation for the Visually Impaired

References

- Ray Smith, Chris Newton, Phil Cheatle, Adaptive Threshold for OCR: A Significant Test, HP Laboratories Bristol, March 1993
- Ray Smith, Tesseract OCR Engine, OSCON Conference 2007
- Tess4J Project Site: http://tess4j.sourceforge.net/
- JOrtho Project Site: http://jortho.sourceforge.net/
- Paul Blenkhorn, A System for Converting Braille to Print, IEEE Transactions on Rehabilitation Engineering, Vol. 3 No., June 1995
- Md. Abul Hasnat, Muttakinur Rahman Chowdhury, Mumit Khan, An open source Tesseract based Optical Character Recognizer for Bangla script, 10th International Conference on Document and Recognition, 2009
- BrailleOCR Project Site: https://code.google.com/p/brailleocr/

Index Terms

Computer Science

Pattern Recognition

Keywords

OCR Tesseract Tess4J JOrtho Phonetic Matching Soundex Braille Braille Translation

Braille ASCII
An Open Source Tesseract based Tool for Extracting Text from Images with Application in Braille Translation for the Visually Impaired