{tag} {/tag}
IJCA Proceedings on International Conference on Distributed Computing and Internet Technology 2014

© 2013 by IJCA Journal

ICDCIT 2014

Year of Publication: 2013

Authors:
Mamata Nayak

Ajit Kumar Nayak

{bibtex}icdcit1306.bib{/bibtex}

Abstract

Development of Optical Character Recognition (OCR) for an Indian script is an active area of research today. The presence of a large number of letters in the alphabet set, their sophisticated combinations and the complicated grapheme's they formed is a great challenge to an OCR designer. There are many application areas where, OCR can be used like, preserving old documents in electronics format, helping visually impaired persons to know the content of a document by transforming into speech, saving document images within limited space, making a electronic dictionary of words, preserving the ancient characters those are not included in the current set of characters of a language and many more. Currently, Tesseract,

an open source OCR engine is considered as one of the most accurate FOSS OCR engines. Tesseract has already been designed to recognizing English, Italian, French, German, Spanish and Dutch and many more [11], as well as for few Indian languages such as Bengali, Tamil, Telugu, Malayalam. Similarly, Tesseract can be made to recognize other scripts if the engine can be trained with the requisite data. The objective of this work is to develop a training process for Tesseract OCR engine such that the engine will be capable of recognizing printed documents of Odia language used in the state of Odisha (formerly known as Orissa), India.

References

- Md. Abul Hasnat, Muttakinur Rsahman Chowdhury, Mumit Khan, "Integrating Bangla Script recognition support in Tesseract OCR", Proceeding of the Conference on Language & Technology, pp-108-112, 2009
- Nick White, " Training Tesseract for Ancient Greek OCR", October 2012 Available: www. eutypon. gr/eutypon/pdf/e2012-29/e29-a01. pdf
- Md. Abul Hasnat, Muttakinur Rsahman Chowdhury, Mumit Khan, "An open source Tesseract based Optical Character Recognizer for Bangla script", 10th Internal Conference on Document Analysis and Recognition (ICDAR), pp. 671-679, ICDAR, 2009. Available: http://ieeexplore. ieee. org/stamp/stamp. jsp?tp=&atnumber=5277476
- Ray Smith, "An Overview of the Tesseract OCR Enigine", Proc. of ICDAR2007, Curitiba, Parana, Brazil, 2007. Available: http://ieeexplore. ieee. org/xpl/freeabs_all. jsp?atnumber=4376991
- Jane Horgan , " Critical review of Tesseract", 4th April 2010 Available: https://muelli. cryptobitch. de/paper/2009-Tesseract-Review. pdf
- Sandip rakshit, Subhadip basu, "Development of a multi-user handwriting recognition system using tesseract open source engine", Proceeding International conference on C#IT, pp 240-247,2009.
- http://code. google. com/p/Tesseract-ocr/
- Md. Abdul Hasnat, " How to train Bangla and Devanagari script for tesseract engine", pp. 1-4, 2008. Available: www. ias. ac. in/sadhana/Pdf2002Feb/pe990. pdf crblpocr. blogspot. com/. . . /how-to-train-bangla-devanagari. html
- B. B. Chaudhuri, U. Pal, M. Mitra, "Automatic Recognition of Printed Oriya Script", pp. 23-34, Sadhana, Vol. 27, part 1,2002. Available:
- Sanghamitra Mohanty, Hemanta Kumar Behra, "A complete OCR Development System For Oriya Script", Proceeding of SIMPLE, Vol. 4, 2004.
- Ray Smith, Daria Antonova, Dar-Shyang Lee, "Adapting the Tesseract Open Source OCR Engine for Multilingual OCR", Proceedings of the International Workshop on Multilingual OCR 2009, Barcelona, Spain July 25, 2009. Available: http://doi. acm. org/10/1145/1577802. 1577804
- http://tesseract-ocr. googlecode. com/svn-history/r719/ trunk/doc/tesseract. 1. html
- http://en. wikipedia. org/wiki/List_of_ISO_639-2_codes
- George Nagy, "At the Frontiers of OCR", Proceeding of the IEEE, Vol 80, No. 7, pp. 1093-1100

Index Terms

Computer Science Pattern Recognization

Keywords

Tesseract Ocr Utf-8 unlv Odia.