Optical Character Recognition

Lecture 6

Qurat-ul-Ain (Ainie) Akram
Sarmad Hussain

Center for language Engineering
Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology, Lahore, Pakistan

Classification and Recognition Using Tesseract

- One of the most popular open source multilingual recognizers
- Initially developed by HP and later in 2006, Google acquired the engine and has released its complete source code at http://code.google.com/p/tesseract-ocr/
- It uses polygonal approximation technique (Smith, 2007)
- Tesseract-based OCR
  - English, French, Italian, German, Spanish and Dutch with 99.25% character recognition accuracy (Smith 2009)
  - Russian with 98.67% for character recognition (Smith 2009).
  - Bangla with 70% accuracy for screen printed text and 93% for cleaned images (Hasnat et.al 2009)
  - Chinese with 84.59% character recognition accuracy (Smith 2009)
  - Davanagri with 96.23% character recognition accuracy (Smith 2009)

Hasnat, M., Habib, M. and Khan, M.. "A High Performance Domain Specific OCR for Bangla Script", Int. Joint Conf. on Computer, Information, and Systems Sciences, and Engineering (CISSE), 2007
Smith, R., *"An Overview of the Tesseract OCR Engine"*, Proc. Ninth Int. Conference on Document Analysis and Recognition (ICDAR), pp. 629-633, 2007.
Smith, R., Antonova, D., Lee, D., *"Adapting the Tesseract open source OCR engine for multilingual OCR"*, In *Proceedings of the International Workshop on Multilingual OCR* (MOCR '09).

Tesseract Training

- Software Requirements

1. **Tesseract**
   1. Download from http://code.google.com/p/tesseract-ocr/downloads/detail?name=tesseract-3.00.1.exe.zip&can=2&q= and run tesseract3.exe setup file
2. **Irfan View**

   Open image in IrfanView and save it in uncompressed TIF format
3. **bbTesseract** used to edit box file

   http://code.google.com/p/bbtesseract/downloads/detail?name=bbT_exe_00_05_38.7z&can=2&q=

https://code.google.com/p/tesseract-ocr/wiki/TrainingTesseract3

Training Steps

- Required Data Files

1. tessdata/eng.unicharset
2. tessdata/eng.inttemp
3. tessdata/eng.pffmtable
4. tessdata/eng.normproto
5. tessdata/eng.config
6. tessdata/eng.unicharambigs
7. tessdata/eng.punc-dawg
8. tessdata/eng.word-dawg
9. tessdata/eng.number-dawg
10. tessdata/eng.freq-dawg

Language codes follow the ISO 639-3 standard see http://en.wikipedia.org/wiki/List_of_ISO_639-2_codes (for Urdu : urd, Tamil : tam, Sinhala: sin, Nepali : nep, Sindhi: snd), but any string can be used

Tesseract Training

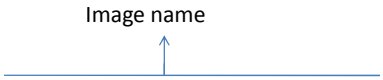
- Step1: Make Box File

tesseract [lang].[fontname].exp[num].tif [lang].[fontname].exp[num] batch.nochop makebox

e.g.

- tesseract eng1.timesitalic.exp0.tif eng1.timesitalic.exp0 batch.nochop makebox

- Step 2: .tr File Generator



tesseract [lang].[fontname].exp[num].tif [lang].[fontname].exp[num] nobatch box.train

tesseract eng1.timesitalic.exp0.tif eng1.timesitalic.exp0 nobatch box.train

Tesseract Training

- Step 3: Character Set Computation

Change directory to …Tesseract-OCR\Training\ and move .tr and .box files here

- unicharset_extractor lang.fontname.exp0.box lang.fontname.exp1.box ...

Box file name

unicharset_extractor eng1.timesitalic.exp0.box eng1.timesitalic.exp1.box …

Tesseract Training

- font_properties.txt

<fontname> <italic> <bold> <fixed> <serif> <fraktur>

nas 0 0 0 0 0

Clustering

- mftraining -F font_properties -U unicharset -O lang.unicharset lang.fontname.exp0.tr lang.fontname.exp1.tr ...

e.g.
•mftraining -F font_properties -U unicharset –O eng1.unicharset eng1.timesitalic.exp0.tr eng1.timesitalic.exp1.tr…

•mftraining -F font_properties.txt -U unicharset –O urd.unicharset urd2.nas.exp0.tr

Output:

1. inttemp (the shape prototypes)
2. pffmtable (the number of expected features for each character)
3. Microfeat (not in used in 3.01)

cntraining lang.fontname.exp0.tr lang.fontname.exp1.tr ...

cntraining eng1.timesitalic.exp0.tr eng1.timesitalic.exp1.tr…

- Output:

1. normproto data file (contains character normalization sensitivity prototypes)

- **Combine training files**
- Rename all the generated data files with a lang. prefix e.g.
  - eng1.normproto
  - eng1.Microfeat
  - eng1.inttemp,
  - eng1.pffmtable
- run combine_tessdata command:

combine_tessdata lang.
e.g. combine_tessdata eng1.

Output:

eng1.traineddata

Tesseract Recognition

- Move **eng1.traineddata** to Tesseract-OCR\tessdata\ directory
- Change directory to …Tesseract-OCR\ and run recognition command

tesseract image.tif output -l lang

e.g.

tesseract eurotext.tif –l eng1

Class Exercise

- Training and Recognition of main bodies of your own language
- Each member will train 5 main bodies types and compute recognition results

Project Deliverables

Deliverable	Urd Group	Nep Group	Sin Group	Sin Group 2	Snd Group	Tam Group
Data Preparation 1. Document Image 2. MBs Real Data 3. Diacritics Real Data						
Binarization						
Line Segmentation						
Ligature/Syllable Segmentation						
Ligature/Syllable Segmentation						
Diacritics Training and recognition						
Main Body Classification and Recognition						
Syllable String Creation						