Optical Character Recognition

Lecture 7

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_ex e_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

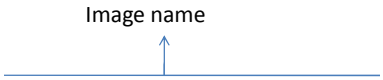
- Step1: Make Box File

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>

Timesitalic 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 output –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

Modifying Tesseract for Nastalique

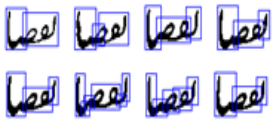
چہ چھو یہ عزیک صا کہا و مز لو فو شو فزمذ کھر ساھ کھا نے

a) Ligature string

حہ حھو ہ عر یک صا کہاو مر لو فو سو فرمد کھر ساھ کھا لے

b) Corresponding Main bodies of ligatures

1. Output ranked list for Main Bodies (MB)
   (separate system for diacritic recognition)
2. Disable chopping (due to inconsistent results)

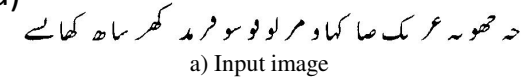


14

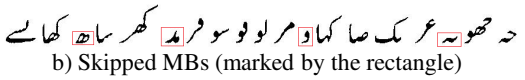
Modifying Tesseract for Nastalique

3. Disable dictionary
   - A statistical word formation model developed to convert ligature sequence into words [10]

4. Change Tesseract pre-processing module
   - Tesseract not able to handle MB size variation in Urdu)



a) Input image



b) Skipped MBs (marked by the rectangle)

15

Modifying Tesseract for Nastalique

5. Merge MB types confused by Tesseract

Main body shape	Ligature string of main body	Similar main body class shape	Ligature string of similar main body class
ملہ	بلہ	ملہ	ملہ
ملے	بلے	ملے	ملے
حصہ	حصہ	حصہ	خفیہ

6. Divide data set and train four Tesseract sub-systems
   - Based on width of MBs
     - extracted using C4.5 algorithm
   - Overlapping for MBs at edges

16

Summary of Systems Developed

System	Base Version for Change	Details
System-1		Tesseract code ver. 3.01
System2	System-1	Modified code to output ranked list
System-3	System-2	Chopping disabled
System-4	System-3	Dictionary disabled
System-5	System-4	Pre-processing changed
System-6	System-5	Similar shapes merged in training data

17

Summary of Data Set

- 1475 unique main bodies
  - Derived from an Urdu corpus of 18 million words [11]
  - 1,490,894 ligatures tokens
  - 7761 ligature types
- Synthesized at 14 and 16 font sizes [12,13] at 300 DPI

Main body classes	Types	Training tokens	Testing tokens
One character	12	10	15
Two characters	61	10	15
Three characters	312	10	15
Four characters	632	10	15
Five characters	458	10	15

18

Results

•14 Font Size Testing Results for MB in ranked list

	One character class		Two characters class		Three characters class		Four characters class		Five characters class	
	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files
System -1	25.00	25.00	85.60	87.30	78.18	79.05	70.81	71.68	63.66	64.90
System -2	97.78	100.00	93.10	95.75	87.86	88.81	87.14	87.88	86.77	87.33
System -3	100.00	100.00	94.71	94.71	89.46	89.82	87.83	88.56	87.19	87.32
System -4	100.00	100.00	93.22	94.83	88.92	89.63	87.91	88.48	86.65	87.30
System -5	100.00	100.00	98.16	98.16	97.03	96.98	96.03	96.30	97.71	97.77
System -6	100.00	100.00	97.70	98.28	96.98	97.11	96.30	96.16	97.77	97.80

19

Results

•16 Font Size Testing Results MB in ranked list

	One characters class		Two characters class		Three characters class		Four characters class		Five characters class	
	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files	Single trained data file	Four trained data files
System -1	25.00	25.00	88.49	89.06	78.95	79.62	70.45	71.61	62.49	63.83
System -2	100.00	100.00	93.64	93.64	88.33	88.87	85.56	86.23	85.61	86.07
System -3	100.00	100.00	94.33	94.56	88.25	88.96	86.24	86.84	85.76	86.23
System -4	100.00	100.00	94.33	94.56	88.25	88.94	85.87	86.19	84.84	85.41
System -5	100.00	100.00	98.70	98.81	96.62	97.35	95.68	96.01	95.94	96.42
System -6	100.00	100.00	98.59	98.81	96.66	97.39	95.93	95.91	96.14	96.47

20

Overall Results

- System-6 with four sets for MB in ranked list

Font size of testing data	Total images	Font size of training data	Accuracy %
14	22125	14	97.87
16	22125	16	97.71
14	22125	16	96.31
16	22125	14	96.71

- System wise Tesseract efficiency results (per 15 MBs)

	Tested on single set (ms)	Tested on four sub-sets (ms)
System-1	170	122
System-2	172	134
System-3	155	105
System-4	158	102
System-5	143	91
System-6	123	84

21

References

[1] R. Smith, D. Antonova and D.-S. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in International Workshop on Multilingual OCR, Barcelona, Spain, 2009.
[2] R. Smith, "An Overview of the Tesseract OCR Engine," in ICDAR, 2007.
[3] M. A. Hasnat, M. R. Chowdhury and M. Khan, "An Open Source Tesseract Based Optical Character Recognizer for Bangla Script," in ICDAR, 2009.
[4] S. Rakhshit, A. Kundu, M. Maity, S. Mandal, S. Sarkar and S. Basu, "Recognition of handwritten Roman Numerals using Tesseract open source OCR engine," in Int. Conf. on Advanced in Computer Vision and Information Technology, 2009.

[5] N. Sabbour and F. Shafait, "A Segmentation Free Approach to Arabic and Urdu OCR," in SPIE, Volume 8658, 2013.
[6] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Mohsin, "Segmentation Free Nastalique Urdu OCR," World Academy of Science, 2010.
[7] G. S. Lehal and A. Rana, "Recognition of Nastalique Urdu Ligatures," in 4th International Workshop on Multilingual OCR, New York, USA, 2013.
[8] S. Tariq and S. Hussain, "Segmentation Based Urdu Nastalique OCR," in CIARP, Havana, Cuba, 2013.
[9] A. Muaz, "Urdu Optical Character Recognition System," Unpublished, MS Thesis Report, NUCES, Lahore, Pakistan, 2010.
[10] M. Akram and S. Hussain, "Word Segmentation for Urdu OCR System," in 8th Workshop on ALR, COLING, Beijing, China, 2010.

[11] M. Ijaz, S. Hussain, "Corpus Based Urdu Lexicon Development," in Conference on Language Technology, University of Peshawar, Pakistan, 2007.

[12] CLE, "CLE Urdu HFL 14 Point Size," CLE. [Online]. Available: http://www.cle.org.pk/clestore/cleurduhfl14pt.htm.
[13] CLE, "CLE Urdu HFL 16 Point Size," CLE. [Online]. Available: http://www.cle.org.pk/clestore/cleurduhfl16pt.htm.
[14] A. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," in CISSE, 2006.

22

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						

Document Image Creation

- Syllable_of_MB1_Samples_1 Syllable_of_MB2_Samples_1 Syllable_of_MB2_Samples_1 Syllable_of_MB3_Samples_1 Syllable_of_MB4_Samples_1 Syllable_of_MB5_Samples_1 ,,, Syllable_of_MB15_Samples_1
- Syllable_of_MB1_Samples_2 Syllable_of_MB2_Samples_2 Syllable_of_MB2_Samples_2 Syllable_of_MB3_Samples_2 Syllable_of_MB4_Samples_2 Syllable_of_MB5_Samples_2 ,,, Syllable_of_MB15_Samples_2
- Syllable_of_MB1_Samples_3 Syllable_of_MB2_Samples_3 Syllable_of_MB2_Samples_3 Syllable_of_MB3_Samples_3 Syllable_of_MB4_Samples_3 Syllable_of_MB5_Samples_3 ,,, Syllable_of_MB15_Samples_3
- Syllable_of_MB1_Samples_4 Syllable_of_MB2_Samples_4 Syllable_of_MB2_Samples_4 Syllable_of_MB3_Samples_4 Syllable_of_MB4_Samples_4 Syllable_of_MB5_Samples_4 ,,, Syllable_of_MB15_Samples_4
- ,
- ,
- ,
- Syllable_of_MB1_Samples_15 Syllable_of_MB2_Samples_15 Syllable_of_MB2_Samples_15 Syllable_of_MB3_Samples_15 Syllable_of_MB4_Samples_15 Syllable_of_MB5_Samples_15 ,,, Syllable_of_MB15_Samples_15

Syllable = MB + Diacritics or Syllable = MB

Example

پر اسے کے ابا بابا پاپا تاپاب اب بابا ہے بات پاپاپار کا کے
میں اب کے کاکا راز کا ہے راد وا کے دیاوار ہے داد ہے
میں کے سے راد وا اسے دیا رب زارا سے پر برا کے ڈڈ ہے
میں کے کاکا پڑ بڑا پرا رب کے رکر کر کر ہے د ہے کا
میں ہات کے ہے سے روکر ڈرہار ہے کاکا سے کر میں ہے
میں کے کی کی کاکی ر کر باب سے میں داد میں ں کان بڑ پر
رں نا ں کاں پر ن برن پڑن بڑرر زار ز زرسے کرنا سے
سے اور و و ہے کر و