April 27, 2012

OCR for Tamil Language

Pradeep Mathesh, CB. EN. U4CSE08131

Guide

Mrs. Hema P Menon, Assistant Professor, CSE
 Dr. K P Soman, Professor, CEN

External Review



Problem definition

Tamil OCR

April 27, 2012

 To develop a semi-automated Optical character recognition (OCR) system for Tamil language



"System Architecture"

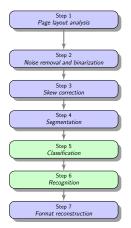


Figure: Architecture of a typical OCR system

4□ > 4□ > 4 = > 4 = > = 90

Tamil OCR
April 27, 2012

Tamil OCR

April 27, 2012

"Implementations"

- Ground truth generating tool train.py
- Blob extraction and preprocessing createb.py
- Recognition engine hyperplane2.py
- Result analyzer confuse.sh

கிக்கடை (Chickadee): வட அமெரிக்காவின் பாரஸ் பேரினத்தைச் சேர்ந்த (பாரிடே குடும்பம்) பல்வேறு பாடும் பறவைச் சிற்றினங்களில் ஏதேனும் ஒன்று. இவற்றின் அழைப்பொலியைப் போலவே இதற்குப் பெயரிடப்பட்டுள் ளது. இவை தமக்கு உணவு கொடுப்பவரிடம் எளிதில் நட்புக்கொள்ளும். கருந்தொப்பி உடைய சிக்கடி (பி. ஆட்ரிகாபிலஸ்) வட அமெரிக்காவில் காணப்படு கருங்கொண்டை சிக்கம கிறது. 5 அங்குல (13 செ.மீ.) (பாரள் ஆட்ரிகாபிலள்).

"Input given to tesseract" நீனமும், கருந்தொப்பி போன்ற आப்பட இச்சுல

Tamil OCR April 27, 2012

Dataset

 $\mathsf{Tamil}\;\mathsf{OCR}$

April 27, 2012

"The dataset is created from two pages of an encyclopedia"

- Training dataset 4201 blobs
- Testing dataset 4067 blobs



Ground truth

Tamil OCR

April 27, 2012

"The ground truth is generated using a handy interface"

- Ground truth helps in automating testing
- It can used along with "grep" to perform random subsampling.

Transliteration

Tamil OCR

April 27, 2012

"For easier processing the ground truth is stored in transliterated form."

- 1 a
- 2 A
- 3 i
- **4** I
- ...

Extraction of blobs

Tamil OCR

April 27, 2012

"Blobs are extracted based on bounding box information given by tesseract"

Junk	Тор	Left	Right	Bottom	Page no
0	865	2893	883	2913	0
Q	867	2875	878	2885	0
W	879	2875	889	2885	0
W	867	2852	881	2862	0

Table: Sample bounding box information for four blobs in a page



 $\mathsf{Tamil}\;\mathsf{OCR}$

April 27, 2012

"Sample output after preprocessing"

Part													
$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$			3.png							10.png		air 12.png	
27 mg 28 mg 29 mg 30 mg 31 mg 32 mg 33 mg 34 mg 35 mg 36 mg 37 mg 38 mg 39 mg 46 mg 41 mg 42 mg 43 mg 44 mg 45 mg 46 mg 47 mg 48 mg 50 mg 50 mg 50 mg 50 mg 54 mg 55 mg 56 mg 57 mg 58 mg 59 mg 60 mg 61 mg 62 mg 63 mg 64 mg 65 mg 66 mg 67 mg 68 mg 69 mg 70 mg 71 mg 72 mg													
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$													
15 mg 15 m													
	(2)	(F)	æ	#	W	r	8	2	ar	al.	as .	96	w
0 0 12, pro 120, pro 121, pro 121, pro 121, pro 121, pro 124, pro 124, pro 125, pro 126, pro 12	2	89	ıb.	А	18	tò	au	etr	n.	#	10	AR.	£.
Tiling 132,ng 133,ng 134,ng 135,ng 136,ng 137,ng 138,ng 138,ng 138,ng 141,ng 142,ng 142,ng 143,ng 144,ng 142,ng 142,ng 143,ng 144,ng 142,ng 143,ng 144,ng 143,ng 143,ng 144,ng 143,ng 143,ng 143,ng 144,ng 143,ng 1	۵	8	-	77	ú		rò.	35	es .	au .	Œ	To.	ab.
A	u	0	Yeu	a		0	Ø	6	6	4è	·S	4	r.
	4	۵	n	10	•	5	ei/	40	ø	eft	a.	u	60
	LO	1		*	9)	và.	æ	-66	60	U		ev.



"Ground truth generation" 11 ta த Tamil OCR
April 27, 2012

Tamil OCR

April 27, 2012

"Model presented to the recognition engine"

er 1.png	2.png	3.png	erfl 4.png	5.png	45 6.png	p) 7.png	8.png	ණ 9.png
. 5 5 11.png	3 12.png	多 14.png	(a) 15.png	(3) 16.png	7; 17.png	2) 18.png	20.png	67 21.png
e37 22.png	₫s 25.png	26.png	24 27.png	സ് 33.png	77 34.png	(⊈) 36.png	5 41.png	வி 42.png
த் 43.png	⊕ 45.png	Afi 47.png	92 ⁄⁄ 49.png	ារ 56.png	57.png	ങ 58.png	64.png	№ 66.png
െ 67.png	12.png	₽ 73.png	74.png	ອນ 76.png	/ir 77.png	92 84.png	89.png	ԱՏ 91.png
5 92.png	-% 96.png	ள் 100.png	90 108.png	9 1 111.png	114.png	பி 125.png	2 143.png	2/ 164.png
. ≠ 168.png	∳ 178.png	® 205.png	₫/ 209.png	6%. 211.png	₽ 213.png	_ த 225.png	店 229.png	234.png
<u>ដ់រ</u> 251.png	<i>ங்</i> 263.png	<i>दर्शवा</i> 288.png	305.png	⊚ 352.png	₽ 388.png	438.png	ஏ 498.png	₽ 520.png
∟ 8 572.png	නාරි 607.png	<u>தீ</u> 609.png	<i>ġ</i> ⁄ 610.png	89 676.png	வ் 883.png	Ω Ω 936.png	9 39.png	5 .7 973.png
4 / 1215.png	1256.png	1287.png	7 1369.png	cafi 1486.png	1712.png	ஹ 1735.png	22 1892.png	ம் 2047.png
മൂം 2100.png	ഖ് . 2140.png	2216.png	മി 2407.png	€ 3333.png	€ 3377.png	S 3381.png	618F 3668.png	



"How model is generated?"

Tamil OCR

April 27, 2012

Figure: Model is generated by copying one instance of each of the classes.



 $\mathsf{Tamil}\;\mathsf{OCR}$

April 27, 2012

"Feature extraction"

The feature extraction is based on four different techniques.

- Active contour model
- Character geometry
- Random Projection
- Hu's Moments

Active contour model

Tamil OCR

- Feature is the number of rings
- Number of points in the contour matrix
- Maximum length of the contour

Character geometry

Tamil OCR

- Euler Number
- Regional area
- Eccentricity
- Zonal Feature

Random Projection

Tamil OCR

- Common dimensionality reduction technique based on Random Projection.
- A form of unsupervised learning.
- Object is classified based on L2-norm.

Hu's Moments

Tamil OCR

- Decomposing object boundary into line segments
- The (p+q)th order of moment is geometric moment M_{pq} of a gray-level image is defined as $M_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy$
- In the case of a digital image, the double integral of the above equation must be replaced by a summation. $m_{pq} = \sum_{i=1}^{n} \sum_{j=1}^{n} i^{p} j^{q} f_{ij}$,
- where N is the size of the image and f_{ij} are the grey levels of individual pixels.



Results

S.No	Class	Confusing class
1	tu	rru
2	vi	li
3	va	ya
4	wa	ta
5	va	na
6	va	la
7	ku	cu
8	mu	zhu
9	ku	ru
10	pa	pu
11	wi	rri
12	ki	ci
13	ka	ca

Tamil OCR
April 27, 2012

Results

Class	Minimum	Maximum
ku	263	294
ru	258	280
tu	256	282
rru	243	262
vi	227	288
li	244	259
ya	186	210
na	239	287
ta	246	256
va	245	264
wa	212	231
ka	212	225
ca	146	200

Tamil OCR
April 27, 2012

Tamil OCR

April 27, 2012

"Work done"

- A Ground truth generating tool has been written
- A recognition engine has been implemented based on i) random projection technique (RPT) and ii) SVM training and classification.
- A result analysing script for RPT has been implemented

Tamil OCR

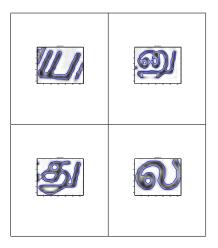
April 27, 2012

"Work done"

- Comparison of SVMTC and RPT is done
- Inspection of confusing classes for hierarchal classification to improve the accuracy
- The least confusing classes were picked
- Possible cases are (க->ச 111),(ச->க 5),(இ->பி 8),(ரு->கு 9)
- Currently, the classification is done solely based on level set method.
- The features that are used to discriminate confusing classes are number of rings and points of the contour matrix.



"Some elusive samples"



Tamil OCR

Tamil OCR

April 27, 2012

"Result"

		Samp	les	,		
S.No	Features	Training	Testing	Training	Testing	
1*	555	4201	4067	75	< 44 (3 c	lasses)
2**	1000	155 classes	4201	98	59	

- Legend
- *Structural and Statistical features (SVM trainer and classifier is used for classification)
- **Random Projection (L2-norm is used for recognition)



"False positives and negatives"

	_	து	ந	ன	ன்	ш	4	ப்	ш	ш	ற
_	92					15					2
து		53				3					
ந			33			1					5
ன				55		1					
ன்	1				56	3					
ш	34					97	2				
4						1	1				
ù	2					5		1			
ம	2					2			80		
ш	1					16				61	
ற			2			2					47

Tamil OCR

Future enhancement

Tamil OCR

April 27, 2012

"Improving the accuracy"

 Rigorous analysis of hierarchal classification can be carried out to improve the accuracy.



Tamil OCR

April 27, 2012

"Reference"



- Raihua Zhang.et.al," Active contours with selective local or global segmentation: a new variational approach and level set method", Image and Vision Computing, 2010.
- Xavier Bresson, A Short Guide on a Fast Global Minimization Algorithm for Active Contour Models, April 22, 2009

Qinfeng Shi, Rapid Face Recognition Using Hashing, Australian National University, and NICTA Canberra, Australia, 2009

Dinesh Dileep, "A feature extraction technique based on character geometry for character recognition", http://www.ece.iisc.ernet.in/ dileep, 2008

Ray Smith, "An Overview of the Tesseract OCR Engine", IEEE Trans., 2007

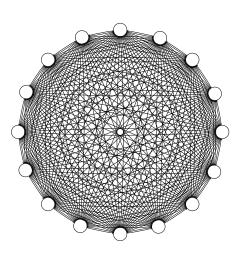
Jan Flusser and Tomas Suk, On the Calculation of Image Moments, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, January,1999

Tamil OCR

Tamil OCR

- V. Krishnamoorthy, "OCR Software for Printed Tamil Text", Proceedings of Tamil Internet 2002, California, USA, 2002
- Anbumani Subramanian and Bhadri Kubendran," Optical Character Recognition of Printed Tamil Characters", http://www.ece.vt.edu/,2000
- Jan Flusser et.al," Moments and Moment Invariants in Pattern Recognition", ISBN 978-0-470-69987-4, Page.no 49,2009

Questions?



Tamil OCR