

OCR for Tamil Language

Pradeep Mathesh,CB.EN.U4CSE08131

Guide

1. Mrs. Hema P Menon, Assistant Professor, CSE
2. 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"

Tamil OCR

April 27, 2012

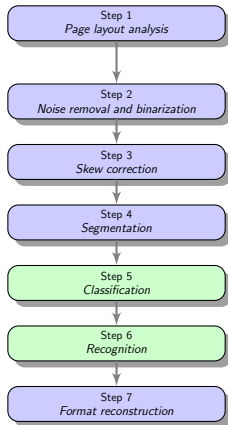


Figure: Architecture of a typical OCR system

"Implementations"

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

சிக்கம (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 சிக்கம
(பி. ஆட்ரிகாபிலஸ்) வட
அமெரிக்காவில் காணப்படு
கிறது. 5 அங்குல (13 செ.மீ.)
நீளமும், கருந்தொப்பி போன்ற



கருங்கொண்டை சிக்கம
(பாரஸ் ஆட்ரிகாபிலஸ்).
WILLIAM D. GRIFFIN

”Input given to tesseract”

"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 l
- ...

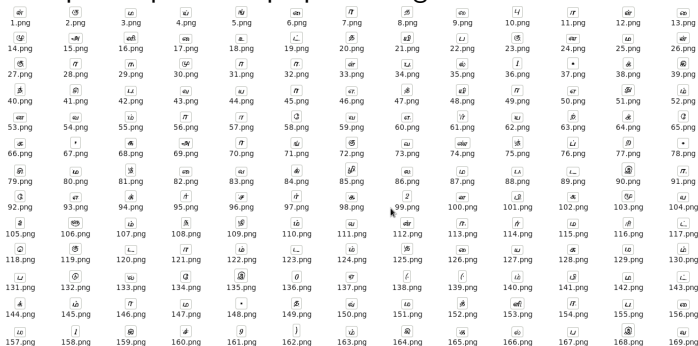
Extraction of blobs

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

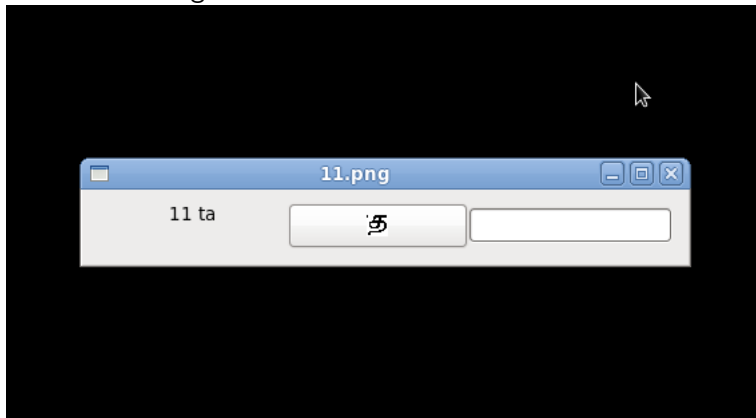
Junk	Top	Left	Right	Bottom	Page no
@	865	2893	883	2913	0
Q	867	2875	878	2885	0
w	879	2875	889	2885	0
w	867	2852	881	2862	0

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

"Sample output after preprocessing"



"Ground truth generation"



”Model presented to the recognition engine”



” How model is generated ?”

Tamil OCR

April 27, 2012



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

"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

April 27, 2012

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

Character geometry

Tamil OCR

April 27, 2012

- Euler Number
- Regional area
- Eccentricity
- Zonal Feature

Random Projection

Tamil OCR

April 27, 2012

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

- 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

Tamil OCR

April 27, 2012

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

Results

Tamil OCR

April 27, 2012

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

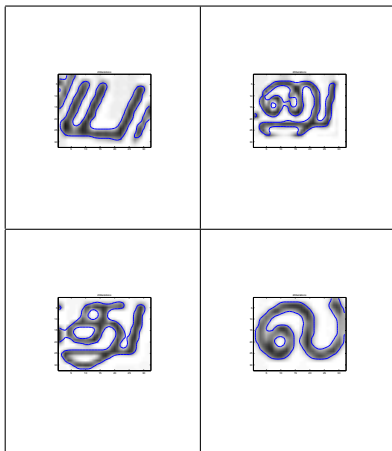
"Work done"

- 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

"Work done"

- Comparison of SVMTC and RPT is done
- Inspection of confusing classes for hierarchical 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

April 27, 2012

"Result"

S.No	Features	Samples		Accuracy	
		Training	Testing	Training	Testing
1*	555	4201	4067	75	< 44 (3 classes)
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)

Tamil OCR

"False positives and negatives"

Tamil OCR


April 27, 2012


	ட	து	ந	ன	ன்	ப	பு	ப்	ம	ய	ற
ட	92					15					2
து		53				3					
ந			33			1					5
ன				55		1					
ன்	1				56	3					
ப	34					97	2				
பு						1	1				
ப்	2					5		1			
ம	2					2			80		
ய	1					16				61	
ற			2			2					47

"Improving the accuracy"

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


"Reference"

 F. Lauer ,MSVMpack: a Multi-Class Support Vector Machine Package,
<http://www.loria.fr/~lauer/MSVMpack,2011>


 Kaihua 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

-  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

April 27, 2012

