## Design and Evaluation of Omnifont Tamil OCR

Tushar Patnaik, CDAC Noida tusharpatnaik@cdacnoida.in Shalu Gupta, CDAC Noida shalugupta@cdacnoida.in CV Jawahar, IIIT Hyderabad jawahar@iiit.ac.in Santanu Choudhury, IIT Delhi santanuc@ee.iitd.ernet.in A G Ramakrishnan, IISc, Bangalore 560012. ramkiag@ee.iisc.ernet.in

IISc Bangalore has developed a recognition engine for Tamil printed text, which has been tested on 1000 document images of pages scanned from books printed between 1950 and 2000. IIIT Hyderabad has developed a XML based annotated database for storing the 5000 images of scanned pages and the corresponding typed text in Unicode. CDAC, Noida has developed an efficient evaluation tool, which compares the OCR output text to the reference typed text (ground truth) and flashes the substitution, deletion and insertion errors in different colours on the screen, so that the design team can quickly identify the issues with the OCR and make corrective steps for improving the performance. IIT Delhi has proposed and developed a novel scheme for segmenting only the text regions from document images containing pictures. The OCRs uses Karhunen-Louve transform (KLT) as features and a support vector machine (SVM) classifier with RBF kernel in a discriminative directed acyclic graph (DDAG) configuration. They assume an uncompressed input image of the document page, scanned at a minimum of 300 dpi with 256 gray levels (not binary or two-level). Tamil OCR currently gives over 94% recognition accuracy at the Unicode level, evaluated on over 1000 printed pages, some of them also containing old Tamil letters.

## The features of the OCRs are:

- 15. Omnifont : Any normal font used by books is handled. We don't say it is font-independent, because ornamental or stylized fonts cannot be handled.
- 16. **Merged Characters:** To a certain extent, the OCR is capable of identifying and segmenting the merger between two adjacent characters in a old, printed book.
- 17. Noise Tolerance: Certain types of breaks in the character are handled successfully.
- 18. Old Tamil or Kannada Script: The pre-1970 (prior to script revision due to E V Ramasamy Naicker) Tamil CV combinations such as NA, RA, nA, Nai, lai, Lai and nai are all recognized, along with the revised representations of the same. Similarly, old Kannada (*halegannada*) characters *of La* and *zha* and their vowel combinations are all handled seamlessly.

- 19. Unicode Output: The output is given in UTF-8.
- 20. **Testing:** Both OCRs have been tested by CDAC, Pune using an annotated corpus of over 1,000 document images of pages scanned from books printed between 1950 and 2002.
- 21. **Consistency:** The OCRs produce consistent and graded performance with font, size and quality variations.
- 22. **Future Enhancement:** The current average performance of 94% for Tamil and 84% for Kannada at the character level is without the use of any language model for postprocessing. Thus, there is a good potential to improve the performance of both OCRs further.

Medical Intelligence and Language Engineering Laboratory has teamed up with Bookshare.org, an International non-profit organization, to provide Tamil and Kannada digital books (copyright free or permitted by authors) online to print-disabled people (visually challenged, old people with vision disabilities and people with other disabilities that make it impossible for holding a book and turn pages of it). A Text-to-speech engine in the respective language will also be provided to the registered user, who can then directly listen to the printed content on their desktop or laptop. We look forward to partners, who can give us copyright free books (hard or soft copies) or direct us to sources of the same. They are also welcome to directly partner with bookshare.org or Worth Trust at Chennai or Enable India at Bangalore.

Figure 1 shows a screen shot showing the performance of the system, as well as the convenience and use of the GUI. Figure 2 shows the confusion matrix shown by the evaluation tool, which helps in identifying common confusion and improve the OCR accordingly. Figure 3 shows the evaluation tool, comparing the XML annotated Tamil text for the page and the OCR's output in Unicode. Use of such convenient tools accelerated the development of the OCR accuracy, and it is currently giving a performance of over 95% for good quality printed pages.

Elle Image Pr		narization Noise Cleaning	elinary Image Analysis CCA Help
-	can Clear	ஒஸிஆர் Tamil	Tamil Ocr P End To End OCR Recognition Engine
Input Windo	w		Active Window
			திவான் லொடபட சிங் பகதூர்
ilvani.odapadaSii 00_Org_Page_00	மி தெ <sup>rgh8</sup> இ பி க ை	ருக்கும் என்பத வானால் ஏற்படு ருக்கிறதென்று த ரஸ்தாபிக்கிறா ளையோ ஜனங்	ளில் வசூலித்தது பிரம்மாண்டமான தொகையா து பரிஷ்காரமாகத் தெரியும் விஷயம். தங்கள் நித்தப்பட்டுள்ள வரிகளோ நிரம்பவும் பளுவாய் ஜனங்கள் உணர்ந்து அதைப்பற்றிப் பலவிதமாகப் ர்கள். திவான் லொடபட சிங் பகதுரரின் வரி ங்கள் அவ்வளவாய் உணர்ந்ததாகவே தெரிய வொன் பட்ட்டைச் கணாட்ட்ச மேகாலி என்றும்
Current Outp	•		Output Editable Window
Current Outp	திவான் லொ கூற்ப இலாக மிருக்கும் எ திவானால் ஏ இருக்கிறதே நேன்தாகேக் களைபோ ஜ னில்லை தங	டபட சில் பக்குரன் ஸ்து மீல்ஷ்கைக்கு தேரியரம் நபடுத்தப்பட டுன்ன வில்வோ ஸ்து ரன்மன் உனந்த அவது புன்ன் ன் திவான் வொட்டியர் ஸ்கன் அவ்வான் உடத்தைக் கரை பாக்குதில் மகை நபுனை என்று என் துபான் கட்டத்தைக் கரை பாக்குதில் மக நபுனை குடிப்புக்கு அருகிய ப் புதன் இப்பான் கட்டத்தைக் கருத்தும்.	ன்படமான தொகையா சேஷயம், தங்கள் நிற்பவழம் பழைவாய் படத்தில் பழைவோல் படத்தில் பழைவை படத்தில் பழைவி கத்தால் தேகில

Fig.1. A screen shot from the OCR, showing the input image and the output text.

4																			•	
6001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
L	0	0	0	•	0	0	•	•	•	0	•	•	0	0	0	0	•	0	0	<u>ली</u> डेक
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28	\$P
86	0	0	0	0	0	0	0	0	0	0	25	0	0	0	1	0	0	255	0	
ы	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1987	0	0	
a	0	0	0	0	0	0	0	•	0	0	0	0	0	0	0	711	0	0	0	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7219	0	1	0	0	
ஓள	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
63 (53	0	0	1	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	
(BS	0	0	0	0	0	0	0	0	0	0	0	268	з	0	0	0	1	0	•	End
83	0	0	0	•	0	0	•	0	•	0	0	•	0	0	0	0	•	0	•	
B	0	0	0	•	0	0	•	•	•	129	•	•	0	0	0	0	•	0	0	Reset
عا	0	0	0	0	1	0	0	•	890	2	0	0	0	0	0	0	•	0	0	Next
륑	0	0	0	•	0	0	0	16	•	0	•	0	0	0	0	0	•	0	•	Previous
4	0	0	0	0	0	0	407	14	0	0	0	0	0	0	0	0	0	0	0	P
Æ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
ଙ୍କ	0	0	2	0	846	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Þ	0	0	0	200	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	
\$	0	0	1404	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
°°	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
•	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A\0	•	%	ሻ	Ŷ	ଙ୍କ	Ħ	٩	196	ध	ø	83	ଞ	<b>G</b> 5	ஒள	16	Ð	щ	80	6	•

*Fig. 2. Display of Confusion matrix for a page recognition.* 

Annotation	OCR Output	•
திவான் லரொடபட சிங் பகதூர	திவான் லரொப்பட சிங் பகத‱	
்அற்ப இலாகாக்களில் வசூலித்தது பிரம்மாண்டமான தரெகயை யிராக்கும் என்பது பிரங்காயமாகக் கடுரியாம் விஷயம் – கங்கள்	அற்ப இலாகாக்களில் வசூலித்தது பிரம்மாண்டமான தரொகயை	ரகமைர
	யிருக்கும் என்பது பரிஷ்காரமாகத் தடுரியாம் விஷயம்.	<u>ร</u> เช่งณ์
ஆரு கைழதலைைற் ஜன்களை உண்ழத்து அதன்பபற்றப் பலவர்தமாகப் பிரஸ்தாபிக்கிறார்கள் , திவான் லரொடபட சிங் பகதூறின் வரி களவைரோ உனங்கள் அவிவளவாய் உணர்க்காகவரே கிரிய	திவானால் ஏற்படுத்தப்படடுள்ள வரிகளரோ திரம்பலும் பளுவாய்	பளுவாய்
வில்லை. தங்கள் திவான் சட்டத்தனக் கரண்டத்த மதோவி என்றும், ராமாங்க மிர்வாகக்கில் (பகா மிராகனர் கண்ரல் கவர்கார் விளைவாம்)	இருக்கிறதவேறு ஜனங்கள் உணர்ந்த அதபைப்பற்றிப் பலவிதமாகப்	ற்கமாகப்
ராதாடாக நரசமாலத்தால் பலா நாடலாற என்ற என்றது. மாகமுப்பட்டவர், அப்படி இராத்தாம், அபியான நமத் திவான் வரோப்ட கிஸ் பாகமிர் சொல் கொடைகாலப்பில் கா	பிரஸ்தாபிக்கிறார்கள் திவான் லரொடபட சிங் பகதூரின் வரி	a <i>ı</i> jî
லைபடபட சாச பதைற ஆன்றன வகை ச வைத்தால் குவைதை , இவர் காரியத்திற்கு ஆகாத ஏட்டச்சுரணக்காய் என்று ரஜப்படத்தி விப் ார் டுவர் எப்பரோசும் சப்பட் சாப் சியம் அகிய	களமைரோ ஜனங்கள் அவ்வளவாய் உணர்ந்ததாகவ தரிய	
இரண்டனை குடைக்கியமாக எண்ணி, உயிரற்ற இயற்றிரம் பரோல விண் கமிர்கிலாரியன் உல்லாலா, உயிரற்ற இயற்றிரம் பரோல வின்ன கமிர்கிலாரியன்றி உற்பாட்டன் வை பயர்க்கிலான வியரோகப்	வில்லன. தங்கள் திவான் சடடத்தனக் கரகைடத்த மதோவி என்றும்,	வி என்றும்,
வலைகை சைவரதாற்களைற், தம்பட்டையாம் நல்தாகாம் படித்து படத்தி முக்தா மாத்தமறித்தா, ஜீவகராணவுமம், பச்சாதாபம் மாதலிய பாலால் என்பட்டால் படிப்படல் படிப்படல் பட்டியால் பட்ட	ராஜாங்க நிரவாகத்தில் மகா திபுணார என்றும் கவர்ன ஜனெரலால்	ஜனோல்
சுணங்களைகை கலந்துகளாணட்சட்டங்கள்டை உபயமாகப்படத்தத் தடுநித்து கரொள்ளாதவர், இவருக்கு நல்ல புத்தி புகட்டவணேட்டம்	புகழப்படடலர் அப்படி இருத்தும், அறூலியான தமது ;	திவான்
எனறா நமதா துவான லவாடபட சாங் பகதுற அ+பயாக இராதது இத்தனை காறியங்களமையம் மகா அற்பாதமாக இயற்றிக் காட்டி	லரொடபட சிங் பகதூர இவரவைகொ சுவபத்தில் ஜயேத்தூ	<i>њ</i> г.,
8 -	இவர காறியத்திற்கு ஆகஷ ஏடடுன்றகைகாய் என்று ராஜுப்படுத்தி	ஜப்படத்தி
கரொண்டும், சட்டத்தில் அகப்பட்டுக்கரொள்ளாமல், எவவளவரோ அபாரமான காரியங்களை எல்லாம் எத்தனை வருஷகாலம்	விட்டார இவர எப் பரோதும் ச1: டம், சாட் சி யம் ஆகி	துகிய
வரமையில் வணேட்டமானாலாம் சமேவது சாத்தியமான விஷயம் என்பது இப்பரோது நன்றாக விளங்கிவிட்டது. ஒருவணுடயை	இரண்டயைமரே முக்கியமஆ எண்ணி , உயிரற்ற இயத்திரம் பரோல	ரம் பரோல
சபெத்தவை இவர்னபொருவன் அக்கிரமமாக அபகரித்துக் கரொள்ளு கிறான். ஆனால் அதற்கு சாட்சிகள் இல்லை. ஆகவடே, சட்டத்தின் படிபார்க்கால் காச கூல்லாசும்கில்லாக சியாயப்பரி பார்க்கால்	▼ வில்லை சடிய்கிறவரியான்றி, தம்முடலய பகுத்தறிவலை உடயிரேகப்	பயரோகப்
0086_DivanLodapadaSinghBagadoor_Img_300_Org_Page_0074	Ţ	4
AnnotationSizeOCRoutputSizeSubstitutionErrorInsertSco15691536105	tution Error insertScore deleteScore Substitution Error-Rate Total Error-Rate 5 38 0.6373% 3.378%	l Error-Rate 8% ◆
Note:- Substitution, Insert, Delete Previous Next Reset	End	<u>ली डेक</u> ्

*Fig. 3. Evaluation Tool showing substitution, insertion and deletion erros.*