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	e Processing Bin	arization Noise Cleaning	Binary Image Analysis OCA Help
D open 5	can Clear	ஒஸி.ஆர் Tamil	Tamil Ocr Find To End OCR Recognition Engine
Input Windo	w		Active Window
			திவான் லொடபட சிங் பகதூர்
livani, odapada Sir 00_Org_Page_00	<i>அ</i>] ගි ඉ ක ක ක ය ග ක ක ක ක ක ක ක ක ක ක ක ක ක ක ක	ற்ப இலாகாக்க ருக்கும் என்பத வானால் ஏற்படு ருக்கிறதென்று த ரஸ்தாபிக்கிறா எனயோ ஜனங் ஸ்லை சம்சன் ச	ளில் வசூலித்தது பிரம்மாண்டமான தொகையா து பரிஷ்காரமாகத் தெரியும் விஷயம். தங்கள் இத்தப்பட்டுள்ள வரிகளோ நிரம்பவும் பளுவாய் ஜனங்கள் உணர்ந்து அதைப்பற்றிப் பலவிதமாகப் ர்கள். இவான் லொடபட சிங் பகதுரரின் வரி மகள் அவ்வளவாய் உணர்ந்ததாகவே தெரிய வொன் சுட்டத்தைக் கணாடங்க மேகாலி என்றும்
•	•		Output Editable Window
Current Outp	திவான் வொ அற்ப இலாக பிருக்கும் என் திவானால் ஏ இருக்கிறதெ இருக்கிறதை பரண்தா69க்கி கனைபோ ஜ வீல்லை தங் ராஜாங்க திரப புகழப்பட்ட வொடபட சிர	சிங் பக்துன் எக்களில் வகுகித்தது நேரம்மாக கூகு பரிக்ஷக்காமர்க்கு தெரியும் நபடுத்தப்பட தேரியா பரிக்கோடி குறாக்கர் தவான் வேன், பட விர தார்கள் தவான் வொட்டா, விர தாக்கர் திவான் சட்டத்தைக் கரை வாகத்தில் மகா திபுனா என்று பா சுத்தில் மகா திபுனா என்று பி சுப்படி இருந்தில், சரூட்வெ ப சுத்தில் மகா திபுனா வேன்று வே	க்கடம்பான தொகையா கிஷயம், தங்கள் இரம்பயற் புதையாய் படத்திய பணதேயாகப் படத்த தேதால் குறிப படத்த தேதால் என்றம், கையால பிருவதற்கால் தம் கொன் தேதாலை பான நடித்த தியான் தும் தோத்த,

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

				•							-	_		-					•	•	• •
	_	-	-	-							-	-		-	-						
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		ीडेक
Ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28		JR (
26	0	0	0	0	0	0	0	0	0	0	25	0	0	0	1	0	0	255	0		
•	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		
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7219	0	1	0	0		
- Foot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
G 5	0	0	1	0	0	0	0	0	•	0	0	0	22	0	0	•	•	0	0		
ക	0	0	0	0	0	0	0	0	0	0	0	268	3	0	0	0	1	0	0		End
83	0	0	0	0	0	0	•	0	•	0	0	0	0	0	0	•	0	0	0		set
6	0	0	0	0	0	0	0	0	•	129	0	0	0	0	0	•	•	0	0		Rec
e	0	0	0	0	1	0	0	0	890	2	0	0	0	0	0	0	0	0	0		Next
8	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0	0		evious
۹	0	0	0	0	0	0	407	14	0	0	0	0	0	0	0	0	0	0	0		Pr
Æ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
9	0	0	2	0	846	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
₽	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		
0		••	ቅ	æ	ଙ୍କ	Ŧ	ما	196	وا	6	83	65	65	per	16	B	њ	8 6	و		
7																				-	

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

Annotation	OCR Output	4
திவான் லரெடபட சிங் பகத [ூ] ர	திவான் லரொடபட சிங் பகத ^ல ன்	
அற்ப இலாகாக்களில் வசூலித்தது. பிரம்மாண்டமான தடொகயை உட்டா	அற்ப இலாகாக்களில் வசூலித்தது. பிரம்மாண்டமான தரெகயைா	
யருக்கும் என்பது பரஷ்களரமாகத் தலர்மும் வாஷ்யம். தங்கள திவானால் ஏற்பட்திப்பட்டுள்ள வறிகளரோ திரம்பவும் பனுவாய் (தி	மிருக்கும் என்பது பரிஷ்காரமாகத் தடுரியும் விஷயம். தங்கள்	
இருக்கிறதனெறு ஜனங்கள உணரத்து அதபைபற்றிப் பலவிதமாகப பிரஸ்தாபிக்கிறார்கள், திவான் லரொடபட சிங் பக்தூரின் வரி	திவானால் ஏற்படுத்தப்படடுள்ள வரிகளரோ திரம்பவும் பளுவாய்	
களமைபோ ஜனங்கள் அவ்வளவாய் உணற்த்ததாகவடு தடுரிய வில்லடை தங்கள் திவான் சட்டத்தனக் கரகைடந்த மதோவி என்றும்,	இருக்கிறதவெற்ற ஜனங்கள் உணர்ந்த அதவப்பற்றிப் பலவிதமாகப்	
ராஜாங்க நிர்வாகத்தில் மகா திபுணர் என்றும் கவர்வர் ஜனேரலால் புகழப்பட்டவர். அப்படி இரந்தும், அபியான நமது திவான்	பிரஸ்தாபிக்கிறார்கள் திவான் லரொடபட சிங் பகத ^க ரின் வரி	
லரொடபட சிங் பகதூற் இவரவை வகொ சுலபத்தில் ஜலியித்து, இவர் காரியத்திற்கு ஆகாத ஏட்டுச்சாரனக்காய் என்றா ரஜாப்படுத்தி	களமைரோ ஜனங்கள் அவ்வளவாய் உணர்ந்ததாகவ தடுரிய	
விட்டார், இவர் எப்பரோதும் சட்டம், சாட்சியம் ஆகிய இரண்டயைமரே முக்கியமாக எண்ணி, உயிரற்ற இயந்திரம் பரோல	வில்லடை தங்கள் திவான் சடடத்தகை கரகைடந்த மதோவி என்றும்,	
வலேனை சமெக்றவரவோனற், தம்முடயை பகுதைறவடை உபயலோகப படுத்தி, யாக்தா யாக்தமறித்து, ஜீவகாராண்யம், பச்சாதாபம் முதலிய	ராஜாங்க நிரவாகத்தில் மகா நிடுணர என்றும் கவர்ன ஜனெரலால்	
காணங்களகை கலந்தாகரொண்டு சட்டங்களவை உபயரோகப்படுத்தத் தடுரிந்தா கரொள்ளாதவர், இவருக்கா நல்ல புத்தி புகட்டவனேஸ்டும்	புகழப்பட்டலர் அப்படி இருத்தும், ஆரூபியான தமது திலான்	
என்று நமது திவான் லரொடபட சிங் பகதூர் அ∔பியாக இருந்து இத்தனை காரியங்களயைும் மகா அற்புதமாக இயற்றிக் காட்டி	லரொடபட சிங் பகதூர இவரவை வகொ சுவபத்தில் ஹமித்து.	
இருக்கிறார் என்பது நன்றாகத் தடுரிகிறது. ஒரு மனிதன் சாட்சிகள் இ		
இலைாமலாம், அலலது, பரெயச் சாட்சிகளடை உண்டாக்கிக் கரொண்டும், சட்டத்தில் அகப்பட்டுக்கரொள்ளாமல், எவ்வளவரோ	இவர காரியத்திறகு" ஆகஷ் ஏட்டுளரகைகாய என்று ரூஜுப்படுத்தி	
அபாரமான காரியங்களை எல்லாம் எத்தனவை வருஷகாலம் 	விட்டார இவர எப் பரோதும் ச1: டம், சாட் சி யம் ஆகி ய	
யரன்பல வலனட்மானாலும் சல்வது சாததுமமான வாஷ்மம என்பது இப்பரோது நன்றாக விளங்கிவிட்டது. ஒருவனுடயை	இரண்டயைமரே முக்கியமஆ எண்ணி, உயிரற்ற இயந்திரம் பரோல	
சரெத்தனை இவர்வரொவன் அக்கிரமமாக அபகரித்துக் கரொள்ளு கிறான். ஆனால் அதற்கு சாட்சிகள் இவ்வை. ஆகவரே, சட்டத்தின் பரி பார்ச்சால் வசு சுற்பமாகிறகில்லை. கியாயப்பரி பார்ச்சால்	வலேனை சமெய்கிறவரமோன்றி , தம்முடனய பகுத்தறிவனை உபயரோகப்	•
0086_DivanLodapadaSinghBagadoor_Img_300_Org_Page_00	-	
AnnotationSize OCRoutputSize Substitution Error insertSco	deleteScore Substitution Error-Rate Total Error-Rate	
1569 1536 10 5	38 0.6373% 3.378%	
Note:- Substitution, Insert, Delete	e	•
Previous Next Reset	End 전다	×

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