# 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.

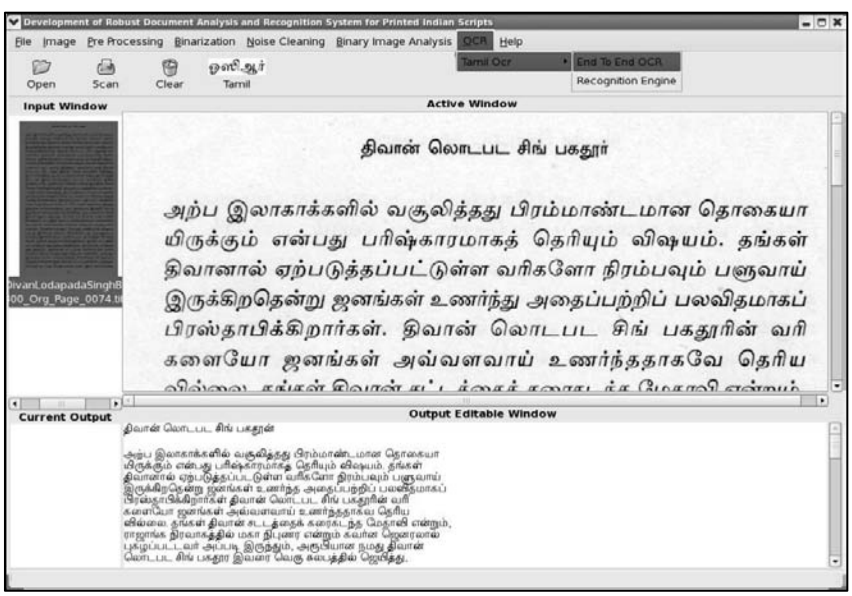
21. **Consistency:** The OCRs produce consistent and graded performance with font, size and quality variations.

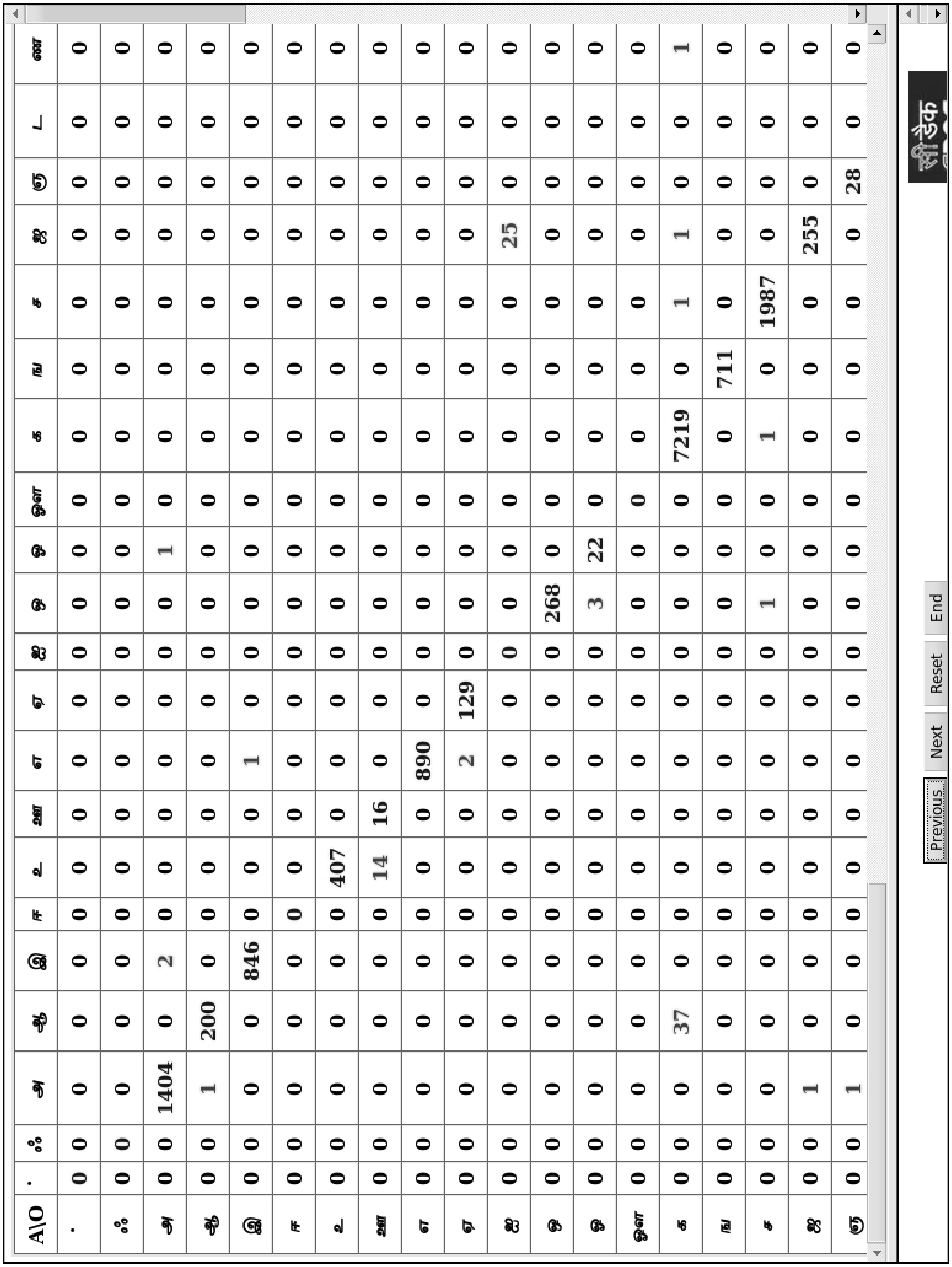
22. **Future Enhancement:** The current average performance of 94% for Tamil and 84% for Kannada at the character level is without the use of any language model for postprocessing. Thus, there is a good potential to improve the performance of both OCRs further.

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.

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





| A\O | · | ஃ | அ | ஆ | இ | ஈ | உ | ஊ | எ | ஏ | ஐ | ஒ | ஓ | ஔ | க | ங | ச | ஜ | ஞ | ட | ண |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| · | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஃ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| அ | 0 | 0 | 1404 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஆ | 0 | 0 | 1 | 200 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| இ | 0 | 0 | 0 | 0 | 846 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஈ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| உ | 0 | 0 | 0 | 0 | 0 | 0 | 407 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஊ | 0 | 0 | 0 | 0 | 0 | 0 | 14 | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| எ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 890 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஏ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 129 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஐ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 |
| ஒ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 268 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஓ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ஔ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| க | 0 | 0 | 0 | 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7219 | 0 | 1 | 1 | 0 | 0 | 1 |
| ங | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 711 | 0 | 0 | 0 | 0 | 0 |
| ச | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1987 | 0 | 0 | 0 | 0 |
| ஜ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 255 | 0 | 0 | 0 |
| ஞ | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 28 | 0 | 0 |



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

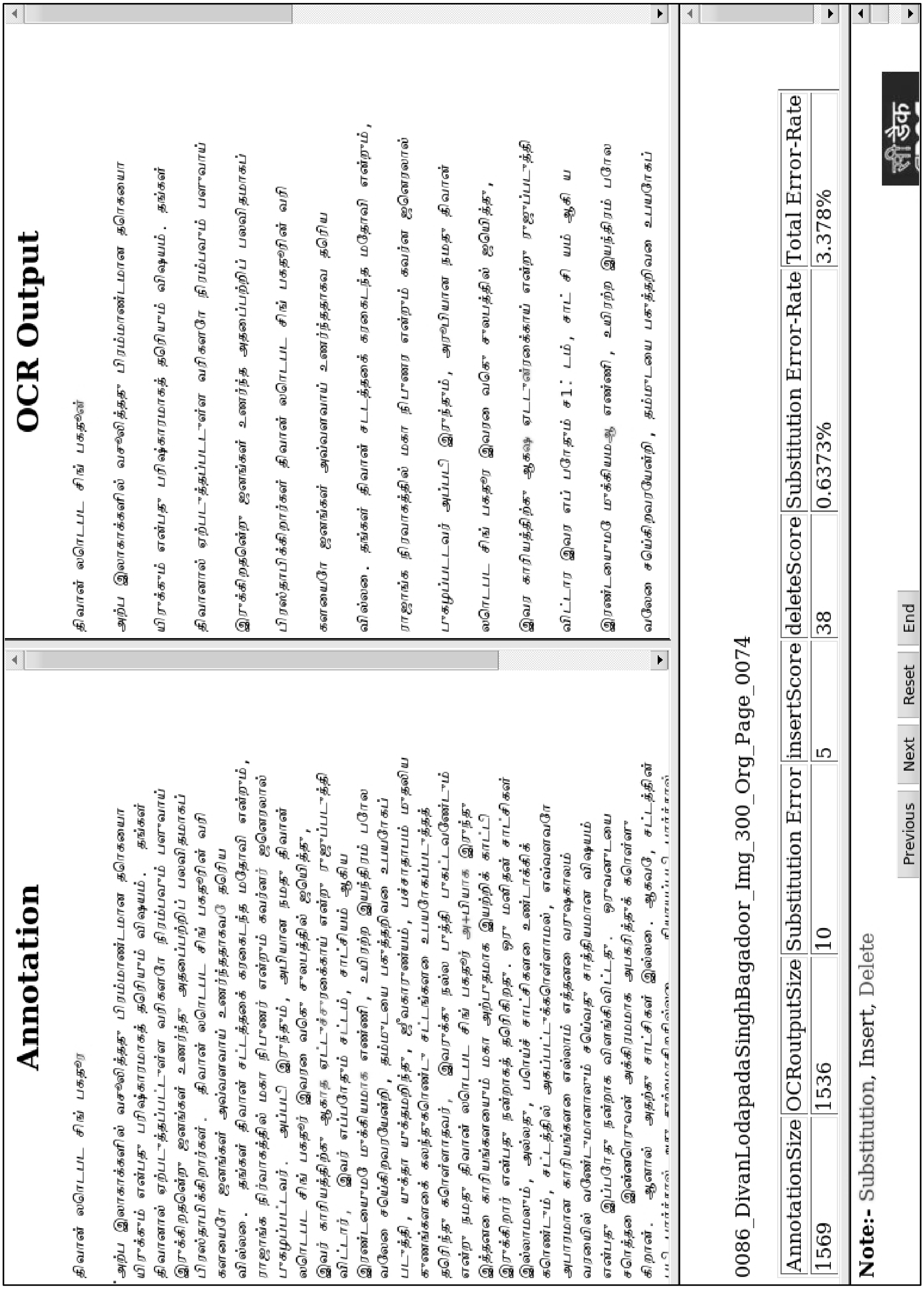



*Fig. 3. Evaluation Tool showing substitution, insertion and deletion erros.*