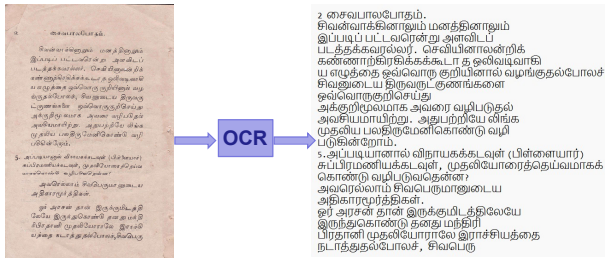
Benchmarking Optical Character Recognition Systems for the Tamil Language

Suthakar **Sivashanth**, Kengatharaiyer **Sarveswaran** and Eugene Yugarajah Andrew **Charles**
Department of Computer Science, University of Jaffna
ssivashanth@univ.jfn.ac.lk

Deutscher Akademischer Austauschdienst
German Academic Exchange Service

Introduction

- Natural Language Processing (NLP) aims to enable computers to comprehend, interpret, infer, and generate human language.
- Significant progress in high-resourced languages (e.g., English, Chinese).
- Limited resources (lexicons, annotation tools, etc) in low-resourced languages like Tamil.
- Rich text corpora are the key source to develop these computational resources/tools.
- Limitation in creating text corpora due to lack of good Optical Character Recognition (OCR) systems, specially to compile text from Tamil books.
- An OCR model for the Tamil language is planned to be developed, with the first step focusing on benchmarking existing OCR systems.



Input image OCR output text

Motivation

- There are several commercial and open-source OCR systems available for Tamil.
- No common benckmaking framework exists to evaluate these OCR systems.
- This research aims to create a benchmark dataset and evaluate existing OCR systems to identify the suitable OCR system for Tamil.

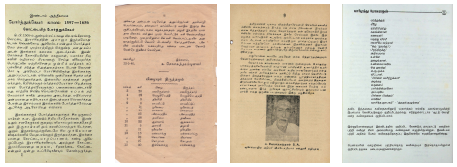
Existing Tamil OCR Systems

OCR Systems	Reported Evaluation
• Aharamariyi Tamil OCR (Not available)	• Character-Level Accuracy 81%
• Tamizhi-Net OCR (Not available)	• Reduced the character-level error rate of Tesseract to 2.61% for Tamil and to 4.74% for Sinhala. • Reduced the word-level error rate to 20.61% for Tamil and to 26.58% for Sinhala.
• OCR_Tamil (Available)	• Character-Level Accuracy 95% for newly printed Tamil books.
• Hybrid Decision Tree based OCR System (Not available)	• Character-Level Accuracy 98.80%

Methodology

Data Collection

- A total of 120 images from the Jaffna University Library and 50 images from the Noolaham Foundation were obtained from 1850, selected in 10-year gaps.
- Different types of pages: cover, imprint, table of contents, preface, plain text, tables, text with tables, text with images, and advertisements.
- Special Categorization: Image Condition (Good, Damaged, Noisy), Number of Languages (Monolingual, Multilingual), Printing Technology (Letterpress, Digital Print).



Examples of Diverse Page Format

Scanning the Images
- Sources: books, pamphlets, and newspapers.
- Resolution: 300 DPI (dots per inch)

Creating the Ground Truth Text
- Recognized texts from scanned images are OCRed using Tesseract and manually corrected.

Benchmarking the Existing OCR Systems

Benchmarking

1. Unicode codepoint-based Levenshtein distance.

$$D(i,j) = \begin{cases} 0, & i=0,\ j=0 \\ i, & j=0,\ i>0 \\ j, & i=0,\ j>0 \\ \min\{ & \\ \quad D(i,j-1)+1, & \\ \quad D(i-1,j)+1, & j>0,\ i>0 \\ \quad D(i-1,j-1)+\mathrm{m}(S_1[i],S_2[j]) & \\ \} & \end{cases}$$

S1, S2: The two strings being compared.

- D(i,j): The Levenshtein distance between the first i characters of S1 and the first j characters of S2.
- i, j: Indices in strings S1 and S2, respectively.
- m(S1[i], S2[j]): The cost of substituting the i-th character of S1 with the j-th character of S2.

2. Graphemes based (To be done)

- A grapheme-based evaluation checks how well the system recognizes Tamil letters, including consonants combined with vowel signs or diacritics.

Experiments

- Cloud Vision API, Google Docs API, and the open-source Tesseract OCR were evaluated with the collected data.

Results

- Overall, the Cloud Vision API produced better results than the others.

Challenges

- Misidentification of Tamil numerals. (0, க, உ, ங, ச, ஞ, சு, எ, அ, கூ)
- Confusion among similar-shaped characters. (வ, ல, க, சு, ன, ள, ஐ, ஜ)
- Complex layout.
- Quality of the image.
- Annotations and Catalog Labels.

Examples of Challenging Images

References

1. Kaundilya, C., Chawla, D., & Chopra, Y. (2019, March). Automated text extraction from images using OCR system. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 145-150). IEEE.
2. Liyanage, C., Nadungodage, T., & Weerasinghe, R. (2015, August). Developing a commercial grade Tamil OCR for recognizing font and size independent text. In *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 130-134). IEEE.
3. Vasantharajan, C., Tharmalingam, L., & Thayasivam, U. (2022, October). Adapting the Tesseract Open-Source OCR Engine for Tamil and Sinhala Legacy Fonts and Creating a Parallel Corpus for Tamil-Sinhala-English. In *2022 International Conference on Asian Language Processing (IALP)* (pp. 143-149). IEEE.
4. Gnana Prasath, D. (2024, January 30). *OCR Tamil - Easy, Accurate, and Simple to Use Tamil OCR*. Medium. https://gnana70.medium.com/ocr-tamil-easy-accurate-and-simple-to-use-tamil-ocr-b03b98697f7b
5. Ramanan, M., Ramanan, A., & Charles, E. Y. A. (2015, August). A hybrid decision tree for printed Tamil character recognition using SVMs. In *2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 176-181). IEEE.
6. Zhu, W., Sokhandan, N., Yang, G., Martin, S., & Sathyanarayana, S. (2022, June). DocBed: A multi-stage OCR solution for documents with complex layouts. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 12643-12649).

Acknowledgement

This research study was carried out under the DigSAL project, supported by the German Academic Exchange Service (DAAD) and funded by the Federal Ministry for Economic Cooperation and Development (BMZ) through SDG Partnerships.