Benchmarking OCR Models for Sinhala and Tamil Document Digitization

Preprint · April 2025

DOI: 10.13140/RG.2.2.20843.25129

2 authors:

Purushoth Velayuthan
University of Nevada, Reno
5 PUBLICATIONS **1** CITATION
SEE PROFILE

Thanuja D. Ambegoda
University of Moratuwa
26 PUBLICATIONS **142** CITATIONS
SEE PROFILE

Engineering Research Unit
University of Moratuwa
Sri Lanka

Benchmarking OCR Models for Sinhala and Tamil Document Digitization

1st Purushoth Velayuthan
Dept. of Computer Science & Engineering
University of Moratuwa
purushothv@cse.mrt.ac.lk

2nd Thanuja D. Ambegoda
Dept. of Computer Science & Engineering
University of Moratuwa
thanuja@cse.mrt.ac.lk

***Index Terms*—Vision Language Models (VLMs), Optical Charater Recognition (OCR), Benchmarks, Low-resource languages, Sinhala and Tamil**

I. Introduction

The digitization of documents in low-resource languages, such as Sinhala and Tamil, presents significant challenges due to the unique complexities of these scripts and the scarcity of high-quality training data. While traditional OCR systems have made strides in converting printed text to digital formats, they struggle with the intricate layouts and linguistic nuances of underrepresented languages. Recent advancements in Vision-Language Models (VLMs), like UDOP [1] and HRVDA [2], have integrated visual and textual data for improved document understanding. However, the application of these models to low-resource languages remains limited, leaving a gap in accurate document digitization.

This research benchmarks several prominent OCR models, including Surya-OCR, TR-OCR [3], EasyOCR [4], and Tesseract OCR [5], focusing on their performance in digitizing documents in Sinhala and Tamil. We evaluate these models using key metrics—Character Error Rate [6] (CER), Word Error Rate [7] (WER), BLEU Score [8], METEOR [9], and Edit Distance [6](ED)—to determine the most effective solutions for low-resource languages.

Our key contributions are:

- A structured benchmarking framework for assessing OCR models on low-resource languages.
- The introduction of five performance metrics that can be applied across various low-resource languages.
- Demonstrating that Surya-OCR sets a new benchmark for document digitization in Sinhala and Tamil.

II. Methodology

Benchmarking Datasets:We benchmark using datasets specifically designed for Sinhala and English document digitization. For Sinhala, we use the Ransaka-Sinhala-Synthetic-OCR and Ransaka-Sinhala-Synthetic-OCR-Large datasets, offering diverse synthetic text images for OCR tasks. For English, we utilize the FUNSD [10] dataset, which focuses on form understanding in noisy document environments. Details on filtration and post-processing are provided below.

Hardware Specifications: All experiments were conducted on a single machine with an Intel i99900K CPU, 64GB of RAM, and an Nvidia Tesla 2X T4 (16GB each).

Software Specifications: All models and training code were developed using the HugginFace(HF) Transformers [11] library. For evaluation, we use CER, WER from fastwer library, BLEU Score from scarbleu library, METEOR from HuggingFace, and Edit distance from nltk library.

Models: For our benchmarking experiments, we use Surya-OCR, Tr-OCR (Ransaka), EasyOCR, and Tesseract OCR. These models were selected to provide a comprehensive comparison of OCR performance for Sinhala and English text digitization. All evaluations were performed on these pre-trained models without additional fine-tuning.

Evaluation Details: We evaluate models using the Hugging Face Transformers API with pre-trained models, employing a consistent batch size of 64. Evaluations focus on accuracy, character error rate (CER), and word error rate (WER) across both Sinhala and English datasets. The benchmarking results are presented in the tables below.

Data Filtration: We describe the filtration methods applied to the datasets used in our study.

*1) **Primary Filtration**:* For sinhala-synthetic-ocr (100 lines) and sinhala-synthetic-ocr-large (6.97k lines), we remove punctuation, whitespace, duplicates, and non-Sinhala text. Sentences with irregular lengths are filtered out. For FUNSD (50 English documents), only printed text blocks are retained by excluding any blocks containing handwritten content.

*2) **Secondary Filtration**:* Using Jensen-Shannon divergence, we filter out high-divergence sentences in the Sinhala datasets, ensuring uniformity in word distributions. For FUNSD, text blocks with abnormal structures are removed, focusing solely on coherent printed content for OCR evaluation.

Post-processing: Before evaluation, we tokenize predicted outputs to align them with the annotated ground truth. This segmentation of dates, special characters, and compound words ensures consistency with the dataset structure. This normalization enhances the accuracy of metrics like CER,

WER, BLEU, METEOR, and Edit Distance, thereby improving evaluation reliability for low-resource languages.

III. Results and discussions

TABLE I
RESULTS ON SINHALA DATASET

Datasets	Model	Metrics				
		WER	***CER***	***BLEU***	***Meteor***	***ED***
A	Surya-ocr	19.57	2.58	81.57	0.91	3.07
B		16.84	2.58	84.55	0.93	3.15
A	Tr-ocr	58.54	25.18	19.41	0.42	25.52
B		61.34	25.72	18.59	0.40	26.47
A	Tesseract	89.23	92.33	7.35	0.13	67.55
B		95.78	97.79	5.67	0.01	74.23

*A = Ransaka-sinhala-synthetic-ocr dataset
*B = Ransaka-sinhala-synthetic-ocr-large dataset

TABLE II
RESULTS ON ENGLISH DATASET (FUNSD)

Model	***WER***	***CER***	***BLEU***	***Meteor***	***ED***
Surya-ocr	76.67	53.99	52.02	0.76	110.90
Tr-ocr	73.23	67.23	12.46	0.22	174.21
Easy-ocr	77.13	49.25	22.64	0.42	135.42
Tesseract	76.35	44.36	34.13	0.55	97.32

TABLE III
OVERVIEW OF GPU POWER CONSUMPTION AND INFERENCE TIMING FOR VARIOUS MODELS

Datasets	Models	Number of Trainable Parameters	Time (hrs)	Power (kWh)
A	Surya-ocr	167	0.32	0.21
	Tr-ocr	300	0.20	0.96
	Tesseract	NA	0.11	NA
B	Surya-ocr	167	0.83	0.69
	Tr-ocr	300	0.56	1.32
	Tesseract	NA	0.34	NA
FUNSD	Surya-ocr	167	0.15	0.68
	Tr-ocr	300	0.10	0.87
	Easy-ocr	6	0.12	0.01
	Tesseract	NA	0.09	NA

The evaluation, as shown in Table I and Table II, highlights a trade-off between computational efficiency and accuracy across models. On the Sinhala dataset, Surya-ocr achieved the lowest error rates (WER: 19.57, CER: 2.58), while Tr-ocr and Tesseract performed poorly, with WERs exceeding 58% and 89%, respectively. On the FUNSD dataset, all models exhibited high error rates, with Surya-ocr leading at a WER of 76.67. Table III reveals that Surya-ocr strikes a balance between power consumption and performance, requiring significantly less power (0.21 kWh for Dataset A) than Tr-ocr, which uses more resources but delivers lower accuracy. While models like Tesseract are more efficient in terms of timing, their high error rates undermine practical usability. Overall, Surya-ocr emerges as the most balanced option for low-resource languages like Sinhala, combining moderate computational demand with relatively high accuracy, though further optimization is needed to improve performance without increasing resource consumption.

IV. Conclusion

This study highlights the challenges of digitizing documents in low-resource languages like Sinhala and Tamil. Our benchmarking of OCR models—Surya-OCR, TR-OCR, Easy-OCR, and Tesseract—demonstrates that Surya-OCR outperforms the others, achieving the lowest WER and CER on the Sinhala dataset while exhibiting greater efficiency. These findings establish Surya-OCR as a promising solution for improving document digitization in underrepresented languages. Our future efforts will focus on the quantization of Surya-OCR to further reduce inference time and computational requirements, optimizing it for low-resource environments.

References

[1] Z. Tang, Z. Yang, G. Wang, Y. Fang, Y. Liu, C. Zhu, M. Zeng, C. Zhang, and M. Bansal, "Unifying vision, text, and layout for universal document processing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 254–19 264.

[2] C. Liu, K. Yin, H. Cao, X. Jiang, X. Li, Y. Liu, D. Jiang, X. Sun, and L. Xu, "Hrvda: High-resolution visual document assistant," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 534–15 545.

[3] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, "Trocr: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 094–13 102.

[4] M. Salehudin, S. Basah, H. Yazid, K. Basaruddin, M. Safar, M. M. Som, and K. Sidek, "Analysis of optical character recognition using easyocr under image degradation," in *Journal of Physics: Conference Series*, vol. 2641, no. 1. IOP Publishing, 2023, p. 012001.

[5] A. Kay, "Tesseract: an open-source optical character recognition engine," *Linux Journal*, vol. 2007, no. 159, p. 2, 2007.

[6] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Proceedings of the Soviet physics doklady*, 1966.

[7] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, 1997, pp. 347–354.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[9] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[10] G. Jaume, H. K. Ekenel, and J.-P. Thiran, "Funsd: A dataset for form understanding in noisy scanned documents," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2. IEEE, 2019, pp. 1–6.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.