Dataset to the tune of Devanagari Document OCR

*Malathi. P, Asst. Prof., Dept of ISE[1][0000−0002−6096−5539] and Dr. Chandrakanth G Pujari, Prof. Head, Dept of MCA[2][0000−0002−1358−2113]

[1]*Dr. Ambedkar Institute of Technology, Mallathahalli, Bengaluru- 560056*
[2] *Dr. Ambedkar Institute of Technology, Mallathahalli, Bengaluru- 560056*
Visvesvaraya Technological University, Belagavi – 590018

Abstract *Many Devanagari OCR solutions offered online have beenobserved to be ineffective on scanned images of analog printed docu- ments and historic printed documents. The primary cause of this inef- ficacy is the treatment of Devanagari OCR as an extension of solutions and methods developed for Latin, Chinese, and other prominent global scripts, without adequate customization for the unique characteristics of Devanagari. Furthermore, these solutions predominantly rely on recent advances in deep learning techniques, which require large datasets to generate accurate machine learning models. However, despite the De- vanagari script forming the foundation for many Indic languages and being spoken by millions of people worldwide, it is still categorized as a low-resource script in the context of OCR solutions, primarily due to the scarcity of available benchmark datasets. The existing public datasetsare limited, typically covering only numerals, handwritten text, or on- line OCR, and the datasets available for scene text OCR or natural language processing have very few background settings that align with the needs of printed document OCR. To address this significant gap, we present a comprehensive dataset comprising 14,000-word images and 400 phrase images, complete with detailed annotations, specifically tailored for printed document OCR. This dataset was rigorously tested using Mi- crosoft Azure Vision OCR, Google Vision OCR, and Easy OCR solutions through their APIs to assess their effectiveness in handling Devanagari script in the context of printed documents. The results highlighted the challenges these mainstream OCR solutions face when applied to De- vanagari, reinforcing the need for specialized datasets and models. Ad- ditionally, we present the frequency distribution of various glyphs in the alphabet set within the dataset, which could serve as a valuable resource for future research and development, enabling more accurate and effec- tive OCR solutions as the situation demands.*

Keywords: **OCR· devanagari script· dataset· structured document · histograms · typefont · low resource · annotation · distribution · data generators**

1. Introduction

Printed document OCR (Optical Character Recognition) solutions play a pivotal role in converting large amounts of offset (analog) and digitally printed text images into editable machine-readable formats. This process is crucial for excavating and utilizing the vast wealth of knowledge present in Indian literature, much of which remains in a dormant state due to the lack of digital accessibility. Additionally, OCR solutions serve as a foundational component in the automation pipeline on digital platforms across multiple domains.

Compared to printed document OCR, scene text OCR, often referred to as "text in the wild" presents a more challenging task. It deals with complex backgrounds and artistic fonts, as well as varying text orientations, making it more difficult to achieve accurate recognition. Online text OCR, which handles handwritten sequences of pixels captured on touch-sensitive screens, is another well-researched area with a wide variety of datasets and significant advance- ments. Offline handwritten text OCR, on the other hand, must contend with challenging and irregular shapes, curvatures, and overlapping alphabets, leading to the development of more complex solutions.

While corpora-based text datasets are available in large numbers on digital plat- forms and are predominantly used for natural language processing (NLP), these datasets are limited in vocabulary and do not have image dataset equivalents that suit document OCR needs. Consequently, the solutions developed for scene text, online text, and handwritten text OCR often prove ineffective for document OCR, which deals with structured text. These solutions tend to overfit or underfit the specific objectives of document OCR.

2. Previous Work

Krzysztof Olejniczak et al.[1] highlighted that state-of-the-art methods for wild text recognition are typically evaluated on complex scenes, yet their performance in the domain of document text recognition is not widely published. This un- derscores the significant differences between scene text databases and document OCR databases.

Michael Arrigo et al.[2] from the Linguistic Data Consortium developed CAMIO, a resource to facilitate the development and evaluation of OCR and related tech- nologies for 35 languages across 24 unique scripts. The Devanagari script is part of this corpus, which contains 2,500 page-level images with an average of 39 lines per image, complete with transcription and detailed metadata. Shailesh Acharya et al.[3] presented an image database of handwritten Devanagari char- acters, consisting of 46 base characters with 2,000 images each.

Subodh Deolekar[4] introduced the SHABD dataset, which includes images of basic consonants and vowels along with their combinations and syllables, cov- ering all aspects of the language. However, models developed on these limited datasets face challenges in recognizing combined alphabets

Prathamesh Zade[5] released a dataset containing synthetic Hindi images and

text pairs of sentences, created using the TextToImageGenerator repository. This dataset includes 80,000 images comprising 630,515 words, generated using eight standard commonly used type fonts

Obaidullah et al.[6]contributed the PHDIndic_11 image database for handwritten Indic scripts, which includes 220 handwritten Devanagari text pages with 2,457 text lines and 23,264 words.

Sayalighodekar[7] presented a collection of approximately 12,000 Marathi word images with corresponding text labels encoded in UTF-8 format, compiled from 12 books across different genres to include diversity in vocabulary and font vari ations.

Nibaran Das[8] et al constructed a dataset of handwritten Devanagari numerals, consisting of 10,000 images with 300 images per class (per digit), as part of a larger dataset catalog that includes Bangla and Telugu datasets.

From these findings, it is evident that most Devanagari databases focus on numerals or handwritten word and sentence-level images, with a scarcity of word- level printed document images. Additionally, challenges related to latency, optimization, and complexity reduction in existing document OCR solutions need to be addressed. To meet these needs, we present the D_w_ph_Db dataset.

3. Dataset Preparation

A refined subset of the Mile Devanagari dataset[9] was utilized during the prepa- ration of our dataset. Text documents were scanned using scanners at 300 dpi resolution and stored in 8-bit grayscale format. The images were selected from books, newspapers, magazines, and other printed documents, ensuring variations in printing style, sizes, and background. A total of approximately 50 pages for the word set and 20 pages for the phrase set were scanned and segmented.

Global-level gross skew correction, thresholding, and binarization were applied using OpenCV, followed by horizontal projection to perform line-level segmentation. Vertical projection was then applied to the resulting intermediate images to obtain word boundary coordinates. For phrase boundary coordinates, a ran- dom number generator was used on vertical projections to extract phrases of varying lengths. The rectangular boundary coordinates obtained for words and phrases were then applied to the original scanned page-level images to construct the image datasets.

To create the annotation files, automated transcription significantly reduced time and costs by using software for the task, followed by rigorous manual transcription. Punctuations were retained, and transcription was performed exactly as in the source images without correcting misspellings, grammatical errors, or other mistakes. The dataset includes 14,000-word images, enabling the development of document OCR models using Convolutional Neural Networks (CNNs) and similar deep learning techniques that favour word-level input images.

Text labels from the image dataset's annotation file were converted to Unicode format to calculate the frequency distribution of the base alphabet

set. The frequency distribution of a total of 41,184 characters across 47 alphabets was normalized on a scale of 1, as shown in Table 1. The word length ranges from 1 to 12, with an average word length of 4-5 characters.

Table 1: Frequency distribution of Devanagari symbols on word dataset

X	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+090X			◌ॅ 1.81	◌ ◌: 0.01		अ 0.79	आ 0.79	इ 0.37	ई 0.37	उ 0.52	ऊ 0.52	ऋ 0.05				ए 0.54
U+091X	ऐ 0.54			ओ 0.11	औ 0.11	क 3.8	ख 0.3	ग 1.26	घ 0.36	ङ 0.19	च 0.8	छ 0.01	ज 1.33	झ 0.05	ञ 0.13	ट 0.41
U+092X	ठ 0.39	ड 0.2	ढ 0.38	ण 0.07	त 3.34	थ 0.4	द 1.43	ध 0.38	न 0.52		प 2.1	फ 0.01	ब 0.25	भ 1.2	म 0.53	य 2.31
U+093X	र 1.9		ल 1.8	ळ 0.01		व 1.64	श 0.73	ष 0.76	स 0.36	ह 2.8						

The dataset also includes 400 phrase images, which can be used to develop document OCR models using Vision Transformers and similar deep learning techniques that favour phrase-level input images. The frequency distribution oftext labels concerning the base alphabet set was calculated using Unicode, withthe distribution of a total of 2,634 characters across 47 alphabets normalized on a scale of 1, as shown in Table 2. The phrase length ranges from 1 to 12, with an average phrase length of 5-6 characters.

Table 2: Frequency distribution of Devanagari symbols on phrase dataset

X	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+090X			◌ॅ 2.36	◌ ◌: 0.03		अ 0.52	आ 0.49	इ 0.58	ई 0.37	उ 0.42	ऊ 0.41	ऋ 0.02				ए 0.58
U+091X	ऐ 0.57			ओ 0.33	औ 0.31	क 4.92	ख 0.26	ग 0.94	घ 0.14	ङ 0.1	च 0.73	छ 0.19	ज 0.93	झ 0.1	ञ 0.03	ट 0.52
U+092X	ठ 0.03	ड 0.31	ढ 0.03	ण 0.52	त 0.78	थ 0.36	द 1.52	ध 0.96	न 3		प 1.99	फ 0.4	ब 0.82	भ 0.44	म 2.25	य 2.57
U+093X	र 5.29		ल 1.85	ळ 0.03		व 2.15	श 0.8	ष 0.56	स 3.48	ह 1.94						

The frequency distribution shown in Table 1 and Table 2 provides guidance for extending the dataset with synthetic images using data generators if required.

4. Testing and Observation

Easyocr api[12] applied to the images resulted in less than 50% text conversion. Google cloud vision api[11] also returned errors for some cases, labeling the data as ”bad.” To improve accuracy, gray images were binarized, thresholded, and thinned before calling the APIs. Many images in the dataset contained punc- tuations along with text. Among the OCR solutions tested. Microsoft azure vision api[10] was the most accurate, while EasyOCR was the least accurate in predicting punctuations. For further analysis, 2,001 sample images without punctuations were manually selected from the 14,000-word dataset. The com- parative effectiveness of the three selected OCR solutions is shown in Table 3.

Table 3: Results

	Accuracy(%)		
Dataset size 2001	Microsoft Azure Vision OCR	Google Cloud Vision OCR	Easy OCR
Word Error Rate	3.05	15.24	26.18

5. Conclusion and Future work

The D_w_ph_db dataset can be leveraged in multiple ways, including fine- tuning large language models (LLMs) to small language models (SLMs) and facilitating machine translation to other languages. The dataset can also be expanded to be part of a larger dataset to support the development of bilingual and multilingual OCR solutions. In the future, we plan to increase the volume of the dataset and extend it to cover printed text images of other major Indic scripts. This will further enhance the utility and applicability of OCR solutions in the context of Indian languages, addressing the unique challenges they present.

6. Data Availability

https://drive.google.com/drive/folders/
1ve-jLphMU95kDrPw_tGvmZIA3z_uguIE?usp=drive_link

References

1. *Krzysztof Olejniczak, Milan Sulc2 (Jan 2023) "Text Detection Forgot About Doc- ument OCR" https://arxiv.org/abs/2210.07903.*
2. *Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, Lisa Mason (June 2022) "CAMIO: A Corpus for OCR in Multiple Languages" 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1209–1216 Marseille, https://aclan- thology.org/2022.lrec-1.129*

3. *Shailesh Acharya, Ashok Kumar Pant, Prashnna Kumar Gyawali (2016)"Deep learning based large scale handwritten Devanagari character recog-nition" https://ieeexplore.ieee.org/document/7400041Databaserepository https://archive.ics.uci.edu/dataset/389/devanagari+handwritten+character+dataset.*
4. *Subodh Deolekar (2024) DOI: 10.34740/KAGGLE/DSV/3211245*
5. *Prathamesh Zade (2024) https://www.kaggle.com/datasets/prathmeshzade/hindi- ocr-synthetic-line-image-text-pair*
6. *Sk Md Obaidullah, Chayan Halder, K. C. Santosh, Nibaran Das, Kaushik Roy(2017) "Development of Document Image Database for Handwritten Indic Script – A State- of-the-art" Multimedia Tools Application DOI 10.1007/s11042-017-4373-y*
7. *GitHub - sayalighodekar/Marathi-OCR-Dataset: A collection of about 12k Marathi word images with corresponding labels, useful for Devanagari Optical Character Recognition.*
8. *Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu "A genetic algorithm based region sampling for selectionof local features in handwritten digit recognition application" Volume 12, Is- sue 5, May 2012, Pages 1592-1606 https://www.sciencedirect.com/science/article/abs/pii/S1568494611004728?via%3Dihub*
9. *http://mile.ee.iisc.ac.in/downloads.html/*
10. *https://portal.vision.cognitive.azure.com/demo/extract-text-from-images*
11. *https://cloud.google.com/vision/docs/ocr*
12. *https://www.jaided.ai/easyocr*

Declarations

Funding and/or Conflicts of Interests/Competing Interests

1. The authors did not receive support from any organization for the submitted work.
2. The authors have no competing interests to declare that are relevant to the content of this article.