LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 23, NO. S5(2025)

KANNADA HANDWRITTEN WORD DATASET FOR OCR VIA SYLLABLE COMPOSITION AND CORPUS AUGMENTATION

Dadapeer[1]*, **Yeresime Suresh**[2]

[1]Department of Computer Science and Engineering, Ballari Institute of Technology and Management(BITM), Ballari, India
Visvesvaraya Technological University, Belagavi, India;
[2]Department of CSE – Artificial Intelligence, Ballari Institute of Technolgy and Management(BITM), Ballari, India
Visvesvaraya Technological University, Belagavi, India;

dpbitm@gmail.com[1]
suresh.vec04@gmail.com[2]

Abstract
Handwritten Text Recognition (HTR) systems require extensive labeled datasets to achieve high accuracy, especially for word-level recognition. While English enjoys rich resources such as the IAM dataset, low-resource Indian languages like Kannada lack publicly available handwritten word-level datasets. In this paper, we present a novel approach for synthesizing a Kannada handwritten word image dataset by combining character-level handwritten images from the Char74K dataset with a real-world Kannada corpus sourced from Kaggle. Our pipeline intelligently segments Kannada words into syllables, maps them to corresponding character images, and stitches them to generate realistic word-level samples. This method bridges the gap in training data availability for Kannada and enables consistent training of word-level OCR systems similar to those available for English. The resulting dataset contains 500+ synthetically generated word images with accurate Unicode labels, and the approach is scalable to thousands of words. This work contributes a reproducible methodology and a valuable resource for the OCR research community.

Keywords: Kannada handwritten Dataset, Char74K; IAM dataset, Optical Character Recognition (OCR); Handwritten Text Recognition (HTR), syllables,

1. Introduction

The Handwritten text recognition in low-resource Indic languages remains a significant challenge due to the scarcity of publicly available, large-scale, word-level datasets [1]. In contrast, English benefits from extensive resources like the IAM dataset [2], Kannada has historically relied on character-level datasets such as Char74K for isolated character recognition a dataset that treats compound vowels/consonants as separate classes and serves as a baseline for many studies [3]. Several recent efforts have attempted to generate or collect Kannada handwritten datasets, but word level corpora remain limited e.g., [1] introduces around 100k word level samples and challenges persist—this aligns with insights from the competition [4], which highlights the scarcity of robust, large-scale word-image datasets for Indian scripts including Kannada. Our proposed approach bridges this gap by synthesizing Kannada handwritten word images via syllable aware composition using Char74K and real-world language data, aiming to enable sequence based OCR models comparable to those trained on high-resource languages [5].

Figure 1 depicts the overall workflow of a handwritten Optical Character Recognition (OCR) system. The process generally begins with preprocessing, which involves tasks such as binarization, noise reduction, and deskewing to enhance the quality of the input image. The next stage is segmentation, where text is separated into lines, words, or individual characters. This is followed by feature extraction, where distinctive numerical representations are derived to capture the structural and contextual aspects of handwriting. In the recognition phase, these features are mapped to corresponding characters or words using machine learning or deep learning approaches. The final stage, post-processing, employs dictionaries or language models to correct errors and improve accuracy. Within this workflow, our proposed Kannada handwritten word dataset fulfills a crucial gap by providing word-level inputs, which are essential for training and benchmarking modern deep learning OCR models.

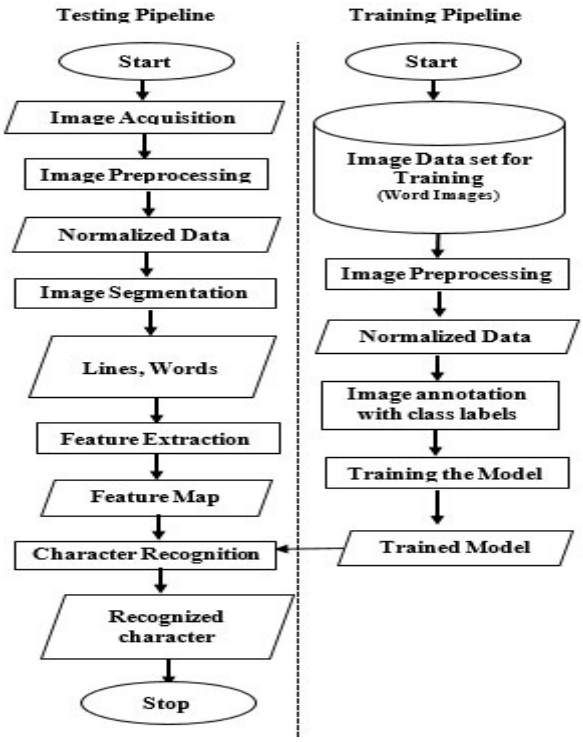


Figure 1. Steps in OCR Process.

2. Related Work

The IAM dataset [2] has long been a benchmark for English handwritten word recognition. The Char74K dataset [3], which includes 657 Kannada grapheme classes treated as individual symbols, has been widely used in Kannada OCR research. Implicit segmentation techniques using HMMs demonstrated improved performance on Char74K by leveraging Kannada's agglutinative nature [6]. More recently, Kannada MNIST was released, providing handwritten digit data and serving as a drop in replacement for MNIST [7]. A larger Kannada word level dataset—IIIT INDIC HW WORDS—contains over 100,000 handwritten Kannada word images with ground truth [1].

OCR engines such as LipiGnani and eAksharayan have been developed specifically for printed Kannada text, achieving better accuracy than general-purpose engines like Tesseract [8]. However, these systems are not designed for handwritten content, and support for handwritten Kannada OCR remains limited. Recent multilingual datasets like MDIW-13 include some handwritten Kannada samples, but these are not sufficient to train robust models [9].Recent advances in synthetic data generation—such as SynthText for scene text localization [10] and compositional techniques for generating synthetic handwritten Indic datasets like Devanagari [11] have influenced our approach. Modern OCR architectures, especially those based on transformers with CTC loss, have substantially improved recognition accuracy. Furthermore, open-source vision-language models like SmolDocling [12] demonstrate strong performance on multi-domain documents that include both printed and handwritten content.

Our work builds on these efforts by developing a syllable composition method to synthetically generate Kannada handwritten word images using genuine character samples and real-word vocabulary.

3. Dataset Resources and Motivation

The Char74K dataset [3], Char based Kannada dataset Kaggle corpus, and Kannada MNIST [7] highlight the need for more realistic, word-level data. The IIIT INDIC HW WORDS Kannada dataset [1] demonstrates feasibility at scale but remains inaccessible to many practitioners. Our method leverages the Char74K dataset as a foundation and enhances it via syllable-level composition and a real-world corpus, producing a dataset format consistent with IAM and other sequence OCR datasets [2].

4. Methodology

The dataset creation pipeline for our Kannada handwritten OCR project comprises several well-defined stages aimed at ensuring robust and reproducible data preparation.

4.1. Environment Setup

To enable persistent storage of both intermediate files and final outputs, Google Drive was mounted within the Google Colab environment. Essential Python libraries—including pandas, matplotlib, PIL, zipfile, re, and kagglehub—were imported to support key functionalities such as CSV parsing, image visualization and manipulation, Unicode character rendering, and seamless dataset retrieval from Kaggle.

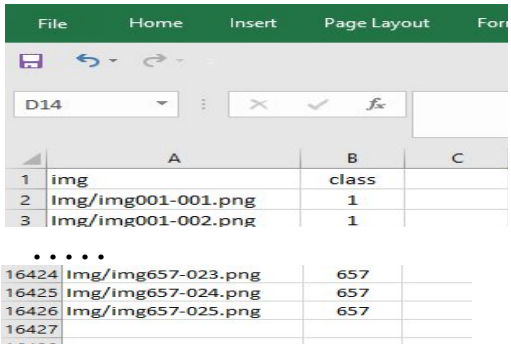
4.2. Data Retrieval

All Multiple resources were sourced from Kaggle to construct the character and word-level datasets. These include:.

- Kannada character dataset (Char74K), which comprises:
  -Img/: a directory containing handwritten character images, with one character per image.
  -kannada.csv: a CSV file mapping image file paths to corresponding numeric class labels representing syllables.
- Real Kannada corpus (kn.txt), which provides authentic Kannada words and was used to generate labels during synthetic word image creation.
- Kannada font files (.ttf), which were essential for rendering Kannada Unicode syllables accurately. These fonts were specifically required in the Colab environment, where default system fonts do not support Kannada script. To ensure consistent rendering of Kannada syllables, we used the NotoSerifKannada TrueType font, which supports the full Kannada Unicode block and complex grapheme shaping [13].

Representative samples from the retrieved resources—namely, the handwritten character images from the Char74K dataset (Img/ folder), the label mappings in kannada.csv, and a few lines from the real Kannada corpus (kn.txt)—are presented in Figures 2a to 2c. These are partial views intended to illustrate the structure and content of the datasets.

(a)



(b)

ಕಂ|| ಅಱಿದುಂಕಂಡುಂನಂಟರ್ಜಂಗ್ಲೆಂಡ್ವರಜೋಸ್ಬಟ್ಲರ್106 , ಬೆನ್ಸ್ಟೋಕ್ಸ್62 ಹಾಗೂರಶೀದ್ಔಟಾಗದೇ33 ರನ್ಬಾರಿಸಿಮಿಂಚಿದ್ರು.ಕಮ್ಮನಹಳ್ಳಿಯಲ್ಲಿನರಾಜ್ಕುಮಾರ್ಪಾರ್ಕ್ಗೆಬಾಲಕನನ್ನುಕರೆದುಕೊಂಡುನಿನ್ನೆಸಂಜೆಪೋಷಕರುಬಂದಿದ್ದಾರೆ. ಹಾಗೆಯೇನಾರ್ವೆ, ಸ್ವೀಡನ್, ಐರ್ಲೆಂಡ್ಮತ್ತುಸ್ವಿಟ್ಜರ್ಲೆಂಡ್ಗಳಲ್ಲಿಕೂಡಜನಸಂಖ್ಯೆಹೆಚ್ಚುತ್ತದೆ.ಅವರುಕೇವಲಮಹಿಳೆಯರಬಗ್ಗೆಹೆಚ್ಚುಕಾಳಜಿಯಿಲ್ಲವೇ, ಅಥವಾಇದುಕೆಲವುರೀತಿಯಮುಂಬರುವಹುಚ್ಚುತನದಮೊದಲಚಿಹ್ನೆಯಾ?ಈಬಾರಿಯಗಣರಾಜ್ಯೋತ್ಸವಪ್ರಶಸ್ತಿ...ಕನ್ನಡದಲ್ಲಿಈ ಮೊದಲೂವೆಬ್ ಸೀರೀಸ್ (ವೆಬ್ಸೈಟ್ ಮೂಲಕಪ್ರಸಾರವಾಗುವಕಾರ್ಯಕ್ರಮ) ರೂಪಿಸುವಕೆಲಸಆಗಿತ್ತಂತೆ.Lamisil ಸ್ವೇತಡೆಗಟ್ಟುವಿಕೆಮತ್ತುಚಿಕಿತ್ಸೆಬಳಸಲಾಗುತ್ತದೆಶಿಲೀಂಧ್ರಗಳಚರ್ಮದಹಾನಿಯನ್ನು.ಕಾಮನ್ವೆಲ್ತ್ಕೂಟನೋಡಲಿರುವಉಸೇನ್ಬೋಲ್ಟ್

ರಾಜ್ಯದವಿವಿಧಭಾಗಗಳಿಂದಹೋರಾಟಗಾರರುಈಅಧಿವೇಶನಕ್ಕೆಬಂದಿದ್ದರು.

(c)

Figure 2. Representative samples from the retrieved resources: (a)Sample images from the Img/ folder of the Char74K Kannada handwritten character dataset. (657 syllables × 25 samples = 16,425 images); (b)Screenshot of the kannada.csv file showing image paths and class labels; (c) Corpus from the kn.txt Kannada corpus used for synthetic word generation.

4.3. Cleaning the Character Dataset

The kannada.csv file from the Char74K dataset [3] provides mappings between handwritten character image file paths and their associated numeric class labels. However, on inspection, several inconsistencies were identified. Notably, multiple numeric class labels were found to correspond to the same syllable, resulting in label duplication, and the label indices were non-sequential and discontinuous.

To address these issues, a label consolidation and remapping procedure was implemented. All duplicate labels representing the same syllable were merged to ensure each syllable had a unique class label. The class indices were then re-indexed sequentially, starting from 1, to maintain consistency for training and decoding purposes. The cleaned and standardized data was stored in a new DataFrame named kan_img_cl_df, which includes the absolute image paths of handwritten characters along with their cleaned, unique class labels.

This refined dataset served as a foundational resource for constructing syllable-to-label dictionaries and for synthesizing word-level images used in later stages of the OCR pipeline.

4.4. Creating the Syllable Set

To establish a consistent and comprehensive mapping between handwritten character images and their corresponding linguistic units, a structured Kannada syllable set was programmatically constructed. This process involved combining fundamental components of the Kannada script to generate all visually and phonetically distinct syllables. To build a structured Kannada syllable inventory, we programmatically combined the core script elements—including 14 vowels, 34 base consonants, vowel modifiers or diacritics, the halant (virama), anusvāra, visarga, and numerals—aligned with Unicode grapheme definitions and Kannada akshara formation rules as detailed in the Unicode Standard [14], the Kannada writing system [15], and documented linguistic structure of agglutinative Indic scripts [16, 17].The core components used were:

1. Vowels (ಸ್ವರಗಳು):

A total of 14 vowels were considered: ಅ, ಆ, ಇ, ಈ, ಉ, ಊ, ಋ, ೠ, ಎ, ಏ, ಐ, ಒ, ಓ, ಔ. Each vowel represents a unique sound and is essential for forming syllabic units. The complete set of vowels used in this work, along with their approximate Romanized pronunciations, is presented in Table 1.

Table 1. Kannada vowels (ಸ್ವರಗಳು) and their corresponding pronunciations

Vowels (Swaragalu)	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ	ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ
Pronunciation	a	aa	i	ii	u	uu	ru	rru	e	ee	ai	o	oo	au

2. Consonants (ವ್ಯಂಜನಗಳು):

A complete set of base consonants in the Kannada script was used for syllable formation. These include: ಕ, ಖ, ಗ, ಘ, ಙ, ಚ, ಛ, ಜ, ಝ, ಞ, ಟ, ಠ, ಡ, ಢ, ಣ, ತ, ಥ, ದ, ಧ, ನ, ಪ, ಫ, ಬ, ಭ, ಮ, ಯ, ರ, ಲ, ವ, ಶ, ಷ, ಸ, ಹ, and ಳ.

Each of these consonants serves as a base character that can be combined with various vowel modifiers to generate distinct syllabic forms. The complete set of base consonants used in this work, along with their Romanized pronunciations, is provided in Table 2.

Table 2. Kannada consonants (ವ್ಯಂಜನಗಳು) and their corresponding pronunciations

Consonants Venjanagalu	ಕ	ಖ	ಗ	ಘ	ಙ	ಚ	ಛ	ಜ	ಝ	ಞ	ಟ	ಠ
Pronunciation	ka	kha	ga	gha	ṅa	ca	cha	ja	jha	ña	ṭa	ṭha
Consonants Venjanagalu	ಡ	ಢ	ಣ	ತ	ಥ	ದ	ಧ	ನ	ಪ	ಫ	ಬ	ಭ
Pronunciation	ḍa	ḍha	ṇa	ta	tha	da	dha	na	pa	pha	ba	bha
ConsonantsVenjanagalu	ಮ	ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ		
Pronunciation	ma	ya	ra	la	va	śa	ṣa	sa	ha	ḷa		

3. Vowel modifiers (Matregalu):

Vowel sounds are applied to consonants through modifiers such as ◌ಾ, ◌ಿ, ◌ೀ, ◌ು, ◌ೂ, ◌ೃ, ◌ೄ, ◌ೆ, ◌ೇ, ◌ೈ, ◌ೊ, ◌ೋ, ◌ೌ and ◌್ (halant). For instance, when the consonant ಕ is combined with these modifiers, it results in forms like ಕಾ (kaa), ಕಿ (ki), ಕೀ (kii), ಕು (ku), ಕೂ (kuu), ಕೃ (kru), ಕೄ(krru), ಕೆ (ke), ಕೇ (kee), ಕೈ (kai), ಕೊ (ko), ಕೋ (koo), ಕೌ (kau), and ಕ್ (halant/suppressed vowel). Table 3 shows the complete set of vowel modifiers (ಮಾತ್ರೆಗಳು) used in this work, along with their effects when applied to the base consonant ಕ. This illustrates how each modifier alters pronunciation and visual representation.

Table 3: Kannada vowel modifiers (Matregalu) and their effects when combined with the consonant ಕ.

Vowel modifiers	◌ಾ	◌ಿ	◌ೀ	◌ು	◌ೂ	◌ೃ	◌ೄ	◌ೆ	◌ೇ	◌ೈ	◌ೊ	◌ೋ	◌ೌ	◌್
Example	ಕ + ◌ಾ	ಕ + ◌ಿ	ಕ + ◌ೀ	ಕ + ◌ು	ಕ + ◌ೂ	ಕ + ◌ೃ	ಕ + ◌ೄ	ಕ + ◌ೆ	ಕ + ◌ೇ	ಕ + ◌ೈ	ಕ + ◌ೊ	ಕ + ◌ೋ	ಕ + ◌ೌ	ಕ + ◌್
Result	ಕಾ	ಕಿ	ಕೀ	ಕು	ಕೂ	ಕೃ	ಕೄ	ಕೆ	ಕೇ	ಕೈ	ಕೊ	ಕೋ	ಕೌ	ಕ್
Pronunciation	kaa	ki	kii	ku	kuu	kru	krru	ke	kee	kai	ko	koo	kau	Halant

4. Special Characters (Vishesha Chinnegalu):

Two special characters were also included—◌ಂ (Anusvara) and ◌ಃ (Visarga). These are used to modify pronunciation or indicate nasal or aspirated sounds. For example, ಕಂ (kam) and ಕಃ (kah) are derived by combining the consonant ಕ with these symbols. Table 4 summarizes the two special Kannada script characters used in our syllable set construction, along with examples and their linguistic roles.

Table 4: Special Kannada characters (Vishesha Chinnegalu) used in syllable formation.

Symbol	**Name**	**Example**	**Meaning**
◌ಂ	Anusvara	ಅಂ	Nasal (am/an)
◌ಃ	Visarga	ಅಃ	Aspirated (aha)

5. Numerals (Sankhegalu):

To support recognition beyond alphabetic characters, Kannada numerals from ೦ to ೯ were included. Each numeral corresponds to a digit (0–9) and has a unique pronunciation, such as ೦ (sonne), ೧ (ondu), ೨

(eradu), and so on. Table 5 lists the Kannada numerals used in this work, along with their Romanized pronunciations and corresponding digit values.

Table 5: Kannada numerals (Sankhyegalu), their pronunciations, and digit values.

Numerals (Sankhyegalu)	೦	೧	೨	೩	೪	೫	೬	೭	೮	೯
Pronunciation	sonne	ondu	eradu	mooru	naalku	aidu	aaru	elu	entu	ombattu
Digit (Value)	0	1	2	3	4	5	6	7	8	9

Using a systematic combinatorial pipeline grounded in Unicode grapheme-cluster definitions, we generated all valid Kannada syllable forms—including consonant vowel combinations, consonants with modifiers, and numerals. In total, 655 unique syllable units were produced and stored in kannada_letter_text_corpus.txt for downstream synthetic word generation and label construction, following approaches similar to those used in Unicode-based grapheme parsing for Indic scripts [18]. Figure 3 provides a visual overview of the complete set of 655 unique Kannada syllables generated by our combinatorial pipeline. These syllables serve as the foundational recognition units in the OCR system.

Note: In this work, the term "syllable" refers to visually distinct units of the Kannada script—often known as grapheme clusters in Unicode—which may include consonant-vowel combinations and are typically rendered as single visual entities. These syllables form the foundational targets in the OCR pipeline.

4.5. Creating Label Dictionaries

To enable consistent mapping between syllables and numerical class identifiers, two key dictionaries were created. The label_dict mapped each unique syllable to an integer label, while the reverse_label_dict provided the inverse mapping—from integers back to syllables. These mappings were crucial both during training, for model input encoding, and during inference, for decoding OCR predictions back into readable Kannada text. The label_dict and reverse_label_dict mappings are analogous to label encoding strategies used in sequence modeling frameworks.

4.6. Mapping Character Labels and Image Paths

Leveraging the cleaned class labels and the previously constructed syllable corpus, each character image in the dataset was linked to its corresponding Unicode syllable. Specifically, the numeric class label in the kan_img_cl_dfDataFrame was converted to its associated syllable using the reverse_label_dict. Additionally, the relative image paths were resolved into absolute file paths by appending the base path obtained via kagglehub. The resulting DataFrame included two essential columns: img, representing the absolute path to each handwritten character image, and kannada_chrs, containing the corresponding Unicode syllable. This enriched and well-aligned dataset enabled seamless visualization and played a vital role in generating synthetic word-level images.

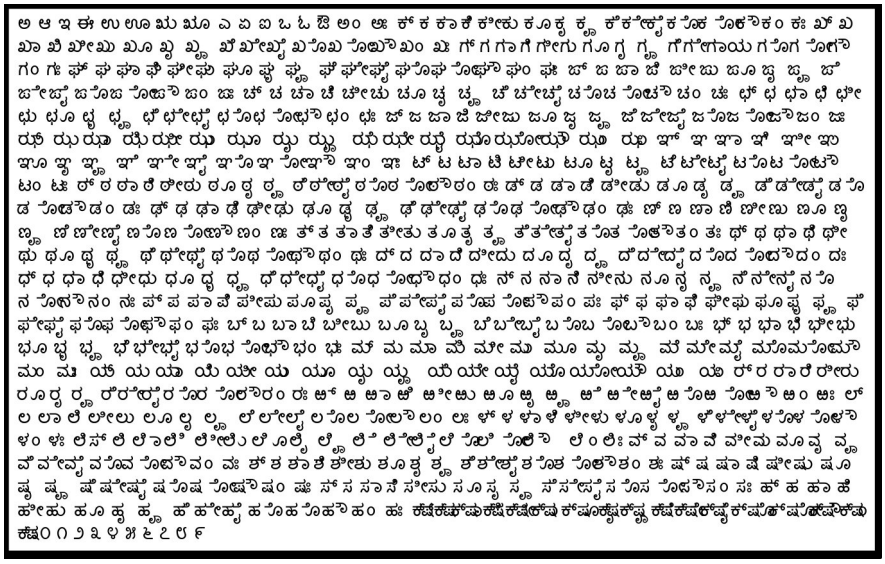


Figure 3. Sample of 655 Kannada syllables produced by combining core components

4.7 Displaying Character Samples

To validate that character images were correctly aligned with their respective syllables, a visualization step was implemented. A Kannada .ttf font was loaded using the matplotlib.font_manager module to ensure accurate rendering of Kannada script in plots. For each unique class, a representative character image was displayed using matplotlib, with its Unicode syllable shown as the plot title. This step served multiple purposes: it provided visual confirmation of the image-label correspondence, facilitated inspection of the dataset's diversity, and helped build confidence in the integrity of the labeled dataset before it was used for model training.

4.8 Loading and Processing the Real Kannada Corpus

Since the original Char74K dataset lacked word-level annotations, a real-world Kannada corpus file named kn.txt was sourced from Kaggle to enable generation of synthetic word images. This file contained naturally occurring Kannada words and sentences. The corpus was read line by line using UTF-8 encoding, and non-Kannada or invalid characters were filtered out using regular expressions. Only clean, visually appropriate words were retained for synthesis. Although a subset of 500 words was used initially for demonstration and debugging, the same pipeline supports large-scale processing of over 100,000 words using the full corpus.

4.9 Mapping Corpus Words to Handwritten Images

To create realistic word-level handwriting samples from individual character images, each word from the cleaned corpus was decomposed into syllables using Unicode-aware segmentation logic. Each syllable was then mapped to its corresponding label ID using label_dict, and a handwritten image corresponding to that label was randomly selected from kan_img_cl_df. These character images were trimmed to remove unnecessary whitespace, resized to a fixed height (e.g., 32 pixels), and then horizontally concatenated to form complete word images. The resulting word images were saved to the kannada_ocr_images/ directory, and their corresponding Unicode labels were stored in a structured CSV file named kan_labels.csv.

4.10 Automating Synthetic Word Image Generation

To streamline and standardize the generation of word-level handwritten images, a modular and automated approach was implemented. This process encapsulated all major steps required for word synthesis, beginning with the segmentation of Kannada words into valid syllables using Unicode-aware logic. Each syllable was then mapped to a corresponding label identifier, and a representative character image was randomly selected from the cleaned dataset. These images were preprocessed through trimming and resizing operations to ensure uniform height (e.g., 32 pixels) while maintaining aspect ratio. The individual character images were then horizontally stitched together to form complete word images. Finally, the generated images were saved to a designated output folder, and their corresponding Unicode labels were recorded in a structured annotation file (kan_labels.csv). This automated pipeline was designed with scalability and reproducibility

in mind, allowing for efficient synthesis of large-scale word-level datasets with minimal manual intervention.

4.11 Generating the Final Dataset

The complete synthetic word dataset was generated by invoking the method processor.process_corpus(target_height=32), which processed real Kannada words into word-level handwritten images. Specifically, a curated set of 500 valid Kannada words from kn.txt was used to generate synthetic samples using the cleaned and labeled character images from Char74K. Each image maintained a fixed height of 32 pixels, with width varying based on word length. All synthesized images and their corresponding labels were saved in a structured format suitable for training OCR models. While this demonstration focused on 500 samples, the same pipeline was extended in subsequent experiments to synthesize over 100,000 word images, showcasing the flexibility and scalability of the approach. To illustrate the output of the final pipeline, Figure 4 displays a few representative synthetic word images generated from the curated Kannada corpus. These examples demonstrate how syllable-wise composition from handwritten character images can produce realistic and legible word-level inputs suitable for OCR training

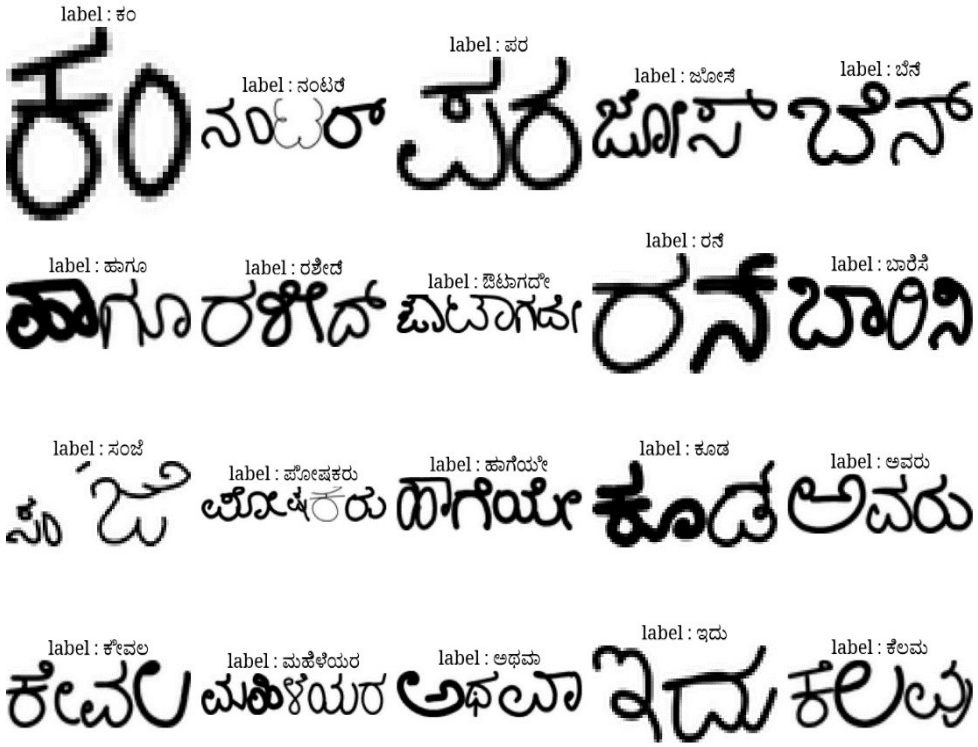


Figure 4. Sample of Sample synthetic handwritten word images generated using syllable-wise composition from the Char74K character dataset and the real Kannada corpus.

Though only 500 samples were generated for demonstration, the same pipeline was used in other experiments to synthesize over 100,000 word images.

5. Summary of OCR Relevance

Our pipeline bridges character-level and word-level datasets, enabling syllable-based modeling, realistic synthesis, and multilingual compatibility akin to IAM [2], LipiGnani [8], and modern vision-language OCR models such as SmolDocling [12]. Its key contributions include:

- Syllable-Level Labeling: By leveraging grapheme clusters (syllables) as labeling units, the approach captures the agglutinative and syllabic structure inherent to the Kannada script.
- Word-Level Supervision: Each synthetic image is paired with a corresponding Unicode label, enabling supervised learning for sequence-based OCR models.

- Realistic Synthesis: The combination of real handwritten characters from the Char74K dataset with natural words from a real-world corpus improves the linguistic and visual authenticity of the generated samples.
- Multilingual Compatibility: The format of output—image paired with UTF-8 encoded text—mirrors established datasets such as IAM (for English), making it readily adaptable for multilingual OCR pipelines.

This methodology facilitates the creation of a high-quality, scalable synthetic dataset tailored for Kannada handwritten OCR, addressing the data scarcity that affects low-resource Indic scripts.

6. Applications

The synthesized dataset and associated pipeline support a wide range of applications in the OCR and language technology domains:

- Training word-level Kannada handwritten OCR models
- Benchmarking performance of syllable-aware sequence learning architectures
- Pretraining and transfer learning across related Indic scripts
- Generating synthetic document lines for research in low-resource script processing

7. Limitations and Future Work

- Unicode-aware syllable segmentation may fail for rare orthographic constructs
- Future work includes sentence-level synthesis, noise augmentation, and human annotations

8. Conclusion

This work presents a scalable and modular pipeline for synthesizing word-level handwritten Kannada images by combining character-level samples from Char74K with real linguistic content from a Kannada text corpus. The methodology addresses a crucial gap in Indian language OCR research, particularly for low-resource scripts like Kannada. The generated dataset aligns with international standards in OCR datasets and paves the way for training and evaluating Kannada OCR systems that can match the performance of those in high-resource languages. The dataset and accompanying tools will be released publicly to support further research in the field.

References

1. Gongidi, S., Jawahar, C.V, IIIT-INDIC-HW-words: a dataset for Indic handwritten text recognition, In: Lladós, J., Lopresti, D., Uchida, S. (eds.) ICDAR Springer, Cham (2021), 2021. LNCS, Part IV, vol. 12824, pp. 444–459.
2. Marti, UV.,Bunke, H., The IAM-database: an English sentence database for offline handwriting recognition,IJDAR5, 39–46 (2002). https://doi.org/10.1007/s100320200071
3. T. E. de Campos, B. R. Babu, and M. Varma.,Character recognition in natural images.,In Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP), Lisbon, February 2009.
4. Harsh Lunia, AjoyMondal, and C. V. Jawahar. , 2024. ICPR 2024 Competition on Word Image Recognition from Indic Scene Images, In Pattern Recognition. Competitions: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XXXIV. Springer-Verlag, Berlin, Heidelberg, 120–133. https://doi.org/10.1007/978-3-031-80139-6_9
5. Roy, Partha&Mohta, Akash& Chaudhuri, Bidyut.,Synthetic data generation for Indic handwritten text recognition., 2018, 10.48550/arXiv.1804.06254.
6. Venkatesh, Manasij, VikasMajjagi, and DeepuVijayasenan. "Implicit segmentation of Kannada characters in offline handwriting recognition using hidden Markov models.,2014, arXiv preprint arXiv:1410.4341.
7. Prabhu, Vinay.,Kannada-MNIST: A new handwritten digits dataset for the Kannada language, 2019, 10.48550/arXiv.1908.01242.

8. Kumar, HR Shiva, and A. G. Ramakrishnan, LipiGnani: a versatile for documents in any language printed in Kannada script., ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)19, no. 4 (2020): 1-23.
9. M. A. Ferrer etal., MDIW-13: a New Multi-Lingual and Multi-Script Database and Benchmark for Script Identification, arXiv 2024.
10. A. Gupta, A. Vedaldi and A. Zisserman, Synthetic Data for Text Localisation in Natural Images, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2315-2324, doi: 10.1109/CVPR.2016.254.
11. Dutta, Kartik& Krishnan, Praveen & Mathew, Minesh &Jawahar C.V., Towards Spotting and Recognition of Handwritten Words in Indic Scripts, 2018, 32-37. 10.1109/ICFHR-2018.2018.00015.
12. Nassar, A., Marafioti, A., Omenetti, M., Lysak, M., Livathinos, N., Auer, C., Morin, L., de Lima, R.T., Kim, Y., Gurbuz, A.S. and Dolfi, M., SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion, 2025,arXiv preprint arXiv:2503.11576.
13. Google Fonts, Noto Serif Kannada, [Online]. Available: https://fonts.google.com/specimen/Noto+Serif+Kannada. [Accessed: Jul. 27, 2025].
14. Unicode Consortium, *The Unicode Standard, Version 15.0 – Kannada*, Unicode, Inc., 2022. [Online]. Available: https://www.unicode.org/charts/PDF/U0C80.pdf
15. W. Bright, Kannada and Telugu Writing, in *The World's Writing Systems*, P. T. Daniels and W. Bright, Eds., New York, NY: Oxford University Press, 1996, pp. 413–419.
16. Venkatesh, Manasij, VikasMajjagi, and DeepuVijayasenan., Implicit segmentation of Kannada characters in offline handwriting recognition using hidden Markov models.,arXiv preprint arXiv:1410.4341(2014).
17. A. Bharath and S. Madhvanath, Online Handwriting Recognition for Indic Scripts, in *Guide to OCR for Indic Scripts: Document Recognition and Retrieval, V. Govindaraju and S. Setlur, Eds., London, U.K.: Springer, 2009, pp. 209–234.
18. Ansary, Nazmuddoha, QuaziAdibur Rahman Adib, TahsinReasat, Asif ShahriyarSushmit, Ahmed ImtiazHumayun, SaziaMehnaz, KanijFatema, Mohammad Mamun Or Rashid, and FarigSadeque, Unicode Normalization and Grapheme Parsing of Indic Languages,arXiv preprint arXiv:2306.01743, 2023.