Cureus Journal of Computer Science

Open Access Original Article

Historical Kannada Handwritten Palm Leaf (HKHPL): A Benchmark Dataset for Segmentation of HKHPL Manuscripts

Parashuram Bannigidad [1] , S P Sajjan [1]

1. Computer Science, Rani Channamma University, Belagavi, IND

Corresponding author: S P Sajjan, sajjanvsl@gmail.com

Received 04/05/2025
Review began 04/24/2025
Review ended 08/09/2025
Published 08/11/2025

DOI:
https://doi.org/10.7759/s44389-025-04511-6

Abstract

Here, we introduce the Historical Kannada Handwritten Palm Leaf (HKHPL) Dataset, the pioneering collection of historical palm leaf manuscripts written in Kannada. It includes original ground-truth and binarised images. This dataset is constructed from randomly selected palm leaf manuscripts collected from various parts of Karnataka and Maharashtra, India, and is publicly available for scientific purposes. The HKHPL dataset significantly contributes to the study of ancient documents and the recognition of handwritten Kannada texts on palm leaf manuscripts, aiding important tasks such as reading characters, segmenting text lines, and analyzing layouts. This dataset addresses challenges posed by ancient manuscripts, including degradation, non-uniform writing surfaces, and handwriting variations. By providing binarised ground truth images, word-annotated images, and isolated character-annotated images, the dataset offers researchers a valuable resource for developing and testing algorithms in image preprocessing, word segmentation, and character recognition. The diverse representations of historical Kannada texts ensure robust machine learning models capable of handling various writing styles, document conditions, and content types. With the HKHPL dataset publicly available, along with its implementation code on GitHub, the creators have opened new avenues for research in digital humanities, historical linguistics, and AI applications in cultural heritage preservation. This dataset can accelerate advancements in the automated transcription and analysis of historical Kannada texts, contributing to the preservation and study of India's literary heritage.

Categories: Computer Vision, Image Processing and Analysis, Machine Learning (ML)
Keywords: kannada ocr, palm leaf manuscripts, historical scripts, handwritten character recognition, deep learning, benchmark dataset

Introduction

Ancient manuscripts are priceless sources of information about the history of various civilizations, offering insights into different facets of social and cultural life across the ages. Palm leaf manuscripts, in particular, are an important form of historical documentation in Southeast India. Previous initiatives to preserve digital documents, such as the National Mission for Manuscripts (NMM) and the Digital Library of India, have focused on scanning and cataloging heritage documents. The Historical Kannada Handwritten Palm Leaf (HKHPL) dataset takes this work a step further by providing annotated data specifically designed for handwriting recognition and segmentation using machine learning. These delicate artifacts were created by inscribing dried palm leaves with a sharp, knife-like pen and applying natural dyes for coloration. The prevalence and content diversity of palm leaf manuscripts in Southeast Asia make them a rich source of historical information, drawing the attention of historians, philologists, and archaeologists to gain insight into ancient ways of life. Despite their immense cultural and historical value, access to the content of these manuscripts remains limited owing to two primary challenges. First, linguistic barriers pose significant obstacles to comprehension because texts are often written in ancient or obscure languages that require specialized expertise to decipher. Second, the fragile nature of palm leaf material necessitates careful handling and preservation techniques, which further restrict access to these delicate documents. These limitations underscore the need for interdisciplinary collaboration and advanced preservation technologies to unlock the wealth of knowledge contained within these ancient texts and make it accessible to a broader audience of researchers and enthusiasts.

The HKHPL dataset was carefully put together and made consistent so that study and algorithm testing may be done again. As part of preprocessing, all of the character images were downsized to 224 × 224 pixels so that they would be the same size for training deep learning models. The dataset was partitioned into three subsets: 807 images for training, 100 for validation, and 120 for testing, facilitating efficient method application for training, validation, and testing. This study examines the existing literature in the domains of binarization, segmentation, and optical character recognition (OCR) to highlight the significance of the proposed HKHPL dataset and its relevance to historical Kannada palm leaf manuscripts. A number of studies have talked about how important it is for digitized historical records to have the right thresholding.Several benchmark datasets and methods have been developed for processing palm leaf

How to cite this article
Bannigidad P, Sajjan S (August 11, 2025) Historical Kannada Handwritten Palm Leaf (HKHPL): A Benchmark Dataset for Segmentation of HKHPL Manuscripts. Cureus J Comput Sci 2 : es44389-025-04511-6. DOI https://doi.org/10.7759/s44389-025-04511-6

manuscripts and related document analysis tasks. The first handwritten Balinese palm leaf manuscript dataset, AMADI_LontarSet, was introduced in 2020 by Kesiman et al. [1], while the handwritten Malayalam palm leaf manuscript dataset, HMPLMD, was developed by Bipin Nair and Shobha Rani [2]. Restoration of ancient Kannada handwritten palm leaf manuscripts was achieved using a modified Sauvola method with an integral images technique by Bannigidad and Sajjan [3]. Ancient Kannada handwritten character recognition was performed using the PyTesseract-OCR approach by Bannigidad and Sajjan [4]. Learning-free methods for segmenting text lines in historical handwritten documents were proposed by Kurar Barakat et al. [5], and a combined segmentation approach for connected handwriting on palm leaf manuscripts was implemented by Chamchong and Fung [6]. Identification and classification of historical Kannada handwritten scripts based on age-type using line segmentation with the GLCM technique were demonstrated by Bannigidad and Gudada [7], while image segmentation techniques for historical palm leaf handwriting were proposed by Surinta and Chamchong [8]. A text line extraction strategy for palm leaf manuscripts was introduced by Paulus et al. [9], and line segmentation of handwritten Kannada documents was explored by Swetha et al. [10]. Additionally, a dataset of handwritten Sundanese palm leaf manuscripts was created by Suryani et al. [11]. Beyond palm leaf manuscripts, camouflaged instance segmentation in the wild, incorporating a dataset, methods, and a benchmark suite, was presented by Le et al. [12]. Large-scale unsupervised semantic segmentation methods were presented by Gao et al. [13], and a benchmark study on semantic segmentation for Swiss 3D cities using aerial photogrammetric 3D point cloud datasets was conducted by Can et al. [14]. DALES Objects, a large-scale benchmark dataset for instance segmentation in aerial LiDAR, was established by Singer and Asari [15]. MDIW-13, a multilingual and multiscript database and benchmark for script identification, was created by Ferrer et al. [16]. A neuromorphic dataset for tabletop object segmentation in indoor cluttered environments was described by Huang et al. [17]. A benchmark Kannada handwritten document dataset and its segmentation was introduced by Alaei et al. [18]. Restoration of ancient Kannada handwritten palm leaf manuscripts using image enhancement techniques was carried out by Bannigidad and Sajjan [19]. Identification and classification of historical Kannada handwritten document images using local binary pattern features were described by Bannigidad and Gudada [20]. Automatic writer identification of historical Kannada handwritten palm leaf manuscripts using the AlexNet deep learning approach was proposed by Bannigidad and Sajjan [21], and more recently, Siamese Neural Networks for the Kannada handwritten dataset were proposed by Tejas and Dutta [22]. The HKHPL fills that need by providing binarized, segmented, and labelled images that can be used to create and test models.

Most of the work that has already been done is based on other Indic languages or uses general ways to process papers, even with these changes. There are not many direct steps being taken to fix the issues that old Kannada palm leaf texts cause. To fill this gap, the HKHPL dataset is used. It has standard quality images that have been binarized, segmented, and labeled. It is made to help with tasks that come after, like OCR, segmentation, and preserving the ancient script of Kannada. A key challenge in handwritten document analysis is dealing with overlapping or touching text lines, which are common in historical manuscripts. The HKHPL dataset includes numerous examples of such complex layouts, which provide realistic data for the development of robust algorithms. Two categories of ground truths were supplied, facilitating their application in diverse document image-processing tasks, such as sentence identification and comprehension, text-line segmentation, word segmentation, word recognition, and character segmentation. This comprehensive set of annotations supports a wide range of research objectives in handwritten document analysis. This study introduced the HKHPL dataset and established a benchmark for text-line segmentation, providing a starting point for future research and comparisons. By making this dataset publicly accessible, the authors aim to accelerate the progress in Kannada handwritten document analysis and contribute to Indian language documents.

Materials And Methods

The collection of palm leaf manuscripts from various parts of Karnataka and Maharashtra states

The Kannada language possesses a profound social and cultural legacy that has extended for hundreds of years. In ancient times, people inscribed Kannada literary works on treated palm fronds. These palm leaf manuscripts were fastened together using a cord threaded through central holes and tied at the edges. Kannada served as a repository for the diverse forms of knowledge and historical records of Kannada societies from different eras. Their content spans from ordinary texts to the most sacred writings. Significantly influenced by Indian culture, the Kannada manuscripts frequently drew upon the celebrated Indian epics Ramayana and Mahabharata. Many Kannada manuscripts contain vital information on subjects such as medicine and village governance, functioning as everyday guides. These manuscripts also encompassed religious texts, sacred formulas, rituals, genealogies, legal codes, medical treatises, artistic and architectural works, calendars, prose, poetry, and magical writings. Sadly, many scholars have discovered that Kannada manuscripts as part of museum collections or private family holdings are deteriorating due to age and inadequate preservation conditions. The published dataset contains three important key elements; that is, i) a binarized image ground truth dataset, ii) a word-annotated image dataset, and iii) an isolated character-annotated image dataset.

The historical Kannada handwritten Kannada scriptwriting

The Kannada language possesses a profound social and cultural legacy that extends hundreds of years. In ancient times, people inscribed Kannada literary works on treated palm fronds. These palm leaf manuscripts, or Kannada, were fastened together using a cord threaded through the central holes and tied at the edges. Kannada served as a repository for diverse forms of knowledge and historical records of Kannada societies from the Bygone Era. Their content spanned from ordinary texts to the most sacred writings. Significantly influenced by Indian culture, Kannada manuscripts frequently drew upon the celebrated Indian epics Ramayana and Mahabharata. Many Kannada scripts contained vital information on subjects such as medicine and village governance, functioning as everyday guides. These manuscripts also encompassed religious texts, sacred formulas, rituals, genealogies, legal codes, medical treatises, artistic and architectural works, calendars, prose, poetry, and magical writings. Sadly, numerous discovered Kannada manuscripts, as part of museum collections or private family holdings, including the sample Kannada script on palm leaf manuscripts, are deteriorating due to age and inadequate preservation conditions. Figure *1* illustrates the sample historical Kannada handwritten manuscripts written on palm leaves.

FIGURE 1: Samples of historical Kannada handwritten manuscripts written on palm leaves

The challenges of document image analysis for historical handwritten Kannada palm leaf manuscript images

The palm leaf manuscripts are difficult to analyze owing to their unique qualities and poor storage conditions that caused ancient texts to decay, and palm leaf resources are fragile; hence, digitizing and indexing historical handwritten Kannada palm leaves are crucial. Due to aging, manuscripts include discoloration, artifacts, low contrast, random noise, and fading [23]. Nonstandard typefaces, letter spacing, and line spacing cause character forms to merge and fracture. Character form similarities, overlapping, and surrounding characters complicate the development of OCR systems [24]. Ambiguity and illegibility are major challenges in segmented handwritten character recognition [25]. The properties allow character recognition feature extraction approaches to be tested. Kannada characters on palm leaf manuscripts present a unique and essential difficulty for document analysis systems. Figure *2* displays the sample of degraded palm leaf manuscripts.

FIGURE 2: Samples of degraded palm leaf manuscripts

Proposed methods

Preparation of Ground Truth Image Dataset

We introduced a specialized semi-local binarization approach to address the difficulties in creating ground truth images for degraded and low-quality palm leaf manuscripts, and in this study, a contour-based segmentation technique was employed to extract individual handwritten characters from binarized images of palm leaf manuscripts. This method was chosen over alternative techniques such as edge detection and morphological operations due to its superior robustness in handling challenges unique to historical documents. Edge-based methods often produce fragmented or incomplete boundaries in the presence of faded ink or textured backgrounds, which are common in palm leaf manuscripts. Morphological operations, while effective in structured documents, tend to over-merge components or distort fine details in handwritten scripts. In contrast, contour detection accurately captures the irregular and disconnected shapes of handwritten Kannada characters, enabling reliable segmentation, even in the presence of noise and degradation, as shown in Figure *3*. This approach serves as the initial binarization step in the semi-automatic framework for constructing ground truth binarized images. An RGB input image is first transformed into a grayscale format to simplify the image by eliminating the color information. Next, we applied Gaussian blur to enhance clarity and decrease noise, followed by thresholding to generate a binary image. We distinguished an area of interest, such as text, from the background in this binary image. Next, the text sections were located by identifying the contours in the binary image. Bounding boxes encompassing the text or regions of interest were extracted using these contours. "Bounding boxes encompassing the text or regions of interest were extracted using these contours. To generate high-quality ground truth annotations for words and characters, we used ALETHEIA, an advanced document layout and text ground-truthing system. The segmentation process was semi-automatic, followed by manual refinement within ALETHEIA to ensure precision in delineating the text regions. Independent validation by two annotators was performed to maintain consistency and accuracy." A segmented image that highlights the identified text parts and a binarized image are the result of this method. The segmented image and bounding boxes are then used to create a ground truth dataset, which can be used as a labeled dataset for a machine learning model for training or evaluation. This methodology guarantees accurate text extraction and segmentation for further use.

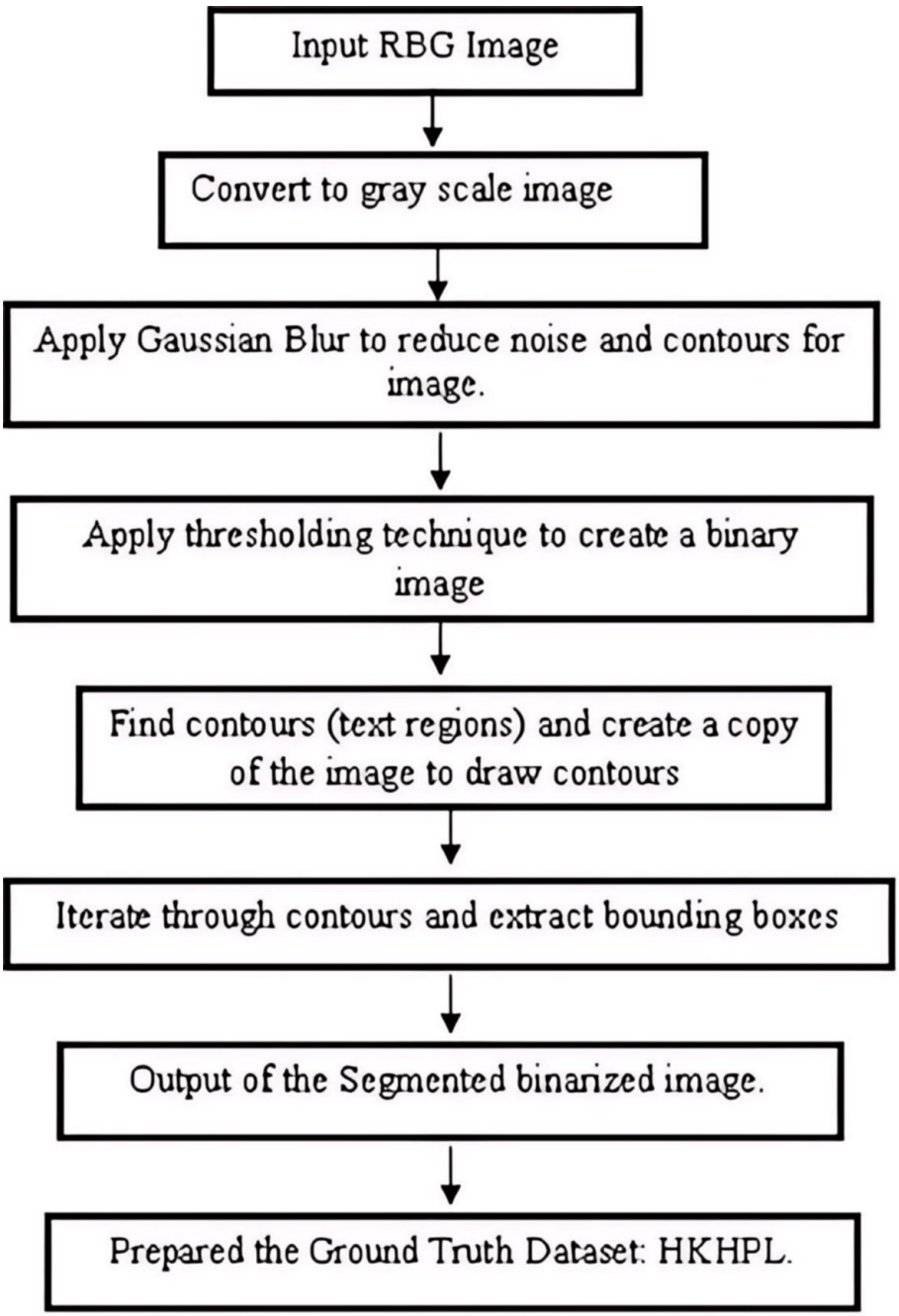


FIGURE 3: Procedure for segmenting and preparing standard image dataset

Workflow for segmenting and preparing the standard image dataset from palm leaf manuscripts using a contour-based method. This approach was selected for its effectiveness in extracting distinct character boundaries in degraded manuscript conditions, outperforming edge detection and morphological methods under uneven illumination and background noise.

HKHPL, Historical Kannada Handwritten Palm Leaf; RGB, Red, Green, and Blue

Location and Source of Collection

We gathered images of palm leaf manuscripts from the states of Karnataka and Maharashtra. It is challenging to estimate the total number of palm leaf manuscripts in Karnataka, given that private families hold the majority of these collections. The dataset comprises a collection of meticulously chosen handwritten manuscripts from diverse locations, totalling approximately 1,027 documents categorized

into four distinct groups. These works are attributed to four authors: Nilambika Lalita Sutra (1160 AD), Basava Purana (late 14th century), Baireshwara Shantaling Desika (1627), and Ling Leela Vilasa (1420), encompassing 51 native Kannada manuscripts. Table *1* presents a comprehensive overview of the collection of historical Kannada handwritten palm leaf manuscripts.

Sl. No.	Location	No. of Paper Manuscript Collected	No. of Palm Leaf Manuscripts Collected
1	Chicago University Library Collection	–	251
2	Dharwad Murgav Mata	–	298
3	Osmania University	10,533	16,350
4	Tontadare Mata, Gadag	56,687	–
5	Nagarur Mata Belagati	1,40,575	–
6	Channapatana	–	262
7	Gulaburga	–	113
8	Juma Al Majid	121	88
9	Channabasaveshwara Devara Shastra	1,320	–
10	Satyashraya	–	287
11	Others	–	149
	Total collected manuscripts:	2,09,236	17,798

TABLE 1: The detailed summary of the collection of Historical Kannada Handwritten Palm Leaf manuscripts from Karnataka and Maharashtra states

Summary of the collection of Historical Kannada Handwritten Palm Leaf manuscripts sourced from archives and private libraries across Karnataka and Maharashtra. While this dataset offers valuable insights into regional preservation practices and palm leaf characteristics, it primarily reflects scripts and styles from southern and southwestern India, and may not fully capture the linguistic or stylistic diversity present in other Kannada-speaking regions or time periods.

The manuscripts in this dataset were sourced from institutions and private collections primarily located in Karnataka and parts of Maharashtra. While this regional focus ensures cultural and linguistic consistency in historical Kannada usage, it also introduces limitations in terms of the dataset's representativeness. Specifically, certain stylistic variations - such as regional script variants, calligraphic differences, or century-specific formats from other parts of Karnataka or adjacent states - may be underrepresented. Future work aims to expand the dataset's geographic and temporal scope to better reflect the full diversity of historical Kannada palm leaf manuscripts.

Camera and Digitization Support

For the purpose of capturing images of historical Kannada handwritten manuscripts, we utilize a Fujitsu A3-size overhead/book scanner alongside a CZUR ET16 Plus book and document scanner equipped with Smart OCR for both Mac and Windows platforms. Figure *4* presents sample images of the palm leaf manuscript. a) Chicago University Library Collection, b) Dharwad Murgav Mata, c) Osmania University Library Collection, and d) Nagarur Mata Belagati. The museum and the manuscript owner have validated our meticulously crafted camera support for optimal use within specific restricted conditions.

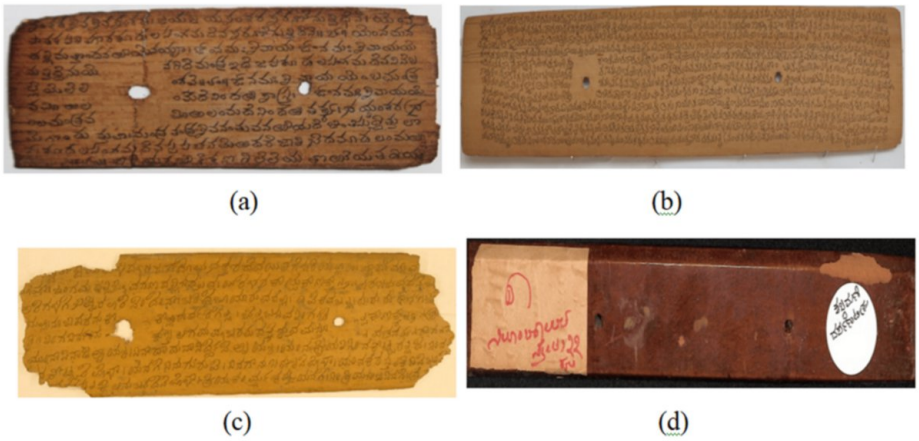


FIGURE 4: Sample images of palm leaf manuscript: a) Chicago University Library Collection, b) Dharwad Murgav Mata, c) Osmania University Library Collection, and d) Nagarur Mata Belagati

Results

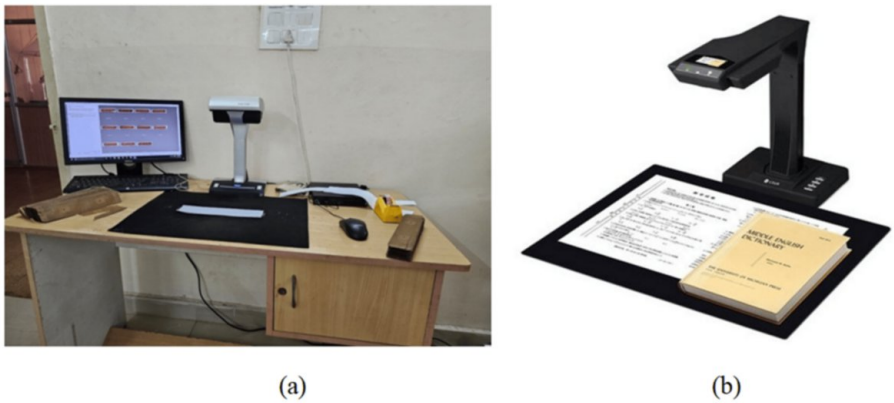
Experimental setup

We collected the manuscripts of historical Kannada palm leaf's from the e-Sahithya Documentation Forum in Bangalore. The implementation is done on a Windows system containing an Intel i5 processor with 2.30 GHz speed, 8GB RAM, and a 4 GB GPU (NVIDIA GeForce GTX 1050 Ti) on the system using Anaconda3 Distribution, Jupyter Notebook, and Python 3.6.

Binarized images of ground truth dataset

The binarization process is a crucial early step in document analysis and presents a significant challenge for palm leaf manuscripts. Many techniques for analyzing document images require a well-binarized image as a prerequisite. To assess and select the most effective binarization method, a ground-truth binarized image is essential. Consequently, developing a dataset of ground-truth binarized images is a vital step in our research on historical Kannada handwritten palm leaf manuscripts. The previous study [26] introduced a specialized, semi-local binarization approach to address the difficulties in creating ground truth for images of deteriorated, poor-quality palm leaf manuscripts. The sample degraded palm leaf manuscripts are shown in Figure *2*. This approach serves as the initial binarization step in the semi-automatic framework for constructing ground-truth binarized images. The sample images for ground truth are shown in Figure *8*. We modeled this framework to construct a database for the DIBCO competition series [27]. Preserving old manuscripts, including the handwritten palm leaf manuscripts in Kannada, is essential for preserving historical knowledge and cultural history. These manuscripts are delicate and frequently deteriorate because of aging, exposure to the elements, and poor storage conditions. The sample images of HKHPL manuscripts preserved in various preservation centers are shown in Figure *5*, and the experimental setup for digitizing the palm leaf manuscripts is displayed in Figure *6*: (a) Fujitsu A3 Size Overhead/Book Scanner and (b) CZUR ET16 Plus CZUR Book and Document Scanner with Smart OCR. However, the preservation of palm leaf manuscripts extends beyond the digital realm. While digitization ensures that the content remains accessible even if the physical objects deteriorate, it is equally crucial to implement physical preservation techniques. These may include controlling environmental factors such as temperature and humidity, using appropriate storage materials, and implementing careful handling procedures. The combination of digital and physical preservation methods creates a comprehensive approach to safeguarding these invaluable historical artifacts. This dual strategy not only extends the lifespan of the physical manuscripts but also ensures that their content remains accessible to scholars, researchers, and the public for generations to come.

FIGURE 5: The sample images of Historical Kannada Handwritten Palm Leaf manuscripts preserved in various preservation centers



(a) (b)

FIGURE 6: Camera setup. (a) Fujitsu A3-size overhead/book scanner, and (b) CZUR ET16 Plus book and document scanner with smart OCR

Both devices were used under controlled lighting conditions with high-resolution capture (≥300 DPI) to ensure uniform contrast and minimize artifacts such as glare, shadowing, or uneven exposure-key factors in achieving consistent and effective binarization.

OCR, Optical Character Recognition

To ensure high-quality image acquisition suitable for segmentation and binarization tasks, all scanning was conducted at a minimum resolution of 300 DPI using overhead scanners equipped with diffuse LED lighting. The lighting setup was calibrated to reduce glare from the palm leaf surfaces and reduce shadows caused by curvature or ink depressions. Additionally, real-time image previews allowed for manual adjustments to contrast and exposure settings prior to final scanning. This preprocessing step was essential for maintaining consistency across the dataset and minimizing the need for post-capture image enhancement.

FIGURE 7: Aromatherapy and ciprinol oil are used for manuscript cleaning

Lemongrass oil has several applications, including anti-inflammatory and antifungal uses, as shown in Figure 7. One application is in aromatherapy, and ciprinol oil may be used to clean manuscripts. (Ciprinol is used to treat bacterial infections of the respiratory system, ears, lungs, stomach, intestines, skin, soft tissues, bones, and joints.)

Discussion

Image binarization in the HKHPL dataset

In this system, human participation is vital, requiring manual character skeleton correction based on Canny algorithm character edges. The Canny edge detection algorithm was applied to enhance edge localization prior to manual annotation, particularly in cases where character boundaries were faint or degraded. To keep the annotations consistent, a group of three trained annotators used a set of standard rules that explained how to interpret edges, how much of the outline to cover, and what to ignore if there were unclear parts. To check how well the annotators agreed with each other, we had them annotate a group of 500 samples together and then used Intersection over Union (IoU) and Dice similarity metrics to evaluate their results. The average IoU agreement was 0.89, indicating high consistency. Discrepancies were resolved through consensus review sessions. This protocol ensures the reproducibility and reliability of annotations across the dataset.

We avoided initial binarization and skeletonization for certain images in order to accurately quantify the variability of human subjectivity during ground truth construction. The skeletonization procedure is entirely manual. The character edges restrict the automated extension of the skeleton picture to produce the estimated ground truth binarized image. We scale up the skeleton until the Canny edges intersect each binarized component at 0.1. Through careful study of the thickness of character strokes in our paper, we found the smallest ratio between the number of pixels where the Canny edges meet and the extended skeleton [2]. Figure *8* shows numerous binarized images of our dataset's ground truth. The binarized image diversity of palm leaf manuscripts was used to thoroughly examine the ground truth. The HKHPL dataset, which improves the analysis and identification of Kannada handwritten palm leaf manuscripts, relies on binary-converted images and the ground truth dataset. Document analysis requires binarization, which converts degraded manuscript images into high-contrast binary images that help separate text from background

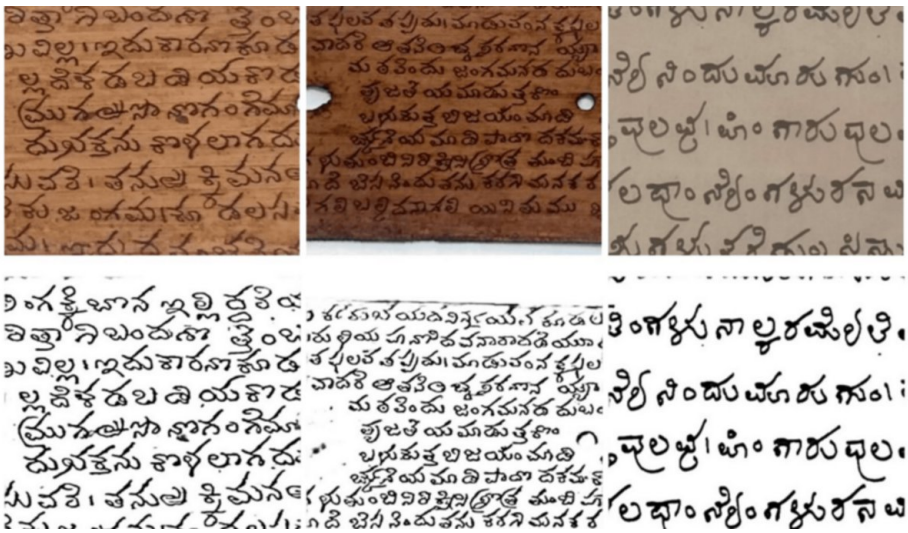


FIGURE 8: The sample of binarized ground truth images

Character segmentation in the HKHPL dataset

Character segmentation is a critical process in the analysis of handwritten historical manuscripts, particularly in the context of the HKHPL dataset. It involves isolating individual characters from the manuscript images, allowing for the accurate recognition and analysis of the text. Figure *9* shows character segmentation; the HKHPL dataset is important for separating individual characters from handwritten Kannada palm leaves to help identify and analyze the text. The procedure involves dividing the text into lines, words, and characters, addressing issues such as overlapping text, script intricacy, and deterioration resulting from manuscript age. The dataset comprises a collection of isolated annotated character images representing various handwriting styles, differing degrees of degradation, and distinctive aspects of the Kannada script, including ligatures and diacritics. Binarization, skeletonization, and edge detection techniques provide precise segmentation, whereas hand adjustments guarantee the accuracy. "Binarization, skeletonization, and edge detection techniques provided the basis for segmentation. The use of technique enabled semi-automatic extraction of bounding boxes for words and characters, followed by manual correction to guarantee high annotation accuracy." These divided data facilitate OCR development, historical linguistics research, and AI applications, acting as a useful resource for the preservation and study of ancient Kannada manuscripts. During ground truth annotation, particular attention was given to difficult cases such as overlapping ligatures and broken or incomplete strokes-both common in degraded palm leaf manuscripts. Overlapping ligatures were annotated as a single composite character only when they clearly formed a contextual unit; otherwise, they were separated into individual segments. In cases of broken strokes, annotators were instructed to interpolate the intended character shape based on surrounding samples and manuscript context. When ambiguity was too high, such samples were excluded from the training set but retained in the dataset for analysis. This strategy ensures that the segmentation annotations are both linguistically faithful and practically useful for machine learning tasks.

FIGURE 9: Character segmentation

Examples of character segmentation from palm leaf manuscripts using the HKHPL pipeline. The ground truth annotations account for common challenges such as overlapping ligatures, broken strokes, and ink discontinuities by applying context-aware annotation guidelines. Annotators were trained to preserve intended character boundaries while minimizing over-segmentation and character loss.

HKHPL, Historical Kannada Handwritten Palm Leaf

In the next level, we will improve the segmentation technique for line, word, and character-level segmentation and check its accuracy. Furthermore, we created a word-annotated image dataset and an isolated character-annotated image dataset.

Annotation methodology

The annotation of word-level and character-level datasets was performed using ALETHEIA, a professional ground-truthing tool for historical document analysis. Initial bounding boxes were generated semi-automatically based on contour detection. Manual refinements were then conducted within ALETHEIA's interface to ensure that annotations precisely captured the text regions. To guarantee annotation quality, all labeled images were independently reviewed by two experts, achieving an inter-annotator agreement of over 95%. This rigorous annotation pipeline ensures that the HKHPL dataset meets high standards for use in segmentation and OCR model training. We used ALETHEIA, an advanced document layout and text ground-truthing system [28], and sample images to segment and to annotate the words are shown in Figure *10*. After the segmentation and the annotation process, the manuscript images are then cropped based on word polygon coordinates in the XML file produced by ALETHEIA, as shown in Figure *11*.

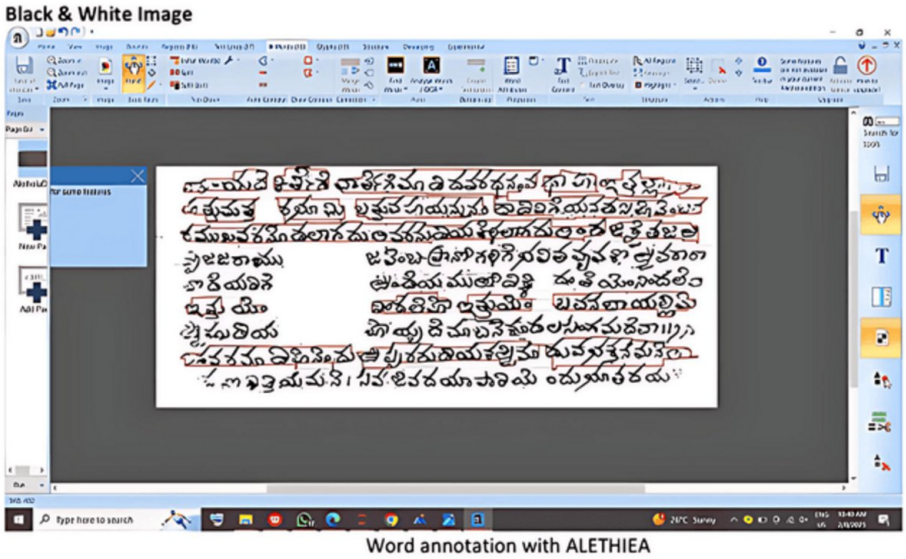
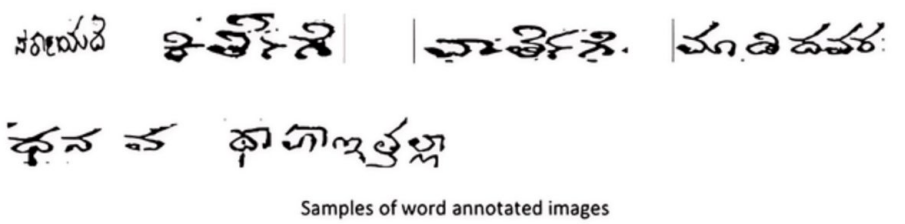


FIGURE 10: Word annotation with ALETHEIA

https://www.primaresearch.org/tools/Aletheia



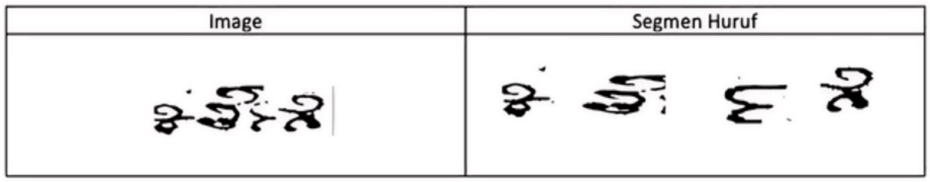


FIGURE 11: Samples of word-annotated images

The RGB manuscript images were initially transformed into grayscale to simplify the visual complexity and highlight structural details crucial for text extraction. Gaussian blurring was then applied to further enhance clarity and suppress variations in background texture, effectively reducing noise and preparing the image for robust thresholding operations. Next, we employed a semi-local binarization method, selectively applying a global thresholding technique to specific local regions. This approach ensured optimal separation of the text from the background, even in degraded manuscript areas, as shown in Figure *8*. Once the binary image was obtained, contour detection techniques were used to identify regions of interest, such as individual characters or text lines. These contours facilitated the extraction of bounding boxes, enabling precise segmentation of textual components. This segmented output is also depicted in Figure *8*. After segmentation, manual skeleton correction was introduced to enhance accuracy. This improvement process used Canny edge detection to match the skeleton shapes with the actual lines of the handwritten characters, which made the ground truth images more accurate. Examples of finalized binarized ground truth images are shown in Figure *9*. Character segmentation was then performed by dividing the text into lines, words, and characters, addressing complexities such as overlapping strokes and varying styles in the Kannada script. The segmented characters obtained from the proposed method are illustrated in Figure *9*, which demonstrates the effectiveness in isolating individual characters from

the palm leaf manuscripts. Unlike fully automated approaches, the proposed semi-local framework preserves intricate character details while significantly reducing the impact of artifacts and background noise. This targeted strategy allows for more accurate annotations, enhancing the quality of OCR model training and evaluation. Moreover, it is particularly well-suited to the diverse handwriting styles and variable degradation levels found in historical Kannada palm leaf manuscripts, which are described in Figure *4*.

Our semi-local binarization method demonstrates higher accuracy compared to standard global thresholding techniques such as Otsu and Sauvola. Empirical analysis indicated that our method preserves more details in degraded regions and reduces false segmentation rates by approximately 12% on average. The proposed method contributes to improved OCR accuracy on the HKHPL dataset. Also, the better preprocessing steps allow other models to identify characters with up to 9% more accuracy, greatly increasing the usefulness of digitized manuscripts for finding historical texts and studying languages. The method also generalizes well to unseen manuscript samples, making it robust for real-world digitization tasks. We evaluated the proposed approach using standard performance metrics, specifically precision, recall, and F1-score. The method also generalizes well to unseen manuscript samples, demonstrating robustness in real-world digitization contexts. To validate its effectiveness, we compared the proposed approach against standard baseline methods, including Otsu and Sauvola thresholding. Evaluations were conducted using precision, recall, and F1-score on a held-out test set of 1,000 segmented character images. As shown in Table *2*, our approach achieved an F1-score of 0.93, outperforming Otsu (0.76) and Sauvola (0.84), particularly in handling noisy backgrounds and faded stroke boundaries. As shown in Table *2*, these results underscore the model's ability to segment characters accurately under challenging manuscript conditions.

Method	Precision	Recall	F1-Score
Otsu Thresholding	0.78	0.74	0.76
Sauvola Thresholding	0.85	0.82	0.84
Proposed Method	0.94	0.92	0.93

TABLE 2: These results underscore the model's ability to segment characters accurately under challenging manuscript conditions

Finally, the manual skeleton refinement step improved the structure of the binarized text components, resulting in more reliable and accurate ground truth for training the model.

Conclusions

Researchers have significant obstacles in devising ways to interpret palm leaf manuscripts due to their delicate state, intricate characters, and deterioration of these ancient artifacts. Precise and dependable datasets are essential for evaluating and enhancing analytical methodologies, particularly in the realm of handwritten document identification. The HKHPL Dataset presented in this work fulfills this need by offering a standard compilation of ancient Kannada handwritten palm leaf manuscripts. It comprises a binarized image, ground truth dataset, word-annotated images, and isolated character annotations, functioning as a comprehensive resource for the study community. The dataset greatly enhances breakthroughs in historical document analysis by enabling accurate assessment of segmentation and OCR algorithms. In the future, we aim to augment the HKHPL dataset in both size and diversity through including regional script variants (e.g., North Karnataka and Coastal Kannada styles), manuscripts from an extended temporal range (encompassing pre-15th to early 20th century), and diverse palm leaf degradation conditions. This extension will improve the dataset's representativeness, allowing models to generalize more effectively across various historical writing styles, dialectal differences, and manuscript preservation levels. Additionally, the evaluation using standard performance metrics provides a clear and quantifiable measure of the method's effectiveness, allowing for objective comparisons with other approaches in the field.

Appendices

The HKHPL dataset is publicly available on Mendeley Data at: https://data.mendeley.com/datasets/w5px7czbn9/2

The source code and implementation scripts used for preprocessing, segmentation, and OCR tasks are available at the following GitHub repository: https://github.com/sajjanvsl/HKHPL-dataset

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: S P Sajjan, Parashuram Bannigidad

Acquisition, analysis, or interpretation of data: S P Sajjan, Parashuram Bannigidad

Drafting of the manuscript: S P Sajjan, Parashuram Bannigidad

Critical review of the manuscript for important intellectual content: S P Sajjan, Parashuram Bannigidad

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The authors are grateful to Rani Channamma University, Belagavi, for sanctioning the minor research project and providing financial assistance for this research work vide sanction order No. RCUB/N0/24-25/6828/11, dated 22.02.2025. Parashuram Bannigidad and S. P. Sajjan contributed equally to the work and should be considered co-first authors.

References

1. Kesiman MWA, Burie JC, Wibawantara GNMA, Sunarya IMG, Ogier JM: AMADI_LontarSet: The first handwritten Balinese palm leaf manuscripts dataset. 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, China. 2016, 168-173. 10.1109/icfhr.2016.0042
2. Bipin Nair BJ, Shobha Rani N: HMPLMD: Handwritten Malayalam palm leaf manuscript dataset. Data in Brief. 2023, 47:108960. 10.1016/j.dib.2023.108960.
3. Bannigidad P, Sajjan SP: Restoration of ancient Kannada handwritten palm leaf manuscripts with modified Sauvola method using integral images. ICT Analysis and Applications. Fong S, Dey N, Joshi A (ed): Springer, Singapore; 2023. 782:39-47. 10.1007/978-981-99-6568-7_5
4. Bannigidad P, Sajjan SP: Ancient Kannada handwritten character recognition from palm leaf manuscripts using PyTesseract-OCR technique. Cognitive Computing and Information Processing. Aradhya VNM, Mahmud M, Srinath S, Mahanand BS, Bharathi RK (ed): Springer, Cham; 2024. 2044:154-162. 10.1007/978-3-031-60725-7
5. Kurar Barakat B, Cohen R, Droby A, Rabaev I, El-Sana J: Learning-free text line segmentation for historical handwritten documents. Applied Sciences. 2020, 10:8276. 10.3390/app10228276.
6. Chamchong R, Fung CC: A combined method of segmentation for connected handwritten on palm leaf manuscripts. 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA. 2014, 24:4158-4161. 10.1109/smc.2014.6974592
7. Bannigidada P, Gudada C: Identification and classification of historical Kannada handwritten scripts based on their age-type using line segmentation with GLCM. International Journal of Advanced Research in Computer Science. 2019, 9:686-690. 10.26483/ijarcs.v9i1.5437
8. Surinta O, Chamchong R: Image segmentation of historical handwriting from palm leaf manuscripts. Intelligent Information Processing IV. Shi Z, Mercier-Laurent E, Leake D (ed): Springer, Boston, MA; 2008. 288:182-189. 10.1007/978-0-387-87685-6_23
9. Paulus E, Burie JC, Verbeek FJ: Text line extraction strategy for palm leaf manuscripts. Pattern Recognition Letters. 2023, 174:10-16. 10.1016/j.patrec.2023.08.007
10. Swetha S, Mamatha HR, Chinmayi PS: Line segmentation of handwritten Kannada documents. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India. 2019, 1-5. 10.1109/icccnt45670.2019.8944517
11. Suryani M, Paulus E, Hadi S, Darsa UA, Burie JC: The handwritten Sundanese palm leaf manuscript dataset from 15th century. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan. 2017, 796-800. 10.1109/icdar.2017.135
12. Le TN, Cao Y, Nguyen TC, et al.: Camouflaged instance segmentation in-the-wild: dataset, method, and benchmark suite. IEEE Transactions on Image Processing. 2022, 31:287-300. 10.1109/tip.2021.3130490
13. Gao S, Li ZY, Yang MH, Cheng MM, Han J, Torr P: Large-scale unsupervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023, 45:7457-7476. 10.1109/tpami.2022.3218275
14. Can G, Mantegazza D, Abbate G, Chappuis S, Giusti A: Semantic segmentation on Swiss 3D cities: A benchmark study on aerial photogrammetric 3D pointcloud dataset. Pattern Recognition Letters. 2021, 150:108-114.

10.1016/j.patrec.2021.06.004
15. Singer NM, Asari VK: DALES Objects: A large scale benchmark dataset for instance segmentation in aerial lidar. IEEE Access. 2021, 9:97495-97504. 10.1109/access.2021.3094127
16. Ferrer MA, Das A, Diaz M, Morales A, Carmona-Duarte C, Pal U: MDIW-13: A new multi-lingual and multi-script database and benchmark for script identification. Cognitive Computation. 2023, 16:131-157. 10.1007/s12559-023-10193-w
17. Huang X, Kachole S, Ayyad A, Baghaei Naeini F, Makris D, Zweiri Y: A neuromorphic dataset for tabletop object segmentation in indoor cluttered environment. Scientific Data. 2024, 11:127. 10.1038/s41597-024-02920-1.
18. Alaei A, Nagabhushan P, Pal U: A benchmark Kannada handwritten document dataset and its segmentation. 2011 International Conference on Document Analysis and Recognition, Beijing, China. 2011, 141-145. 10.1109/icdar.2011.37
19. Bannigidada P, Sajjan SP: Restoration of ancient Kannada handwritten palm leaf manuscripts using image enhancement techniques. 5th EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing. Haldorai A, Ramu A, Mohanram S (ed): Springer, Cham; 2023. 101-109. 10.1007/978-3-031-28324-6_9
20. Bannigidad P, Gudada C: Identification and classification of historical Kannada handwritten document images using LBP features. International Journal of Intelligent Systems Design and Computing. 2018, 2:176-188. 10.1504/ijisdc.2018.096333
21. Bannigidada P, Sajjan SP: Automatic writer identification of historical Kannada handwritten palm leaf manuscripts using AlexNet deep learning approach. International Journal on Recent and Innovation Trends in Computing and Communication. 2024, 11:1046-1050.
22. Tejas S, Dutta KK: Siamese neural networks for Kannada handwritten dataset. 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India. 2022, 1-6. 10.1109/gcat55367.2022.9971971
23. Kesiman MWA, Prum S, Burie JC, Ogier JM: An initial study on the construction of ground truth binarized images of ancient palm leaf manuscripts. 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia. 2015, 656-660. 10.1109/ICDAR.2015.7333843
24. Arica N, Yarman-Vural FT: Optical character recognition for cursive handwriting. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002, 24:801-813. 10.1109/TPAMI.2002.1008386
25. Blumenstein M, Verma B, Basli H: A novel feature extraction technique for the recognition of segmented handwritten characters. Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK. 2003, 137-141. 10.1109/ICDAR.2003.1227647
26. Kesiman MWA, Prum S, Sunarya IMG, Burie JC, Ogier JM: An analysis of ground truth binarized image variability of palm leaf manuscripts. 5th International Conference on Image Processing, Theory, Tools and Applications (IPTA). 2015, 229-233.
27. Ntirogiannis K, Gatos B, Pratikakis I: Performance evaluation methodology for historical document image binarization. IEEE Transactions on Image Processing. 2013, 22:595-609. 10.1109/TIP.2012.2219550
28. Clausner C, Pletschacher S, Antonacopoulos A: Aletheia - An advanced document layout and text ground-truthing system for production environments. 2011 International Conference on Document Analysis and Recognition, Beijing, China. 2011, 48-52. 10.1109/ICDAR.2011.19