INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Advancements in Segmentation of Unconstrained Malayalam Handwritten Documents f or Enhanced OCR

A T Anju[1]*, Binu P Chacko[2], K P Mohamed Basheer[3]

1 Research Scholar, Department of Computer Science, Sullamussalam Science College, Calicut University, Malappuram, Kerala, India
2 Associate Professor, Department of Computer Science, Prajyoti Niketan College, Calicut University, Puthukkad, Kerala, India
3 Associate professor, Department of Computer Science, Sullamussalam Science College, Calicut University, Malappuram, Kerala, India

OPEN ACCESS

Received: 17-07-2024
Accepted: 30-08-2024
Published: 18-09-2024

Citation: Anju AT, Chacko BP, Basheer KPM (2024) Advancements in Segmentation of Unconstrained Malayalam Handwritten Documents f or Enhanced OCR. Indian Journal of Science and Technology 17(36): 3763-3775. https://doi.org/10.17485/IJST/v17i36.2220

* **Corresponding author**.

anjuat89@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop an effective method for segmenting handwritten document images into individual words by focusing on Malayalam language documents and also to aid in creating a benchmark dataset. **Method:** This study adapts an approach that integrates Canny edge detection, morphological operations, and contour analysis. The dataset used comprises 500 unconstrained Malayalam handwritten scanned document images, specifically police's First Information Statement (FIS) collected from the Department of Kerala Police. Parameters considered include grayscale conversion, noise removal techniques, edge detection, thresholding (Otsu method), dilation, and contour extraction for word segmentation. The proposed method was compared with gold standards by manually segmenting the documents and calculating accuracy, precision, recall, and F1 score, yielding superior performance metrics across various instances. **Findings:** The method achieved accuracies of 95.08% for Document 1, 91.67% for Document 2, and 100% for Document 3, resulting in an average accuracy of 95.58%. Additionally, the method yielded an average Precision of 100%, an average Recall of 95.58%, and an average F1 Score of 97.72%. **Novelty:** This study addresses the scarcity of Malayalam handwritten word datasets by introducing a robust segmentation method that is not only effective for Malayalam but also adaptable to other scripts. The novelty of this work lies in its comprehensive approach to segmentation, validated through experiments on a preliminary subset of the dataset. While the dataset is still in the creation phase, the current work provides significant insights and methodologies that will contribute to the final benchmark dataset. The benchmark dataset will be made publicly available upon completion, ensuring that this research lays the foundation for future studies in handwritten document analysis.

Keywords: Handwritten Document Segmentation; OCR; Malayalam; Canny Edge Detection; Morphological Operations; Contour Analysis; Benchmark Dataset

1 Introduction

Segmenting Malayalam handwritten documents presents unique challenges due to the complex spatial arrangement of characters, diverse character set, and varying individual writing styles. Despite advances in segmentation techniques across various languages, a high-quality dataset with ground truth annotations in the Malayalam language remains unavailable, significantly impeding progress in this area. Existing methods have predominantly focused on text line and word segmentation, often utilizing projection histograms. For example, a method employing horizontal and vertical projection histograms based on midpoint detection achieved notable accuracies for Meitei Mayek and English handwritten documents [(1)]. However, this approach's reliance on projection histograms limits its effectiveness for documents with intricate layouts and variable spacing.

An unsupervised statistical method for segmenting handwritten and degraded machine-printed documents used gap distribution to frame segmentation as a two-class classification problem [(2)]. Despite its robustness across various datasets, this method's simplicity may not handle highly complex or irregular document structures effectively. Other techniques, such as a binary quadratic process combined with structured support vector machines for word segmentation, demonstrated state-of-the-art performance on public datasets [(3)]. Nonetheless, these methods require substantial computational resources and may struggle with documents featuring extreme variability in handwriting styles.

Combining horizontal projection profiles with seam carving for line and word segmentation showed promise on English and Bangla datasets but proved computationally intensive due to dynamic programming and energy matrices [(4)]. For ancient handwritten documents in Devanagari and Maithili scripts, segmenting text lines into words and characters by dividing text lines into horizontal zones achieved high accuracies [(5)]. However, evaluations on self-generated datasets may not generalize well to other scripts or diverse handwriting styles. In Gujarati, conventional methods like morphological operations and connected components analysis were explored for word segmentation. While results were encouraging, traditional methods might not scale effectively with increasing document complexity and variability [(6)].

A novel encoder-decoder deep model for historical Arabic manuscripts involved text segmentation followed by line and word segmentation using smoothing and Chamfer distance. However, this model's performance heavily relies on the availability and quality of training data, which may be inadequate for other languages [(7)]. Density-based clustering for word segmentation in printed and handwritten documents, including cursive handwriting, demonstrated reasonable effectiveness. However, the approach's success is contingent on preprocessing steps, which may not be robust for all types of document degradation and handwriting styles [(8)]. Gaussian filters for line and pseudo-word segmentation in historical documents achieved excellent results compared to other methods but struggled with non-uniform text arrangements [(9)]. Techniques addressing touching and crossing words through junction branch association showed promise but were complex and specifically focused on crossing words, limiting their applicability to generalized segmentation tasks [(10)].

Previous work on text line segmentation utilized morphological operations, contour extraction, and Hough line transform [(11)]. Building on this foundation, the current study advances the field by integrating these morphological techniques specifically for word segmentation, addressing the unique challenges of handwritten Malayalam documents. Unlike previous machine learning-based approaches, which require extensive training data often unavailable for low-resource languages, the proposed method combines Canny edge detection, morphological operations, and contour analysis. This

combination enhances segmentation accuracy and robustness across diverse document layouts. However, it is important to note that the method does not introduce a novel segmentation technique but rather optimizes existing methods to suit the specific requirements of Malayalam handwritten documents.

Contributions

- **Optimized Segmentation Approach:** The study combines well-established techniques, such as Canny edge detection, morphological operations, and contour analysis, to enhance the segmentation of Malayalam handwritten documents, addressing the challenges posed by the script's complex and diverse nature.
- **Addressing Dataset Scarcity:** By focusing on optimizing segmentation for Malayalam documents, this work aims to facilitate the creation of a high-quality benchmark dataset, which remains a significant gap in the current research landscape.
- **Scalability and Applicability:** The combined approach is designed to be adaptable to other scripts with similar complexities, offering a scalable solution that could be applied beyond Malayalam.

2 Methodology

Previous research often followed a sequential approach where line segmentation was performed prior to word segmentation. In contrast, the proposed method eliminates the need for line segmentation before word segmentation, showcasing its script-independent nature and applicability across various languages.

2.1. Data Collection and Description

The proposed method was evaluated using a dataset of unconstrained Malayalam handwritten scanned document images, specifically First Information Statements (FIS) obtained from the Kerala Police Department in Kozhikode District, Kerala, India. This dataset comprises 500 PNG-format images, featuring diverse sizes, line counts, and word counts per line. These images were selected to represent a broad spectrum of handwriting styles, ensuring that the dataset captures the inherent variability of real-world Malayalam handwritten documents. The dataset will be made publicly available upon project completion to support further research in this field. Figure 1. Shows the workflow of the proposed method.

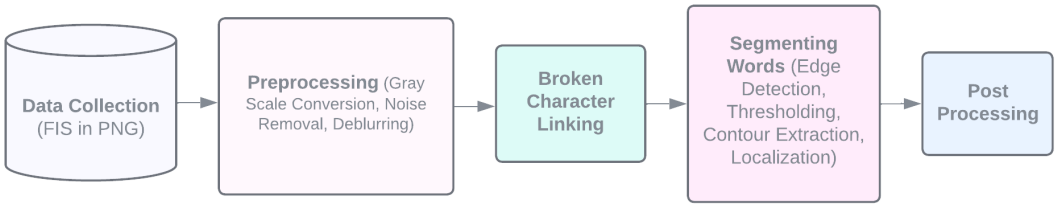

Fig 1. Work Flow of the Proposed Method

2.2. Input and Output Variables

Input Variables:

- **Image Data:** Scanned document images of handwritten Malayalam text in PNG format
- **Preprocessing Parameters:** Grayscale conversion, noise removal techniques (morphological operations, median filtering), and Gaussian filter application.
- **Segmentation Parameters:** Parameters for edge detection, thresholding (Otsu method), dilation, contour extraction, and bounding box localization.

Output Variables:

- **Segmented Words:** The primary output is the set of word segments, each localized within a bounding box in the original document image.

2.3. Data Preprocessing

Preprocessing is a crucial step to ensure high-quality segmentation results in the proposed methodology. The preprocessing pipeline includes the following steps:

2.3.1. Grayscale Conversion

The preprocessing begins with converting the PNG format image to grayscale. This transformation reduces the image from a three-dimensional color space to a one-dimensional intensity space. Grayscale conversion minimizes computational complexity and reduces intensity variance due to different writing instruments and colors. This variance reduction is vital for effective binarization, which is crucial for further image analysis. In mathematical terms, the grayscale image is represented as a matrix where each element denotes the intensity value of a pixel. This matrix format simplifies mathematical operations and enhances alignment with various algorithms. Additionally, grayscale conversion reduces information entropy by condensing the image to a single intensity channel, thereby improving the efficiency of subsequent processing steps.

2.3.2. Noise Removal

Noise removal is performed to eliminate insignificant parts of the image, improving output quality. Morphological operations, such as dilation and erosion, are employed for this purpose. Dilation expands handwriting strokes to preserve structural details, as described by the grayscale dilation Equation (1).

$$(f \oplus b)(x) = \sup_{y \in E} [f(y) + b(x - y)] \tag{1}$$

where f is the image, b is the structuring function, E is euclidean space.

Erosion, using the same kernel size, removes insignificant parts and dark spots from the image. The grayscale erosion is defined by Equation (2).

$$(f \ominus b)(x) = \inf_{y \in B} [f(x + y) - b(y)] \tag{2}$$

where f is the image, b is the structuring element, and B is the space that b is defined.

2.3.3. Deblurring

Following noise removal, median filtering is applied to address blurring from isolated noise while preserving sharp edges of the handwriting. Morphological operations are used to link broken characters, which may have been fragmented during noise removal.

2.4. Broken Character Linking

The challenge of broken edges significantly impacts pattern recognition accuracy[12]. Addressing this issue, previous research, such as[13], introduced a swarm intelligence-based approach using the Ant Colony Optimization (ACO) algorithm. This method leverages the foraging and self-organizing behaviors of ant colonies to address discrete optimization problems. A distance-based edge linking technique was employed to connect the nearest edge segments, utilizing a binary edge image map (BEIM) generated by the Canny edge detection algorithm.

In this study, the issue of broken character linking, which often arises due to noise removal and ink inconsistencies in handwritten documents, is systematically addressed using morphological operations. Specifically, we employ the morphological closing operation, a fundamental technique in image processing that is particularly effective for connecting disjointed components in binary images. The closing operation is mathematically defined as the dilation of an image followed by the erosion of the result.

Let $I(x, y)$ be the binary image, and B be the structuring element. The dilation $I \oplus B$ expands the boundaries of the foreground regions by the shape of B effectively filling small gaps and holes. Subsequently, the erosion $(I \oplus B) \ominus B$ reduces the boundaries of the expanded regions, restoring the original size while maintaining the newly filled connections. The closing operation $I.B$ is then expressed as $I.B = (I \oplus B) \ominus B$.

This operation is particularly relevant in our context as it effectively links fragmented character components by closing gaps between broken segments. By restoring continuity in the handwritten text, the closing operation enhances the robustness of subsequent segmentation and recognition processes, ensuring that characters are accurately represented in the processed images. This approach is crucial for dealing with the inherent challenges of Malayalam script, where characters are often complex and prone to fragmentation due to inconsistencies in writing and ink distribution.

2.5. Word Segmentation

The process of word segmentation from the image involves several key steps: edge detection, thresholding, morphological operations, contour extraction, and localization.

2.5.1. Edge Detection

Edge detection is a critical preprocessing step in image analysis, aimed at identifying significant discontinuities in intensity within the input image. These discontinuities correspond to the edges, which represent the boundaries of objects or features within the image. By detecting these edges, the algorithm reduces the volume of data to be processed in subsequent stages while preserving essential structural information.

For Malayalam handwritten documents, edge detection is particularly effective in maintaining the integrity of the script's intricate details. Prior to the application of the edge detection algorithm, a Gaussian filter is applied to the image to smooth the data and reduce noise, which can otherwise lead to spurious edges. This filtering process enhances the edge quality by attenuating high-frequency components, thereby improving the accuracy and reliability of the subsequent edge detection.

Once the image is smoothed, an edge detection algorithm, such as the Canny edge detector, is employed. Which involves several mathematical steps to identify significant edges in the image. Initially, the gradient of the image intensity function $I(x,y)$ is computed using convolution with gradient filters, yielding G_x and G_y for the x and y direction where $G_x = \partial I/\partial x$ and $G_y = \partial I/\partial y$. The gradient magnitude G and direction θ are then calculated as $G = \sqrt{G_x^2 + G_y^2}$ and $\theta = arctan(Gy/Gx)$.

Non-maximum suppression thins the edges by retaining only local maxima in G along θ. Double thresholding uses high $T_H$and low $T_L$thresholds to classify pixels as strong edges, weak edges, or non-edges. Finally, edge tracking by hysteresis connects weak edges to strong edges if they are contiguous, ensuring coherent edge detection for accurate segmentation.

2.5.2. Thresholding

After detecting edges, binarization is applied to clearly distinguish text from the background. The Otsu method is used for this purpose. It automatically determines the best threshold value T to separate the image into two classes: text and background.

Mathematically, Otsu's method calculates the optimal threshold by minimizing the within-class variance σ_w^2, which measures how much the pixel values within each class vary. The method finds a threshold T^* that best separates the two classes, resulting in a binary image where pixels are either part of the text (value 1) or the background (value 0).

In the context of handwritten document segmentation, employing the Otsu method is crucial. It ensures a clear separation between text and background, enhancing the accuracy of subsequent steps such as contour extraction and word localization. By improving the binarization process, the Otsu method helps in better delineating text regions, which is essential for effective segmentation and recognition of handwritten Malayalam documents.

2.5.3. Morphological Dilation

Morphological dilation is a crucial image processing technique used to enhance the connectivity of word components in a binarized image. Specifically, horizontal dilation is applied to bridge gaps between character segments within the same line. This operation uses a structuring element to expand the foreground pixels (text) of the binary image. Mathematically dilation is defined as $(A \oplus B)(x,y) = \max_{(b_x, b_y \in B}$ A $\left(x - b_x, y - b_y\right)$, where A represents the binary image, B is the structuring element, and (x,y) are the coordinates of the pixel in the output image. The operation expands each foreground pixel of A by the shape of B, connecting nearby components.

In the context of handwritten document segmentation, applying horizontal dilation is pivotal. It helps in linking disconnected word segments that may have been separated due to noise or variations in handwriting. By enhancing the connectivity between characters in a line, dilation facilitates more accurate contour extraction and word localization. This preprocessing step is essential for improving the effectiveness of subsequent segmentation stages, ensuring that words are correctly identified and isolated for further analysis.

2.5.4. Contour Extraction and Localization

In the process of isolating word segments from a handwritten document, contour extraction is employed to identify potential word boundaries within the dilated image. The contours are detected by applying a contour-finding algorithm to the binary image, which results from prior dilation operations designed to enhance connectivity between word components.

Mathematically, a contour can be defined as a curve joining all the continuous points along the boundary, having the same intensity or color. If $I(x,y)$ denotes the intensity function of the image, a contour C is a set of points satisfying the equation

$C = \{(x, y) : I(x, y) = constant\}$

Once contours are identified, each contour is encapsulated within a minimal bounding box, which serves to localize the potential word segment. The bounding box B for a given contour C is defined by the minimum and maximum coordinates of the contour in both the x and y directions is $B = \left[\min_{(x,y)\in C} x, \max_{(x,y)\in C} x, \min_{(x,y)\in C} y, \max_{(x,y)\in C} y\right]$.

The bounding box B provides a rectangular enclosure that isolates each word from the rest of the document. This step is crucial for ensuring that subsequent processing stages, such as feature extraction and recognition, can operate on accurately localized word segments. The precise extraction of contours and their corresponding bounding boxes is vital in preserving the structural integrity of the handwritten text, thereby enhancing the overall accuracy of the segmentation process.

2.5.5. Post Processing

Area-based filtering refines the segmentation results. This involves calculating the area of each contour and removing segments that fall below a predefined threshold, based on empirical analysis, to eliminate noise.

3 Results and Discussion

3.1. Experimental Results

For the experiment, Malayalam handwritten document image were collected from Kerala Police Department, representing unconstrained real-world scenarios. Firstly, the input image was converted to grayscale image. Here, cvtColor() method in OpenCV was used to convert the color image to grayscale. Figure 2 shows the image after grayscale conversion. In our experiments, for getting a better edge detection result, a median filter was applied on the image to reduce the noises through smoothing. The kernel size of the median filter was estimated through a trial and error method, and here for the handwritten scanned images, a filter of kernel size 3 was found to be sufficient and applied on the image for smoothing. In this case, smoothing reduced the noise in a significant level and it preserved the sharp features of the handwritten images. Figure 3 shows the blurred image after median filtering.

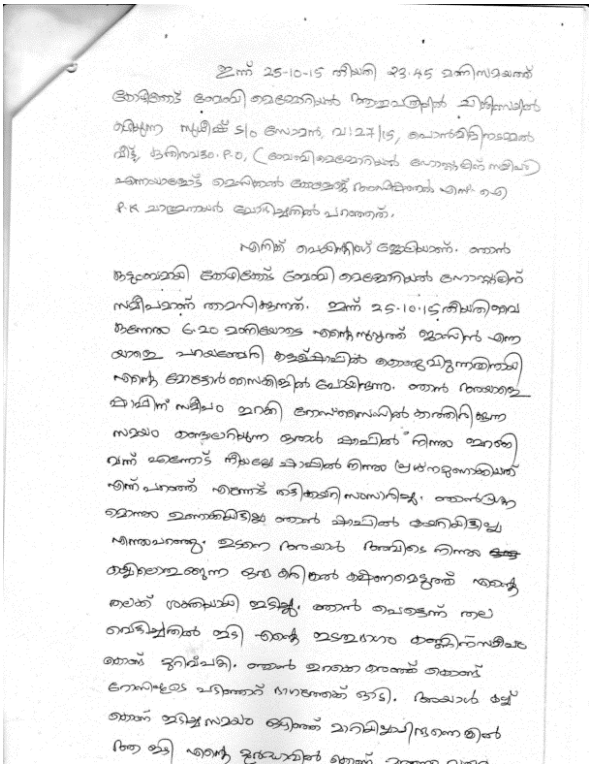

Fig 2. Gray Scale Image

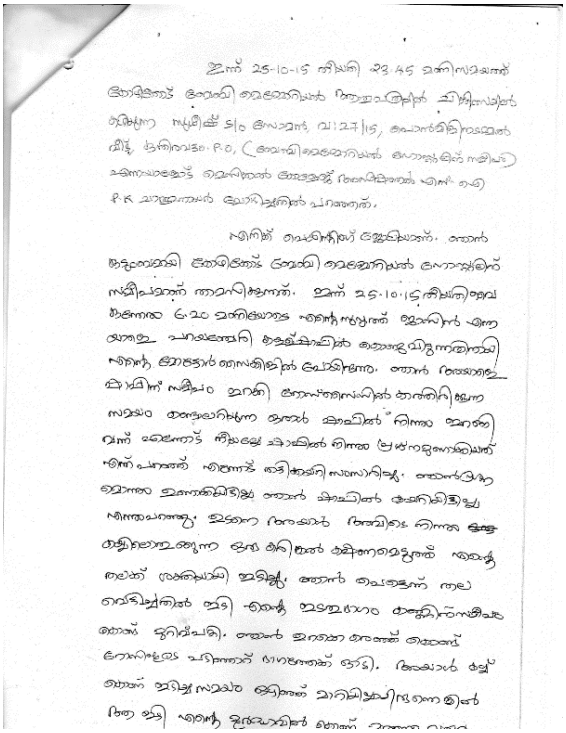

Fig 3. Blurred Image

After the application of the Gaussian filter, in this method, edge detection was achieved. And again for the estimation of the kernel size of the Gaussian filter, a trial and error method was equipped and found that the kernel size of (3, 3) is sufficient in the case of handwritten images. The Equation (3) denotes the Gaussion filter of kernel size (2k+1)×(2k+1).

$$H_{ij} = \frac{1}{2\pi\sigma^2} exp\left(-\frac{(i-(k+1))^2+(j-(k+1))^2}{2\sigma^2}\right); 1 \le i,j \le (2k+1) \tag{3}$$

For the edge detection of the handwritten image, we experimented some edge detection techniques like Sobel, Prewitt, Laplacian and Canny edge detection. Among the other edge detection techniques canny edge detector performed well even though the computational complexity was a little higher.

In the proposed method, binarization was really difficult due to the varying nature of intensities found in the input image. As a result, we carried out binarization after edge detection and it produced a good result. By using trial and error method, we concluded that, using Otsu method, most of the images were binarized correctly. The morphological operation, dilation was performed on the input image to dilate the handwritten words horizontally to make them connected for better contour extraction. The kernel size was set to (1, 30) to dilate the words to make them connected, and a gap was there in between the words in each line.

Then the shape of each word was extracted equipping contour extraction by using the findContours() function in OpenCV. The bounding box method was applied on the image for localization. Here, the rectangle coordinates of each contour were estimated by applying the boundingRect() function in OpenCV2 and the Figure 7 shows the localization result. Finally, by the exploitation of a postprocessing method, insignificant segments were removed from the segmented output image. Figure 8 shows the final output of the word segmentation.

While considering the image from Figure 8, the proposed method is providing 116 segmented words from that document image; however, 122 segments are there when we segment it manually. The accuracy obtained in this case is 95.08%. There is no restriction on the number of lines and words present in the input image. Therefore, for another two instances, the proposed method provided 22 and 123 word segments; nevertheless, 24 and 123 were the actual word segments obtained in these cases while performing manual segmentation. The obtained accuracies are 91.67% and 100% for the other two instances and the results are shown in Table 1.

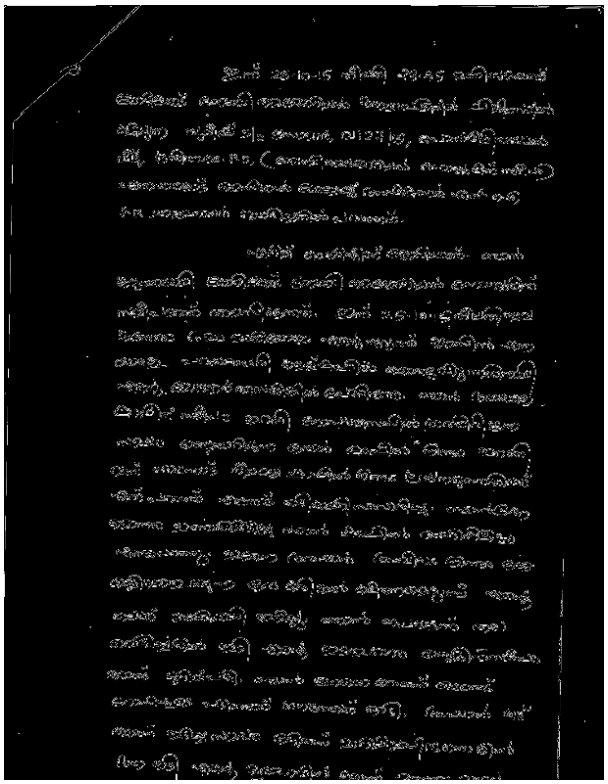

Fig 4. Edge Detected Image

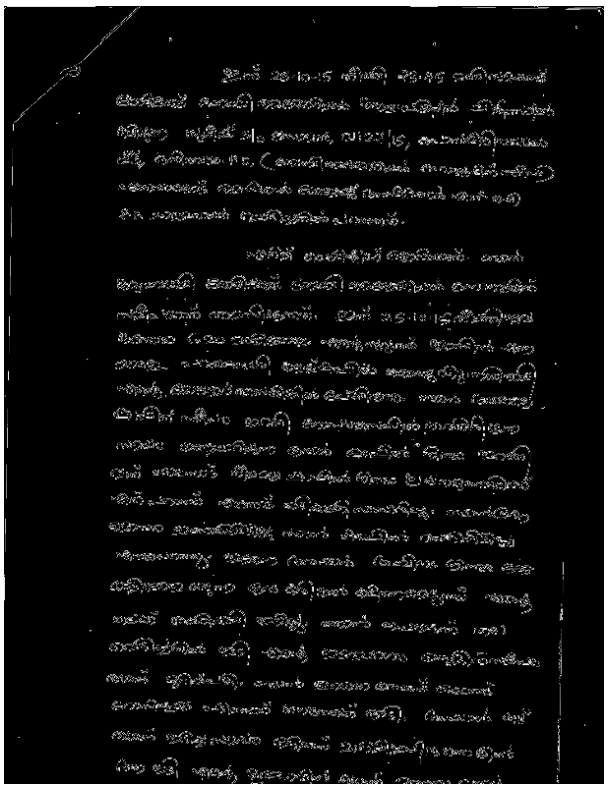

Fig 5. Image after Thresholding

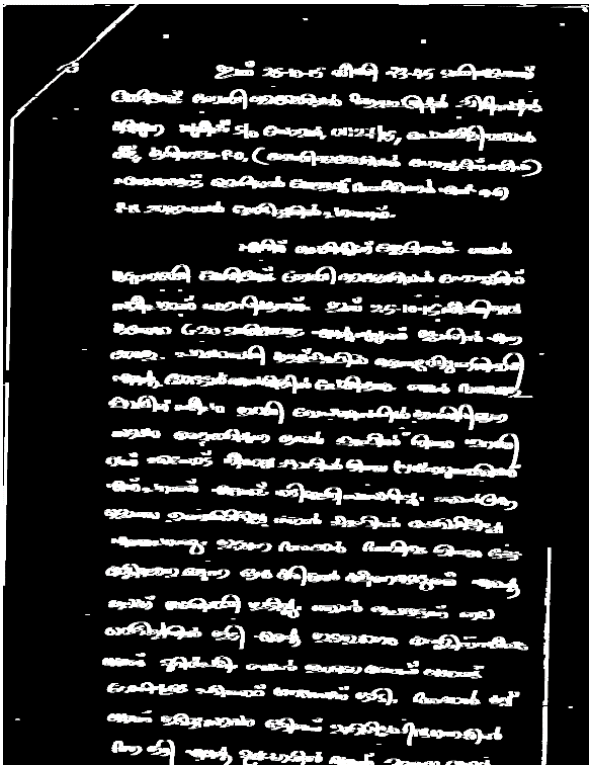

Fig 6. Image after dilation

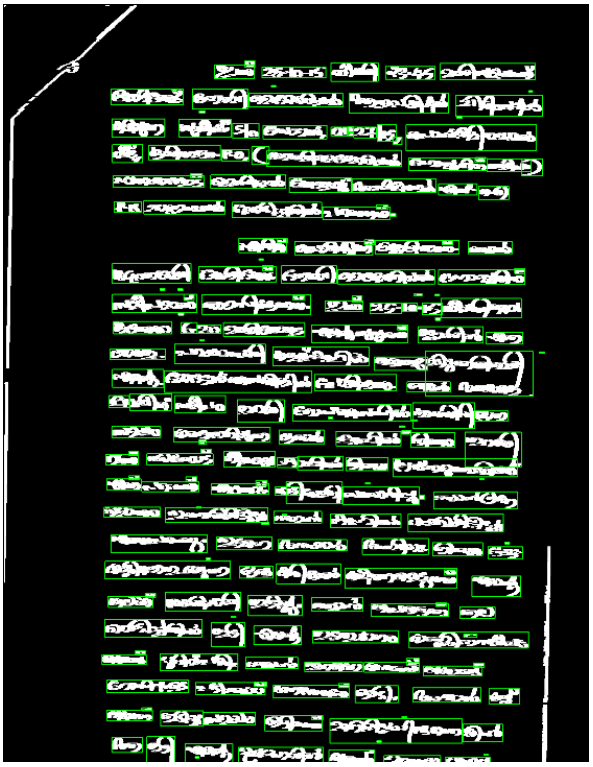

Fig 7. Imageafter Localization

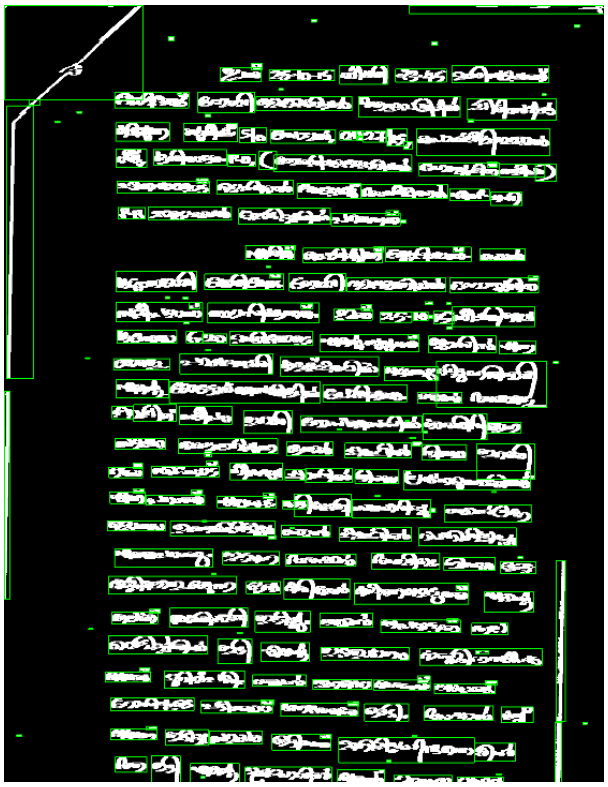

Fig 8. Image after Post processing

Table 1. Malayalam word segmentation experimental results

Instances	Obtained Result (No.of Words)	Actual Result (No.of Words)	Accuracy
Document 1	116	122	95.08%
Document 2	22	24	91.67%
Document 3	123	123	100%
	Average Accuracy		95.58%

To further evaluate the effectiveness of the proposed method, we used Precision, Recall, and F1 Score as the primary performance metrics. These metrics are defined as follows:

- **Precision:** Precision measures the accuracy of the positive predictions made by the model. It is calculated using Equation (4).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

- **Recall:** Recall, also known as Sensitivity or True Positive Rate, quantifies the model's ability to identify all relevant instances. It is defined by Equation (5).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

- **F1 Score:** The F1 Score provides a balance between Precision and Recall by calculating their harmonic mean. It is expressed using Equation (6).

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

where, TP represents the number of True Positives (correctly identified words), FP represents the number of False Positives (incorrectly identified words), FN represents the number of False Negatives (words that should have been identified but were not). Results are shown in Table 2.

Table 2. Precision, Recall, and F1 Score

Document	True Positives (TP)	False Positives (FP)	False Negatives (FN)	Precision	Recall	F1 Score
Document 1	116	0	6	1.0	0.9508	0.975
Document 2	22	0	2	1.0	0.9167	0.9565
Document 3	123	0	0	1.0	1.0	1.0
		Average		1.0	0.9558	0.9772

Due to the unavailability of ground truth data in low-resource languages like Malayalam, a comparison of the proposed method was conducted against several state-of-the-art image processing-based segmentation techniques. These included Connected Component Analysis, morphological operations, and the Projection Profile method. Among these, Connected Component Analysis and morphological operations yielded better results compared to the Projection Profile method. However, the segmentation accuracy of these techniques remains significantly inferior to that of the proposed method. The results obtained through Connected Component Analysis and morphological operations are illustrated in Figure 9, highlighting the limitations of these traditional approaches in handling the complexities of handwritten Malayalam documents.

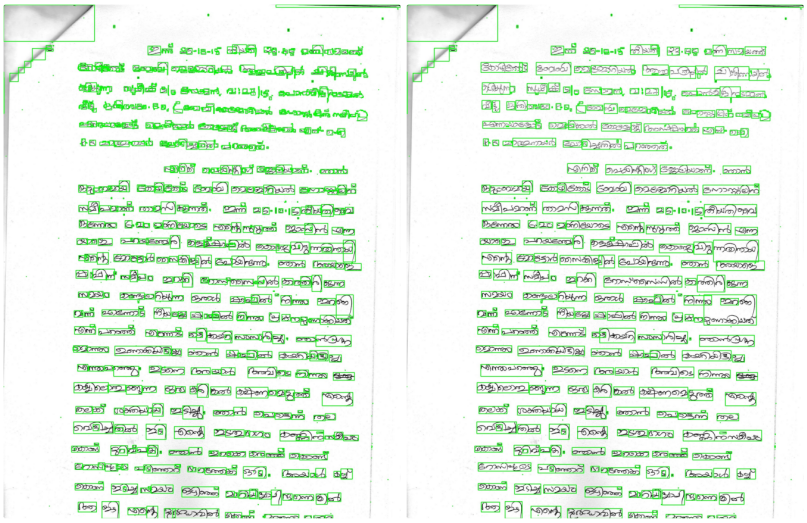

Fig 9. Results obtained through Connected Component Analysis and morphological operations

3.2. Comparison with Other Scripts

The performance appraisal of the method on different scripts was evaluated by considering 8 other Indian scripts taken from a public handwritten document image dataset PHDIndic_11 [14]. Figure 9 shows the handwritten word segmentation result.

Table 2 shows the experimental results of word segmentation of other Indian scripts. The results have shown that the method is ideal for other scripts as well.

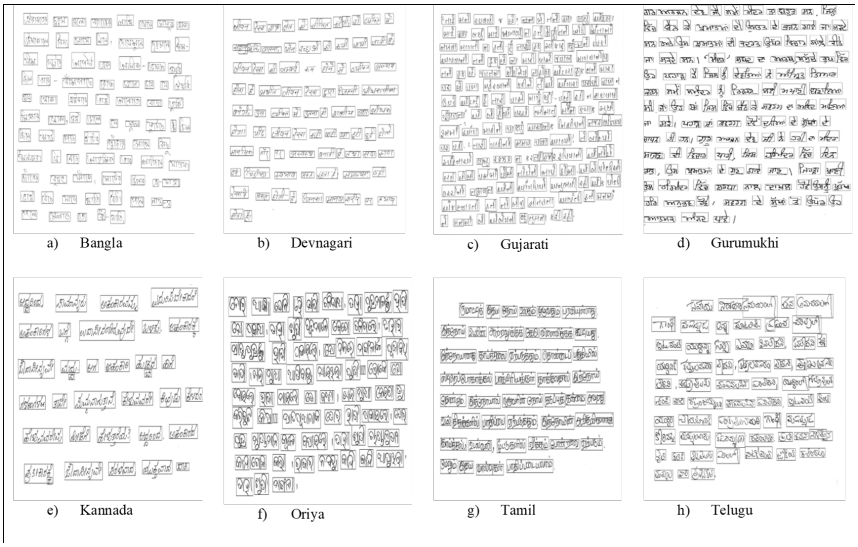

Fig 10. Segmentation results ofother Indian Scripts

Table 3. Other Scripts word segmentation experimental results

Script	Obtained Result (No. of Words)	Actual Result (No. of Words)	Accuracy
Bangla	80	79	98.75%
Devnagari	83	84	98.80%
Gujarati	165	159	96.36%
Gurumukhi	125	129	96.89%
Kannada	31	31	100%
Oriya	66	66	100%
Tamil	44	44	100%
Telugu	61	59	96.72%

3.3. Comparison with Previous Work

In this study, a comparison is made with the work of Bhatia et al. [6], which focuses on word segmentation in Gujarati handwritten documents. This study was selected for comparison due to its use of a similar image processing approach, aligning with the non-machine learning methodology of the current study. The method in [6] demonstrated varied accuracy based on document type, with high accuracy for well-aligned documents and reduced accuracy for those with uneven spacing or misalignment. Specifically, it achieved up to 97% accuracy for 3-length words and 100% for 4-length words in well-aligned documents, but accuracy dropped to as low as 23% for single-length words in documents with spacing issues. In contrast, the proposed method consistently delivered higher accuracy across different document types, with an average precision of 100% and an overall average accuracy of 97.72%. This performance underscores the robustness of the proposed method in managing document variations effectively.

4 Conclusion

This study developed a robust, script-independent method for word segmentation of handwritten Malayalam documents using conventional techniques due to the lack of training data. This method, integrating canny edge detection, morphological operations, and contour analysis, achieved an accuracy of 98.95% for Malayalam and high accuracy rates across multiple scripts: 98.75% for Bangla, 98.80% for Devnagari, 96.36% for Gujarati, 96.89% for Gurumukhi, 100% for Kannada, 100% for Oriya, 100% for Tamil, and 96.72% for Telugu. The novel outcomes include a flexible, efficient framework for word segmentation, reducing manual annotation efforts. Despite its effectiveness, our method's reliance on conventional techniques may struggle with extremely complex or degraded documents. The absence of ground truth data for training and testing impacts the evaluation of the segmentation method, as there is no benchmark to fully validate its performance. This limitation raises concerns about the generalizability of the results to other datasets or real-world scenarios. Future work could involve creating a benchmark labeled dataset to apply deep learning techniques, enhancing accuracy and robustness, and extending the method to character segmentation and comprehensive document recognition.

References

1) Sanasam I, Choudhary P, Singh KM. Line and word segmentation of handwritten text document by mid-point detection and gap trailing. *Multimed Tools Appl*. 2020;79(41):30135–50. Available from: https://doi.org/10.1007/s11042-020-09416-1.
2) Mukherjee J, Parui SK, Roy U. An unsupervised and robust line and word segmentation method for handwritten and degraded printed document. *Trans Asian Low-Resour Lang Inf Process*. 2021;21(2):1–31. Available from: https://doi.org/10.1145/3474118.
3) Sharma MK, Dhaka VS. Segmentation of handwritten words using structured support vector machine. *Pattern Anal Appl*. 2020;23(3):1355–67. Available from: https://doi.org/10.1007/s10044-019-00843-x.
4) Das M, Panda M. Seam carving, horizontal projection profile and contour tracing for line and word segmentation of language independent handwritten documents. *Results Eng*. 2023;18:101110. Available from: https://doi.org/10.1016/j.rineng.2023.101110.
5) Jindal A, Ghosh R. Word and character segmentation in ancient handwritten documents in Devanagari and Maithili scripts using horizontal zoning. *Expert Syst Appl*. 2023;225:120127. Available from: https://doi.org/10.1016/j.eswa.2023.120127.
6) Bhatia D, Goswami MM, Mitra S. Word Segmentation for Gujarati Handwritten Documents. In: Intelligent Information and Database Systems: 13th Asian Conference, ACIIDS 2021;vol. 13. Springer International Publishing. 2021;p. 569–80. Available from: https://doi.org/10.1007/978-3-030-73280-6_45.
7) Elharrouss O, Al-Maadeed S, Jm, Hassaine A. A robust method for text, line, and word segmentation for historical Arabic manuscripts. *Data Analytics for Cultural Heritage: Current Trends and Concepts*. 2021;p. 147–72. Available from: https://doi.org/10.1007/978-3-030-66777-1_7.
8) Guo H, Ding Q. A Framework for Word Segmentation in Images using Density-based Clustering. *CATA*. 2020;p. 187–96. Available from: https://hguo5.github.io/files/CATA-20-Segmentation.pdf.
9) Makhfi NE. Handwritten text segmentation approach in historical Arabic documents. In: Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019. 2020;p. 645–54. Available from: https://doi.org/10.1007/978-981-15-0947-6_61.
10) Omayio EO. Word segmentation by component tracing and association (CTA) technique. *J Eng Res*. 2022. Available from: https://doi.org/10.36909/jer.15207.
11) Anju AT, Chacko BP, Basheer KP. Text line segmentation of offline Malayalam handwritten document. *AIP Conference Proceedings*. 2024;2965(1):211999. Available from: https://doi.org/10.1063/5.0211999.
12) Kamble PM, Hegadi RS, Hegadi RS. Distance Based Edge Linking (DEL) for Character Recognition. In: Recent Trends in Image Processing and Pattern Recognition;vol. 1037. Springer. 2019;p. 233–277. Available from: https://doi.org/10.1007/978-981-13-9187-3_23.
13) Dorigo M, Birattari M, Stutzle T. Ant colony optimization. *IEEE Comput Intell Mag*. 2006;1(4):28–39. Available from: https://doi.org/10.1109/MCI.2006.329691.
14) Obaidullah SM, Halder C, Santosh KC, Das N, Roy K. PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification. *Multimed Tools Appl*. 2018;77(2):1643–78. Available from: https://doi.org/10.1007/s11042-017-4373-y.