INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Tri-level handwritten text segmentation techniques for Gujarati language

Bhargav Rajyagor[1]*, **Rajnish Rakholia**[2]

1 Gujarat Technological University, Gujarat, India
2 S. S. Agrawal Institute of Management and Technology, Navsari, Gujarat, India

OPEN ACCESS

Received: 28.11.2020
Accepted: 27.02.2021
Published: 03.03.2021

Citation: Rajyagor B, Rakholia R (2021) Tri-level handwritten text segmentation techniques for Gujarati language. Indian Journal of Science and Technology 14(7): 618-627. https://doi.org/10.17485/IJST/v14i7.2146

* **Corresponding author**.

rajyagorbhargavp@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 improve the efficiency of tri-level segmentation tasks for handwritten Gujarati text. **Methods**: Using hybrid methods for tri-level segmentation, we have used line, word and character segmentation from the image. This study presents a segmentation paradigm that works with touching characters, slop of the line written on the page, character overlapping, etc. It evaluated on the dataset of 500+ images created by us on different writing sentences by different people. We have used the Horizontal projection technique for line segmentation, Scale-space technique for word segmentation and the Vertical projection technique for character segmentation**. Findings**: The experimental results show that the proposed method is more efficient for handwritten Gujarati text with diacritics. We have obtained the accuracy for character level segmentation is 82%, word-level is 90% and for the line-level segmentation is 87%**. Novelty**: We have designed a methodology to segment Gujarati handwritten text with diacritics at all three levels including characters, words and lines**. Applications**: We have proposed tri-level segmentation which is pre-processing task that can be used in any character recognition systems i.e. OCR.

Keywords: Deep learning; trilevel segmentation; handwritten Gujarati text

1 Introduction

In the recent era of computer digital evaluation, Natural Language Processing (NLP) is getting more obligatory in our day-to-day life. To educate and enhance the scope of technology we have to reach a root level of the population. It requires thoughtful hard work for NLP. Character recognition of printed documents has achieved a great accomplishment in this field. Gujarati is the 7th most spoken language in India. Gujarat government and local persons are also used Gujarati as their communication medium either verbal or written. From the literature review, many studies that focused on online and word segmentation have not deeply focus on word segmentation. Many authors have a focus on the different segmentation methods but due to the difficulty of the writing style, they are not enough able to get 100% line, word, and character segmentation accuracy.

Segmentation can be described as a method of separating or isolating a document into smaller sectors or small useful region. Segment partitioned the whole document into standardized units like line, word, and character. The segmentation approaches phases can be divided into line segmentation, word segmentation, and character segmentation (tri-level segmentation). The important entity as a text line segmentation is the toughest task in Gujarati handwritten script. In Guajarati language, character recognition is challenging due to : (1) it has more curves, holes, and strokes, different writing style for individual persons (2) a little difference between some characters like;[ઘ(gh),ધ(dh)], [ખ(kh),અ(a)]. (3) presence of joint characters like;[કઙ્યું,લાવ્યા]. (4) Characters are written by different people with different shapes and sizes. (5) A character wrote in a variety of styles. (6) Overlapping of the characters like in [નિશ્ચય [શ-શ and ચ etc. The developments in Gujarati documents make segmentation is an inspiring task.

In [1] author has improved the methods for lie segmentation and develop a novel method for word segmentation. They have used Cartesian space calculation with the Hough transformation method for line segmentation and achieved 98.9% accuracy. For the word segmentation, they have used two different stages 1. Distance computation for calculating the distance of neighbor character and stage 2. Gap classification is used to identify the word interclass gap. In this stage, the author has used the Gaussian mixture method for universal clustering. They have achieved 96.8% accuracy for word segmentation.

In [2] author has noted that every natural language processing system has different requirements of segmentation with unique writing styles. They also suggest making hybrid segmentation methods for the segmentation of line, word, and characters from a scanned image. In their comparative study of different segmentation techniques for a variety of languages, they suggest neural network, HMM, or SVM for the segmentation phase. In [3] authors discussed and concluded that the use of SVM and CNN for the deep learning approach and concluded that SVM provides more accurate results for segmentation and character recognition. In [4] authors have used SVM and BLSTM decision tree and dynamic programming for character segmentation. With the help of the mention methods, they achieved 98.81% accuracy.

In [5] authors have presented the segmentation using horizontal and vertical projection for the line and character segmentation respectively. They also present a novel concept for overlapping characters like 'matras' in the Gujarati language and apply the split of each overlapping character into multiple points using projection and then re-merge all characters. With help of projection, they execute segmentation for 112 documents having 7724 characters and find the accuracy for 96.72% correctness. In [6] authors have also noted that histogram projection techniques for the character segmentation have use ON pixels in each row to identify the line from the images same they also use the projection methods for word and character segmentation. They have achieved significant results and suggested improving the method available for segmentation. In [7] authors have used modified horizontal and vertical projection methods for line and character segmentation. They have used orthogonal projection towards the x-axis. They have experimented on over 550 images and get a segmentation accuracy of 96%.

In [8] authors have implemented methods like zone determination for line segmentation also used zone boundary detection, segmentation line generation, segmentation line confirmation, and lower zone component separation. They also present the methods for overlapping character segmentation. With the help of this implementation, they achieved almost 85% accuracy. Another author in [9] has used the layout projection method also divides their script into multiple zones and creates N*N blocks. They have used this technique for the Tibetan language. With help of this method, they have achieved 76% accuracy with 5844 images of the database.

In [10] have used a modified header and based line method for line segmentation. This method completely depends on the pixel value of the image. They can get accurate results of 98.1% for the handwritten line segmentation. In [5]authors have particularly focused on an overlapping character with vowels of Lanna language uses in Thailand. With this method they have used the histogram method for splitting, rotating, and margining the processed characters.

In [11] authors have used Bangla OCR with Hough transform for the line segmentation along with this they have also used color filling for the non-text area in the line. After segment a line they work on each word and used the connected component (CC) analysis and with the help of this they can easily segment a word from the line and they have used zone segmentation for the character segmentation. Authors achieved accuracy like 90.46% in line segmentation, 90.06% in word segmentation, and 75.97% in character segmentation. In [12] authors also, perform Hough-based projection for line segmentation and they have segment lines from the scanned image. They have used the vertical histogram method for character segmentation. They achieved almost 90.866% accuracy in printed Guajarati characters. In [13] authors have implemented a Hough transformation technique for text line segmentation used in Arabic language script. For line segmentation, they have achieved 98.9% accuracy.

In [14] the author has used projection and adaptive thresholding algorithms for line segmentation. They tested with their dataset with almost 2500 lines and IAM public handwriting dataset. They have achieved 97.70% accuracy for their dataset.

In [15] the author has used a deep neural network window approach with the right and left context of the target syllable for the Dzongkha language. They have also experimented with using pre-trained syllable ending and others have not used pretrained syllable ending. The author has achieved 94.40% accuracy for line and word segmentation. In [16] authors in this methodology

use another neural network model. Authors have used deep fully convolutional networks (FCNs). They have used this network to identify x-height as a line representation, and they can segment lines from the image. With the help of this method, the author has achieved 91.3% segmentation accuracy.

2 Process of segmentation

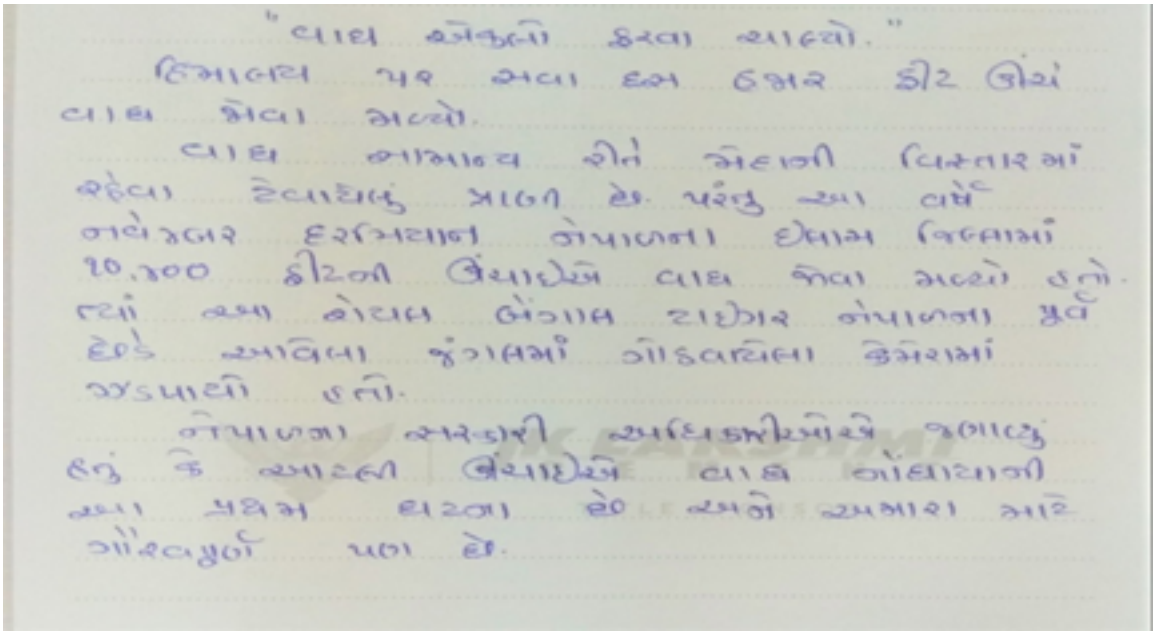

Fig 1. Original scanned image for the use of segmentation

1. Line segmentation

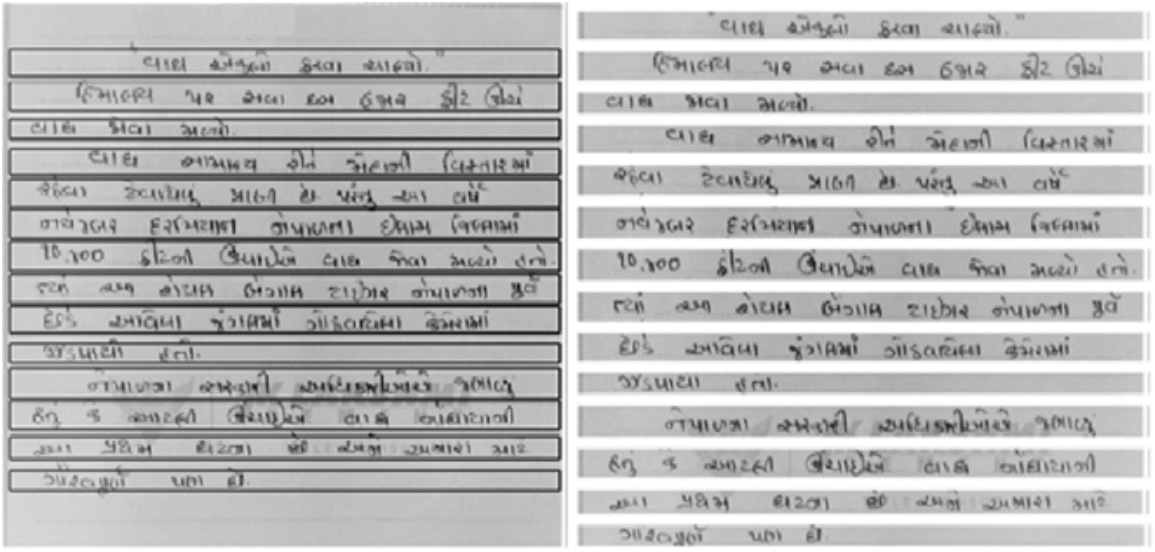

Fig 2. Divide a whole image into lines.

2. Word segmentation

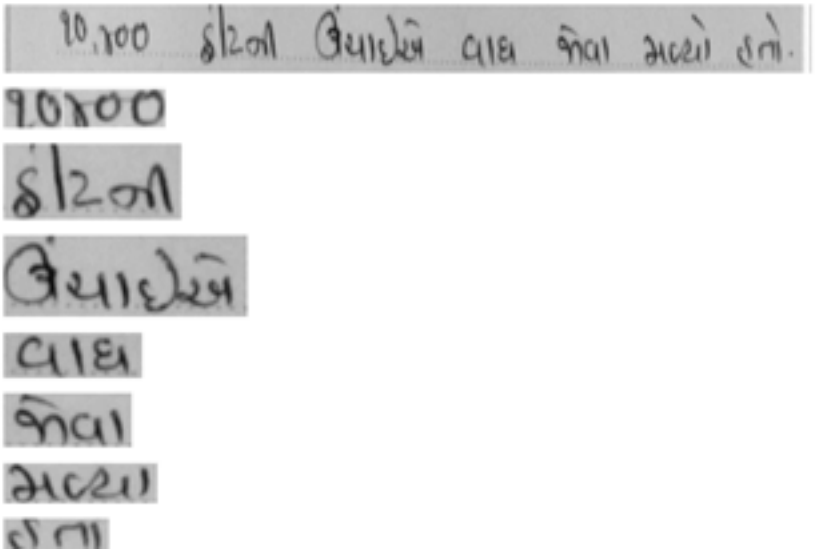

Fig 3. Divide a whole line into words.

3 Character segmentation

Fig 4. Divide a whole word into characters.

In this study, we are trying to summarize the latest techniques available for line, word, and character segmentation and also discuss the proposed method that finds use within our research. For the implementing different segmentation methods author have used own dataset consist of 500+ images of different handwritten style with 1000+ lines and approximate 5000+ different characters.

4 Methodology

We have used the hybrid algorithm to achieve maximum accuracy and uniformity in the segmentation technique. Due to diacritics, writing pattern, slop of characters, character overlapping, etc. segmentation is more difficult in the Gujarati language. In this methodology we have used the **horizontal projection** method for line segmentation, we have used the **scale-space** method for word segmentation, and the **vertical projection** for character segmentation. In this paper, we have focused on the preprocessing and segmentation of Gujarati handwritten scanned images. A complete segmentation process we have divided into different levels as below.

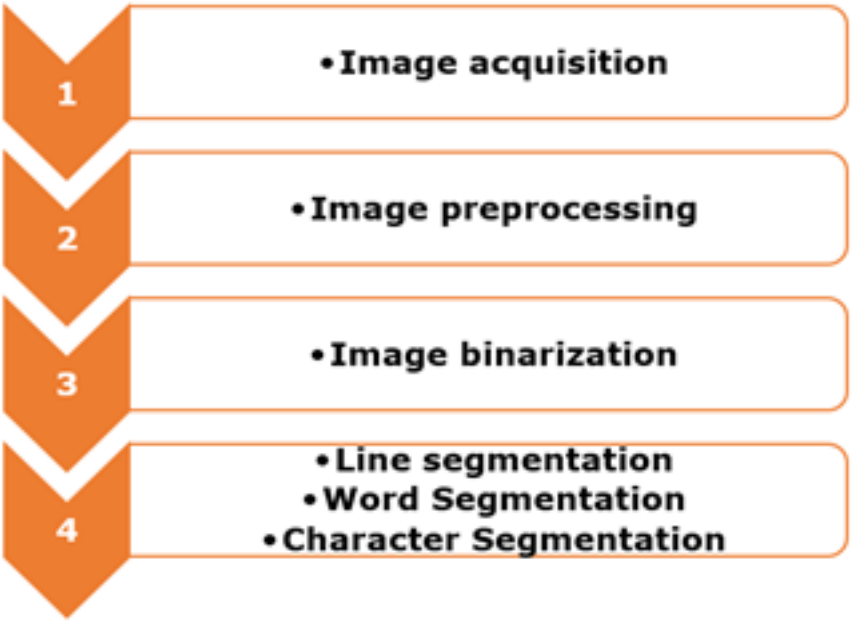

Fig 5. Process flow up to the segmentation.

For segmentation we have followed the below mentions algorithms.

Horizontal projection technique for line segmentation

- **Step 1:** Identify and set an appropriate threshold value for converting an image into a binary value. It also converts the image into a grayscale format. We have used 120 pixels and THRESH_BINARY constant value.

cv.adaptive threshold(src, maxValue, adaptive method, threshold type, blockSize, C)

- **Step 2:** Remove all image borders. So that we can easily identify the inner region of the image.

$$dst(x,y) \quad = \frac{min}{(x',y') : element(x',y') \neq 0} Src(x + x', y + y')$$

- **Step 3:** Create a vector, which can store each pixel value for a particular line. Here we have used vector z[] with initial value as 0 and size as per the height of an image.

$$\mathbf{z} = [0]^* \text{ height}$$

- **Step 4:** Using horizontal projection we fill the value of image height pixels into z[] vector.
- **Step 5:** Identify the appropriate starting point from where we can split the line. At this step, we are also storing the line spilt position using that we can segment a line. Here we will omit the pixels that have a height value that is less than the thresh value. Here we are using tValue as the thresh value.

$$lim\, 0 < i < height\ f(i)$$

$$where\ f(i) = \begin{cases} start = i & Where \\ and\ inline = 0 & inline = 1\ and \\ & z[i] >= tValue \\ \\ inline = 1 & Where \\ hfg[j][0] = start - 2 & inline = 0\ and \\ hfg[j][1] = i + 2 & (i - start) > 3\ and \\ j = j + 1 & z[i] < tValue \end{cases}$$

- **Step 6:** Draw and cut the line with the desire position.

Scale-space technique for word segmentation

- **Step 1:** Create kernel with the kernel size, sigma that is the standard deviation of Gaussian function used for filter kernel, theta used for approximated width/height ratio of words, the filter function is distorted by this factor and minArea: ignore word candidates smaller than specified area.

$$H(x,y) = \sum_{i=0Mi-1} \sum j = 0Mj - 1I\,(x+i-a_{i,}y+j-a_j)\,K(i,j)$$

- – Place the kernel anchor on top of a determined pixel, with the rest of the kernel overlaying the corresponding local pixels in the image.
  – Multiply the kernel coefficients by the corresponding image pixel values and sum the result.
  – Place the result to the location of the anchor in the input image.
  – Repeat the process for all pixels by scanning the kernel over the entire image.
- **Step 2:** The Filter2D operation convolves an image with the kernel. And make all pixels as an average of the other pixels.
- **Step 3:** convert the image into binary value using a threshold value. Here we have also applied the THRESH_BINARY and THRESH_OTSU for the Binarization of an image.
- **Step 4:** find connected components. We have used the method to find contours to identify the components of a line.
- **Step 5:** Create a resource vector and append all the components from the line here we will skip all small components. Append bounding box and image of the word to result from the list.
- **Step 6:** List of words, sorted by x-coordinate is ready to use.

Vertical projection technique for character segmentation

- **Step 1:** Identify and set an appropriate threshold value for converting an image into a binary value. It also converts the image into the grayscale format. We are using 20 pixels and THRESH_BINARY constant value.
- **Step 2:** Remove all image borders. So that we can easily identify the inner region of the image.
- **Step 3:** Create a vector, which can store each pixel value for a particular word. Here we have used vector v[] with initial value as 0 and size as per the height of an image.

$$\mathbf{v} = [0]^*\ \text{width}$$

- **Step 4:** Using vertical projection we fill the value of image width pixels into v[] vector.
- **Step 5:** Identify the appropriate starting point from where we can split the word. At this step, we are also storing the line spilt position using that we can segment a word into character. Here we will omit the pixels that have a width value that is less than the thresh value. Here we are using cValue as the thresh value.

$$\lim 0 < I < \text{ width } f(i)$$

$$where\ f(i) = \begin{cases} \text{start} = \mathbf{i} \\ \text{and incol} = 0 \\ \\ \text{incol} = 1 \\ \text{Ifg}[j][0] = \text{start} - 2 \\ \text{Ifg}[j][1] = i + 2 \\ \text{j} = \text{j} + 1 \\ \text{i} = \text{I} + 2 \end{cases} \begin{array}{l} Where \\ incol = 1\ and \\ v[i] >= cValue \\ \\ Where \\ inline = 0\ and \\ (i - start) > 4\ and \\ v[i] < cValue \end{array}$$

- **Step 6:** Draw and cut the word with the desire position. And create images for each character.

5 Data Collection

The database is of utmost significance for any research or experimental task. In the Gujarati language there are data available for individual handwritten characters but with the multiple lines with paragraph is not available online. We have downloaded the character dataset from the Indian government portal. This data may use to compare individual segmented characters as a trained dataset.

6 Dataset Generation

We used the dataset that has been created on our own with the help of people with different handwriting styles, different age groups, and a different gender. Our dataset includes printed as well as handwritten Gujarat images. We have also included the dataset with the isolated and allied characters for experimental purposes. This dataset includes 1000+ handwritten documents with 10 lines, 4-5 words, and almost 12-15 characters each.

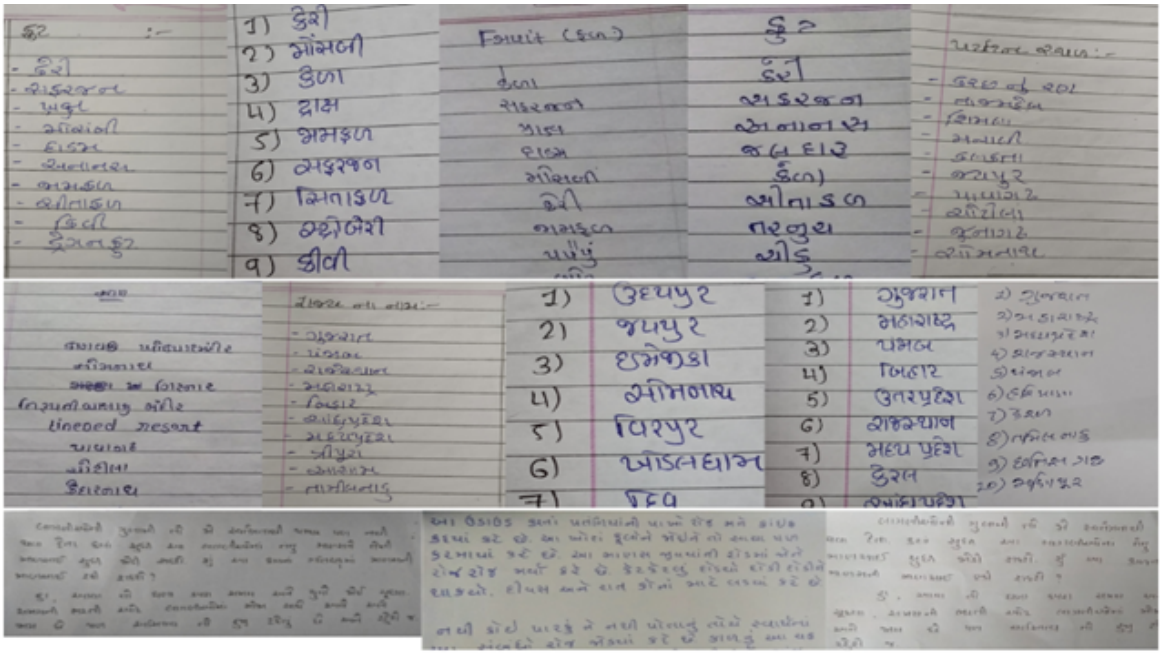

Fig 6. Available datasets for training and testing

7 Model Architecture

For the enhancement of the line, word, and character segmentation and also for the improvement of character recognition we have used projection and scale methods for segmentation. This model architecture represents the proposed workflow of the complete paradigm it includes image scanning, image prepossessing, binarization, segmentation. With this study, we have focused on tri-level segmentation. We have received fruitful results.

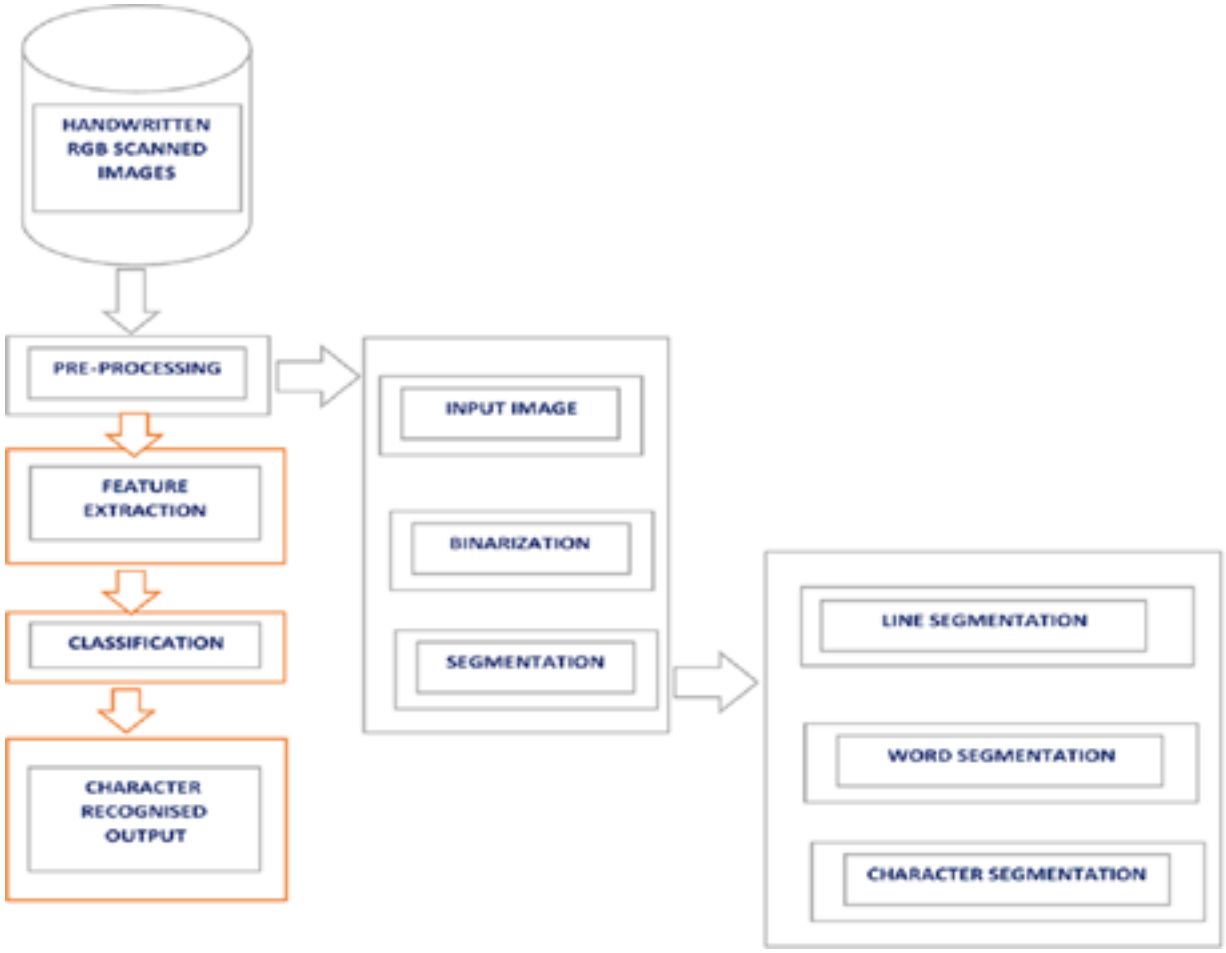

Fig 7. A workflow for the process of segmentation

8 Experimental results and discussion

The tri-level algorithm was performed on the own created dataset with almost 1000+ images. From this dataset, we had select 600+ images for the testing purpose. In the Gujarati language, there are 11 vowels and 36 consonants. Each character in Gujarati has a special appearance and Gujarati handwritten script is irregular in style due to many connected characters, overlapping words, slop of the line, etc. with the above methods we used projection methods for line and character segmentation while we have used scale space method for word segmentation.

From the research review[(2)], we have derived that a combined method for all 3 level segmentation is required. In[(5)] also used the projection method for the segmentation. It is understood that an individual method is not sufficient for complete segmentation. Hence we have combined the methods and we can have better results.

After implementing the above mention tri-level segmentation we have achieved the following results;

Table 1. Experimental results

Sr. No.	Segmentation Method	Images for testing (No)	Derived Segmented images (No)	Success (%)
1	Horizontal projection Line segmentation	642 – Images	5580 - Lines	~87%
2	Scale-space Word segmentation	550 – Lines	2790 - Words	~90%
3	Vertical projection Character segmentation	2700 - Words	8842 - Characters	~82%

With the above mention algorithm, we have executed the horizontal projection method for line segmentation we have a successfully segmented line from a scanned image. With the help of different images with different writing patterns, we can deeply test our line segmentation algorithm. We have passed a total of 642 images and each image carries an average of 10 lines. We have a deficiency for ~13% of the accuracy due to the writing patterns and slop of the lines.

In the case of word segmentation, we have achieved more accurate results with our implementation. We have used 550 lines produced by our previous algorithm having 6 words for each line. We have achieved almost ~90% accuracy for the word segmentation from the lines. Some lacking is here due to very diminutive space between two words. We can generate almost all the words from the inputted line to our algorithm.

Character segmentation is the most difficult task in the segmentation process due to connected characters in the writing style and some characters have an irregular shape. With the help of 2700 different words with an average of 4 characters in each word image our algorithm successfully segmented the character images. We achieved almost ~82% success for the same.

In this study, we have present different methods for segmentation. Tri-level segmentation is also implemented with hybrid methods like projection and scale-space technique. In comparison to the existing work specifically for the Gujarati language, we have achieved major success results in the handwritten scripts. With this, we also attempt to get the accurate segmentation from the script with the vast variant in the writing pattern, paper quality, ink color, etc.

A Combined accuracy for our implementation is as under.

Table 2. Tri-level segmentation accuracy

Sr. No.	Segmentation Region	Success Accuracy (%)	Combined Accuracy (%)
1	Line	87%	
2	Word	90%	86%
3	Character	82%	

9 Conclusion and future work

This study introduces the segmentation methods that can be applied with the handwritten image segmentation for line, word, and characters. This study also presents a novel database that is not available publicly for the Gujarati language. It presents the multiple combinations of segmentation methods for a common handwritten image. We have used the horizontal and vertical projection method for line and character segmentation respectively. Also used the scale-space method for word segmentation. Hence the authors can achieve tri-level segmentation with hybrid segmentation methods.

With the novelty of a hybrid algorithm, we may achieve good results but still, the improvement of the tri-level segmentation is required in concern with words and characters. Also, with these unique practices, a prototype must be implemented to combine all the tri-level segmentation for further enhancement.

With the help of our segmentation methods, we will implement a combined model for image segmentation. These segmented images will be passed as an input of the next phase and that will be our learning model. Further improvements can be made after comparing our image with the existing dataset and identifying a complete handwritten character in Guajarati language with diacritics.

References

1) Ali A, Ali A, Suresha M. An Efficient Character Segmentation Algorithm for Recognition of Arabic Handwritten Script. In: and others, editor. 2019 International Conference on Data Science and Communication;vol. 2019. 2019;p. 1–6. Available from: https://doi.org/10.1007/978-3-030-12385-7_11.
2) Dobariya AR. "A Comparative Study of Various Techniques and Challenges for Hand Written Document Processing of Indian Script." : 309–13. Available from: Proceedings of the 13th INDIACom; INDIACom-2019; IEEE Conference ID: 46181 2019 6th International Conference on "Computing for Sustainable Global Development", 13th - 15th March, 2019 Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), New Delhi (INDIA). 2019.
3) Pratikshaba J, Nehal C. Machine Learning and Deep Learning Approaches for Sanskrit Character Recognition. 2019. Available from: https://doi.org/10.1109/DAS.2018.50.
4) Volkova V, Deriuga I, Osadchyi V, Radyvonenko O. Improvement of Character Segmentation Using Recurrent Neural Networks and Dynamic Programming. In: and others, editor. Proceedings of the 2018 IEEE 2nd International Conference _1on Data Stream Mining and Processing;vol. 2018. 2018;p. 218–222. Available from: https://doi.org/10.1109/DSMP.2018.8478457.
5) Kosarat R, Hiransakolwong N. Segmentation of overlapping characters in lanna using mixed algorithm. 2020. Available from: https://www.xisdxjxsu.asia/c928ace96906187d791a7ef2f1c4f349.pdf.
6) Pareek J, Singhania D, Kumari RR, Purohit S. Gujarati Handwritten Character Recognition from Text Images. *Procedia Computer Science*. 2020;171:514–523. Available from: https://dx.doi.org/10.1016/j.procs.2020.04.055.
7) Bal A, Saha R. An Improved Method for Text Segmentation and Skew Normalization of Handwriting Image. *Advances in Intelligent Systems and Computing*. 2018;518:181–196. Available from: http://dx.doi.org/10.1016/j.procs.2016.07.227.
8) Malakar S, Sarkar R, Basu S, Kundu M, Nasipuri M. An image database of handwritten Bangla words with automatic benchmarking facilities for character segmentation algorithms. *Neural Computing and Applications*. 2021;33(1):449–468. Available from: https://dx.doi.org/10.1007/s00521-020-04981-w.
9) Ma L, et al. Segmentation and Recognition for Historical Tibetan Document Images. *IEEE Access*. 2020;8:52641–51. Available from: http://dx.doi.org/10.1109/ACCESS.2020.2975023.
10) Ahmed SM, Muazzam M, Farhan A, Muhammad FK. An Efficient Segmentation Technique for Urdu Optical Character Recognizer (OCR). Springer International Publishing. 2020. Available from: http://link.springer.com/10.1007/978-3-030-12385-7.
11) Rakshit P, Halder C, Ghosh S, Roy K. 666 Advances in Intelligent Systems and Computing Line, Word, and Character Segmentation from Bangla Handwritten Text-A Precursor toward Bangla HOCR. Singapore. Springer. 2018. Available from: http://dx.doi.org/10.1007/978-981-10-8180-4_7.
12) Patel H. 2020. Available from: https://www.academia.edu/download/64584430/IJCST-V8I5P9.pdf.
13) Bajaj S, Amali DGB. Distribution Modeling for Panthera Tigris Tigris ' Royal Bengal Tiger ' Using Machine Learning. Singapore. Springer. 2019. Available from: http://dx.doi.org/10.1007/978-981-13-5953-8_22.
14) Suleyman E, Hamdulla A, Tuerxun P, Moydin K. An Adaptive Threshold Algorithm for Offline Uyghur Handwritten Text Line Segmentation. 2020. Available from: https://doi.org/10.1007/s11276-019-02221-1.
15) Jamtsho Y, Muneesawang P. Dzongkha Word Segmentation Using Deep Learning KST 2020 - 2020. *12th International Conference on Knowledge and Smart Technology: 1-5*. 2020;p. 1–5. Available from: https://doi.org/10.1109/KST48564.2020.9059451.
16) Renton G, et al. Fully Convolutional Network with Dilated Convolutions for Handwritten Text Line Segmentation. *International Journal on Document Analysis and Recognition*. 2018;21(3):177–186. Available from: https://doi.org/10.1007/s10032-018-0304-3.