Character Segmentation from Offline Handwritten Gujarati Script Documents

Mit Savani, Dhrumil Vadera, Krishn Limbachiya, and Ankit Sharma

Abstract Modern innovations make great impacts on the human lifestyle and their way of working. It boosts the efficiency and productivity of people by reducing efforts, which help to handle several tasks at a time. Nowadays, all the government offices, banks, businesses, and education systems are influenced by paperless technology. It improves documentation and the secure sharing of information by saving space and resources. Paperless technology works on Optical Character Recognition (OCR) to convert all physical documents into machine-based documents. OCR consists of mainly two steps: segmentation, and recognition. The success rate of character recognition depends on the segmentation of required regions of interest. This paper introduced an algorithm that utilized projection profile, bounding box, and connected component labeling techniques for the development of Gujarati handwritten dataset and segmentation of handwritten text from Gujarati documents into line, word, and characters.

Keywords Handwritten document segmentation · Bounding boxes · Connected component labeling · Projection profile technique

M. Savani (✉) · D. Vadera · K. Limbachiya · A. Sharma
Department of Instrumentation and Control Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India
e-mail: 17bic046@nirmauni.ac.in

D. Vadera
e-mail: 17bic060@nirmauni.ac.in

K. Limbachiya
e-mail: 20ftphde43@nirmauni.ac.in

A. Sharma
e-mail: ankit.sharma@nirmauni.ac.in

A. Joshi et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2021)*, Lecture Notes in Networks and Systems 400,
https://doi.org/10.1007/978-981-19-0095-2_7

1 Introduction

People working at educational institutes, government officials, and private businesses having many difficulties dealing with physical documents. Finding particular details from the paper-based document and updating those details are tough and time-consuming tasks. Even these physical documents require large space to store them appropriately. It is a wastage of time and resources like paper and money. Optical Character Recognition (OCR) is a solution for this problem, which helps to store, read, and modify data easily. The OCR system converts all physical documents into machine-based documents. Users can easily read, write; update the details from machine-based documents in a short time. Optical Character Recognition-based natural language processing analyzes documents by utilizing several steps such as preprocessing, segmentation, feature extraction, recognition, and postprocessing to process documents.

Processing of machine-printed documents is easier than handwritten documents because of uniformity in characters' spacing and writing style and size. Irregularities in handwritten characters create a big issue of false segmentation, which reduces the recognition accuracy of the overall system. Segmentation of the region of interest can boost the performance of the system. Many segmentation methods are popular, but not so efficient to find a correct region of interest. Here, we modified several parameters from popular techniques to get a region of interest accurately to achieve a good recognition rate.

2 Literature Survey

The earlier work for Gujarati OCR worked on printed Gujarati documents, which is quite easy due to the uniform size of characters and proper spacing between them. Working with a handwritten document is difficult due to handwriting variation. Handwriting presents skew variation, size variation, and incorrect spacing between lines and characters. Here, we provided a comprehensive survey for different segmentation techniques proposed for different languages.

Rajyagor and Rakholia [1] suggested a segmentation methodology that considers connecting characters, line slop, character overlapping, and other factors. Line segmentation was done using the horizontal projection approach, word segmentation was done using the scale-space technique, and character segmentation was done using the vertical projection strategy. Over a sample of 500 + photos, they attained an accuracy of 82% for character segmentation, 90% for word segmentation, and 87% for line segmentation. According to Tamhankar et al. [2], the separation of individual characters is the most challenging module. They believed that the rate of character recognition is highly influenced by character segmentation accuracy. To reduce the character segmentation error on the MODI script, which has no separation between two consecutive words, they used a dual thresholding criterion. The algorithm was evaluated on over 15,000 characters, with over 10,000 of them correctly segmented, yielding 67% accuracy. Dahake and Sharma [3] explained an approach to deal with word segmentation from online handwritten Gurmukhi statements. For the separation of words from a sentence, the thresholding method is utilized. The vertical gaps between the strokes are first identified, and then the word is retrieved from the text using the maximum threshold value. The proposed method was put to the test on 200 sentences. With the use of the algorithm, an accuracy of 91.0% was achieved in segmentation. Jindal and Jindal [4] proposed the midpoint detection-based approach for segmenting lines and words from the handwritten Gurumukhi script, which features skewed lines, overlapped lines, and connected components between adjoining lines. The approach relies on the detection of gaps between two lines or words. The accuracy was found to be 95% in the line and word segmentation. Saha and Bal [5] introduced a technique using modified horizontal and vertical projection for segmentation of line and word from a document having multiple skewed text lines and overlapping of characters. A baseline recognition and normalization approach based on orthogonal projections was combined with a pressure detection method to determine a writer's personality. The method was tested over more than 550 text images of the IAM database and obtained segmentation accuracy of 95.65% over lines and 92.56% over words. With a very low error rate, the technique correctly normalizes 96% of lines and words. James et al. [6] demonstrated a unique approach for segmentation of overlapped, non-uniformly skewed, and touching text lines. For text line extraction, the method employed a piece-wise water flow methodology. The Spiral Run Length Smearing Algorithm is used to separate words from lines. Each word's skew is determined, and skew correction is done at the word level. The Connected Components Labeling method is used to extract the constituent characters. In Malayalam handwritten character recognition, the technique outperforms previous methods. Method of segmentation of Devnagri characters was proposed by Sahu and Sahai [7] using a bounding box, rectangle function, and region props operation on MATLAB. They used Otsu's threshold for the conversion of an image into binary format. The authors proposed a segmentation approach for line, word, and characters and achieved 100% accuracy for line and word segmentation, where 90%accuracy for character segmentation. Tan et al. [8] described a nonlinear clustering-based handwritten character segmentation technique. The technique divides a text line into strokes and computes a similarity matrix based on the stroke gravities. The cluster labels for these strokes are obtained using nonlinear clustering techniques. They utilized two nonlinear clustering methods: segNcut, which is spectral clustering based on Normalized Cut (Ncut), and segCOLL, which is kernel clustering based on Conscience On-Line Learning (COLL). They applied the SegNcut method on IAM, KAISTHanja1, HCL2000, and HIT-MW and achieved a correct rate of 85.4%, 89.2%, 88.9%, 89.7%. They applied another approach called segCOLL to all those databases and achieved accuracies of 88.1%, 85.2%, 89.3%, 87.7%. Arefin et al. [9] suggested the distance-based segmentation (DBS) approach for segmentation problems in lines, words, and characters in Bangla handwritten character recognition. To eliminate the shadows from the input image, they used the adaptive thresholding approach. They utilized segmented characters as ROI to extract HOG features and sent them to SVM for classification. The technique achieved an accuracy of 94.32% for character recognition but it is incapable to segment joint and italic characters. Alaei et al. [10] developed a unique technique for unconstrained handwritten text segmentation into lines. They utilized a painting method that improves the separation of the foreground and background sections, making text lines easier to identify. Applied morphological dilation and thinning operations subsequently with some trimming operations to obtain several separating lines. They analyzed the distances between the starting and ending points by using these separating lines. The suggested technique was tested upon documents in English, German, Greek, French, Persian, Bangla, and Oriya with remarkable results. The existence of the header line, overlapping of characters in the middle zone, touching characters, and conjunct characters complicate the segmentation process, according to Gurjar, Deshmukh, and Ramteke. They experimented with segmenting handwritten Marathi text documents into lines, words, and characters depending on content and non-content. The results showed an accuracy of 83%, 90%, and 86% for the segmentation of lines, words, and characters. The proposed approach only works for isolated characters and not for touched characters [11]. Swetha et al. [12] have proposed two algorithms the Vertical Strips and the Connected Components techniques over a database consisting of approx 200 images to evaluate the performance of algorithms. The dataset is classified into three categories: easy, medium, and complicated. They showed an overall line recognition rate on three different types of images, each with a different level of complexity. For the easy and medium categories of photos, the Connected Components approach outperforms the Vertical Strips method. For a Complex category of images, the Vertical Strips method outperforms the Connected Components method. The algorithm is limited to data having uniform line spacing and word size.

3 Proposed Method for Segmentation

The method introduced in this paper is used for two aspects such as dataset generation for Gujarati language and extraction of text in terms of characters from handwritten Gujarati form. As we discussed that specific part of an image that we can call "Region of Interest" is quite important to obtain the best quality for any object recognition or identification-based applications.

3.1 Dataset Generation

Here, we used handwritten forms filled by different people in different writing styles for dataset generation as shown in (see Fig. 1). Due to the presence of grids, the segmentation task was difficult by normal Contour Formation techniques. Even when these grids are removed, the quality of Characters is degraded. We performed several

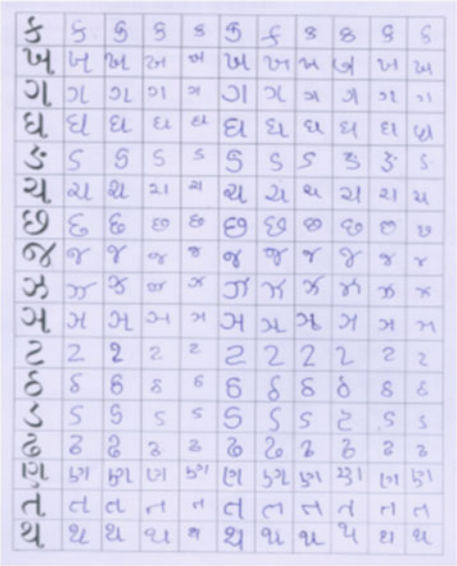

Fig. 1 Dataset form

logics using different techniques like Connected Components, Hough Transforms, or Kernel Generation for data segmentation. Not all those were enough to give the results we expected. Finally, we tried a fusion approach of projection profiles with vertical and horizontal profiles and bounding boxes to form contours and to extract the data inside of them. We applied vertical and horizontal profiles separately on forms and combined their result to get grid line contours.

Our algorithm for dataset generation works as per the following steps:

- Scan the image into binary format to make it ready for image preprocessing (see Fig. 1).
- Apply preprocessing techniques such as canny edge detection to detect edges and Gaussian Blur filter to reduce noise (see Fig. 2).
- Generate horizontal and vertical kernels separately with minimum width to detect horizontal and vertical lines by neglecting characters present in forms.
- Apply opening operation of image morphology to remove noise from detected straight horizontal and vertical lines.
- Addition of results from horizontal and vertical kernels to obtain an image in form of grids (see Fig. 3).
- Find an area of contours using specific limits and draw bounding boxes around the characters.
- Sort all detected contours from top to bottom vertically and left to right horizontally.
- Use the pixel address of each bounding box to cut out the area of interest with the best quality from the test image.
- Store cut-outs of characters in separate folders according to class (see Fig. 4).

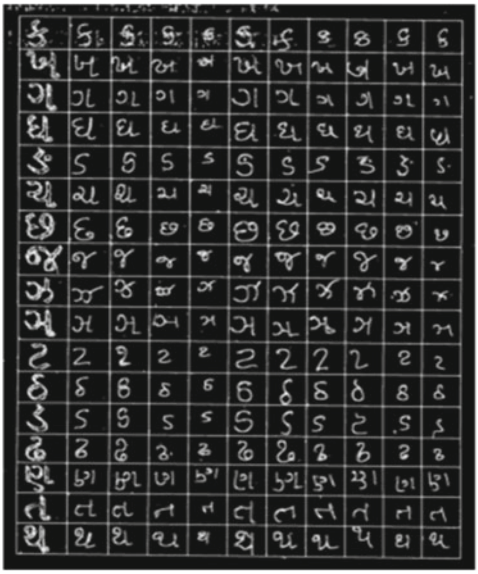

Fig. 2 Form preprocessing

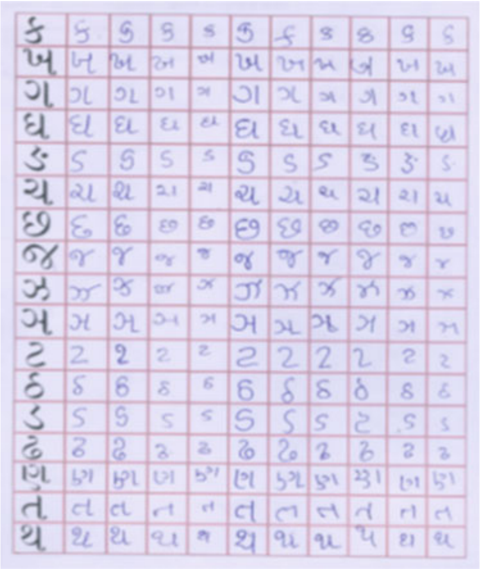

Fig. 3 Contour detection

3.2 Character Segmentation of Gujarati Document

Segmentation of characters from a handwritten document can be done in the same way as the dataset generation procedure. Here, we scanned a handwritten Gujarati text document and converted it into a digital image (see Fig. 5). Digital image of text

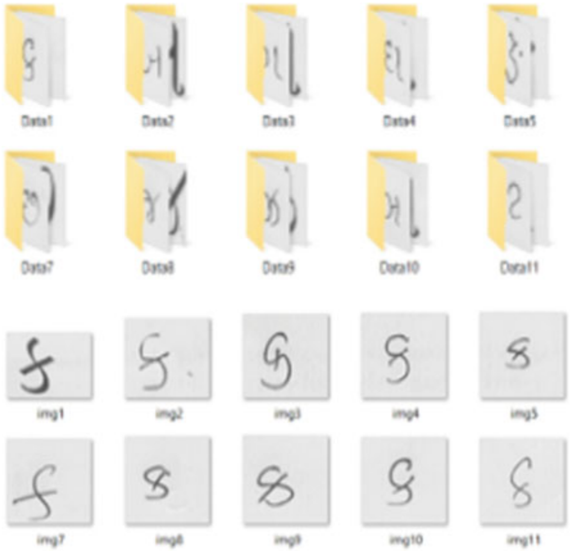

Fig. 4 Store image cut-outs

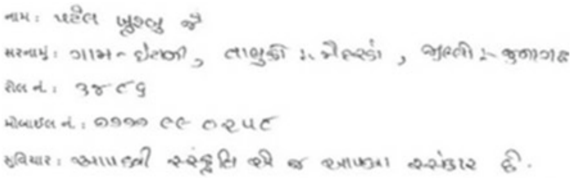
નામ: પટેલ ખુશ્બુ જે
સરનામું: ગામ- ઇટાળી, તાલુકો:- મેંદરડા, જીલ્લો:- જુનાગઢ
રોલ નં.: ૩૪૮૬
મોબાઈલ નં.: ૭૭૭ ૯૯ ૦૨૫૮
સુવિચાર: આપણી સંસ્કૃતિ એ જ આપણા સંસ્કાર છે.

Fig. 5 Input scanned document

document was processed through the OpenCV library to convert into a grayscale image, to correct the skew angle of the image, and to apply thresholding operation. Thresholding converts pixel values above a certain threshold into white (i.e., 255 value) and the rest are converted into black (i.e., 0 value). If vertical lines are present in the form image, it can generate noise.

Morphological opening operation with vertical kernel applied to remove noise generated through vertical lines. Finally, we obtained a black and white image in which the bright regions are the handwritten or printed words or characters considered as foreground images, and the dark region in space is considered as background in the pre-processed image (see Fig. 6).

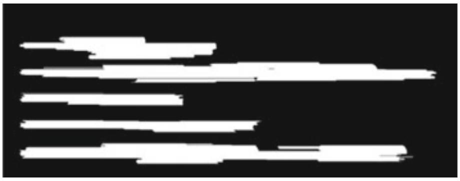

Fig. 6 Dilated image of a scanned document

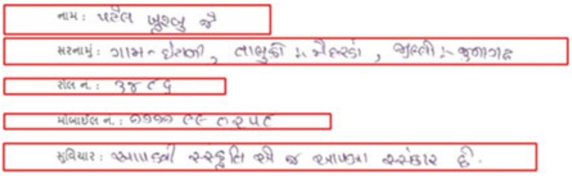

Fig. 7 Detected lines from the document

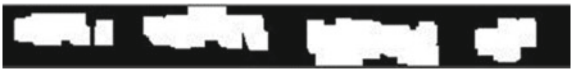

Fig. 8 Dilated image of a single line

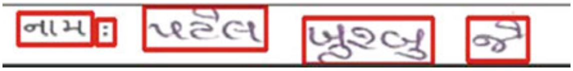

Fig. 9 Detected words in a single line

We utilized several functions available in the OpenCV library such as Find Contour to extract the contours from the binary image, RETR_TREE, and CHAIN_APPROX_SIMPLE as contour approximation methods to compress and remove all redundant points from the contour. In this way, we segregated the lines from handwritten document images (see Fig. 7).

The process of word segmentation from detected lines was done as similar as line segmentation but we used a loop over the segmented lines sequentially (see Figs. 8 and 9). Applied resize operation on segmented word images to maintain the accuracy of the algorithm.

Finally, characters were segmented using the discussed method for line and word segmentation. We applied gray scaling, binarization, thresholding, dilation, contours detection along the bright region, and obtained the individual characters from handwritten forms (see Figs. 10 and 11).

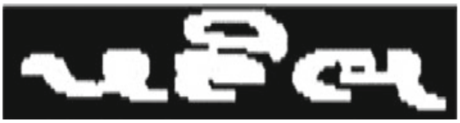

Fig. 10 Processed word image

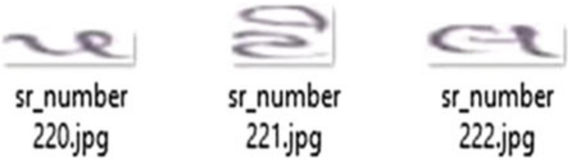

Fig. 11 Characters segmentation

4 Result and Discussion

The proposed algorithm can be used in areas of Banking, Education, Defense, Information Technology, etc. as a solution for Gujarati Character Segmentation. The algorithm can be clubbed with OCR models and various Digital formatting tools that convert the Physical Documents to Digital Format. In addition, the algorithm can be used to develop a Dataset for different languages. Here, we processed an approach for image segmentation that is useful for dataset generation as well as document and form segmentation. We utilized several functions from the OpenCV library to develop it. We tried our algorithm for segmentation of 50 grid forms (see Fig. 1) to generate a dataset for the Gujarati language and the results are as expected and satisfactory. Once we tried this algorithm for 50 images of handwritten Gujarati documents (see Fig. 5) for segmentation, it worked efficiently for characters with uniform sizes presented in a horizontal line. Sometimes it caused an error to detect details available in slant lines and segment a few portions of characters because of discontinuity in characters' shape.

5 Conclusion

Handwritten forms have great complexity because of variation in handwriting from person to person, variation in the use of pencil/pen with different ink colors, and text width variation. The proposed method is used for two different tasks such as dataset generation from handwritten or machine-printed grid-based forms and segmentation of handwritten or machine-printed details from actual forms available at schools, colleges, banks, etc. We used a cascaded way, where outputs of line segmentation are provided to word segmentation and followed by character segmentation to extract essential details. In the proposed algorithm, we used the projection profile technique with vertical and horizontal profiles and the bounding box method to get the region of interest. Overall the method worked accurately to segment details for stated purposes. This segmentation work can be applied to the development of the Gujarati language OCR system in the future to boost performance.

References

1. Rajyagor B, Rakholia R (2021) Tri-level handwritten text segmentation techniques for Gujarati language. Indian J Sci Technol 14(7):618–627. https://doi.org/10.17485/ijst/v14i7.2146
2. Tamhankar PA, Masalkar KD, Kolhe SR (2020) A novel approach for character segmentation of offline handwritten Marathi documents written in MODI script. Procedia Comput Sci 171(2019):179–187. https://doi.org/10.1016/j.procs.2020.04.019

3. Dahake D, Sharma RK, Singh H (2017) On segmentation of words from online handwritten Gurmukhi sentences. In: Proceedings of the 2017 2nd international conference on man and machine interfacing (MAMI) 2017, vol 2018, pp 1–6. https://doi.org/10.1109/MAMI.2017.8307870
4. Jindal P, Jindal B (2016) Line and word segmentation of handwritten text documents written in Gurmukhi Script using mid point detection technique. In: 2015 2nd international conference on recent advances in engineering and computational sciences (RAECS) 2015. https://doi.org/10.1109/RAECS.2015.7453388
5. Bal A, Saha R (2016) An improved method for handwritten document analysis using segmentation, baseline recognition and writing pressure detection. Procedia Comput Sci 93:403–415. https://doi.org/10.1016/j.procs.2016.07.227
6. Haji SAB, James A, Chandran S (2016) A novel segmentation and skew correction approach for handwritten Malayalam documents. Procedia Technol 24:1341–1348. https://doi.org/10.1016/j.protcy.2016.05.140
7. Sahu N, Sahai M (2016) Segmentation of Devanagari character. In: 2016 3rd international conference on computing for sustainable global development (INDIACom), 2016, pp 4005–4008
8. Tan J, Lai JH, Wang CD, Wang WX, Zuo XX (2012) A new handwritten character segmentation method based on nonlinear clustering. Neurocomputing 89:213–219. https://doi.org/10.1016/j.neucom.2012.02.026
9. Arefin N, Hassan M, Khaliluzzaman M, Chowdhury SA (2018) Bangla handwritten characters recognition by using distance-based segmentation and histogram oriented gradients. In: 5th IEEE region 10 humanitarian technology conference (R10-HTC), vol 2018, pp 678–681, 2018. https://doi.org/10.1109/R10-HTC.2017.8289049
10. Alaei A, Pal U, Nagabhushan P (2011) A new scheme for unconstrained handwritten text-line segmentation. Pattern Recognit 44(4):917–928. https://doi.org/10.1016/j.patcog.2010.10.014
11. Ramteke S, Gurjar AA, Deshmukh DS (2016) Automatic segmentation of content and noncontent based handwritten Marathi text document. In: Proceedings of the international conference on global trends in signal processing, information computing and communication ICGTSPICC 2016, pp 404–408. https://doi.org/10.1109/ICGTSPICC.2016.7955335
12. Swetha S, Chinmayi P, Mamatha H (2019) Line segmentation of handwritten Kannada documents. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT) (2019):1–5