Int. j. inf. tecnol. (June 2021) 13(3):905–909
https://doi.org/10.1007/s41870-021-00664-4

ORIGINAL RESEARCH

Shirorekha based character segmentation for medieval handwritten Devnagari manuscript

Nikita Mehta[1] · **Jyotika Doshi**[2]

Received: 27 May 2020 / Accepted: 29 March 2021 / Published online: 15 April 2021

Abstract In the process of optical character recognition (OCR), segmentation is always a crucial phase. Here, segmentation refers to all types of segmentation—page segmentation, line segmentation, word segmentation and character segmentation. The character recognition rate of any OCR system is largely depending on correct and accurate segmentation. This paper addresses the character segmentation for medieval handwritten Devnagari manuscripts. These manuscripts are hundreds of years old. In recent Devnagari, shirorekha (upper horizontal line) is placed on each word; whereas in medieval Devnagari, a separate shirorekha is placed on each character. Using this unique feature as a key, a novel Shirorekha Based Character Segmentation (SBCS) method is proposed. In this technique, first the shirorekha is identified to separate characters. The shirorekha is examined horizontally to find breaks in it. Wherever there is a break in shirorekha, it is assumed to be a possible segmentation point for a character. Thereafter, possible segmentation points are scanned for vertically spacing between two characters. According to the gap between characters, the segmentation points are finalized. Using this approach, segmentation accuracy achieved is 88.28%. This accuracy is better as compared to many existing approaches applied on recent Devnagari script. As per our knowledge no research work for character segmentation for medieval Devnagari script is found. This is the first attempt of its kind.

Keywords Segmentation · Character segmentation · Shirorekha based Character Segmentation (SBCS) · Projection profile · Medieval Devnagari manuscript · Image processing · Optical character recognition (OCR)

1 Introduction

India is a land of culture, civilization, art, craft, music, dance, and literature. India possesses a dynamic culture which is spanning back to the beginning of mankind. Indian literary tradition is also one of the oldest in the world [1]. Indian literature is always been a reason to feel proud for every Indian. It is written in many languages using different scripts.

Language is a medium for communication consisting of vocabulary and grammar. While script is a collection of graphical symbols which are used to denote the characters used by languages. For example, Hindi is a language, and it is written in Devnagari script. Devnagari script is also used to write other languages like Sanskrit, Marathi etc.

Handwritten documents written between twelfth to sixteenth century are known as medieval manuscripts. These manuscripts are written in many languages and different scripts. A large number of medieval manuscripts are found in Devnagari script.

Many researchers have proposed different segmentation techniques for different languages and scripts. It is observed that, techniques working well for English language characters (Roman script), are not well suited for Indian languages. Most of Indian scripts including Devnagari use modifiers (matras) above, below and at sides of

✉ Nikita Mehta
nikipshah@gmail.com

Jyotika Doshi
jyotikadoshi@gmail.com

[1] School of Doctoral Research and Innovation, GLS University, Ahmedabad, India

[2] Faculty of Computer Technology, GLS University, Ahmedabad, India

the characters. These scripts also use joint characters. Modifiers and joint characters make the segmentation process difficult. Handwritten text also produces some complexities such as overlapped characters, skewed text, touching characters, uneven gaps between lines and characters etc. [2–5]. In medieval scripts, many decorative marks and symbols are used which makes segmentation process very complex.

Very less research work is done for character recognition from ancient manuscripts. As per our knowledge, this is the first attempt of its kind to segment characters from handwritten medieval Devnagari manuscripts.

This paper proposes a novel approach for character segmentation using the unique writing style of the medieval Devnagari script. This approach is referred as 'Shirorekha Based Character Segmentation' (SBCS) in rest of the paper. In recent Devnagari script (which is used nowadays), a space is placed between two words and the whole word is covered with a single continuous shirorekha above it. In unique writing style of medieval Devnagari scripts, there is no space between two words (may be to save the valuable paper). The text written in medieval Devnagari is a continuous sequence of characters. There is no space between any words; each character has a separate shirorekha above it (as shown in Fig. 1).

Segmentation is very important phase in optical character recognition process. Most recognition errors occur because of segmentation error [6]. Proposed SBCS method scans horizontal lines along with the shirorekha from preprocessed image. Wherever there is a break in shirorekha, or a minor touch between two characters, it is considered as segmentation point for character. The method is discussed in detail in Sect. 3.

2 Literature review

Below Table 1 shows some existing approaches proposed by different researchers for various languages.

3 Shirorekha Based Character Segmentation (SBCS)

The proposed Shirorekha Based Character Segmentation (SBCS) method is used for character segmentation for medieval Devnagari manuscript. There are certain operations performed on manuscript pages before applying the SBCS method. These operations are:

- Grayscale conversion of colored images
- Binarization using OTSU global thresholding algorithm
- Line segmentation [10]

The segmented text lines are taken as input for SBCS method. A sample text line image is shown in Fig. 2.

As mentioned earlier, medieval Devnagari has a unique writing style that is no space between words and separate horizontal bar (shirorekha) on each character. The shirorekha on the character also covers the modifiers attached to that particular character.

The SBCS technique works using this unique writing style of the script. SBCS algorithm is described in detail:

Step 1: Create a row histogram for entire line image. Consider the row with having maximum number of black pixels in the upper half of the image as shirorekha row. Figure 3 shows the shirorekha form the row histogram chart.

Step 2: As the characters written in manuscript are thick, a shirorekha cannot be a single pixel width. So, a couple of rows above and below of the identified shirorekha row is considered as shirorekha span.

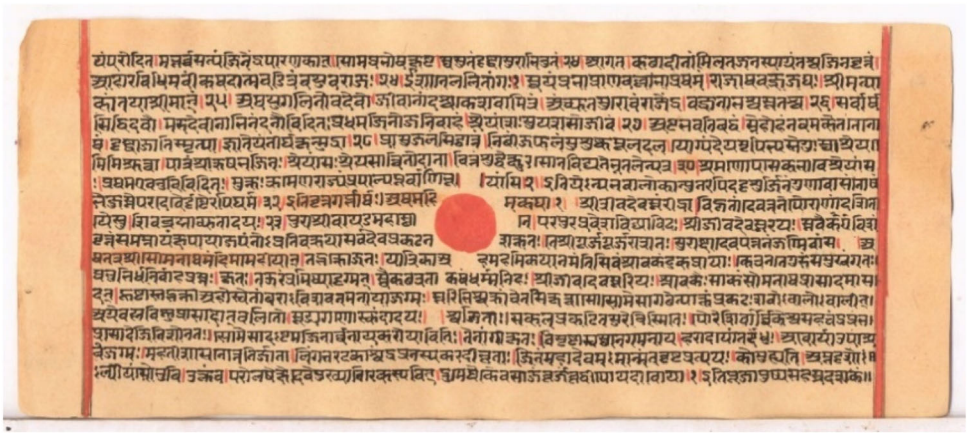

Fig. 1 Sample medieval Devnagari manuscript

Table 1 Existing character segmentation method study

Sr. No	Authors	Language	Method	Accuracy (%)
1	Tamhankar et. al. [7]	Handwritten MODI script for Marathi documents	Vertical projection profile is used for character segmentation	67
2	Kohli et. al [8]	Handwritten Devnagari joint characters	Continuous Pixel Technique (CPT) is used to identify the zones and cropping technique to segment the characters	80.94
3	Gupta et. al. [2]	Handwritten Hindi character segmentation	Polygon approximation is used to obtain approximate segment points and then graph traversal to segment the characters	95.70
4	Pramanik et. al. [9]	Handwritten Hindi word segmentation	Fuzzy and contour based segmentation is used to identify header line first. Then upper, middle and lower zones are identified. Outer contour, upper contour and lower contour is generated. Upper and lower contour is used to segment upper and lower modifiers	93.89
5	Palakollu et. al. [4]	Overlapping characters of handwritten Hindi	A new technique is proposed where the header line and upper modifiers are separated from the word first. Then for character segmentation, it generates stripes after every 5 pixels, if any character is in path it converts its path and comes down with characters boundary	89.90

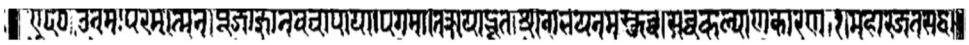

Fig. 2 A segmented text line image

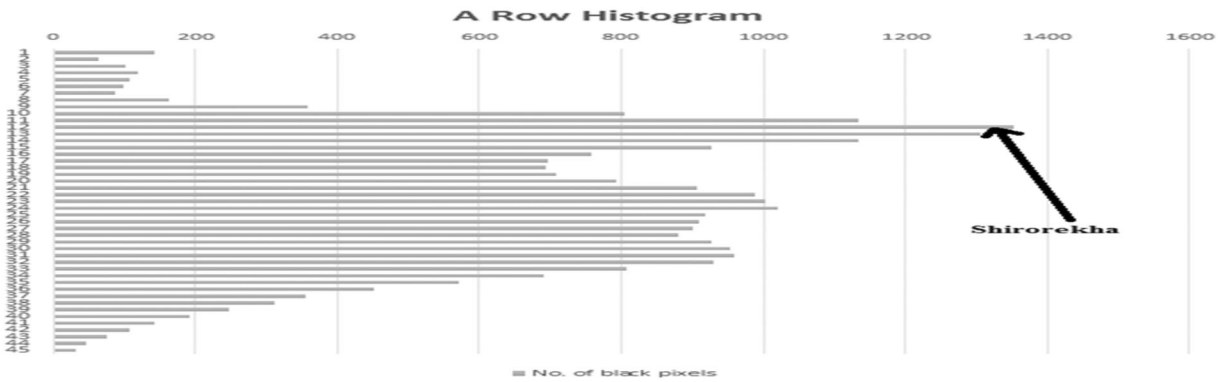

Fig. 3 Shirorekha from the row histogram

Step 3: This shirorekha span is analyzed to find all breaks in shirorekha line. Breaks in shirorekha is identified by creating a column histogram for the shirorekha span. A threshold value is to be assumed here (it should be according to character thickness). If the column histogram contains intensity less than assumed threshold, it is considered as possible segmentation point.

Step 4: In some characters, the character body can be wider than the shirorekha or shirorekha can be longer than the character body. In these cases, if segmentation done only by considering break in the shirorekha, it can lead to loss of partial character as shown in Fig. 4 (it can lead to partial character loss for 'ka').

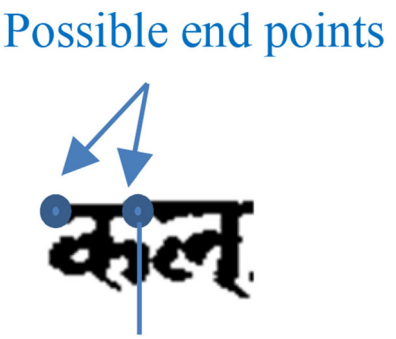

Fig. 4 Identified possible end points

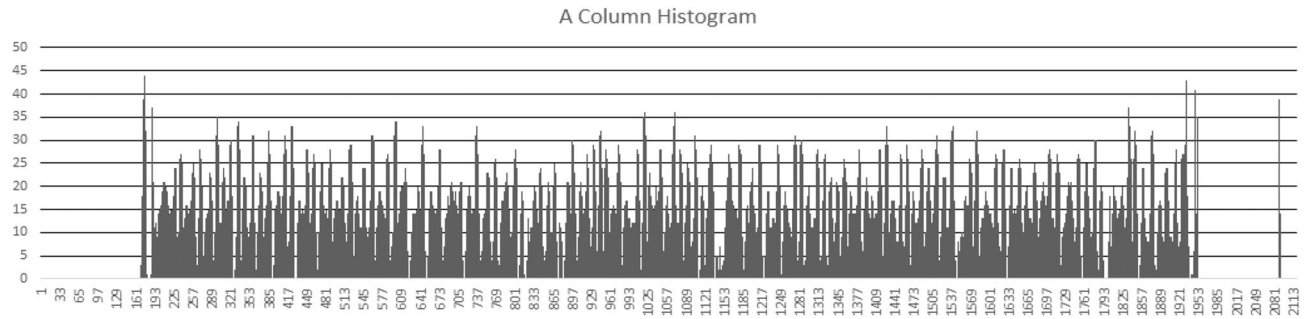

Fig. 5 Column histogram for text line image

Fig. 6 Segmented characters with SBCS algorithm

To resolve this, a column histogram for whole image is created as shown in Fig. 5.

Step 5: Column histogram values some columns before and after the possible end point is observed. The column with the minimum histogram value among them is the final segmentation point for the character. Figure 6 shows the segmented characters using SBCS algorithm.

Table 2 Experiment results on various prats

Image	Total number of characters	Correctly segmented characters	Percentage of accuracy (%)
Prat 1	463	409	88.34
Prat 2	630	565	89.68
Prat 3	633	525	82.94
Prat 4	709	628	88.58
Prat 5	719	621	86.37
Prat 6	807	681	84.39
Prat 7	663	576	86.88
Prat 8	388	350	90.21
Prat 9	706	628	88.95
Prat 10	345	321	93.04
Prat 11	243	223	91.79
Overall	6306	5527	88.28

4 Experiment results

The proposed novel technique SBCS is implemented using Visual Basic.Net 2013 under Microsoft Windows environment with × 86 based PC, 2.27 GHz processor, 3 GB RAM. The experimental results are analyzed for performance evaluation.

Here, segmented lines from eleven manuscript pages (prats) are selected for an experiment. Table 2 depicts the experimental results and computed accuracy. Accuracy percentage for each prat is calculated as the percentage ratio of total characters and correctly segmented characters.

It is observed that percentage of accuracy varies in the range of 82.94% to 93.04% with an overall average of 88.28%. The SBCS method is suitable for segmenting whole characters, numbers, characters with modifiers and joint characters.

As this is initial research of its kind. There are no results for medieval Devnagari character segmentation available to compare. The proposed research is achieving a promising result compared to many results for recent Devnagari scrip.

5 Conclusion

The proposed Shirorekha Based Character Segmentation (SBCS) method works very well with all kind of modifiers as well as joint characters used in the Medieval Devnagari Script. The method shows character segmentation results with the accuracy 88.28%. This accuracy rate is very good

as compared to many methods proposed for recent Devnagari script.

6 Future scope

The proposed method can be improvised to achieve better accuracy and results. It can be extended to segment special symbols and decorative marks used in ancient manuscripts.

References

1. www.knowindia.gov.in. Accessed 15 Mar 2020
2. Gupta D, Bag S (2018) An efficient character segmentation approach for handwritten Hindi text. In: 2018 5th international conference on signal processing and integrated networks (SPIN). https://doi.org/10.1109/spin.2018.8474047
3. Bathla AK, Gupta SK, Jindal MK (2016) Challenges in recognition of Devanagari Scripts due to segmentation of handwritten text. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE.93, pp 2711–2715
4. Palakollu S, Dhir R, Rani R (2012) Handwritten Hindi text segmentation techniques for lines and characters. In: Proceedings of the world congress on engineering and computer science, vol 1, pp 24–26
5. Thakral B, Kumar M (2014) Devanagari handwritten text segmentation for overlapping and conjunct characters-A proficient technique. In: Proceedings of 3rd international conference on reliability, infocom technologies and optimization. IEEE002E, pp 1–4
6. Bansal V, Sinha RMK (2002) Segmentation of touching and fused Devanagari characters. Pattern Recogn 35(4):875–893
7. 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 171:179–187. https://doi.org/10.1016/j.procs.2020.04.019
8. Kohli M, Kumar S (2018) Improved zoning and cropping techniques facilitating segmentation. In: International conference on advanced informatics for computing research. Springer, Singapore, pp 651–657.
9. Pramanik R, Bag S, Kumar R (2018) A fuzzy and contour-based segmentation methodology for handwritten Hindi words in legal documents. In: 2018 4th international conference on recent advances in information technology (RAIT). https://doi.org/10.1109/rait.2018.8389031
10. Mehta N, Doshi J (2020) Text line segmentation for medieval Devnagari manuscript. In: Proceedings of international conference on communication and computational technologies. Springer, Singapore, pp 405–412