A Solution for Line Segmentation Problems in Sindhi Character Recognition System

Shanky Goel, Gurpreet Singh Lehal

Abstract*: The Sindhi language uses extended and modified set of the Arabic and Persian alphabets. It is the largest extension of the Arabic alphabet. Thus, Arabic or any other Arabic script based language character recognition system cannot recognize all characters of Sindhi. From character recognition point of view, Sindhi is a tough script. Sindhi's cursive, context-sensitive characteristics, a large set of characters, highly similar shapes of the basic character, font-type variations, and size variations create high challenges for Sindhi character recognition research. In addition, line segmentation is a hard task as we have non-uniformity in the line heights. In this paper, we present an algorithm for segmenting the Sindhi text image into lines. The proposed algorithm solves the over-segmentation and under-segmentation problems in the line segmentation for Sindhi documents. The algorithm is tested on 100 text images of different Sindhi books and it has successfully segmented 99.95% lines correctly.

Keywords: Sindhi, Under-Segmentation, Over-Segmentation, Segmentation Algorithm.

I. INTRODUCTION

In today's world, there is massive demand to digitize the printed and handwritten information so that it can be preserved, share and re-utilize electronically. But what one can do if the published information is not available in the editable text? One way is to convert any printed, handwritten or historical documents into electronic form is to type the whole document in a computer. This is a traditional method of converting the paper-based information into computer text files through the keyboard. It should not be a big problem for an experienced stenographer, but on the other hand, many computer users are not so fast in typing and find it is a boring and time-consuming task. The other solution to this problem is optical character recognition (OCR). It is an automatic process to convert the hard copy information into editable text. In general, the optical character recognition system recognizes a natural language and converts it into machine-readable text. OCR provides numerous benefits to any industry with a lot of paperwork such as less storage space, easy management, text editing, etc. It is an old area of research in Pattern Recognition field. In spite of the early start of OCR research, recognition of cursive or degraded text still remains a challenging task. Character recognition generally involves various phases to convert the image into editable text such as Pre-processing, Segmentation, Feature Extraction, and Classification. In the segmentation phase, the image is first segmented into separate lines of text. These lines are then segmented into words and finally words into characters. It is a crucial phase in the development of Arabic OCR because correct segmentation ensures the success of the feature extraction and classification phases. Poor segmentation leads to the misrecognition of the character and impacts the recognition rate of the text [1]. Very less research work has been reported in the field of Arabic script based Sindhi OCR system. The first ever stated work in Sindhi character recognition system is the segmentation of Sindhi ligature into characters which is based on the height profile vector of the thinned primary strokes [2]. Nizamani and Janjua [3] have proposed a recognition system to recognize isolated Sindhi characters of the specific font "MB Lateefi". The proposed system used Back Propagation neural network approach for the characters to be recognized in a drawing panel. A neural network based approach is also used to recognize isolated Sindhi characters [4] . The authors have achieved accuracy of 93% for the machine-printed characters. The properties of the Sindhi language, methods, and techniques used by various researchers for OCR systems have presented in [5]. Hakro *et al.* [6] presented the recognition issues in Sindhi OCR and developed a large database having Sindhi word images and characters in different fonts [7]. The main reason behind less work in the domain of Sindhi character recognition system is due to the challenges in the recognition process. The cursiveness and context-sensitivity are the two major problems in the development of Sindhi OCR. To the best of our knowledge, line segmentation in Sindhi OCR has not been reported in the literature which takes care of over and under segmentation problems. Sindhi is an Indo-Aryan language and the official language of the Pakistani province of Sindh. It is also one of the scheduled languages in Indian constitution. Two scripts, Arabic and Devanagari are used for writing the Sindhi language. Sindhi based on the Arabic alphabet is used in Pakistan and it follows Naskh writing style. In Naskh writing style, characters are connected on a horizontal imaginary line called baseline. This language has a total of 52 basic characters, 24 more characters than the Arabic language. The Sindhi language is cursive and context-sensitive i.e. characters in a word/ligature change its shape depending upon their position and former or subsequent characters. In Sindhi, each character has 2 to 4 different shapes. The shape of the character varies according to its position (initial, middle, final and isolated) in the word/ligature (see Table I).

Table- I: Different shapes of

Revised Manuscript Received on August 05, 2019.

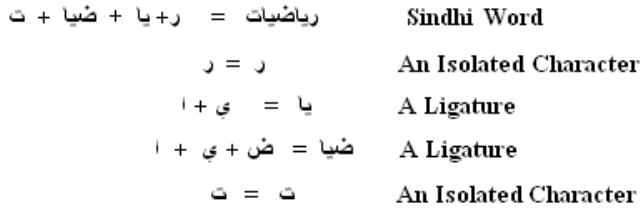
Shanky Goel*, Department of Computer Science, Punjabi University, Patiala, India. Email: sgoel9803415203@gmail.com

Dr. Gurpreet Singh Lehal, Department of Computer Science, Punjabi University, Patiala, India. Email: gslehal@gmail.com

Sindhi character shown in Red color

Character	Initial	Middle	Final	Isolated
ع	عمع	عمع	عمع	ع
ٹ	ٹٹٹ	ٹٹٹ	ٹٹٹ	ٹ

In Sindhi, the character shapes join together to make ligatures or words. A ligature or sub-word is a part of a word or sometimes a whole word, in which all character shapes must be connected together. A word in Sindhi is composed of ligatures and isolated characters, for example, the word "رياضيات" is formed by two ligatures "يا", "ضيا", and two isolated characters "ر", "ت".



Many times primary and secondary components are recognized separately in Arabic script OCR. The primary component represents the basic shape of the ligature, while the secondary connected component corresponds to the dots and diacritics marks and special symbols associated with the ligature (fig. 1).

Fig. 1. (a) Sindhi Ligature (b) Primary ligature/component (c) Secondary components.

The Sindhi language is very rich in its literature and history. Thus, it is important to digitize printed Sindhi books/documents to preserve its rich literature.

II. LINE SEGMENTATION PROBLEMS

Extraction of lines from text image is a crucial step in Sindhi character recognition system. If the line segmentation is not done correctly in Sindhi OCR then the whole recognition process may go wrong. Usually, the horizontal projection profile (HPP) method is used for text lines extraction of printed documents. In HPP method, the valleys of the horizontal projection are computed to segment the text image into lines. The position where histogram height between two consecutive horizontal projections is near to zero marked as the splitting point. Using these points, the text image is segmented into text lines. But, it has been observed that the page image of the Sindhi books often contains less inter space between lines and has non-uniformity in line heights. Such text lines could not be extracted precisely with the state of art method and several times this technique results in over and under segmentation of the text lines. In over-segmentation, a single text line is mis-segmented into two or more separate lines because many times, there is white space between the primary components and the secondary components (dots and diacritics) below and/or above the primary components, as shown in fig.2.

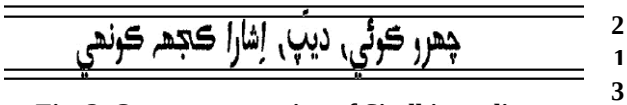


Fig. 2. Over segmentation of Sindhi text lines.

In under-segmentation, multiple text lines get merged in a single line due to less spacing between lines or non-uniformity in line heights as a result consecutive lines overlapped horizontally. As seen in fig. 3, three consecutive lines get merged in one line.

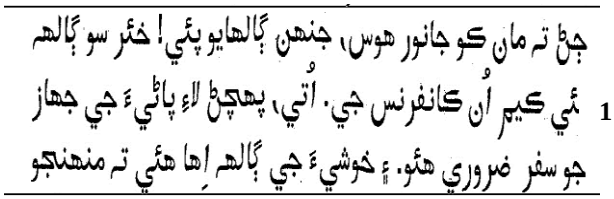


Fig. 3. Under segmentation of Sindhi text lines.

The problem using Horizontal projection profile (HPP) is pictorially shown in Fig. 4, where a binary input text image is resulted in incorrect segmentation of lines (over and under segmentation of the lines).

Fig. 4. Sindhi text lines using HPP.

III. LINE SEGMENTATION ALGORITHM

To effectively extract the text lines from an image, we propose the following algorithm for line segmentation in Sindhi OCR:

Step1. Different types of lines in the input binary document are identified using the horizontal projection profile. Let us assume there are 1, 2,…..,n strips/ lines (L) obtained using HPP. Height of the ith strip is calculated by

Height(i)=lastrow(i)-firstrow(i)+1; for i=1,2,….n.

Step2. We have estimated the Average row height and Gap row height by using

$$\text{AvgHeight}=\frac{1}{n}\sum_{i=1}^{n} Height(i),$$

GapHeight(j)=firstrow(i+1)-lastrow(i)-1; for j=1,2,……n-1.

AvgGapHeight =

$$\frac{1}{n-1}\sum_{j=1}^{n-1} GapHeight(j)$$

Step3. To solve the over-segmentation problem occurs in Sindhi text image segmentation as shown in Fig. 2. Identify the lines whose height is less than half of the AvgHeight i.e. (Height (L) < AvgHeight/2). This type of lines may contain either diacritics/dots, small font sized text or underlines.
Check the distance of this type of line (L) with its preceding (L1) and succeeding line (L2). Select the nearest line to L and merge these two lines, such that
a) If distance between these lines is less than AvgGapHeight.
b) Black pixel density of line L is less than the half of the black pixel density of selected nearest line. Black pixel density represents the number of black pixels per row. This is to ensure that L does not contain underline or text in smaller font.

Step4. To solve the under-segmentation problem arises in line segmentation as shown in Fig. 3. Identify the lines whose height is greater than 1.5 times the AvgHeight i.e. (Height(L) > (1.5)*AvgHeight).
This type of lines may contain multiple text lines or large-sized single text line. This type of line say L is a candidate for splitting if it meets the following criteria:
a) If there exist valleys in the line L and height of the proposed line (PL) is greater than the AvgHeight.
b) The black pixel density of proposed line (PL) to be split should be greater than fifty percent of the original line (L).
The flowchart for the proposed line segmentation method is given in fig. 5 for a better understanding of the algorithm.

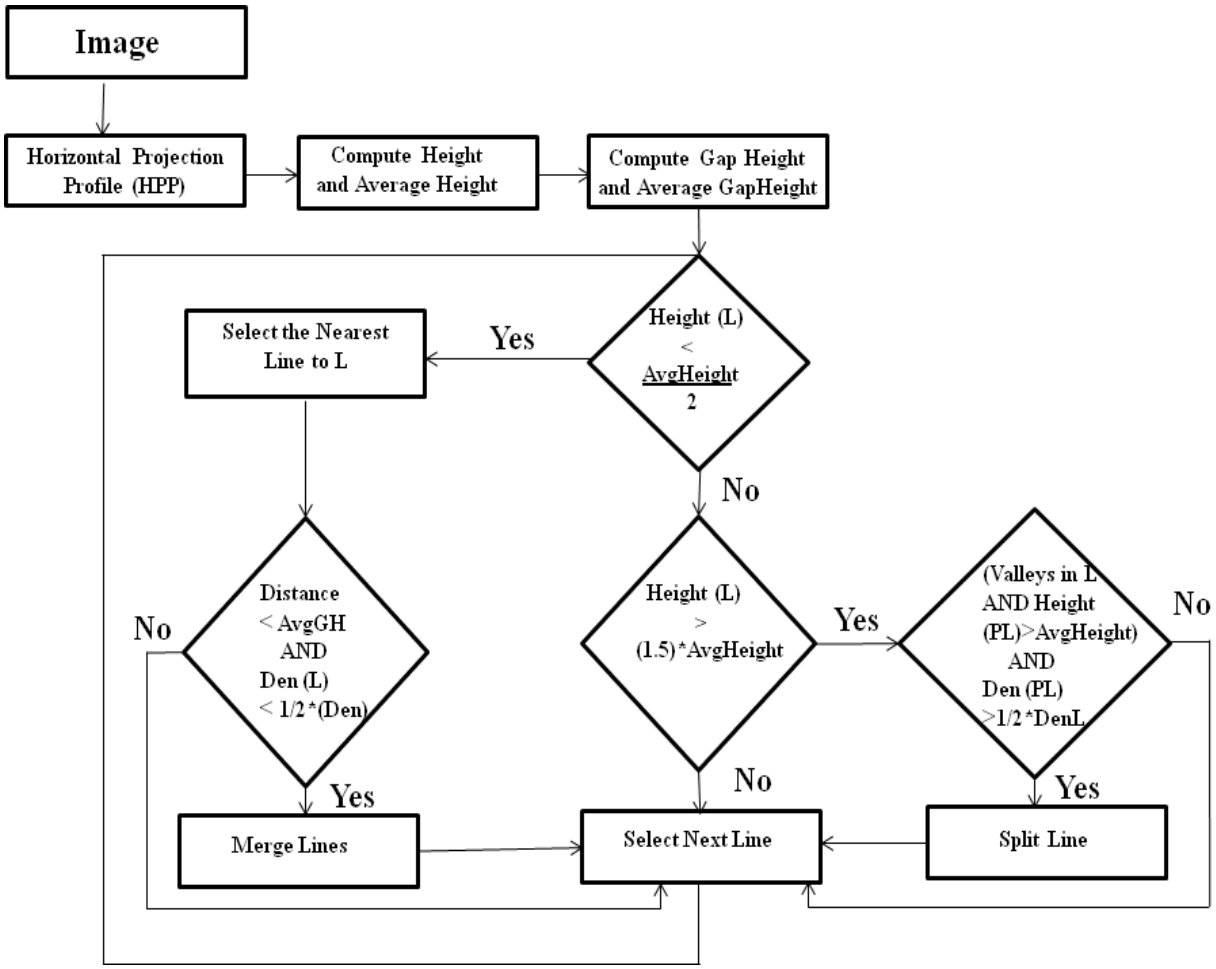


Fig. 5. Flowchart of Line segmentation algorithm in Sindhi text recognition system.

IV. TEST RESULTS ON SINDHI TEXT PAGE IMAGES

The proposed line segmentation method is evaluated against real images of the Sindhi text. To the best of our knowledge, there is no dataset of real images exists for the research purpose in Sindhi OCR. So, we have created the dataset of real Sindhi text images in this domain. The input text image can be acquired through a scanner or camera. Usually, scanner based digitization of books or any other text document is considered as a standard approach in the OCR field. Thus, a total of 100 text page images are acquired through scanning 8 Sindhi books using HP scanner. Most of the images are taken from different old Sindhi books. These books differ in contents, font type, and font size. All page images are scanned with a resolution of 300 dpi and cover almost all text layout variations. The variations in the text layout of Sindhi books are due to the nature of contents. The contents can be categorized as novel, poetry, religion, etc. However, the scanned images do not contain any graphics and are input to the proposed technique as they are. To apply the proposed method for the image which contains graphics, we suggest first to segment the non-text blocks from the image. On average 12 pages are selected from each book. Sample text page images from different books are shown in fig. 6. At first, input text image is acquired in RGB format and then it is converted to the gray scale image. The Otsu binarization technique is applied to binarize the gray scale image. Median filter algorithm is commonly used for removing noise in both printed and handwritten Arabic

text images. Therefore, after binarization, we apply a median filter algorithm on input Sindhi text image to remove noise. The testing details of Sindhi page images are provided in table II.

Fig. 6. Sample text page images of Sindhi books.

Table- II: Testing details of proposed line segmentation method on Sindhi text images

Images	Actual Lines	Segmented Lines	Remarks
Image_1	9	9	
Image_2	23	23	
Image_3	22	22	
Image_4	21	21	
Image_5	22	22	
Image_6	24	24	
Image_7	24	24	
Image_8	22	22	
Image_9	18	18	
Image_10	20	20	
Image_11	21	21	

Image_12	24	24	
Image_13	25	25	
Image_14	20	20	
Image_15	22	22	One secondary component of a line gets merged into its next line.
Image_16	14	14	
Image_17	9	9	
Image_18	12	12	
Image_19	23	23	
Image_20	23	23	
Image_21	19	19	
Image_22	23	23	
Image_23	23	23	
Image_24	23	23	
Image_25	18	18	
Image_26	13	13	
Image_27	10	10	
Image_28	19	19	
Image_29	23	23	
Image_30	23	23	
Image_31	14	14	
Image_32	25	25	
Image_33	16	16	
Image_34	31	31	Some secondary components of lines get merged into its previous line.
Image_35	28	28	Some secondary components of lines get merged into its previous line.
Image_36	28	28	
Image_37	30	30	One secondary component of a line broke between two lines.
Image_38	27	27	
Image_39	15	15	
Image_40	30	30	Some secondary components of a line get merged into its previous line
Image_41	28	28	
Image_42	29	29	
Image_43	18	18	
Image_44	22	24	One line gets segmented into 3 lines.
Image_45	24	24	
Image_46	21	21	
Image_47	26	26	
Image_48	27	27	
Image_49	27	27	
Image_50	14	14	
Image_51	26	26	
Image_52	21	21	
Image_53	25	25	
Image_54	25	25	
Image_55	20	20	
Image_56	15	15	

Image_57	22	22	
Image_58	30	30	
Image_59	23	23	
Image_60	17	17	
Image_61	20	20	
Image_62	19	19	
Image_63	19	19	
Image_64	21	21	
Image_65	22	22	
Image_66	22	22	
Image_67	22	22	
Image_68	19	19	
Image_69	14	14	
Image_70	19	19	
Image_71	23	23	
Image_72	18	18	
Image_73	23	23	
Image_74	12	12	
Image_75	19	19	
Image_76	12	12	
Image_77	27	27	
Image_78	29	29	
Image_79	28	28	
Image_80	10	10	
Image_81	29	29	
Image_82	28	28	
Image_83	23	23	
Image_84	30	30	
Image_85	24	24	
Image_86	25	25	
Image_87	28	28	
Image_88	25	25	
Image_89	30	30	
Image_90	26	26	
Image_91	21	21	
Image_92	26	26	
Image_93	27	27	
Image_94	22	22	
Image_95	28	28	
Image_96	26	27	One line of noise
Image_97	24	24	
Image_98	24	24	
Image_99	29	29	
Image_100	28	28	
Total	**2217**	**2220**	

The total correct text lines are 2217. Our proposed method extracts total 2220 text lines as only one text line image is incorrectly segmented. This line segmentation algorithm can handle Sindhi as well as Arabic documents. As an example the result of text line segmentation of the image_34 (Table II) using HPP and proposed algorithm has been shown in fig. 7 and fig. 8 respectively.

Fig. 7. Line segmentation of Image_34 (Table II) with horizontal projection profile approach

Fig. 8. Line segmentation of Image_34 (Table II) with proposed method

REFERENCES

1. Y. M. Alginahi, "*A survey on Arabic character segmentation*," International Journal on Document Analysis and Recognition, vol. 16, no. 2, 2013, pp. 105–126.
2. N. A. Shaikh, G. A. Mallah, Z. A. Shaikh, "*Character Segmentation of Sindhi, an Arabic Style Scripting Language using Height Profile Vector*," Australian Journal of Basic and Applied Sciences, Vol. 3, 2009, pp. 4160-4169.
3. A. M. Nizamani, N. H. Janjua, "*Sindhi OCR using Back propagation Neural Networks*," International Journal of Advanced Computer Science, Vol. 3, 2013, pp. 113-117.
4. D. N. Hakro, M. Memon, S. A. Awan, Z. A. Bhutto, M. Hameed, "*Isolated Optical Character Recognition*", Sindh Univ. Res. Jour. (Sci. Ser.), Vol. 48, no. 4, 2016, pp. 839-844.
5. D. N. Hakro, A. Z. Talib, Z. Bhatti, G. N. Mojai, "*A Study of Sindhi Related and Arabic Script Adapted languages Recognition*," Sindh Univ. Res. Jour. (Sci. Ser.), Vol. 46, no. 3, 2014, pp. 323-334.
6. D. N. Hakro, I. A. Ismaili, A. Z. Talib, Z. Bhatti, G.N. Mojai, "*Issues and Challenges in Sindhi OCR,"* Sindh University Research Journal (Science Series). Vol. 46, no. 2, 2014, pp. 143-152.
7. D. N. Hakro, A. Z. Talib, "*Printed Text Image Database for Sindhi OCR*," ACM Transactions on Asian and Low-Resource Language Information Processing, Vol. 15, Issue 4, 2016, pp. 1-18.

AUTHORS PROFILE

Shanky Goel received her Bachelor's degree in mathematics and Post Graduate degree in Computer Science from Punjabi University, Patiala, India. She is pursuing Ph.D. from Punjabi University, Patiala, Punjab, India. Her research interests include Character Recognition and Natural Language Processing.

Gurpreet Singh Lehal received undergraduate degree in Mathematics from Punjab University, Chandigarh, India, and Post Graduate degree in Computer Science from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, India. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. He is Director of Research Centre for Punjabi Language Technology and Dean of Faculty of Computing Sciences, Punjabi University, Patiala. His areas of research are- Natural Language Processing and Optical Character Recognition. He has published more than 100 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration.