Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 50 (002) 205-208 (2018)
http://doi.org/10.26692/sujo/2018.06.0036
SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCESERIES)

Segmentation of Sindhi Handwritten Text

S. A. AWAN, D. N. HAKRO* Z. H.ABRO, A. H. JALBANI

Department of Information Technology, QUEST, Sindh, Pakistan

Received 20th October 2017 and Revised 24th April 2018

Abstract: Optical Character Recognition (OCR) and Intelligent Character Recognition (ICR) are two emerging areas to understand and convert document text into editable text. The change of language script on a text image pose various challenges and demand challenging algorithms and approaches to overcome these challenges especially in Arabic script and its adopting languages like Sindhi, Urdu, Pashto and Farsi. Sindhi is a very rich literature language and needs some powerful OCR and ICR systems to manage the level of advances with other languages having perfection in these areas such as English , Latin, Russian and Korean. This study presents a segmentation algorithm for the segmentation of lines, words and characters. The input images written by various subjects are scanned and preprocessed and tested on segmentation algorithm. The segmentation of lines produced 100% accuracy along with words accuracy of 95%. The characters segmentation level also produced and acceptable accuracy of 81%.

Keywords: Character Recognition, Handwritten, Sindhi, segmentation, feature recognition.

1. INTRODUCTION

The least advanced field of Artificial Intelligence (AI) is Natural Language Processing (NLP) due to the varying challenges posed by changing in script. The change in language or script needs change of algorithms and merely the same algorithms are to be used. Optical character recognition changes the image form of text into editable text. The text represented in terms of pixels are converted into form of codes called editable and this text can be edited and used in other softwares (Hamid, 2001).Data entry forms recognition is one of the important and useful applications of OCR (Assabie, and Bigun, 2011). Intelligent character recognition (ICR) is an advanced version of OCR in which the handwritten characters are recognized with an extra layer of user or subject identification, the actual writer of the text. After the preprocessing steps applied the input image is segmented to divide input image into number of segments and the process is called segmentation. Segmentation is applied on images to extract region of interest and textual information from text image. In text images, the lines are extracted and the extracted lines are further divided into characters.

2. RELATED MATERIAL

OCR and ICR researchers follow one of the two approaches for the segmentation and recognize various scripts by selecting any one suitable for the script. Segmentation free OCR and ICR recognize a limited words or ligatures whereas the counterpart segmentation based OCR and ICR segment up to the unit level of segmentation such as characters and letters of the scripts. Some of the studies Li *et al*., (2010) for Uyghur language, Shaikh *et al*., (2009) for Sindhi and Akram and Hussain (2010) for Urdu have selected segmentation approach for their research experiments. Some of the researchers used segmentation free approach for their OCRs (Lehal and Rana, 2013).

Li *et al*., (2010) segmented the Uyghur handwritten characters on mobile. The segmentation process initiated by converting words into strokes. Black pixels are the indication of formation of strokes. The location, width and height identification is the next step. The found strokes are analyzed and then the affiliation of strokes with concerned word is decided and the affiliation is recorded. The segmentation points are identified by using two approaches. Detection of Harris corner and high point have been used to identify the potential segmentation points. After The segmentation points detection, the unwanted or unused strokes are removed.

Shaikh *et al*., (2009) have experimented word segmentation using height profile vector for Sindhi language. The Sindhi words have been thinned by applying thinning algorithm followed by various steps so that a sub-word can be obtained. The horizontal projection is the method to obtain text lines from the text image. The base detection has been performed so that the sub words can be segmented using connected components. The study is limited to six structures of

++Corresponding emails: dill.nawaz@gmail.com, shafique.neduet@gmail.com, zhussain@quest.edu.pk
*Institute of Information and Communication Technology, University of Sindh, Jamshoro,Pakistan

Sindhi language whereas the remaining were considered as future directions.

Akram and Hussain (2010) proposed a segmentation system for Urdu script. The ligature sequences have been used to segment by applying boundaries identification and determination. The OCR is working on cleaned corpora which has been already segmented. Based on the lexical look up the words are ranked accordingly. The valid words are processed further. The final process is equipped with statistical approach to decide. The number of sentences use for their analysis and experiments are 150 containing 6075 ligatures and 2156 words.

Lehal and Rana (2013) presented a segmentation free approach for the recognition of Urdu Nastliq script. The writing style of Nastaliqposes more challenging task to recognize. They prefer next unit of recognition ligature rather than the character. A sum of 9272 ligatures have been used as the unit of recognition selected from 2207 primary and secondary components. The study reports that many of the classifiers were used for test of accuracy and 98% of accuracy have been achieved through the pre-segmented ligatures.

2.1 Peculiarities in Sindhi Character Recognition:

An in-depth and comprehensive study of the peculiarities can be found in Hakro *et al.*, (2014) and it is one of the challenging script adopted from Arabic script. It is the largest extension of 52 characters compared 28 of Arabic script. Sindhi is cursive, written from right to left possessing dots in vertical and horizontal positions. These dots and cursive nature of Sindhi script impose challenging task to the researchers of OCR and ICR. The characters in Sindhi change shape according to position in a sentence.

2.2 Segmentation:

Segmentation of text image is the core step of an OCR system in which text lines, words and characters are segmented so that appropriate feature extraction can be applied. The segmentation of lines, words and characters are needed in segmentation based OCR whereas segmentation free does not need segmented images of words or characters. The segmentation free OCR recognizes limited number of words and recognizes limited numbers. For the purpose of segmentation input image is converted from color image to grayscale. The grayscale image is converted into binary so that the image can be easily segmented as shown in **(Fig.1)**. For extracting the information segmentation is necessary and in this regard text image has been segmented into text lines using horizontal histogram as shown in **(Fig.2)**. The text lines segmented in **(Fig.2)** have been segmented into words using vertical histogram approach as shown in **(Fig-3)**. The words have been segmented and the next is to segmented characters from words and this stage poses the challenges as some of the languages scripts do not need character segmentation.

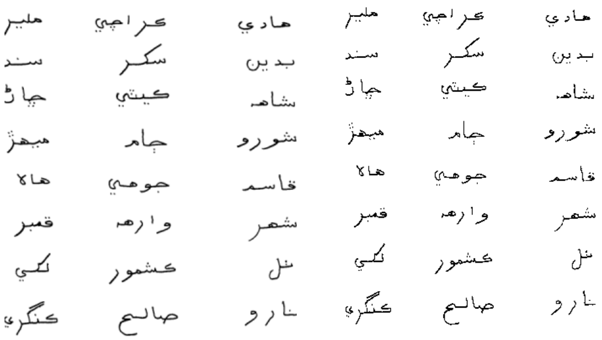


Fig 1: Conversion from color to binary image

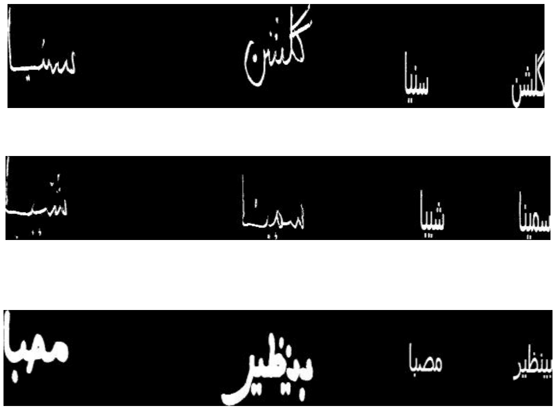


Fig 2: Segmentation of text lines of Sindhi Language

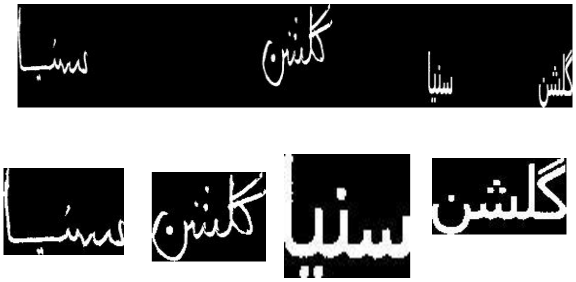
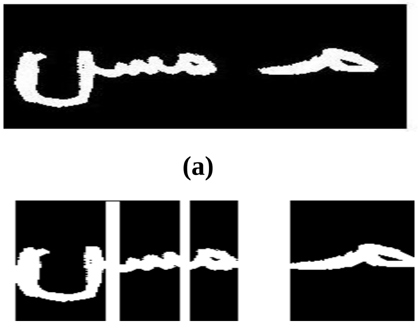


Fig 3: Segmentation of words from text lines of Sindhi Language

2.3 Character Segmentation:

For the segmentation of Sindhi handwritten characters we used the approach of edge identification achieved by setting a threshold and checking the values in column and rows. The number of columns and rows are counted and if the desired number of rows and columns are less than the threshold then the white lines are drawn between the characters as shown in **(Fig.4) (a)** shows the original word and ligature image and **(Fig.4)(b)** image after identifying edges. The value can be increased or decreased according to the size of the font for segmentation. In the case of handwritten characters, it can be finetuned according to the nib of the pen or strokes of the characters. The lines will differentiate between the segments of the characters where the large segments will be recorded and the smaller segments are dropped. These small segments are also tuned according to the value of the threshold set.



(a)

(b)

Fig4: Identifying edges: (a) before (b) after

The complete process of image segmentation form input image to character segmentation is shown in **(Fig-5)**.

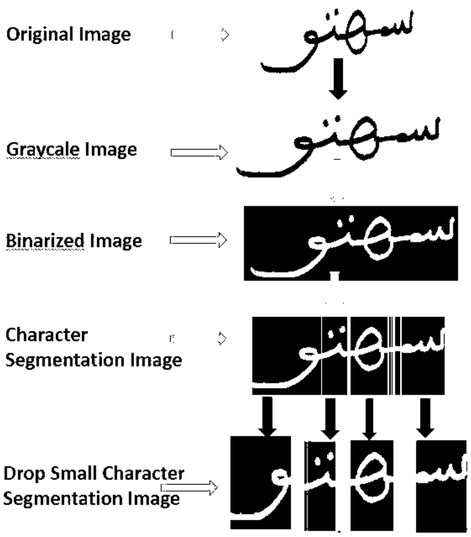

Original Image
Graycale Image
Binarized Image
Character Segmentation Image
Drop Small Character Segmentation Image

Fig 5: Complete process of handwritten character segmentation

3. RESULTS AND DISCUSSION

The segmentation algorithm based on projection and depth has been tested on various handwritings written by various subjects. Some of the images have been presented in this paper. More than 600 pages were filled by various users (Awan *et al*., 2017) and tested for the proposed algorithm in which line segmentation resulted in 100% accuracy as the writing forms were prepared carefully to avoid line segmentation errors (Awan *et al.*, 2017). The word segmentation averaged more than 90% accuracy as the space between the words produced promising results. The character segmentation is a challenging task and needs to fine tune as the reduced accuracy of 81% is due to the difficult glyphs of characters “س“ and “ش” and needs some improvements in future.

4. CONCLUSION

An efficient and intelligent segmentation algorithm for segmenting Sindhi handwritten characters has been presented with a high accuracy for line segmentation and word segmentation. An acceptable accuracy for character segmentation has also been presented which needs some improvements so that the character segmentation errors can be reduced for the sake of increasing accuracy. This study is a part of Sindhi handwritten character recognition system resulting more challenging steps of feature extraction and recognition which is beyond the scope of this study but their importance cannot be ignored. The refinement and increased accuracy will result the increased recognition rate can be considered as the future work of this study.

REFERENCES:

Akram, M., S. Hussain, (2010). Word segmentation for Urdu OCR system, Proceedings of the 8th Workshop on Asian Language Resources, Beijing, China, pp. 88–94.

Assabie, Y. J. Bigun, (2011), 'Offline handwritten Amharic word recognition ', Pattern Recognition Letters 32(8), 1089 - 1099.

Awan, S. A., D. N. Hakro, I. A. Lashari, A. H. Jalbani, M. Hameed, (2017), ‘A Comprehensive Database for Offline Sindhi Handwritten Text Recognition’, Case Studies Journal ISSN (2305-509X) – Volume 6, Issue 3 72-82.

Bhatti, Z., I. A. Ismaili, D. N. Hakro, A. Waqas, (2014), 'Unicode Based Bilingual Sindhi-English Pictorial Dictionary for Children', American Journal of Software Engineering, 2 (1), 1-7.

Bhatti, Z., I. A. Ismaili, W. J. Soomro, D. N. Hakro, (2014), 'Word Segmentation Model for Sindhi Text', American Journal of Computing Research Repository2(1), 1--7.

Bhatti, Z., A. Waqas, I. A. Ismaili, D. N. Hakro, W. J. Soomro, (2014), 'Phonetic based SoundEx&ShapeEx algorithm for Sindhi Spell Checker System', arXiv preprint arXiv:1405.3033.

Hakro, D. N. and I. A. Ismaili, (2014), 'Issues and Challenges in Sindhi OCR', Sindh University Research Journal(Science Series) 46(2).

Hakro, D. N., I.A.Z. Talib, Z. Bhatti, G. N. Mojai, (2014), A Study of Sindhi Related and Arabic Script Adapted languages Recognition', Sindh University Research Journal (Science Series) 46(3), 323-334.

Hakro, D. N., Z. Talib, G. N. Mojai, (2015), 'Multilingual Text Image Database for OCR ', Sindh University Research Journal (Science Series) 47(1), 181-186.

Hakro, D. N., A. Z. Talib, (2016), 'Printed Text Image Database for Sindhi OCR', ACM Trans. Asian Low-Resour. Lang. Inf. Process.15(4), 21:1--21:18.

Hakro, D. N., (2015) "Enhanced Segmentation and Feature extraction approaches for Sindhi Optical Character Recognition", PhD thesis Dissertation submitted to Universiti Science Malaysia (USM), Malaysia.

Hamid, A., R. Haraty, (2001) "A Neuro-Heuristice Approach for Segmenting Hand written Arabic Tex", ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 110-113.

Lehal, G. S., A. Rana, (2013). Recognition of Nastalique Urdu ligatures, Proceedings of the 4th International Workshop n Multilingual OCR, MOCR' 13, ACM, Washington, DC, USA, 7:1–7:5. URL: http://doi.acm.org/10.1145/2505377.2505379

Li, J., Z. Lu, A.Yimiti, F. Tan, (2012). Handwritten Uyghur character segmentation and performance evaluation, Proc. SPIE 8349, Fourth International Conference on Machine Vision (ICMV 2011): Machine Vision, Image Processing, and Pattern Analysis, 83491E (January 11, 2012); doi: 10.1117/12.920349, International Society for Optics and Photonics, Singapore, 83491E–83491E.

Pal, U., A. Sarkar, (2003), Recognition of printed Urdu script, in 'Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)', Computer Society, Edinburgh, Scotland, Edinburgh, UK, 1183 -1187.

Shaikh, N., G. Mallah, Z. Shaikh, (2009). Character segmentation of Sindhi, an Arabic style scripting language, using height profile vector, Australian Journal of Basic and Applied Sciences 3(4): 4160–4169.

Zaafouri, A., M. Sayadi, F. Fnaiech, (2012), Printed Arabic character recognition using local energy and structural features, in 'Communications, Computing and Control Applications (CCCA), 2012 2nd International Conference on', 1-5.