International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2390657

Generalized Segmentation Algorithm for Dissimilar Script Languages

[1]Abdul Majid, [1]Qinbo, [2]Dil Nawaz Hakro, [1] Muhammad Owais Khan

[1] Department of Computer Science and Technology, Faculty of Information Science and technology, Ocean University of China

[2]Faculty of Engineering and Technology (FET) University of Sindh, Jamshoro, Pakistan

Corresponding Author: Abdul Majid, Abdul.majid827@yahoo.com

ARTICLEINFO

Article History:

Accepted: 10 Dec 2023

Published: 27 Dec 2023

Publication Issue

Volume 9, Issue 6

November-December-2023

Page Number

303-309

ABSTRACT

Optical Character Recognition is considered one of the fastest methods of data entry. OCR converts the text image representation of x and y coordinates representing pixel information to be converted into text data in a particular language. OCR as a field of pattern recognition and document image understanding, OCR requires a challenging job once a different language text is available on the image. Difference in language script will pose different challenges for OCR which requires entirely different approaches and algorithms. Latin scripts require a different approach whereas the Arabic adopted language scripts require a different approach. In this regard, various solutions have been proposed for different languages. Segmentation is considered one of the important tasks in the process of OCR. A good segmentation will definitely increase the accuracy of an OCR. Segmentation includes the segmentation of text lines from text images which are further divided into words. These segmented words are further divided into characters which are to be recognized. A single segmentation algorithm to segment various scripts of the languages is proposed in this study which checks the script and then segments the text image for the further processing in OCR. The proposed generalized algorithm will check the style, direction and other properties of the script and then adopts the segmentation process to segment text lines, words and characters of the language. The proposed algorithm segments more than ten languages of three scripts and segments for their OCRs. These images can be further processed for feature extraction and classification further. The process of OCR for selected languages will be made easier to recognize. Multiple scripts, languages and images were experimented, and the proposed algorithm successfully segmented 32,833 images of text line, words and character image. The algorithm provides 97% accuracy

while segmenting these images and can be extended to further languages as well as scripts.

Keywords: Multiscript, Languages, Scripts

I. INTRODUCTION

Due to advance in Technology data entry has been made easy because of the invent of Optical Character Recognition (OCR). The OCR converts the text image into editable text. The editable text can be further processed and hence the much of the typing effort is removed. In OCR, after removal of noise and the correction of skewness, the scanned document would be ready for the feature extraction and the segmentation. Segmentation step is of three types the text image divided into text lines called line segmentation, another type of segmentation is the word segmentation in which the text lines are segmented into words. The final type of the segmentation is called character segmentation and the final type of character segmentation approach is available in only those scripts where the characters are connected such as Arabic scripts. The segmentation of roman script is considered as easy task because the characters are not connected where the best examples of these scripts are English, German and Spanish. In some of the scripts like Arabic and its adopting scripts, the character segmentation is very much necessary as the characters inside such scripts are changing their shapes according to preceding and following characters connected. A typical method for the line and word segmentation is the projection method. The vertical and horizontal projection methods are used for character and word segmentation. The common projection method has been used by the Pal and Sarkar (2003) for the segmentation of Urdu language text and Urdu is one of the adopting languages from Arabic script as shown in Figure 1.

The number of black or white pixels are found to segment the lines or the block of the lines and the black pixels typically build the valleys of projection profile. Inside the text image, the free space is utilized to indicate the boundary between two text lines (Pal and Sarkar, 2003). The component labeling and the vertical projections have been used for the character segmentation. The number of pixels available in a line becomes less than two or equal to two is converted to zero and if the pixels are more than the defined number then the pixels are recorded in the column. The vertical projection works in the same manner as the horizontal projection except for the direction (Pal and Sarkar, 2003). The projection approach has been utilized for the Sindhi isolated characters as shown in Figure 2 where the segmentation of Sindhi standalone characters is illustrated (Hakro, 2015).

Figure 1: Projection Method in Urdu (Pal and Sarkar, 2003)

Figure 2: Projection Method used in Sindhi (Hakro, 2015)

1.1 Word Segmentation

The word segmentation is necessary to divide the words from the lines of the text and it is one of the factor which improves the accuracy of the OCR. Height Profile Vector approach has been used by Shaikh et al. (2009) for the Sindhi segmentation. The cursive nature of Sindhi script has been shown as the basic concept of study and for this purpose a thinning of the basic strokes have been used for the segmentation purpose. The sub words from the word have been obtained after multiple phases. The thinning algorithm produces the skeleton of the word and then the text lines are segmented using the projection method employed horizontally. The connected components are detected and on the basis of these components the text has been extracted using extraction method. The final character images are obtained in the form of strokes and these strokes are in the form of thinned binary images. Many of the primary experiments were based on basic six patterns of Sindhi Language.

1.2 Character Segmentation

A segmentation-free approach has been employed by the Ozdil and Vural (1997) in which the segmentation has been avoided and the four staged feature extraction algorithm has been used. The extracted features have been used for the training purpose. The following step is the probability matrix where inverse vector has been used optionally. The classification and the matching process has been performed through matrix vector. The final step is the identification of gaps and the resolution of these gaps.

II. LITERATURE REVIEW

A segmentation algorithm has been proposed for six types of fonts and the images used for experimentation have been scanned in 300 dpi. These images are in binary format (Omidyeganeh et al. ,2005). The text lines have been segmented via projection profile. The contour curvatures have been used. Manually segmented corpora have been used for the segmentation of Urdu script by Akram and Hussain (2010). The boundaries have been identified using statistical model. The lexical lookup has been used for the word sequence. The final decision has been made based on the valid words and the probability processing. Li et al. (2012) proposed the intelligent mobile systems for the Uyghur language and stated as the robust segmentation algorithm. The black connected pixels have been used as the factor of stroke identification followed by the width and height and location of the pixels. The strokes are analyzed, checked for the connections and the affiliation of the strokes to the words. High point detection, local high point and another method Harris corner detection has been used for the identification of potential segmentation areas. The last step is the removal of unwanted strokes. Word Segmentation is typically followed by the character segmentation in segmentation based OCR systems where the text lines are divided into words.

Some of the segmentation approaches are summarized in Table 1.

Table 1: Segmentation approaches and properties

Authors	Approach	Remarks
Erlandson et al. (1996)	48,200 words	A word level segmentation which recognizes only words rather than characters.
Tolba and Shaddad (1990)	Sliding window	A sliding window based on threshold sliding right to left to calculate segmentation parameters.
Cheung et al. (2001)	Recognition based	A recognition based Arabic word segmenation in which feedback link is sent from classification to input stage.
Cavalin et al. (2006)	HMM based	An Arabic system for handwritten numeral strings. The system works only for standalone numerals.
Razak et al. (2007)	Region of interest	Overlapped Jawi character segmentation for a chip. The overlapping problem between the lines is addressed.
Alaei et al. (2010)	Baseline tracing	Persian handwritten character segmentation by applying baseline tracing. Baseline is detected by using projection.
Aghbari and Brook (2009)	Projection	An Arabic holistic approach for document retrieval

Generalized Segmentation Algorithm for multiscript

A single segmentation algorithm for multiple scripts is a challenging task and requires to understand the peculiarities of the multiple scripts especially while segmentation of the characters. In order to improve the accuracy of the segmentation process and overcome some issues and challenges in segmenting Sindhi text, an enhanced multi script segmentation algorithm that segments various script text is proposed. The enhanced and generalized segmentation algorithms take an input image (synthetically created or scanned image) and converts it into grey level format and then into binary format so that image can be represented in only two types of pixels, i.e. white and black pixels. The text image is selected from the database created by using the custom-built application or scanning text images (Hakro and Zawawi,2016) and it is noise free and skew corrected. This step is called preprocessing. The text image is scanned for horizontal moments or projections. The free space is an indicator for segmentation of the text lines. If there is no black pixel or very few black pixels in a row, then the text line is segmented and stored in an array. The image array is created at this stage and these extracted lines are stored in an image array. These stored lines are retrieved one by one and they are then processed in the next steps, i.e. segmentation, feature extraction and recognition.

After preprocessing, the standard algorithm (Ahmad, 2009) segments Urdu characters in one font size only and for a single line whereas the proposed enhanced algorithm solves this problem by using horizontal projection to segment text page into lines, and lines are segmented into words or ligatures and stored in an image array. The white space between the lines are used as a clear demarcation for segmenting the text into lines. The lines are segmented by calculating the number of white pixels and number of black pixels in every row. The horizontal moments are calculated with the help of the tools available in MATLAB 2022b. Conventional approaches are employed so that algorithm can work for more than one line or a paragraph of text. The lines are segmented and stored in an image array so that these lines can be processed line by line for segmentation of words and ligatures in the next step. For this, an image array is created and all the available lines are extracted from a text page and stored in this array. The next step of the segmentation is to process these stored lines one by one for further segmentation. As mentioned, the standard algorithm (Ahmad, 2009) can segment only one font and one size of Urdu text and thus, a very simple if-else rule is required to decide whether a character or a small segment should be dropped. On the other hand, The generalized algorithm proposes the following steps so that it can work on multiple lines, fonts and font sizes

Edge identification: The enhanced algorithm firstly, searches for the edges i.e. the location where a character may start or end. This step involves drawing lines at the start and end of a character in a word or text line. This step is not available in the standard algorithm (Ahmad, 2009) and helps to identify the boundaries of the characters and ease the process of stroke finding and segmentation of strokes.

Finding strokes and segmentation: This step finds the strokes and segments of all of the edges detected by the previous step (Edge identification:). The white lines drawn in edge identification are used in this step in

which characters are found based on the drawn lines. This step checks if there is a character or segment based on the energy level and these segments are segmented as probable characters and the next step (dropping small segments) will decide whether the probable segments are to be merged or dropped based on their size.

Dropping small segments: This step drops or merges the segment of characters based on a specified threshold, which were segmented in the previous step (Finding strokes and segmentation). The threshold is specified based on some preliminary experiments. Two segments are merged to form a single character and some of the segments which are less than the specified threshold is discarded or dropped.

Storing on disk: The segments which were not discarded are connected and merged to store on the disk. The final step has been added just for the sake of analysis and checking of the step-by-step mechanisms of the algorithm. These stored segments can easily be used for the other experiments.

III. RESULTS AND DISCUSSION

The proposed algorithm has produced a promising result as pr standards. The character size plays a vital role in the accuracy of the proposed segmentation algorithm. The Seen character (one of the character in Sindhi alphabet) has challenged the algorithm and produced many challenges for the segmentation algorithm and this challenge was because of the cursiveness. The results comprising the number of correctly segmented characters and accuracy of the segmentation are given in subsequent sections. The algorithm was applied on a wide variety of fonts and various font sizes. Some of the visual results are presented here. The text image page is segmented into text lines. Figure 3 (a) shows the text image containing multiple lines in MB Latifee font while Figure 3 (b) shows the text lines extracted from the image of Figure 3 (a).

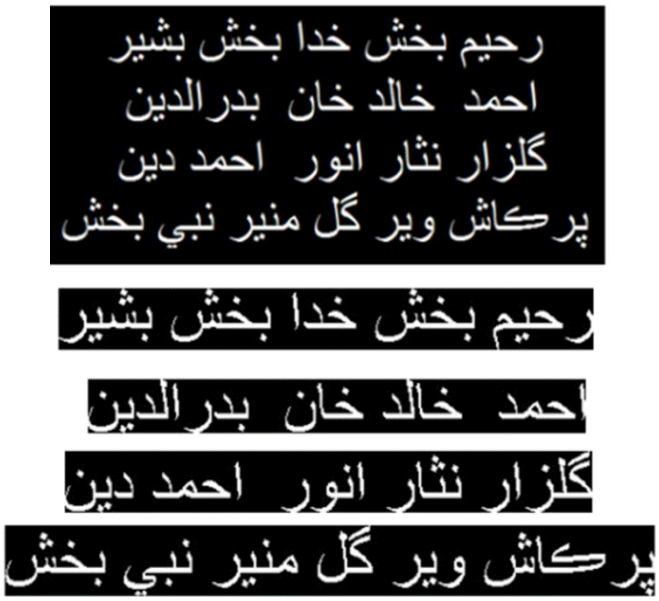

Figure 3 : (a) Text image in Sindhi language (b) segmented lines from the text image

Arabic Script

Figure 4 presents the Arabic script font titled as Thabit. Many of the other fonts have been used for the experimentation including Karbala, courier New, Unicode MS, Arial, Arabic Casting, Globatec1, Siddiqua and many others. Figure 5 is the illustration of text image segmentation on Persian script

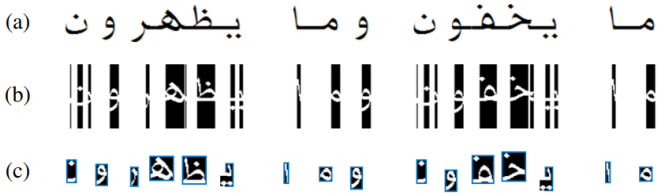

Figure 4: Segmentation of Arabic characters from Arabic text line (Thabit Font) (a) Original text image (b) Edge identification (c) Segmented characters

Figure 5: Segmentation of Persian characters from Persian text line (a) Original image (b) Edge identification (c) Segmented characters

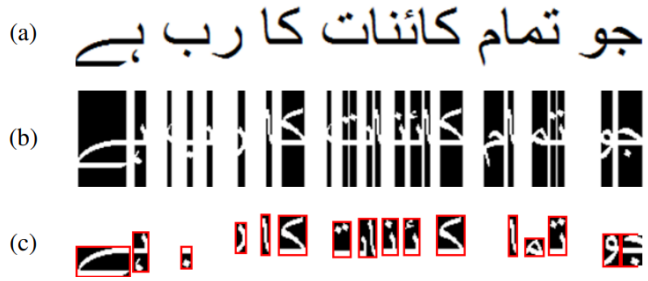

Figure 6: Segmentation of Persian characters from Persian text line (a) Original image (b) Edge identification (c) Segmented characters

Figure 7: Segmentation of Uyghur characters (a) Original image (b) Edge identification (c) Segmented characters

Latin Script

The Latin alphabet contains 26 basic characters which are extended to some extent by various writing systems. The experiments performed on Latin script languages and their results are presented next. Figure 8 depicts the results for French script segmentation, Figure 9 shows the segmentation results for Swedish script segmentation Figure 10 shows the segmentation results for Greek Characters. As discussed earlier in this thesis, Latin OCRs are mature and near to perfection. So, there is no problem at all in segmenting Latin script characters because these characters are isolated and there are no overlapping characters. The isolated character segmentation accuracy can be seen in Latin script-based languages. Hence, these language OCRs can also achieve higher recognition rates.

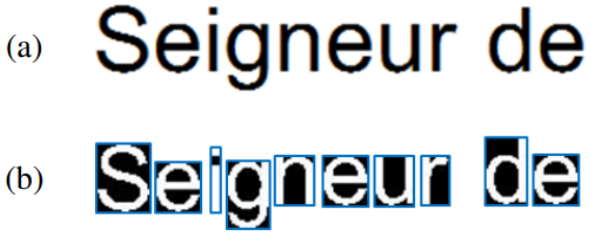

Figure 8: Segmentation of French characters (a) Original image (b) Segmented characters

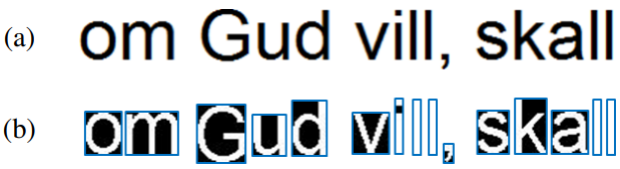

Figure 9: Segmentation of Swedish characters (a) Original image (b) Segmented characters

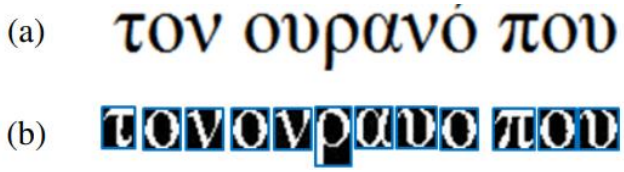

Figure 10: Segmentation of Greek from text line(a) Original image (b) Segmented characters

Indian Scripts

There are various names for the Indian scripts such as Indic or Brahmic. Most of the Indian languages adopt such type of script that's why it is called as Indian script. These group of languages are mostly spoken in India. Malayalam, Bengali, Kannada, Assamese, Gujrati, Punjabi, Marathi, Tamil and Telugu are the examples of Indian scripts.

The languages such as Assamese, Gujarati, Oriya, Kannada, Bengali, Malayalam, Marathi, Punjabi, Telugu and Tamil are some examples of the Indian scripts. Some experiments were performed on the Indian scripts and some of the results are presented in Figure 11, and Figure 12 are the results of Kannada, and Tamil scripts segmentation respectively.

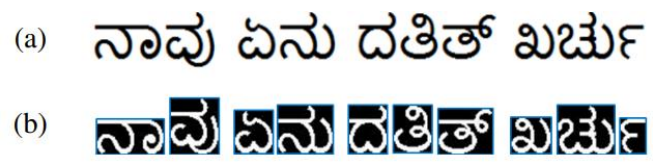

Figure 11: Segmentation of Kannada characters (a) Original image (b) Segmented characters

Figure 12: Segmentation of Tamil characters (a) Original image (b) Segmented characters

All of the characters in Figure 11 are correctly segmented and no problem is encountered due to the nature of the isolated script. The Tamil characters are also segmented properly as shown in Figure 12.

IV. REFERENCES

[1]. Bag, S., & Harit, G. (2011). An improved contour-based thinning method for character images. Pattern Recognition Letters, 32(14), 1836-1842.

[2]. Cavalin, P. R., de Souza Britto, Jr., A., Bortolozzi, F., Sabourin, R. and Oliveira, L. E. S. (2006). An implicit segmentation-based method for recognition of handwritten strings of characters, Proceedings of the 2006 ACM Symposium on Applied computing, SAC '06, ACM, Dijon, France, pp. 836–840. URL: http://doi.acm.org/10.1145/1141277.1141468

[3]. Cowell J. and H. Fiaz (1992). "Thinning Arabic character feature extraction", IEEE Transactions on Pattern Analysis Machine Intelligence, Vol. 14, No.11, 869-885,

[4]. Fan, X. and Verma, B. (2001). Segmentation vs. non segmentation based neural techniques for cursive word recognition: an experimental analysis, Computational Intelligence and Multimedia Applications, 2001. ICCIMA 2001. Proceedings. Fourth International Conference on, IEEE, Yokusika City, Japan, pp. 251–255.

[5]. Hakro (2015), ENHANCED SEGMENTATION AND FEATURE EXTRACTION FOR SINDHI OPTICAL CHARACTER RECOGNITION, PhD thesis, Submitted to University science Malaysia (USM), Malaysia.

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

[7]. Premaratne, H. and Bigun, J. (2004). A segmentation-free approach to recognise printed Sinhala

[8]. Script using linear symmetry, Pattern recognition 37(10): 2081–2089.

[9]. Shang, L. and Z.Yi, (2007)."A class of binary images thinning using two PCNNs", Neurocomputing, Vol.: 70, 1096-1101,

[10]. Zhang T. Y. and C. Y. Suen, (1984). "A fast Parallel Algorithms for Thinning Digital Patterns", Research Contributions, Communications of the ACM. 27 (3): 236-239

Cite this article as :

Abdul Majid, Qinbo, Dil Nawaz Hakro, Muhammad Owais Khan, "Generalized Segmentation Algorithm for Dissimilar Script Languages", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 6, pp.303-309, November-December-2023. Available at doi : https://doi.org/10.32628/CSEIT2390657

Journal URL : https://ijsrcseit.com/CSEIT2390657