Research Article

Enhancing OCR: A Novel Segmentation Approach for Pashto Text Images into Characters

Arbab Waseem Abbas[1]*, Waseem Ullah Khan[1], Huda Nauman[1], Khalid Saeed[2] and Syed Mujtaba Abbas Zaidi[3]

[1]Institute of Computer Science and Information Technology, Faculty of Management and Computer Sciences, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan; [2]Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Upper Dir, 18200, Pakistan; [3]FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.

Abstract: This paper presents a novel approach to segmenting typed Pashto text images into individual characters, addressing a critical challenge in Optical Character Recognition (OCR) for this language. Pashto, a right-to-left, highly cursive language similar to Arabic and Urdu, poses unique segmentation difficulties due to the variable shapes and forms of its characters depending on their position in a word. The segmentation of Pashto characters remains an underdeveloped area in language processing, significantly hindering OCR performance. To tackle this, an image database of isolated Pashto characters was created. Pashto text samples were generated in Microsoft Word, with images saved in Bitmap (BMP) format for processing. These images were preprocessed, converting them to binary form and removing noise. These preprocessed images were then segmented into their constituent characters by the proposed algorithm. The proposed algorithm measure pixels strength to segment words into characters. The algorithm achieved a segmentation accuracy of 84.6%, verified through manual analysis, although some new and unwanted characters (garbage) were also generated. This work contributes a significant step toward improving OCR for the Pashto language, offering a reliable method for character segmentation, which is fundamental to the development of an accurate Pashto OCR system.

Received: October 02, 2024; **Accepted:** November 15, 2024; **Published:** November 26, 2024
***Correspondence:** Arbab Waseem Abbas, Institute of Computer Science and Information Technology, Faculty of Management and Computer Sciences, The University of Agriculture, Peshawar, Khyber Pakhtunkhwa, Pakistan; **Email:** aristocratarbab@aup.edu.pk
Citation: Abbas, A.W., W.U. Khan, H. Nauman, K. Saeed and S.M.A. Zaidi. 2024. Enhancing OCR: A Novel segmentation approach for Pashto text images into characters. *Journal of Engineering and Applied Sciences,* 43: 54-64.
DOI: https://dx.doi.org/10.17582/journal.jeas/43.54.64
Keywords: Segmentation, Pashto text, OCR, Characters

Introduction

Segmentation of characters of a written language has for a long time been the most crucial area of Optical Character Recognition (OCR) process in language analysis. In this process, images of a sequence of characters, connected with each other to give a semantic signal to a machine, are broken up into sub-images of individual alphabetic symbols (Choudhry, 2014). One of the main fields of pattern

recognition is OCR in Artificial Intelligence. The basic process is composed of scanning and converting handwritten, typed or printed text material into machine recognizable symbols. Efforts are underway to develop OCR for different languages of the world. Many different OCR systems have been suggested and tried for Chinese, Japanese English and other similar languages that employ non-cursive standalone characters for word formation (Choudhary *et al.*, 2017). Segmentation of text into characters is a crucial decision method for optical character recognition (OCR). Detaching single alphabetic symbols in the script image is important for the success of the whole system. If the words are segmented properly into their constituent characters, the overall recognition rate will be very high otherwise poorly segmented words results in a low recognition rate.

Pashto language is a member of the Indo-European group of languages, and it is directly descended from its East-Iranian branch. It is the official language of Afghanistan and is widely used in Khyber Pakhtunkhwa and northern Baluchistan province of Pakistan. It has got a rich record of poetry and folklore extending back about 700 years from modern times (Kanaparthi and Raja, 2022). The written literature available in this language mainly deals with different fields like poetry, religion, history, biography and culture. Because of the lesser political and economic importance of the language there has been very little attention paid to this language in the field of modern technology like computer translation of the text, speech recognition text analysis and so on. As the western world and especially USA gets more and more involved in the power politics of the geographical area populated by the people speaking Pashto, so an interest has awakened in the production of OCR for Pashto. As the nature of Pashto is such that the number of phonemes in it is more than Arabic, Persian or Urdu so the production of OCR may prove more complex and problematic than them. The characters of Pashto script are connected along a horizontal line extending from right to left whether it is a printed script or a handwritten one (Memon *et al.*, 2020). Pashto follows the method of Arabic, Persian and Urdu languages in characters joined to each other in sub-words, words and sentences on a baseline with diacritic marks appearing above and below the sub-words. An example is the joined-up script of Latin in handwriting which is also cursive in nature. This natural aspect of connectivity on a line presents the greatest challenge to OCR designing and producing segmentation algorithms. Figure 1 depicts a sample of Pashto text taken from BBC news website (Al-Yahya *et al.*, 2022).

د ایالتونو متحده د ته نوموري کي خبرو ټلیفوني په سره مودي نرېندر وزیر لومړي له هند د ټرمپ ډونلډ ولسمشر امریکا د
.ده کړي ور بلنه لیدنې

ولسمشر ټرمپ په نړۍ کي د مخ په وراندې ننګونو د لرې کولو لپاره هند ته د امریکا د "د سپینې ماڼۍ بیانیه وایي،
"رېښتوني ملګري او ملاتړي په سترکه ګوري

Figure 1: *Pashto script sample.*

Many Pashto characters have sub parts in the shape of diacritic marks appearing over or under main part of the character. The script that is handwritten has a lot of variation in the positioning of these marks. There are 45 letters in Pashto language and one dochashmi hye as distinct characters. Instead of vowels, it has 4 diacritic marks which may occur at different positions in a word influencing its semantic value as shown in Table 1.

Table 1: *Pashto characters.*

ح	چ	ج	ث	ټ	ت	پ	ب	ا
ژ	ز	ړ	ر	ډ	د	څ	خ	خ
ع	ظ	ط	ض	ص	ښ	ش	س	ښ
ڼ	ن	م	ل	ګ	ک	ق	ف	غ
۱	ئ	ی	ې	ی	ي	ء	ه	و
۰	۹	۸	۷	۶	۵	۴	۳	۲

The four forms of characters according to their location in the sub-word i.e., isolated, initial, middle, final is shown in Figure 2.

Character	S	I	M	F
ALEF	ا	ا	---	ا
BEH	ب	بـ	ـبـ	ـب
JEEM	ج	جـ	ـجـ	ـج
SEEN	س	سـ	ـسـ	ـس
QAF	ق	قـ	ـقـ	ـق
KAF	ک	کـ	ـکـ	ـک
KEHEH	ک	کـ	ـکـ	ـک

Figure 2: *Four forms of Pashto characters.*

The characters, their names and forms according to their location in the sub-word i.e., isolated, initial, middle, final are shown in Table 2.

Table 2: Pashto characters forms and names.

S. No.	Character حرف	Forms شکلونه	Name نوم	Name
1	ا	ا	الف	alif
2	ب	ببب	بي	be
3	پ	پپپ	پي	pe
4	ت	تتت	تي	te
5	ټ	ټټټ	ټي	tay
6	ث	ثثث	ثي	ce
7	ج	ججج	جيم	jim
8	چ	چچچ	چي	che
9	ح	ححح	حي	he
10	خ	خخخ	خي	khe
11	څ	څڅڅ	څی	cey
12	ځ	ځځځ	ځي	zey
13	د	د	دال	dal
14	ډ	ډ	ډال	dhal
15	ذ	ذ	ذال	zal
16	ر	ر	ري	re
17	ړ	ړ	ړي	rhe
18	ز	ز	زي	ze
19	ژ	ژ	ژي	zhe
20	ږ	ږ	ږی	ghe
21	س	سسس	سين	seen
22	ش	ششش	شين	sheen
23	ښ	ښښښ	ښين	kheen
24	ص	صصص	صواد	swad
25	ض	ضضض	ضواد	dwad
26	ط	ططط	طوي	twe
27	ظ	ظظظ	ظوي	zwe
28	ع	ععع	عين	ain
29	غ	غغغ	غين	gain

When characters set of Pashto language is studied it shows that most of the character's primary stroke are similar but are differentiated from each other by using dots or symbols above or below these forms. Table 3 shows groups of similar characters.

Table 3: Groups of similar characters.

S.	Characters	Groups	S.	Characters	Groups
1	ا	1	24	ص	7
2	ب	2	25	ض	
3	پ		26	ط	8
4	ت		27	ظ	
5	ټ		28	ع	9
6	ث		29	غ	
7	ج	3	30	ف	10
8	چ		31	ق	
9	ح		32	ک	11
10	خ		33	ګ	
11	څ		34	ل	12
12	ځ		35	م	13
13	د	4	36	ن	14
14	ډ		37	ڼ	
15	ذ		38	و	15
16	ر	5	39	ه,ﻫ	16
17	ړ		40	ھ	17
18	ز		41	ء	18
19	ژ		42	ي	19
20	ږ		43	ى	
21	س	6	44	ې	
22	ش		45	ۍ	
23	ښ		46	ئ	

From Table 3, it can be observed that there are actually nineteen groups of forty-five Pashto characters. The study of other forms (initial, middle and final) of these characters shows that ein might be confused with hamza similarly wow might be mixed up with ze resembles noon and mem can be confused with middle form of ein and with standalone ghe.

Segmentation of Pashto script is a very complex process because characters often merge into each other and there are vertically overlapping words. Pashto script is based on Arabic alphabet, with some borrowing from Persian and some indigenous additions is highly cursive. It proceeds from right to left in a diagonal line and its being very highly context sensitive makes it very hard for OCR segmentation. Segmentation of Pashto script images is necessary

for recognition purposes. The process of isolating, in a script image, its various individual characters is a very significant step towards the successfulness of the whole system (Chooi and Asb, 2021). In this research work segmentation of typed Pashto image text into characters for the first time up to author knowledge will be performed.

The main *contribution* of this paper is as follows:

- To propose an algorithm for segmenting Pashto text images into its constituent characters.
- To collect and develop database from samples of Pashto typed image text and noise removal in preprocessing stage.
- To segment images of typed Pashto sentences into its constituent characters using proposed algorithm.
- To assess the performance/accuracy of the proposed algorithm manually in percentage.

The rest of the paper is structured as follows: Section II provides the existing work done related to segmentation in different languages. Section III presents the proposed algorithm for the segmentation of Pashto script. The results and performance analysis of proposed algorithm is described in Section IV. Section V is about the concluding remarks and future work.

Related work

This section provides reviews of different work done and techniques applied for segmentation on different languages around the world. The algorithm tried by (Al-Wajih and Ghazali, 2023) for the process of segmentation was to look for existence of an angle which was produced by connecting two characters. This angle is produced at the baseline. This baseline is the horizontal profile made up of the highest number of black pixels. Cursive handwritten script of Bengali language is segmented in this paper. Bengali handwritten characters, in contrast to English characters, and its components frequently surround the central character, causing the traditional segmentation procedures unsuitable. The connected component-labeling algorithm is implemented to isolate each preliminary segment of words by (Akbari *et al*., 2017). A new method for Arabic character segmentation 'ACSA', has been presented which is based on morphological analysis of word contours. Topological characteristics of Arabic text were used to find out morphological rules. These rules can then be utilized to identify accurate segmentation points. Through this method new and trustworthy techniques have been developed to segment characters. This method assumes that each segment contains only one character (Sarwar *et al*., 2022).

In Humayun *et al*. (2024), authors used different approaches of taking into consideration the vertical and horizontal segmentation of text in order to segment a page into individual lines and further on into sub-words, respectively. Al-Hamad *et al*. (2024) have come up with another solution. They point out that there are 2 basic approaches in segmentation; they are segmentation followed by recognition and recognition based on segmentation. This latter approach can be adapted to various situations because it is more inclusive. Also, there are 2 different approaches within the segmentation-based approach. They are called explicit segmentation and implicit segmentation. It is found that implicit approach is more subtle to the variation of fonts and is specially characterized by increase of complexity in large computation. In Preethi *et al*. (2023), authors have suggested a different probabilistic segmentation model. First step is to perform a contour-based over-segmentation, which cut the word image into small sized images. The images are arranged into 3 queues, which are main part and sub-part characters diacritics above or below main parts, respectively. The probabilistic model calculates the confidence for each character by taking into consideration the output of the recognizer and geometric confidence with all the constraints. Global optimization is then employed to obtain optimal cutting path (Qaroush *et al*., 2022; Khan *et al*., 2022). Weighted average of character confidence is taken as objective function. Trials with Arabic handwritten scripts with various styles have given encouraging results showing that the method is more effective.

In Alghyaline (2022), authors provide a concise survey of different segmentation techniques used for words and sub-words segmentation into its constituent characters in Arabic like cursive scripts. Different approaches were explored that may assist scholars in OCR to discover new concepts and provide new answers to issues of the languages like Persian or Urdu text segmentation. Analytical approach was found to be difficult one owing to over and under segmentation problems which yield incorrect sub-words recognition. Some scholars suggest

segmentation free approaches, and some are in favor of segmentation at recognition time. Segmentation is one of the crucial phases of any OCR system where a scanned text image is segmented into its constituent characters. The proposed segmentation algorithm first preprocesses the image by correction of skew angle and thinning process to achieve the single pixel stroke width. After detection of the ligatures individual characters are cut from the cursive text by vertical segmentation technique (Nguyen *et al.*, 2022; Khaliq *et al.*, 2023; Abbas *et al.*, 2024). In Sakshi and Kukreja (2023), authors suggested a system consisting of a pipeline of three neural networks for Arabic OCR. The first neural network models calculate the font size of Arabic text and then change it to standard 18thpt font size. The next model segment the cursive text into characters. Third and last convolutional neural network model takes as an input the segmented characters and produces the recognized Arabic characters with an accuracy of 94.38%.

Material and Methods

This paper will take Pashto text images written in MS Word software and segment the cursive letters. Segmentation of Pashto numerals has not been considered in this research work as they are standalone symbols. The proposed process of segmentation consists of several structured modules (Ahmad *et al.*, 2015). The workflow diagram of these modules has been shown in Figure 3. The following subsections discuss these modules in detail.

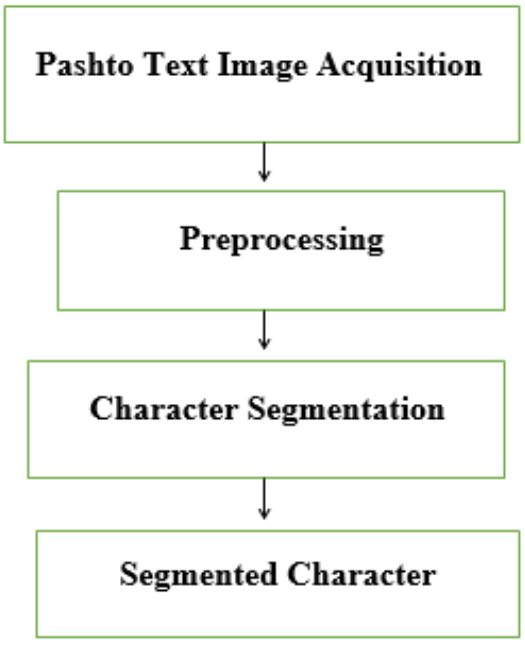

Figure 3: *Workflow of the segmentation process.*

Pashto text image acquisition

Pashto text will be written in Microsoft Word application. Text font Arial of size 36 will be used to write Pashto text. Pashto sentences will be cropped by snipping tool and saving them in bmp format as shown in Figure 4.

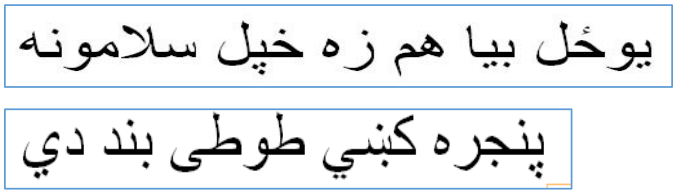

Figure 4: *Images of Pashto text.*

Preprocessing and proposed segmentation algorithm

Several pre-processing steps will be performed on the text image before actual segmentation process starts. Firstly, the input typed text image will be converted into its corresponding binary equivalent data or gradient image as shown in Figure 5. A binary image is one in which there are only two colors, normally with a value of zero corresponding to a black pixel and a value of one corresponding to a white pixel. Area above and below the text will be discarded. It is assumed that there will be no noise and skewing in the text image as the images will be directly taken from MS Word in BMP format. This format is chosen because it is machine independent.

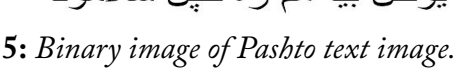

Figure 5: *Binary image of Pashto text image.*

Binary image will be inverted to remove noise as shown in Figure 6.

Figure 6: *Inverted binary image.*

The inverted image will be inverted back for segmentation as shown in Figure 7.

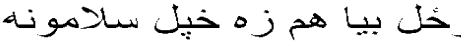

Figure 7: *Pashto image inverted back for segmentation.*

The proposed general segmentation algorithm is given as follows and details can be found in the upcoming sections.

Step 1: Take Pashto image sentence from database

Step 2: Do preprocessing
Step 3: Calculate seam strength and seam vector: Scan last row from right to left for pixel==0 when found turn into the (i,j)"th pixel, search in the (i-1)"th row for the pixel==0 in the 3 adjacent pixels of (i, j)"th pixel and save the one found first to the seam path. Set the trail of pixels that constitute the seam to 1.
Step 4: Character segmentation: If seam energy is less than average energy of the columns perform vertical segmentation else not consider. If more than one contiguous seam found, segment middle of these seams.
Step 5: Character test: If (segment< threshold) merge with earlier segment, else perform horizontal segmentation output character image when found in knowledgebase.

Character segmentation
Those seams will be given precedence for words segmentation which is vertically straight. Vertical seams will be given priority for character segmentation but if a segment of larger size than that of a threshold value is encountered then in the same segment horizontal seams will be employed for further segmentation. For words segmentation zero level energy is selected whereas for character segmentation energy of the seam is calculated and compared with the average energy of the columns for vertical segmentation. If the energy is less than the seam is chosen for segmentation else not considered. If more than one contiguous seam is found, the middle of these seams is selected for segmentation. Segmentation of the image for words or characters is performed by the selection of these seams as shown in Figure 8.

Figure 8: *Segmented characters.*

Garbage characters
During the process of segmentation, a keen examination is carried out to segment a complete character. If the character is segmented in parts these are merged with its other parts, if two are more than two characters are segmented as a single character these are further segmented to get single characters. This results in the production of some garbage characters. These are unessential and unwanted sections of a characters, which the algorithm is incapable to combine with its pertinent sections and treated as characters till they are pronounced as garbage characters in the recognition phase. Figure 9 shows a line of Pashto text (top) and segmented characters (bottom) from the above Pashto sentence image. Correctly segmented characters and garbage characters yielded during the line of action can be seen clearly.

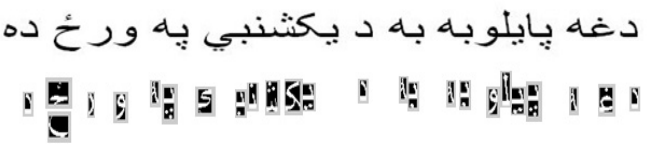

Figure 9: *Line of Pashto text (top) segmented characters (bottom).*

The Pashto sentence image in Figure 9 consists of 26 characters. The total number of correctly segmented characters from the image sentence is 24. Character (ا) at 5th position and ghe at the end in the image is not output as a character after segmentation process. The reason is its small segment size. Character zey at 24th position in the image sentence is segmented into two images. As character zey is larger than the threshold value so it went through horizontal segmentation and results into one required character image and other is garbage character. Segmentation accuracy for the above given sample image is 92.3% which is quite promising.

Results and Discussion

In this section, segmentation algorithm results have been discussed in detail. Successfully and unsuccessfully segmented characters, some new characters and unwanted segments known as garbage characters have been discussed in detail in this section.

Algorithms applied for segmentation produce good results. The algorithm produces segmentation accuracy of about 84.6% when examined through human eye. Segmentation of character family of bee, pee, tee, tay, cee and fee are around 80% similar is the case of character family of kaf and gaf. Character noon when used in the beginning it looks like ze and re and thus produces low results. Similarly, character ghee is mostly misunderstood as noon and be. The character lam act like alif when it is used in middle of a word. However, alif is not mistaken as lam in most cases.

Garbage characters generates after the segmentation of seen, sheen, swad, dwad and noon which usually

clears the character test during segmentation phase, whereas bee, pee, tee, tay, cee and fee also generates garbage characters as shown in Table 4 but for most of the time they are identified as garbage characters. Also, these characters produce garbage only when they are located at the end of a word.

Table 4: *Characters and garbage generated.*

Characters	Garbage\Part
be, pe,te, tayand say (ٹ ټ ، ب)	
ghee (ږ)	
seen sheen, swad, dwad (ض،ص ښ ، ش س)	
Noon (ڼ)	
yee ( )	

The combination of alif and lam as used in words like and results in a new characters, in the segmentation phase as shown in Figure 10. This needs careful treatment.

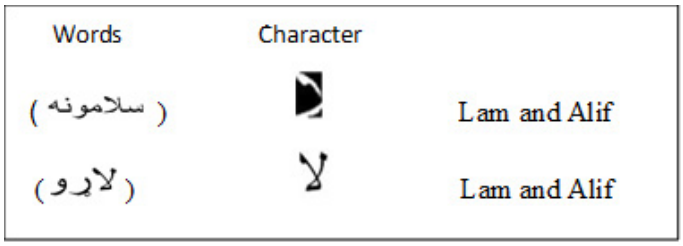

Figure 10: *Lam and Alif after combining resulting in new characters.*

A total of 25 sample image sentences were segmented by the proposed algorithm. Out of 501 total characters in all samples 424 has been correctly segmented whereas 77 characters were either not or incorrectly segmented. Ratio of correctly segmented characters by the given algorithm is 84.63%. Table 5 shows total number of characters in each sample along with accurately and inaccurately segmented characters in each sample image. Some characters are not segmented as their size may be too small and are sometimes ignored by segmentation algorithm e.g., like alif and ghe. Some characters may look like other characters and are inaccurately segmented e.g., Character ghee as usually is segmented as be and noon.

Accuracy results of each Pashto characters by the proposed segmentation algorithm in each of twenty-five samples is shown with the help of column chart in Figure 11.

Table 5: *Accurately and inaccurately segmented characters in each sample text image.*

Sample No	Total Char	Correct	Incorrect	problems in characters
Sample No	Total Characters	Correctly Segm	Incorrectly Segm	Problematic characters
sample 1	21	17	4	yee, alif,lam
sample 2	9	8	1	wow
sample 3	17	13	4	hee, lam, alif
sample 4	17	14	3	tee, lam, alif
sample 5	21	18	3	pee, re, yee
sample 6	11	10	1	alif
sample 7	13	12	1	alif
sample 8	19	12	7	pee,yee,alif,dhaal
sample 9	26	21	5	wow,alif,pee
sample 10	17	15	2	alif
sample 11	11	10	1	pee
sample 12	17	14	3	pee,alif,dhaal
sample 13	17	15	2	pee, hee
sample 14	27	22	5	zwe,mem,alif,pee
sample 15	30	25	5	alif,hee
sample 16	23	21	2	alif
sample 17	25	23	2	ye,noon
sample 18	27	24	3	ye,pe
sample 19	25	18	7	noon, lam, ye, ze,kaf
sample 20	14	11	3	noon,tee, kaf
sample 21	12	11	1	lam
sample 22	16	15	1	jeem
sample 23	26	24	2	te,lam
sample 24	30	28	2	te ,noon
sample 25	30	23	7	pe,noon,be,ye,lam
Total	501	424	77	
Ratio of correctly segmented charaters				84.63073852

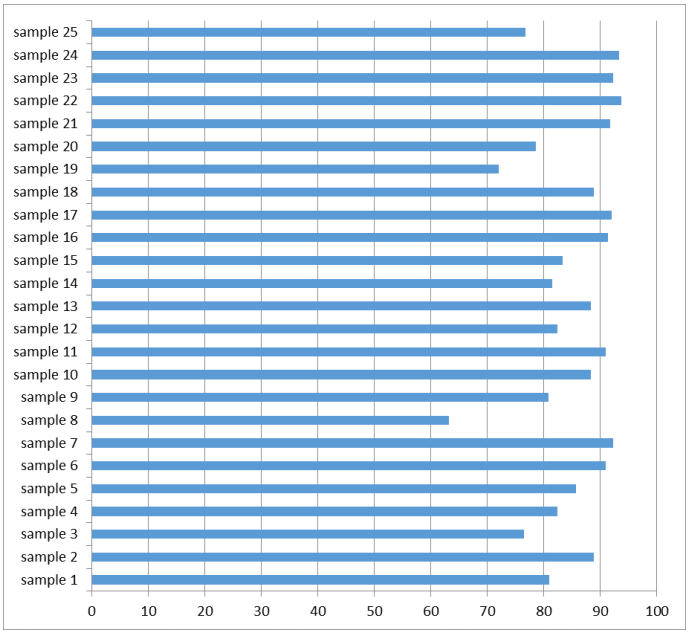

Figure 11: *Sample wise segmentation % age.*

Accuracy results of each Pashto characters by the proposed segmentation algorithm is shown with the help of a column chart in Figure 12.

The software has been simulated on different text image samples. Total number of characters, accurate and inaccurate characters, new and garbage characters yielded after segmentation of few samples typed text images has been discussed in detail as follows.

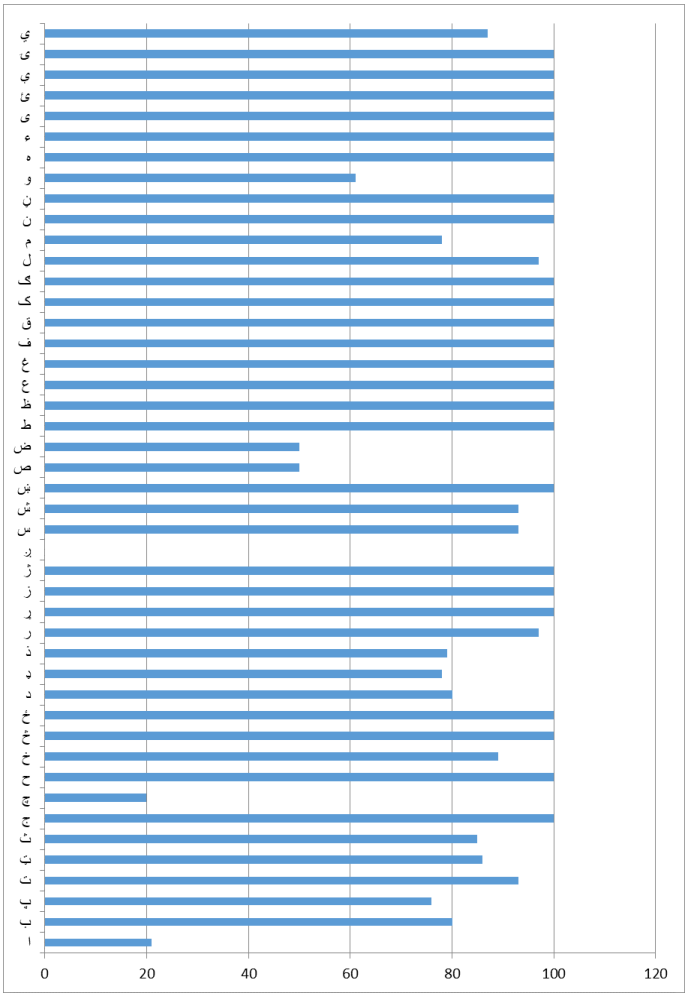

Figure 12: *Character-wise segmentation % age.*

Segmentation results of the sample Pashto text image given in Figure 13 have been discussed below.

Total number of characters in sample image: 30
Total number of accurately segmented characters from the sample image: 26
Total number of inaccurately segmented characters from the sample image: 4
Total number of new characters yielded after segmenting the sample image: 0
Total number of garbage characters yielded after segmenting the sample image: 4
Ratio of correctly segmented characters by the given algorithm: 86.6%

Figure 13: *Pashto sample image (top) and segmented characters (bottom).*

It can be seen clearly that character at 3rd and 18th position in the sentence is completely ignored in the segmentation process. Similarly, when it is followed by noon produces garbage character.

Segmentation results of the sample Pashto text image given in Figure 14 have been discussed below:
Total number of characters in sample image: 18
Total number of accurately segmented characters from the sample image: 15
Total number of inaccurately segmented characters from the sample image: 3
Total number of new characters yielded after segmenting the sample image: 1
Total number of garbage characters yielded after segmenting the sample image: 3
Ratio of correctly segmented characters by the given algorithm: 83.3%

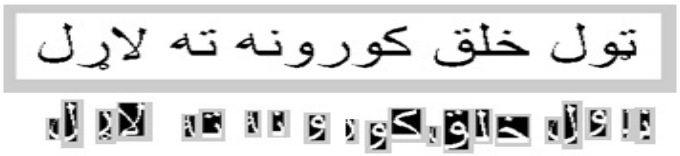

Figure 14: *Pashto sample image (top) and segmented characters (bottom).*

It can be observed that a new character and garbage character which is part of lam results from the segmentation process.

Segmentation results of the sample Pashto text image given in Figure 15 have been discussed below:
Total number of characters in sample image: 26
Total number of accurately segmented characters from the sample image: 24
Total number of inaccurately segmented characters from the sample image: 2
Total number of new characters yielded after segmenting the sample image: 0
Total number of garbage characters yielded after segmenting the sample image: 1
Ratio of correctly segmented characters by the given algorithm: 92.3%

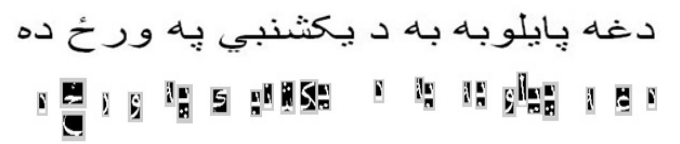

Figure 15: *Pashto sample image (top) and segmented characters (bottom).*

Segmentation process produces only one garbage character which is part of the character.

Segmentation results of the sample Pashto text image given in Figure 16 have been discussed below:
Total number of characters in sample image: 20
Total number of accurately segmented characters from the sample image: 15
Total number of inaccurately segmented characters from the sample image: 5
Total number of new characters yielded after segmenting the sample image: 0
Total number of garbage characters yielded after segmenting the sample image: 2
Ratio of correctly segmented characters by the given algorithm: 75%

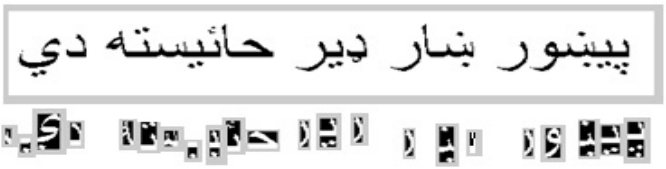

Figure 16: *Pashto sample image (top) and segmented characters (bottom).*

Segmentation process produces garbage character which is part of the character which is part of character. Character ي at 15th and 20th position in the sentence image is inaccurately segmented.

Segmentation results of the sample Pashto text image given in Figure 17 have been discussed below:
Total number of characters in sample image: 14
Total number of accurately segmented characters from the sample image: 10
Total number of inaccurately segmented characters from the sample image: 4
Total number of new characters yielded after segmenting the sample image: 0
Total number of garbage characters yielded after segmenting the sample image: 0
Ratio of correctly segmented characters by the given algorithm: 71.4%

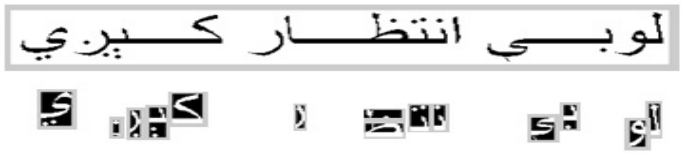

Figure 17: *Pashto sample image (top) and segmented characters (bottom).*

Character if it is in either in start or last position is not segmented from the image. The reason may be because of its small size. Character ghee as usually is segmented as be and noon.

The objective of construction of a knowledge base has been successfully achieved by collection of one hundred images against 57 classes of Pashto characters. In this section the results of proposed segmentation algorithm for Pashto text image into characters is discussed in detail. Segmentation of character family of bee, pee, tee, tay, cee and fee are around 80 % similar is the case of character family of kaf and gaf. Character noon, ghee is mostly incorrectly segmented. The combination of alif and lam as used in words like results in a new character. Garbage characters generate after the segmentation of seen, dwad, noon, bee and ghee family characters. The proposed algorithm provided segmentation accuracy of 84.6% when examined through the human eye. Segmentation accuracy against each sample and characters has been shown by tables and column charts.

Conclusions and Recommendations

This research successfully addressed the segmentation of typed Pashto text images into individual characters, an essential step in developing Optical Character Recognition (OCR) for Pashto. The segmentation accuracy of the proposed algorithm was 84.6%, with notable success in character families such as bee, pee, tea, tay, cee, and fee, as well as kaf and gaf. However, certain characters like noon and ghee were frequently missegmented, and the combination of alif and lam resulted in the creation of new characters. Additionally, garbage characters were generated during the segmentation of certain characters such as seen, dwad, noon, bee, and ghee. The algorithm's performance can be further enhanced by expanding the character database with more diverse samples. Although the system was developed for static Pashto text images in a single font and size, it lays the foundation for future improvements. A dynamic system capable of handling different fonts, sizes, and handwritten text could be developed, paving the way for a fully functional and accurate Pashto OCR system.

Acknowledgement

The authors are grateful to the University of Agriculture Peshawar to give the opportunity for this research work.

Novelty Statement

This research introduces a novel segmentation approach specifically tailored for Pashto text images, addressing the unique complexities of Pashto script by improving character-level recognition accuracy, which has not been adequately explored in existing OCR methodologies.

Author's Contribution

Arbab Waseem Abbas, Waseem Ullah Khan and Khalid Saeed:Conceptualization;
Arbab Waseem Abbas, Waseem Ullah Khan, Huda Nauman and Syed Mujtaba Abbas Zaidi: Methodology;
Arbab Waseem Abbas, Waseem Ullah Khan, Khalid Saeed and Syed Mujtaba Abbas Zaidi: Formal analysis;
Waseem Ullah Khan, Huda Nauman and Syed Mujtaba Abbas Zaidi: Data management;
Arbab Waseem Abbas, Waseem Ullah Khan and Huda Nauman: Writing --original draft.
Arbab Waseem Abbas, Khalid Saeed and Syed Mujtaba Abbas Zaidi: project administration;

All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Conflict of interest
The authors have declared no conflict of interest.

References

Abbas, A.W., Khan, W.U., Marwat, S.N.K., Ahmed, S., Saeed, K. and Arfeen, N.U., 2024. Image compression exploration using discrete wavelets transform families and level. Int. J. Innov. Sci. Technol., 6(2): 366-379.

Ahmad, R., Afzal, M.Z., Rahsid, S.F. and Liwicki, M., 2015. Recognizable units in Pashto language for OCR. IEEE 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1246-1250. https://doi.org/10.1109/ICDAR.2015.7333963

Akbari, Y., Nouri, K., Sadri, J., Djeddi, C. and Siddiqi, I., 2017. Wavelet-based gender detection on offline handwritten documents using probabilistic finite state automata. Image Vision Comput., 59: 17-30. https://doi.org/10.1016/j.imavis.2016.11.017

Alghyaline, S., 2022. Arabic optical character recognition: A review. Comp. Model. Eng. Sci., 135(3): 1825-1861. https://doi.org/10.32604/cmes.2022.024555

Alhamad, H.A., Shehab, M., Shambour, M.K.Y., Abu-Hashem, M.A., Abuthawabeh, A., Al-Aqrabi, H., Daoud, M.S. and Shannaq, F.B., 2024. Handwritten recognition techniques: A comprehensive review. Symmetry, 16(6): 681. https://doi.org/10.3390/sym16060681

Al-Wajih, E. and Ghazali, R., 2023. Threshold center-symmetric local binary convolutional neural networks for bilingual handwritten digit recognition. Knowl. Based Syst., 259: 110079. https://doi.org/10.1016/j.knosys.2022.110079

Alyahya, S., Khan, W.U., Ahmed, S., Marwat, S.N.K. and Habib, S., 2022. Cyber secure framework for smart agriculture: Robust and tamper-resistant authentication scheme for IoT devices. Electronics, 11(6): 963. https://doi.org/10.3390/electronics11060963

Chooi, S.L. and Asb, A.G., 2021. Handwritten character recognition using convolutional neural network. J. Phys. Conf. Ser., 1918(4): 042152. https://doi.org/10.1088/1742-6596/1918/4/042152

Choudhary, A., 2014. A review of various character segmentation techniques for cursive handwritten words recognition. Int. J. Inf. Comp. Technol., 4(6): 559-564.

Choudhary, A. and Kumar, V., 2017. A robust technique for handwritten words segmentation into individual characters. Speech and language processing for human-machine interactions, pp. 99-106. https://doi.org/10.1007/978-981-10-6626-9_11

Humayun, M., Siddiqi, R., Uddin, M., Kandhro, I.A., Abdelhaq, M. and Alsaqour, R., 2024. A novel methodology for offline English handwritten character recognition using ELBP-based sequential (CNN). Neural Comp. Appl., 36: 19139-19156. https://doi.org/10.1007/s00521-024-10206-1

Kanaparthi, S.K. and Raja, U., 2022. Content-based image retrieval on big image data using local and global features. Int. J. Inf. Technol., 14(1): 49-68. https://doi.org/10.1007/s41870-021-00806-8

Khaliq, F., Shabir, M., Khan, I., Ahmad, S., Usman, M., Zubair, M. and Huda, S., 2023. Pashto handwritten invariant character trajectory prediction using a customized deep learning technique. Sensors, 23(13): 6060. https://doi.org/10.3390/s23136060

Khan, W.U., Marwat, S.N.K. and Ahmed, S., 2022. Cyber secure framework for smart containers based on novel hybrid DTLS protocol. Comp. Syst. Sci. Eng., 43(3): 1297-1313. https://doi.org/10.32604/csse.2022.024018

Memon, J., Sami, M., Khan, R.A. and Uddin, M., 2020. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8: 142642-142668. https://doi.org/10.1109/ACCESS.2020.3012542

Nguyen, V.D., Bui, N.D. and Do, H.K., 2022. Skin lesion classification on imbalanced data using deep learning with soft attention. Sensors, 22(19): 7530. https://doi.org/10.3390/s22197530

Preethi, P. and Mamatha, H.R., 2023. Region-based convolutional neural network for segmenting text in epigraphical images. Artif. Intell. Appl., 1(2): 119-127. https://doi.org/10.47852/bonviewAIA2202293

Qaroush, A., Awad, A., Modallal, M. and Ziq, M., 2022. Segmentation-based, omnifont printed Arabic character recognition without font identification. J. King Saud Univ. Comp. Inf. Sci., 34(6): 3025-3039. https://doi.org/10.1016/j.jksuci.2020.10.001

Sakshi, V. and Kukreja, V., 2023. Image segmentation techniques: Statistical, comprehensive, semi-automated analysis and an application perspective analysis of mathematical expressions. Arch. Comp. Methods Eng., 30: 457-495. https://doi.org/10.1007/s11831-022-09805-9

Sarwar, A., Alnajim, A.M., Marwat, S.N.K., Ahmed, S., Alahya, S. and Khan, W.U., 2022. Enhanced anomaly detection system for IoT based on improved dynamic SBPSO. Sensors, 22(13): 4926. https://doi.org/10.3390/s22134926