Handwritten Pashto Characters Dataset for Optical Character Recognition

Wahidullah Mudaser
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
wahidullah.mdr@gmail.com

Jonathan H. Chan
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
jonathan@sit.kmutt.ac.th

***Abstract*—In this work, we introduced a Pashto character dataset of handwritten scanned images, and we have made the database freely accessible for use in science, research as well as for application of Pashto Optical Character Recognition (OCR) systems. Pashto language is used by over fifty million citizens for both oral and written communications, but there is still no effort being made to the Pashto language for (OCR) system.**

***Index Terms*—Optical Character Recognition, OCR, Pashto Character Recognition**

I. Introduction

Character Recognition or Optical Character Recognition may be defined as a system that converts machine written and handwritten scanned images to editable form [1]. OCR has been extensively used as the basic application of different learning methods in machine learning [2]. The importance of the OCR apparent from the fact that a paper will become out of date in the age of the digital computers. Thus, old books and papers could be archived and stored again in digital formats. This technique leads important information to re-usable form.

However, most of the OCR systems have been built to recognize Latin, Japanese, Chinese and other characters, while comparatively Pashto text recognition is unseen in the area of language research. Pashto is national language of Afghanistan and spoken in most part of Pakistan as well. Pashto is spoken by around 50 million people around the globe [3]. Pashto language has rich literature and diversity. There is verity in terms of words of written material available, which covers very diverse topics such as education, politics, religion, poetry, and much more. Apart of all these, Pashto language still needs some improved and advance technology called OCR. There are so many reasons for such unseen and lake of advance system (OCR) for Pashto language, such as its cursive language written from right to left-hand side and very little variation occurs in characters' shape for non-cursive script languages. Unlike non-cursive script languages, characters in the Pashto language have significant variations.

Most of research done in Arabic OCR, Persian OCR and Urdu OCR were focusing on the recognition of handwriting scripts. However, the recognition of Pashto character remains a challenging task.

we have a long-term study policy, which will give Pashto full OCR system. Since there is no standard dataset for OCR development, we are going to prepare Pashto character dataset as in the primary stage. We will also introduce deep learning method for these characters. In short, this study contribution is the development of new handwritten characters dataset for character recognition of Pashto language.

II. Related Work

Characters of Arabic and Persian are similar to the Pashto language, : thus, there is lack of specific research about Pashto OCR, However, we reviewed some of the Arabic and Persian prior works and summarize them as follow.

OCR has been identified through two popular methods, namely holistic and analytical methods [4]. we are discussing these two approaches as we proceed. Holistic methods do not have specific rules governing typography. As can be generalized to any language, such approaches are common. An image with text is considered to be a vector of one dimension, and features are extracted from the image [4]. For such methods, no segmentation is required. One of the major drawbacks of these methods is that a large amount of training data is needed. These algorithms are robust in size, and rotational changes. Furthermore, it requires a rich set of features to build a model.

BBN Byblos OCR system is a common OCR system based on holistic approaches [5]. Multiple languages have been tested on this system. With these methods, a very low error rate was recorded for synthetic data. When applied to a comparatively larger database, these approaches fail to perform since very little training data was used during the development stage.

For Pashto text, a method developed on the holistic algorithm is reported in [6]. In this work the paper's authors used Noori Nastaliq language script. The synthetic database evaluated this OCR system. Some methods developed for OCR can be explored in the references [7]–[12].

The second class of OCR methods is analytical methods, which are advanced methods and are constructed through specific grammatical rules for the respective language. A unique set of features are used to identify a character. Segmentation at atomic level is performed for these methods. The performance of these methods is better when results of the prior segmentation is easy. For non-cursive script languages

boundary of a character can easily be located; hence results are much better [4]. For getting acceptable performance for these algorithms, better segmentation is mandatory, which is itself a big challenge in analytical methods. For the Pashto language, still, no algorithm has been developed, which is based on analytical methods. Some methods which are based on Hidden Markov Models and Neural Networks are reported in [13], [14] for other cursive script languages.

A database for Pashto ligatures is also reported in [20]. Authors of the paper used Recurrent Neural Networks to develop a Pashto OCR. Tests are performed on a limited set of images in [20]. Authors named their introduced database KPTI. The KPTI consists of 17, 015 images of Pashto text. To the best of our knowledge, this [20] is the best research work reported particularly for Pashto language. Some other works which used deep learning-based methods for cursive script languages can be explored in references [16]–[19].

A medium size database for Pashto OCR has been developed [20]. In the same work, they have also reported the development of an OCR system for recognition of isolated Pashto characters. The classification is performed at two levels, i.e. High level classification and Low-level classification, and the K nearest neighbor (K-NN) classifier has been utilized for low level feature classification. Beside a few reported researches on Pashto OCR, the research on Pashto OCR is still in the initial stage and a lot of research work is needed to develop a Pashto OCR system deployable for practical applications.

III. METHODOLOGY

A. Pashto Language

Pashto is written in Arabic script and by comparing its character-set, we can conclude that all Arabic and Persian characters are subsets of the Pashto language. While 36 characters of the Urdu language are also available as a subset of the Pashto character set. There are 44 basic Pashto alphabets, as shown in Fig 5.

A textual analysis of the Pashto web corpora is reported [1]. The study shows the most frequent words and ligatures in Pashto text along with the complexities caused by breaker characters. Similarly, the count and frequency information regarding Pashto's unique ligatures and primary ligatures are also presented [1]. The following section describes the important aspects of dataset proposed in this work.

B. Pashto Dataset

An appropriate dataset shall hold almost all possible word/shapes with respect to a target language. In general, Pashto literature contains a variety of text layouts. These variations mainly exist due the contents of text materials. The contents of Pashto literature are classified as poetry, essay, novel, reports, news, and religion. Thus, this work attempts to create a novel real Pashto handwritten characters dataset [21]. Steps for providing Pashto character dataset are as follow:

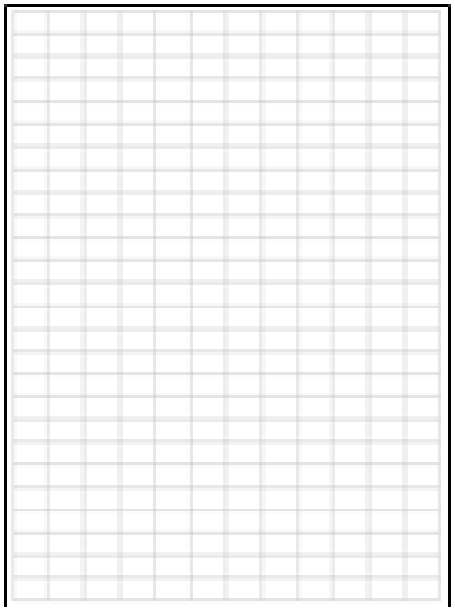

Fig. 1: Structure of paper given to participants

1) collection of Pashto characters: The real instances of Pashto character images are collected data from various regions in order to bring differences to the writing style. These images were collected by faculty members, teachers and students from two universities, such as Benawa University and Kandahar University. Furthermore, classmates, afghan students in KMUTT and some volunteer in Kandahar also shared their handwritten Pashto characters. The total number of participants in the data collection was 650. Collecting of the data per participant is shown in Figure 5.

TABLE I: Participants in various regions

S/No	Regions	Participants
1	Benawa University	158
2	Kandahar University	230
3	Afghan Students in KMUTT	12
4	Volunteers	250
	Total	650

For database creation of Pashto handwritten characters, a blank paper of A4 size was designed with 112 rows and 26 columns. As the total number of Pashto character is 44, two blank pages were distributed amongst each person. All papers were collected and scanned with resolution of 300dpi (dots per inch). Sample of the white paper is shown in Fig 1, and Fig 2.

2) Preprocessing on characters: Character segmentation is an operation that seeks to decompose an image of a sequences of characters into sub images of individual symbols. It is one of the decision processes in a system for (OCR). The segmentation of different characters from the scanned image is a puzzling work. we used OpenCV python library

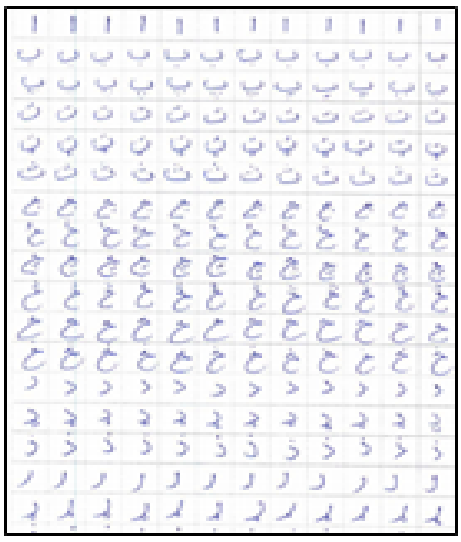

Fig. 2: Sample of handwritten characters dataset

for segmentation of each character that is explained in the following steps:

a) Thresholding: Thresholding is a technique in OpenCV, which is the assignment of pixel values in relation to the threshold value. In thresholding, each pixel value is compared with the threshold value. In this work threshold is applied for edge detection for better accuracy and used binary images. Threshold has two regions on its either side with the lower threshold the upper threshold being selected as 127 and 255, respectively.

b) Shape Analysis: Contours come handy in shape analysis, finding the size of the object of interest, and object detection. OpenCV has findContour() function that helps in extracting the contours from the image. It works best on binary images, so we should first apply thresholding techniques. in the current work we used findContours() function for character detection and then we extract each character and save in a separate class. Sample of extracting the characters are in Fig 3.

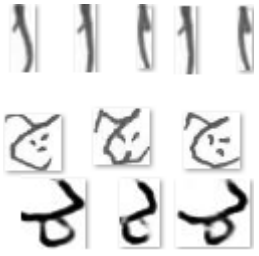

Fig. 3: Extracted Pashto handwritten characters

c) Data Augmentation: Data augmentation is the process of increasing the amount and diversity of data. We do not collect new data, rather we transform the already present data. We augmented the data to increase the diversity of each character for training models. An image is shown in Fig 4, where the characters are augmented through different shapes. There are 4 different shapes (a, b, c, d).

Fig. 4: Pashto character augmentation

IV. Conclusion and Future Work

In this paper, we introduced a Pashto character database [21]. The training set per each character contains of 400 images on average. For example, there are 611 train set image for AYN and 397 train set for ALIF. The test set per each character contains 140 images on average. For example, there are 145 test set images for character ALIF. We prepared the dataset freely accessible for downloading to use for research purpose or apply any convolution neural models.

This research is part of our long-term cursive-script language analysis strategy. The future work of this study is to develop a baseline deep learning Neural Network model and then fine-tuning it in order to evaluate and achieve better results.

References

[1] S. Taha, Y. Babiker, and M. Abbas, "Optical character recognition of arabic printed text," in 2012 IEEE Student Conference on Research and Development (SCOReD), 2012.

[2] Zand, M., Nilchi, A. N., Monadjemi, S. A. (2008, February). Recognition-based segmentation in Persian character recognition. In Proceedings of World Academy of Science, Engineering and Technology (Vol. 28, pp. 183-187).

[3] F. B. J. Kuiper, "A grammar of Pashto, a descriptive study of t the dialect of Kandahar, Afghanistan," Lingua, vol. 7, pp. 103–104, 1957.

[4] "Pashtu numerals recognition through convolutional neural networks," Journal of Applied and Emerging Sciences, pp. 91–96, 2019.

[5] M. Decerbo, E. MacRostie, and P. Natarajan, "The BBN Byblos Pashto OCR system," in Proceedings of the 1st ACM workshop on Hardcopy document processing - HDP '04, 2004.

[6] [6] S. A. Husain, "A multi-tier holistic approach for Urdu Nastaliq recognition," in International Multi Topic Conference, 2002. Abstracts. INMIC 2002, 2002.

[7] Mostefa, D., Choukri, K., Brunessaux, S., Boudahmane, K. (2012). New language resources for the Pashto language. LREC 2012.

[8] R. Ahmad, S. H. Amin, and M. A. U. Khan, "Scale and rotation invariant recognition of cursive Pashto script using SIFT features," in 2010 6th International Conference on Emerging Technologies (ICET), 2010.

[9] M. Wahab, H. Amin, and F. Ahmed, "Shape analysis of Pashto script and creation of image database for OCR," in 2009 International Conference on Emerging Technologies, 2009.

[10] Khan, K., Ullah, R., Khan, N. A., Naveed, K. (2012). Urdu character recognition using principal component analysis. International Journal of Computer Applications, 60(11).

[11] K. Khan, R. Ullah, N. Ahmad Khan, and K. Naveed, "Urdu character recognition using principal component analysis," Int. J. Comput. Appl., vol. 60, no. 11, pp. 1–4, 2012.

[12] K. Khan, R. U. Khan, A. Alkhalifah, and N. Ahmad, "Urdu text classification using decision trees," in 2015 12th International Conference on High-capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015.

Class	Label	Pashto Character
0	0	ا
1	1	آ
2	2	ب
3	3	ې
4	4	ت
5	5	ټ
6	6	ټ
7	7	ح
8	8	خ
9	9	چ
10	10	څ
11	11	څ
12	12	ج
13	13	د
14	14	ذ
15	15	ډ
17	17	ر
18	18	ز
19	19	ړ
20	20	ږ
21	21	س
22	22	ش
23	23	ښ
24	24	ص
25	25	ض
26	26	ط
27	27	ظ
28	28	ع
29	29	غ
30	30	ف
31	32	ق
32	32	ک
33	33	ل
34	34	م
35	35	ن
36	36	ڼ
37	37	و
38	38	ه
39	39	ي
40	40	ی
41	41	ئ
42	42	ې
43	43	ۍ

Fig. 5: Classes and labels of Pashto Characters.

[13] Saad Ali Hussien Al-Qahtani, "Recognizing cursive Arabic script using Hidden Markov Models", University of King Saud, 2004.

[14] Gillies, A., Erlandson, E., Trenkle, J., Schlosser, S. (1999, April). Arabic text recognition system. In Proceedings of the Symposium on Document Image Understanding Technology(pp. 253-260).

[15] A. Cheung, M. Bennamoun, and N. W. Bergmann, "An Arabic optical character recognition system using recognition-based segmentation," Pattern Recognit., vol. 34, no. 2, pp. 215–233, 2001.

[16] M. Z. Alom, P. Sidike, T. M. Taha, and V. K. Asari, "Handwritten Bangla Digit Recognition using deep learning," arXiv [cs.CV], 2017.

[17] H. M. Najadat, A. A. Alshboul, and A. F. Alabed, "Arabic handwritten characters recognition using convolutional neural network," in 2019 10th International Conference on Information and Communication Systems (ICICS), 2019.

[18] S. Ram, S. Gupta, and B. Agarwal, "Devanagri character recognition model using deep convolution neural network," J. Stat. Manag. Syst., vol. 21, no. 4, pp. 593–599, 2018.

[19] "Recognition of handwritten characters using deep convolutional neural network," Special Issue, vol. 8, no. 6S4, pp. 314–317, 2019.

[20] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, T. Breuel, and A. Dengel, "KPTI: Katib's Pashto Text Imagebase and Deep Learning Benchmark," in 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2016.

[21] https://github.com/mudaser37/pashtoCharacterDataset