PMDJAS
ISSN 2789-2611
PAKISTAN'S MULTI DISCIPLINARY JOURNAL FOR ARTS & SCIENCE
https://pmdjas.com
| December. 2020, VOL. 1 (Issue, 02) Page. 25 – 31 |

IMAGES DATABASE FOR THE RECOGNITION OF PASHTO ISOLATED PRINTED CHARACTERS

Abdul Ali [1] Abdul Basit[1] Arfa Sidique[1]

[1]*Department of Computer Science & Information Technology, University of Balochistan, Quetta*

Accepted on: December: 02/2020. Published on: December: 10/2020.

Abstract december

A study to assess the various primary tourism resources of Cavite as baseline data in marketing and promotion of the province as tourism destination. Qualitative method was used in the study and interview was conducted to the 23 municipal/city tourism officers of Cavite. There are eminence and accessible infrastructures available around the area hence, no latest tourism investments were intended to the province. There is a total of 157 primary tourism resources and 25 potential attractions identified on the area. Most of the attractions are maintained by the private sectors while others are under the liability of LGUs. Marketing and promotion are mostly printed materials due to lack of funds, but is supported by their marketing partner such as the Provincial tourism office. It is recommended that tourism officers should maintain and enhance the existing and potential attractions in their area. The linkage between the public and private sector should be strengthened more and the government should invest in the marketing and promotions of the tourism resources for sustainable tourism development.

KEYWORDS: Optical, Character, Recognition, Image Database, Pashto, Dataset

1. INTRODUCTION

The images are in two formats printed or handwritten text, the text is later converted into images through a camera or scanner which brings them in pixel format and called raster images. The machine cannot process and the text from them directly. There is a technique to convert the handwritten or printed images into a format that machine can read them called poetical character recognition. The document that is converted into the image is an official document, passport receipt, or bank receipt. We use optical character recognition for different purposes of license plate recognition, data warehouses, and also natural language processing.

Optical character recognition works in three stages, the first stage to scan the printed or handwritten document through a camera or scanner, which produces the image in digital

format second recognize the printed or handwritten isolated characters, third it stores the characters in a computer understandable format. Optical character recognition extended its area around the entire word for writing and spoken languages which include the right to left languages such as Urdu, Persian, and Arabic, and Pashto language.

Pashtun's circle major language is Pashto which is also pronounced Paktho, Pushto, or Pukhto in our country Pakistan and the Pashto language is the official language of Afghanistan. Persian recognizes the Pashto as Afghani Pashto which is similar to languages written from right to left like Urdu and Arabic language. These languages are cursive in nature having dots and diacritics. Due to cursive in nature and dots, it is difficult for the recognition system.

Pashto language having 44 characters which are similar to Urdu language characters few are different from the Urdu language which is called Pashto language special characters that are shown in red color in the images show billow, See Fig 1.

ا	خ	ژ	ع	ڼ
ب	ح	ډ	غ	و
پ	خ	س	ف	ه
ت	د	ش	ق	ي
ټ	ډ	ښ	ک	ۍ
ث	ذ	ص	ګ	ى
ج	ر	ض	ل	ے
څ	ړ	ط	م	ئ
چ	ز	ظ	ن	

Fig. 1. Pashto language 44 characters

The images contained in the document in two forms printed and handwritten. Printed images are produced by a machine such as a printer and it is correctly and well formatted. The scan printed documents for different purposes like bank and passport receipts. The OCR algorithm enables the computer to recognize the text in the scanned document and convert it into a digital format. The Pashto language also contains six special characters that are different from Urdu and the Arabic language are shown in fig 2.

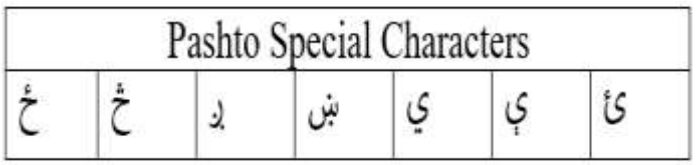

Fig. 2. Pashto language Special characters

The Afghani Pashto contains characters different from the Pashto used in the Frontier of Pakistan. The character set is different from each other. In this paper, we focus on the Afghani Pashto and its character dataset. We made it pubic to carry further research on this dataset.

1.1. Languages from right to left

In this section, we discuss the scripts written from the right to left generally used in South Asia. The languages that are written from right to left include Arabic, Persian, Urdu, Pashto, and Sindhi. These languages have connected character, cursive in nature, and case sensitive.

Connected characters mean that they have no space between the characters to make words many characters are connected with each other to make a word segmentation is the challenging problem in the connected character Arabic Urdu and Pashto characters are connected in nature and having dots and diacritics.

1.2. Pashto language

Pashto is the mother's tongue of Pashtuns which Pronounced as Pashto/Pukhto/Pakhto and the Pashto language is the major language of the Pashtun circle in Pakistan whereas Pashto is the official language of Afghanistan. In Hindi and Urdu literature Pashto is called Pathani, and in Persian literature, it is called afghani Pashto. The language is spoken in Pakistan by about 15.42% and it is the primary language of federally administered tribal Areas (FATA) and Khyber Pakhtunkhwa.

Whereas it is as well-spoken in some districts of Pakistan’s Punjab province that are Attock and Mian wali. Pashtuns traveled to other cities like Lahore and Karachi where Pashto is understood and spoken. Pashtoons also found in Iran (South Khorasan province) and Tajikistan. Pashtun communities also in Jammu and Kashmir and in India.

1.3. Limitation of dataset

We create dataset which was the initial work for Pashto dataset but Machine learning required large dataset we are created dataset which is not sufficient for machine learning in this case we use augmentation technique for machine learning.

2. RELATED WORK

In this part, we discuss the current work on the image dataset for the recognition system of cursive and right to left languages like Arabic, Urdu, and Pashto [1]. Electronic Article Management technology is extensively using the OCR which makes the process efficient to retrieve the document. Character recognition is a front end for electronic document management. For the Latin language there exist many OCR technologies but few systems exist for reading the Arabic language. For Arabic optical character recognition, it is important to create a database. That we use for many purpose tastings and training a recognition system this paper presents a comprehensive study about the study of existing database and also describes the development of database using Arabic words (6 million) this paper also work on the pieces of Arabic word with our diacritics background information is also presented.

The optical character recognition (OCR) development for a cursive script like the Pashto language required knowledge of their shape and various fount within Pashto script. For training and testing of optical character recognition, database development is important [2] This paper describes the image database development for recognition Optical character recognition development for a cursive script like Urdu and the Pashto language required high knowledge of their shape and various fount within Pashto script. For training and testing of optical character recognition, database development is important. This paper describes the image dataset development for optical character recognition.

In this paper, the writer describes the regularization of research words on recognition in the Farsi language [3] He describes the development of a novel normal handwritten dataset that having isolated digits, numerical strings, letters, and dates. With our internet search, there is no publicly available Farsi dataset was available. It was decided to create many Farsi datasets that are useful for optical character recognition and also do some experiments for recognition of handwritten isolated Farsi digits.

This paper present dataset for isolated offline handwritten Arabic and Farsi characters and numbers For suing optical character recognition [4]in research this dataset

is available freely for academic no such datasets are freely available for Farsi language which includes 17740 numerals and images of 52380 is included in every image was scanned at 300 dpi this has a restriction to write ever character in rectangular box simples are not uniform in each class for comparison purpose and real-life distributions.

This paper deals about the overview of the database used for optical recognition system and also focus on their application [5] the main purpose of this paper compact with Arabic handwritten database for optical recognition system result and computation are given finally this paper is deal with Arabic handwritten dataset and also describe their usability finally compared the strategy of Arabic handwritten dataset and computations is presented.

The writer proposed two different methods for character segmentation one is cursive Latin characters and the other is Arabic handwritten characters [6] The problem of upper and lower overlapping is solved using two methods one is junction segments second detects the upper contour of character these two methods are compared. In this paper, the writer explains that automatic recognition of the text on the document that contains scanned images allowed many applications which include automatic sorting of email, editing of previously document which was printed and search for a word in document different method are applied to a different type of images this is the initial survey to emphasis on Arabic handwritten recognition which is firstly applied on Arabic character recognition this includes a discussion of field background of field and research directions on future.

The writer presents a dataset for Arabic off-line handwritten recognition, and with processing procedures [7] The writer developed a novel dataset for the store, collection, and access of Arabic language handwritten text this dataset is advanced in the size and different writers involved which is easily abstract bitmaps of words. The writer constructs also processing class that having processing operations the writer identified the most popular words through the associated programs.

The writer describes the formation of a broad database of Arabic handwritten signature, words, and number and uses them for recognition which is related to the Arabic language [8] The was no commercially and freely dataset are available. We built the Pashto language isolated printed character dataset. Make it free for academic use to work in the future.

In artificial intelligence text recognition enable the machine with to recognize spoken languages with ti interpretation of them [9,10] Recognition is sub domain of artificial intelligence. Which convert the scan image into computer editable format the researcher suggests different recognition techniques for cursive and connected language

script that are written from right to left. Balochi language is a curse script. No research on this language in Pakistan in this paper the writer proposes convolvationl Neural network model for balochi language character recognition they computer the VGGNet with baseline LeNet Model the result show the VGGNet model propose method improved over baseline method with precision 96% they collected the balochi character database and make them public for future work

2.1. Dataset

There are different techniques to learn machine-like deep learning you need to have a dataset. For training, data is the most vital aspect machine learning algorithm tries to extract and identify the patterns for the data. So data is important for extracting the patterns. Your data is useless even the data is more harmful if a machine cannot find or identify the pattern data has flaws that are an issue which is the reason that the data preparation is an important step to word machine learning we have established a framework for the preparation of data set to summarize and other NLP (natural language processing) this preparation have a set of procedures and also have guidelines that help us to make the suitable datasets. This framework contains a collection of data cleaning.

We do not have any standard image dataset for the Pashto language to apply the recognition system. We make an image dataset [11] for the Pashto language isolated characters. The Pashto language has no standard image dataset for isolated character research on character recognition. We need a standard dataset to check the performance of the OCR .in our paper dataset has been developed. The data set containing 44 isolated characters with 50 to 55 variations of each character making the total images in our data set are 2480. The data set used in this paper is our own development. The size (width and height) of each alphabet in the dataset is 32 × 32. Using Inkscape for making the images. See Fig. 3

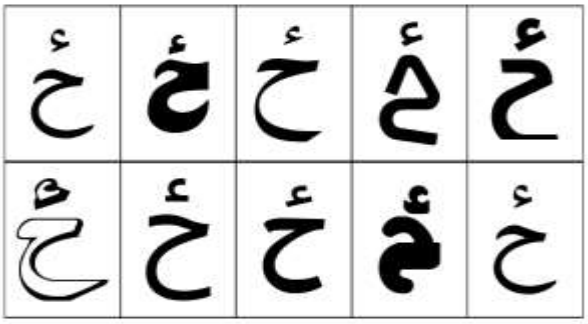

Fig. 3. Dataset images

3. DATA COLLECTION

Data collection is the procedure of collecting data from different sources, it is challenging and needs extra efforts to build the dataset. An incomplete and inaccurate collection of data could lead your research to poor results. Which effect the result of our study in order to collect more data with less resource. In the Pashto language, there was no enough isolated printed characters dataset for machine learning, we used the graphical software for the creation of the dataset. The existing literature lacks a dataset and also not available online to carry further research.

3.1. Tools use for data collection

Inks cape is an open-source and free vector graphics editor for creating images, it is used for a wide range of computer graphics to develop desktop applications, it is a designing tool. So, we can use Inkscape for different purposes such as making diagrams, web graphics, logos, and paper scrapbooking. It is also used for game scraping; we used this graphic software for creating images dataset for the Pashto language characters.

3.2. Creation of image dataset for the Pashto language

For the image dataset, we have required a synthetic method each image in the dataset having a different shape. We type those images in Inkscape software which uses for a graphic. We created the images in the following way install the Pashto language 55 to 60 different font and created images of each isolated character of different fonts of size 32 × 32 pixels we created 44 different folders on the name of Pashto characters each character have different 55 to 60 images then these images are store in sprit folder these images are in monochromatic which means only two colors black and white. There is no need for color images text is always in black and white background because monochrome takes less memory the color images. Our dataset contains 44 different folders each folder contains 60 images of a single character and the dataset contains more than 2500 images. Which is freely available for academic use.

4. CONCLUSIONS

We created the image dataset for the Pashto language printed characters and made them available online to carry further research. We removed and correctly designed the shape of

every character in the Pashto Language. This dataset is the baseline for creating optical character recognition for Pashto isolated printed character recognition.

4.1 Future work

In the future, we will also create a handwritten and connected words dataset. Additionally, we can also now correctly apply the machine learning techniques to this dataset to build the optical character recognition algorithm for the Pashto character recognition.

5. REFERENCES

A. Ali, Pashto Script Isolated Character Dataset, https://drabasit.herokuapp.com/ali accessed: 2019

A. Raouf, A., Higgins, C. A., & Khalil, M. A database for Arabic printed character recognition. In International Conference Image Analysis and Recognition, Springer, Berlin, Heidelberg. (pp. 567-578) June 2008

Al-Ma'adeed, S., Elliman, D., & Higgins, C. A. A data base for Arabic handwritten text recognition research. In Proceedings eighth international workshop on frontiers in handwriting recognition (pp. 485-489). IEEE. August 2002

Kharma, N., Ahmed, M., & Ward, R. "A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing." In Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No. 99TH8411) (Vol. 2, pp. 766-768). IEEE. May 1999

Lorigo, L. M., & Govindaraju, V. Offline Arabic handwriting recognition: a survey. IEEE transactions on pattern analysis and machine intelligence, 28(5), 712-724. March 2006

Märgner, V., & El Abed, H. Databases and competitions: strategies to improve Arabic recognition systems. In Summit on Arabic and Chinese Handwriting Recognition (pp. 82-103). Springer, Berlin, Heidelberg. September 2006

Mozaffari, S., Faez, K., Faradji, F., Ziaratban, M., & Golzan, S. M. A comprehensive isolated Farsi/Arabic character database for handwritten OCR research. October 2006

N Gul. J., Basit, A., Ali, I., & Iqbal, A. "Balochi Non Cursive Isolated Character Recognition using Deep Neural Network."

Romeo-Pakker, K., Miled, H., & Lecourtier, Y. A new approach for Latin/Arabic character segmentation. In Proceedings of 3rd International Conference on Document Analysis and Recognition (Vol. 2, pp. 874-877). IEEE. August 1995

Solimanpour, F., Sadri, J., & Suen, C. Y. Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in Farsi language. October 2006

Wahab, Mahreen, Amin, H., & Ahmed, F. Shape analysis of pashto script and creation of image database for OCR. International Conference on Emerging Technologies (pp. 287-290). IEEE. October 2009.