Hindawi
Security and Communication Networks
Volume 2021, Article ID 3543816, 7 pages
https://doi.org/10.1155/2021/3543816

Research Article

Offline Pashto Characters Dataset for OCR Systems

Sulaiman Khan,[1,2] **Habib Ullah Khan**,[2] **and Shah Nazir**[1]

[1]*Department of Computer Science, University of Swabi, Swabi, Pakistan*
[2]*Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar*

Correspondence should be addressed to Habib Ullah Khan; habib.khan@qu.edu.qa

Received 19 May 2021; Accepted 17 July 2021; Published 27 July 2021

Academic Editor: Marimuthu Karuppiah

In computer vision and artificial intelligence, text recognition and analysis based on images play a key role in the text retrieving process. Enabling a machine learning technique to recognize handwritten characters of a specific language requires a standard dataset. Acceptable handwritten character datasets are available in many languages including English, Arabic, and many more. However, the lack of datasets for handwritten Pashto characters hinders the application of a suitable machine learning algorithm for recognizing useful insights. In order to address this issue, this study presents the first handwritten Pashto characters image dataset (HPCID) for the scientific research work. This dataset consists of fourteen thousand, seven hundred, and eighty-four samples—336 samples for each of the 44 characters in the Pashto character dataset. Such samples of handwritten characters are collected on an A4-sized paper from different students of Pashto Department in University of Peshawar, Khyber Pakhtunkhwa, Pakistan. On total, 336 students and faculty members contributed in developing the proposed database accumulation phase. This dataset contains multisize, multifont, and multistyle characters and of varying structures.

1. Introduction

A significant work has been reported for the automatic recognition of text or characters in images using different classification and recognition techniques in the last decade. Consequently, in any OCR system, a systematic database must be created by collecting and scanning written documents, which is easier nowadays due to the availability of high definition cameras, mobile phones, and many more. In spite of this abundance, many datasets are developed in many languages including Arabic, Urdu, English, Japanese, Chinese, and many others except Pashto language. Pashto is followed as the state language of Afghanistan and a major language in the northern areas of Khyber Pakhtunkhwa province of Pakistan. During the census of 2009, it was concluded that approximately 60 millions of people around the globe can speak or understand this language.

Urdu, Persian, and Arabic are considered as sister languages of Pashto language, as the latter has either borrowed or modified the characters of these languages in order to form an indigenous dataset of 44 characters. Multiple characters including machine printed and handwritten formats databases are developed in those languages except Pashto language. Bouressace and Csirik developed a dataset that is named printed Arabic characters for OCR systems [1]. This dataset is developed by slicing text from ten different Arabic newspapers comprising 2954 varying characters images. This dataset contains letters varying in font and styles and contains blurring effects and varying resolutions. Abdalkafor performed a survey study on different datasets developed in Arabic language [2]. Some researchers like Khan et al. [3–5] and Ma'adeed et al. [6] have developed datasets for their research, but these datasets are limited to their work and not available online. Pechwitz et al. [7] developed a database for Arabic languages known as the IFN/ENET database. This dataset contains more than 2200 binary images collected from 411 different scholars. It comprised of 26000 isolated words images. It contains the information of Tunisian cities and villages.

This research work presents the first handwritten Pashto characters database for the research community in the computer science field. This dataset consists of 14784 binary images of carved samples with labels for the automatic recognition systems in Pashto language. In the proposed research work, the dataset and database represent the same terminologies. Rest of the study is organized as follows. Section 2 represents the background and literature relevant to the proposed work in other languages. Section 3 gives details about the proposed database that is collecting the handwritten Pashto character samples and developing a database for the future simulation purposes. Section 4 provides information about the proposed database followed by the conclusion in Section 5.

2. Materials and Methods

Pashto is followed as the language of northern areas in the Pakistan and the state language of Afghanistan. Pashto language contains some distinguishing features as compared to other languages spoken in the world. Slight changes in the character's shape can make the recognition process even difficult. In order to tackle this problem, an automatic OCR system for the recognition of handwritten Pashto characters is proposed [3]. Cursive nature and variations in character's shape at different positions in the word are the additional hurdles for the researchers in the proposed recognition task. Table 1 presents the change in character's shape with respect to its position in the word.

Figure 1 shows the full character set of the Pashto language. It consists of 44 characters. The Pashto character set is classified into three categories: (1) the letters with no diacritics/dots (13 characters out of 44 characters) that are ا, د, ر, س, ص, ط, ع, ک, ل, م, و, ه, and ی, (2) the characters that contains a circle below or beneath the basic shape that are ڼ, ګ, ړ, and ټ, and (3) the characters with diacritics (remaining 27 characters except these two categories). This can further be classified based on the characters that share the same basic shape that only differs by number of dots or some additional variations like س and ص or ح and چ, and many more are the additional hurdles in developing an optimal OCR system for the Pashto script identification.

Arabic language consists of 28 characters and uses Naskh writing style [8]. Persian language has encapsulated the 28 characters of Arabic language plus additional four characters specific to Perso language to make a 32 characters dataset. Urdu language that is considered as a version of Perso language has borrowed all the 32 letters of Persian language plus six more special characters specifically for Urdu language to form a dataset of 38 characters [9]. Urdu language follows Nasta'liq writing style. Pashto has encapsulated all the 38 characters of Urdu language with the modification of Urdu specific characters as given in Table 2. After borrowing the 37 characters from Urdu language, the Pashto language add seven more special characters to the borrowed character to develop a character dataset of 44 characters as shown in Figure 1. Table 3 presents the Pashto specific characters. Pashto follows Naskh writing styles while Nasta'liq writing style to an extent.

3. Available Characters Datasets

The availability of an enriched character dataset is the key task in any recognition system. Several datasets are developed in many languages by multiple researchers in many languages such as Arabic, Urdu, English, and many more, but there is no handwritten characters dataset available for the Pashto language. Although a dataset is developed for the simulation of the recognition algorithm by Khan et al. [3] and Ahmad et al. [10], these are not available online. Some of the datasets that are developed by the researchers in other languages and available online are

Table 1: Character's shape in words with respect to its position.

Isolated character's shape	Starting shape	Middle shape	Ending shape
ک	کـ	ـکـ	ـک
ل	لـ	ـلـ	ـل
س	سـ	ـسـ	ـس
ح	حـ	ـحـ	ـح
ع	عـ	ـعـ	ـع

Pashto character set

S/No.	Alphabet	S/No.	Alphabet	S/No.	Alphabet
1	أ	16	ر	31	ق
2	ب	17	ړ	32	ک
3	پ	18	ز	33	ګ
4	ت	19	ژ	34	ل
5	ټ	20	ږ	35	م
6	ث	21	س	36	ن
7	ج	22	ش	37	ڼ
8	ځ	23	ښ	38	و
9	چ	24	ص	39	ه
10	څ	25	ض	40	ي
11	ح	26	ط	41	ې
12	خ	27	ظ	42	ی
13	د	28	ع	43	ی
14	ډ	29	غ	44	ئ
15	ذ	30	ف		

Figure 1: Pashto character set.

Table 2: Urdu language characters modified in Pashto language.

S/No.	Urdu characters	Pashto equivalent
1	ٹ	ټ
2	ڈ	ډ
3	ڑ	ړ
4	گ	ګ
5	ے	ۍ

Table 3: Pashto specific characters.

ځ	څ	ڼ	ښ	ي	ې	ئ

discussed in the following. Abed et al. [11] developed the ADAB dataset for Arabic language. This dataset is developed by Tunisian writers for online handwritten text recognition. This dataset consists of the names of Tunisian cities, villages, and towns. Each word is prelabelled using UPX file format based on postcode that is a sequence of numeric character references (NCR). This dataset consists of 20,000 words from 170 different writers.

Sabbour and Shafait [12] developed the Urdu printed text image database (UPTI) for Urdu language. UPTI dataset consists of 10,063 words and ligature images. Lamghari et al. [13] developed the DBAHCL dataset for Arabic language. DBAHCL stands for "Database of Arabic Handwritten Characters and Ligature," and it consists of 3400 characters. The IBN SINA dataset for Arabic language is developed by Moghaddam et al. [14]. This dataset, known as Iben Sina, contains 60 pages containing more than 25000 words, where every individual page consists of about 500 subwords. CALAM is a dataset for Urdu text developed by Choudhary and Neeta [15]. It consists of 1200 text images, 3043 lines, 64664 subwords, and 101180 ligatures. CENIP-UCCP characters dataset is developed for Urdu language. CENIP-UCCP abbreviated for "Center for Image Processing-Urdu Corpus Construction Project" is developed by Raza et al. [16]. This dataset encapsulates 400 digitized forms. During the development phase of this dataset, 200 different writers distributed.

The IESK-ArDB handwritten Arabic text database was developed by Elzobi et al. [17]. This dataset is developed by the Institute of Electronics, Signal Processing and Communication (IESK) at Otto-von-Guericke University, Magdeburg in Germany. IESK-ArDB is the name proposed to this dataset, where "ArDB" abbreviates "Arabic Database." It consists of 285 pages of historical manuscripts that belong to fourteenth century. Also, it contains 6000 handwritten word samples and more than 8000 isolated character images. Especially, this dataset contains information about parts of speech noun, pronoun, verbs, city/region names, security terms, and words written for amount in bank cheques.

Based on the literature study, multiple datasets (both offline and online formats) are available for Urdu and Arabic languages, but there is no dataset available for the handwritten Pashto language (a few number of datasets are developed for Pashto language, but these are only for printed format). After analyzing the literature for a range of years (2010–2019), a number of datasets are available for cursive languages as shown in Figure 2. The proposed research work presents the first handwritten Pashto characters dataset for the researchers in the natural language processing field.

4. Dataset Development Process

Eventhough, a significant research has been reported globally by many researchers to address the problems of recognizing an unconstrained handwriting. Jehangir et al. [18] developed a medium-sized Pashto characters database for the simulation and experimental work. They performed research work on OCR development for handwritten Pashto characters. Huang et al. [19] developed a small-sized handwritten characters database for the automatic recognition of the handwritten Pashto characters. Rai and Li developed visualization in smart devices for generating associative images of Chinese characters [20]. Du and Duan developed the English phrase recognition system using continuous speech recognition and word tree algorithms [21]. Ali et al. [22] developed a database of 103965 labelled sentences in Urdu language. These sentences are classified using the convolution neural network, recurrence neural network, and deep neural network to extract event information from these languages. Zhang et al. [23] proposed a hybrid dataset of objects and text to analyze the recognition capabilities of the YOLOv3 classifier for both text and object detection purposes.

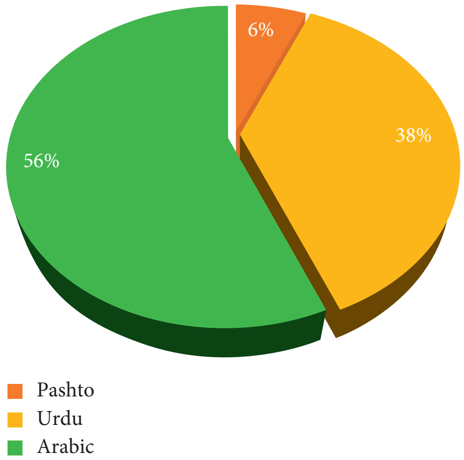

Figure 2: Number of benchmark datasets developed for cursive languages.

From the literature, it is concluded that an effective solution to this problem is still a main hurdle for the researchers. Without a dataset of handwritten characters, the development of an efficient solution for the recognition of inconsistent handwritten text is unfeasible. To address this problem of recognizing handwritten characters in Pashto language, this research work presents the development of a database. The proposed database development process is divided into the following three phases: the data collection phase, scanning and characters extraction phase, and the final labelling and organization phase.

Figure 3 shows the lifecycle of the handwritten characters database development process.

4.1. Data Collection. For collecting the handwritten samples of Pashto characters, we visited the Department of Pashto in University of Peshawar, Pakistan. A4-sized paper is selected for the collection phase, where each paper is divided into 6 columns to collect variant samples (varying in font, shape, and styles) from different and even from the same student or faculty members. Table 4 presents the contribution of students and faculty members based on gender and age. Three hundred and thirty-six varying handwritten samples are collected for each individual character in Pashto script.

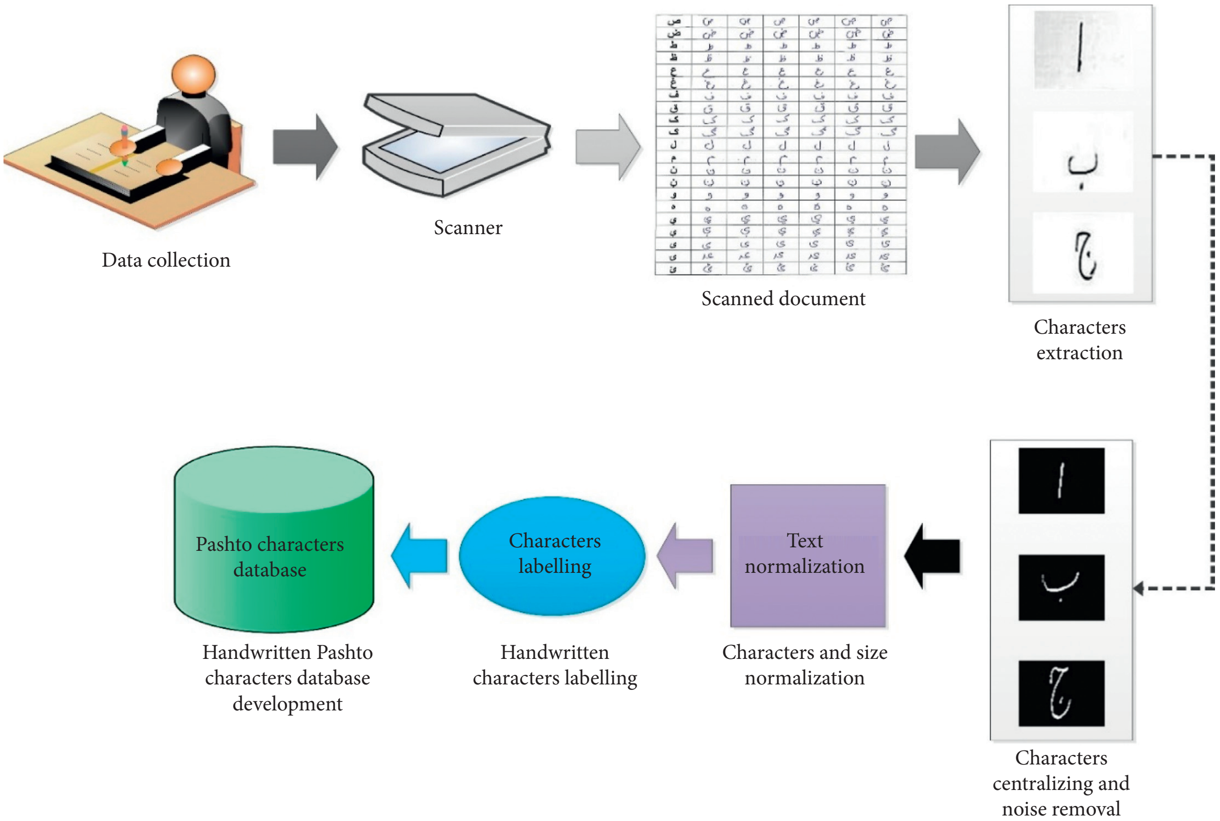
Data collection
Scanner
Scanned document
Characters extraction
Characters centralizing and noise removal
Text normalization
Characters and size normalization
Characters labelling
Handwritten characters labelling
Pashto characters database
Handwritten Pashto characters database development

Figure 3: HPCID development process.

Table 4: Agewise and gender distribution in the collection phase.

Age (years)	Males	Females	Percentage	Samples
18–30	67	53	35.7	120
32–40	57	59	34.5	116
45–60	45	51	28.6	96
Total samples collected				**336**

4.2. Scanning and Character's Extraction. After the data collection, the next phase is to scan the handwritten documents, in order to convert it to the machine readable format. Figure 4 shows the scanned form of the handwritten characters.

After scanning, the next step is character's extraction that is discussed in detail in the following.

4.2.1. Character's Extraction. After scanning the documents, the characters are extracted from the rectangular box shown in Figure 1. These characters are extracted based on the coordinates of each box. Figure 5 shows the sliced characters.

During the accumulating process of characters from different people, a predefined small-sized rectangular window is selected, till extraction of letters from tilted and nontilted scanned images causes elimination of letter's some portion. In order to avoid this problem, morphological operations such as erosion and dilation are followed to fix letter's shape for proper feature extraction and accurate classification. Also, the sliced characters shown in Figure 5 contain black spots (noise in image processing terminologies). For better feature's collection, it is pretty good to remove this noise.

4.2.2. Thresholding. After zooming the extracted letter, these dark spots can be easily seen as shown in Figure 6. In order to remove these black spots, different mechanisms are handy in this problem like average filtering and thresholding. In our case, thresholding is prominent as averaging not only removes the dark spots but it also removes the diacritics (dots) with some letters in Pashto script. In our case, thresholding not only removes the black spot but it also retains the information relevant to the letter in our case.

Mathematical formula for thresholding is

$$T(x) = \left\{ \begin{matrix} 0, & \text{if } n \le \epsilon \\ n, & \text{otherwise} \end{matrix} \right\}. \tag{1}$$

A threshold value of 140 is considered as an optimal value in our case. A threshold value below 140 does not remove the noise completely, while a threshold value above 140 removes the potential information in the letters as shown in Figure 7.

Figure 8 shows the selected thresholded image.

4.2.3. Morphological Operations and Centralizing of Characters. After thresholding the sliced image, some imperfections occurred in the sample's character in the form of

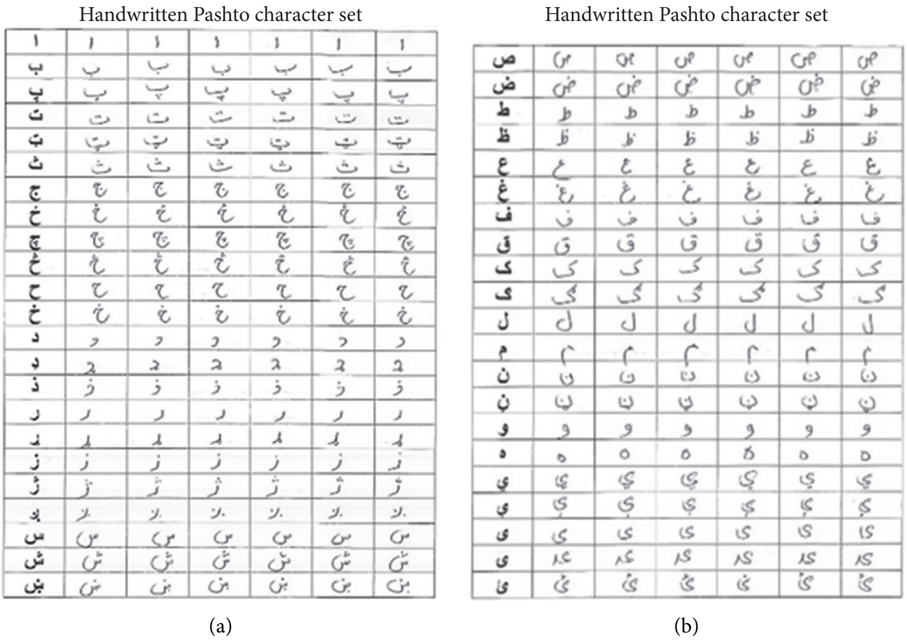
Handwritten Pashto character set
Handwritten Pashto character set
(a)
(b)

Figure 4: Scanned handwritten documents.

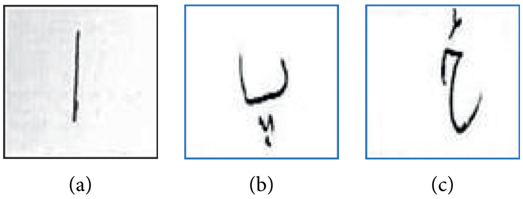

Figure 5: The sliced handwritten characters.

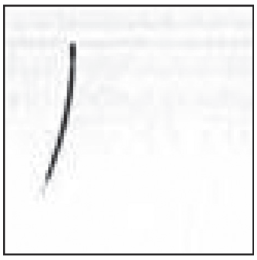

Figure 6: Zoomed character affected with noise.

holes or cutting edges. Also, during the data collection phase, some of the volunteers wrote characters at the top of the rectangular box, some at the centre, and vice versa. For accurate classification and feature's extraction, it is mandatory to remove these imperfections from the sliced images. To address this problem, all the characters are centralized in our case.

All the characters are centralized by finding the centre of the image, and then, the character is centralized based on the centre of the image. In our case, the central point is 32 × 32 as the sliced character image size is 64 × 64. After applying this process, the character shape is fixed and centralized as shown in Figure 9. All these steps are performed for all sliced characters to develop a medium-sized database of 14784 handwritten samples for Pashto script, where 336 samples are accumulated for each 44 letters in Pashto script. This database is used for further experimental results of automatic handwritten Pashto characters recognition. To the best of our knowledge, this is the first database developed for handwritten Pashto characters.

4.3. Labelling and Organization. After scanning, characters extraction and centralizing the next phase is to develop a database of the handwritten samples and label it in order to identify each handwritten character. A fixed size of 64 × 64 is selected for each character. Individual folder is created for each handwritten character where all the 336 samples are enclosed. The same process is repeated for all the 44 characters. An overall database is with 44 directories and a total of 14784 handwritten Pashto characters. The characters are labelled in each individual repositories based on total number of count (1–336), and JPEG format is selected for the character's images.

5. Discussion

During the development of the proposed dataset, multiple factors are considered in order to develop a standard dataset for the experimental and research purposes. These factors include the following.

(i) Multisize
(ii) Multifont
(iii) Multistyle
(iv) Varying structure
(v) Age of the selected personnel
(vi) Gender

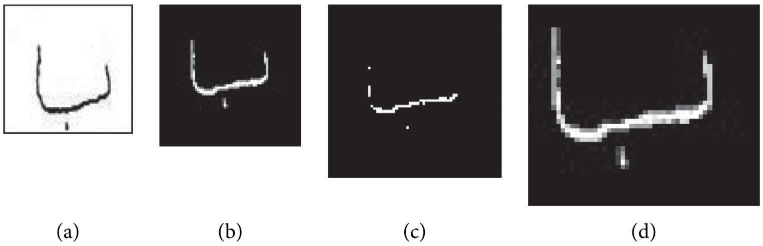

Figure 7: Resultant image at different threshold values. (a) Original character image. (b) Thresholded at 140. (c) Thresholded at 150. (d) Thresholded at 130.

Figure 8: Thresholded image.

Figure 9: Centralized character image.

Considering these factors during the accumulation phase, we developed an optimum medium-sized dataset for the handwritten Pashto characters.

6. Conclusions

This study presents the pioneer Pashto characters database that is available for free online, namely, the handwritten Pashto characters image database. In the proposed database, the handwritten samples are collected from the students and scholars of the Pashto department in University of Peshawar, Khyber Pakhtunkhwa, Pakistan. Three hundred and thirty-six volunteers contributed in the accumulation of handwritten samples. The HPCID database consists of a wide range of varying font size, shape, style, and structure. A main directory is created for the database, where subfolders are created for each individual Pashto characters, where each subfolder encapsulates 336 varying samples of each character. These characters are labelled based on its counting (1–336). The database is freely available for research purposes, and we hope it will assist the Pashto handwritten characters recognition research community.

Data Availability

The data used to support the findings of this study are available from the principal author upon request.

Additional Points

Implications. This developed dataset is fully explained (by assigning labels, folders, and other required details) and is intended for research community to test and evaluate their techniques for an improved performance.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] H. Bouressace and J. Csirik, “Printed Arabic text database for automatic recognition systems,” in *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, pp. 107–111, Istanbul, Turkey, April 2019.

[2] A. S. Abdalkafor, “Survey for databases on Arabic off-line handwritten characters recognition system,” in *Proceedings of the 2018 1st International Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–6, Riyadh, Kingdom of Saudi Arabia, April 2018.

[3] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, “KNN and ANN-based recognition of handwritten Pashto letters using zoning features,” *International Journal of Advanced Computer Science and Applications*, vol. 9, pp. 570–577, 2018.

[4] S. Khan, A. Hafeez, H. Ali, S. Nazir, and A. Hussain, “Pioneer dataset and recognition of handwritten Pashto characters using convolution neural networks,” *Measurement and Control*, vol. 53, no. 9-10, pp. 2041–2054, 2020.

[5] S. Khan, S. Nazir, H. U. Khan, and A. Hussain, “Pashto characters recognition using multi-class enabled support vector machine,” *Computers, Materials & Continua*, vol. 67, no. 3, pp. 2831–2844, 2021.

[6] S. A. Ma’adeed, D. Elliman, and C. A. Higgins, “A data base for Arabic handwritten text recognition research,” in *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, pp. 485–489, Niagara on the Lake, Ontario, Canada, August 2002.

[7] M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, and H. Amiri, “IFN/ENIT-database of handwritten Arabic words,” in *Proceedings of the Francophone International Conference on writing and Document, CIFED’02*, pp. 127–136, Hammamet, Tunisia, October 2002.

[8] W. contributors, “Arabic alphabet. In wikipedia, the free encyclopedia,” 2019.

[9] W. contributors, “Urdu alphabet. In wikipedia, the free encyclopedia,” 2019.

[10] S. Khan, K. Ahmad, M. Murad, and I. Khan, "Waypoint navigation system implementation via a mobile robot using global positioning system (GPS) and global system for mobile communications (GSM) modems," *International Journal of Computational Engineering Research (IJCER)*, vol. 3, 2013.

[11] H. E. Abed, V. Märgner, and A. Alimi, "On-line Arabic handwriting recognition competition—ADAB database and participating systems," *IJDAR*, vol. 14, pp. 15–23, 2011.

[12] N. Sabbour and F. Shafait, "A segmentation-free approach to Arabic and Urdu OCR," *Document Recognition and Retrieval XX*, vol. 8658, Article ID 86580N, 2013.

[13] N. Lamghari, M. E. H. Charaf, and S. Raghay, "Hybrid feature vector for the recognition of Arabic handwritten characters using feed-forward neural network," *Arabian Journal for Science and Engineering*, vol. 43, no. 12, pp. 7031–7039, 2018.

[14] R. F. Moghaddam, M. Cheriet, M. M. Adankon, K. Filonenko, and R. Wisnovsky, "IBN SINA: a database for research on processing and understanding of Arabic manuscripts images," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp. 11–18, Boston, MA, USA, June 2010.

[15] P. Choudhary and N. Nain, "A four-tier annotated Urdu handwritten text image dataset for multidisciplinary research on Urdu script," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 15, no. 4, pp. 1–23, 2016.

[16] A. Raza, I. Siddiqi, A. Abidi, and F. Arif, "An unconstrained benchmark Urdu handwritten sentence database with automatic line segmentation," in *Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition*, pp. 491–496, Bari, Italy, September 2012.

[17] M. Elzobi, A. A. Hamadi, Z. A. Aghbari, and L. Dings, "IESK-ArDB: a database for handwritten Arabic and an optimized topological segmentation approach," *Document Analysis and Recognition*, vol. 16, 2012.

[18] S. Jehangir, S. Khan, S. Khan, S. Nazir, and A. Hussain, "Zernike moments based handwritten Pashto character recognition using linear discriminant analysis," *Mehran University Research Journal of Engineering and Technology*, vol. 40, no. 1, pp. 152–159, 2021.

[19] J. Huang, I. U. Haq, C. Dai, S. Khan, S. Nazir, and M. Imtiaz, "Isolated handwritten Pashto character recognition using a *K*-NN classification tool based on zoning and HOG feature extraction techniques," *Complexity*, vol. 2021, Article ID 5558373, 8 pages, 2021.

[20] L. Rai and H. Li, "MyOcrTool: visualization system for generating associative images of Chinese characters in smart devices," *Complexity*, vol. 2021, Article ID 5583287, 14 pages, 2021.

[21] H. Du and H. Duan, "English phrase speech recognition based on continuous speech recognition algorithm and word tree constraints," *Complexity*, vol. 2021, Article ID 8482379, 11 pages, 2021.

[22] D. Ali, M. M. S. Missen, and M. Husnain, "Multiclass event classification from text," *Scientific Programming*, vol. 2021, Article ID 6660651, 15 pages, 2021.

[23] F. Zhang, J. Luan, Z. Xu, and W. Chen, "DetReco: object-text detection and recognition based on deep neural network," *Mathematical Problems in Engineering*, vol. 2020, Article ID 2365076, 15 pages, 2020.