IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 October 2022, accepted 19 October 2022, date of publication 25 October 2022, date of current version 2 November 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3216881*

RESEARCH ARTICLE

# PHTI: Pashto Handwritten Text Imagebase for Deep Learning Applications

**IBRAR HUSSAIN[1,2], RIAZ AHMAD[2], SIRAJ MUHAMMAD[2], KHALIL ULLAH[3], HABIB SHAH[4], AND ABDALLAH NAMOUN[5], (Member, IEEE)**

[1]Department of Computer Science and Information Technology, University of Malakand, Khyber Pakhtunkhawa 18800, Pakistan
[2]Department of Computer Science, Shaheed Benazir Bhutto University (SBBU), Sheringal, Upper Dir, Khyber Pakhtunkhawa 18050, Pakistan
[3]Department of Software Engineering, University of Malakand (UOM), Khyber Pakhtunkhawa 18800, Pakistan
[4]Department of Computer Science, King Khalid University, Abha 61421, Saudi Arabia
[5]Faculty of Computer Science and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia

Corresponding author: Ibrar Hussain (ibrar@sbbu.edu.pk)

The author extends their appreciation to the Deanship of Scientific Research at King Khalid University through Large Group under Grant RGP 2/212/1443.

**ABSTRACT** Document Image Analysis (DIA) is one of the research areas of Artificial Intelligence (AI) that converts document images into machine-readable codes. In DIA systems, Optical Character Recognition (OCR) plays a key role in digitizing document images. The output of an OCR system is further used in many applications including, Natural Language Processing (NLP), Sentiment Analysis, Speech Recognition, and Translation Services. However, standard datasets are an essential requirement for the development, evaluation and comparison of different text recognition techniques. Pashto is one of such low resource languages that lacks availability regarding standard dataset of handwritten text. This paper therefore, addresses the unavailability of standard dataset for the Pashto handwritten text by developing a dataset named Pashto Handwritten Text Imagebase (PHTI). The PHTI is created by collecting handwritten samples from diverse genre of the Pashto language including poetry, religion, short stories, articles, novels, sports, culture and news. The dataset consists of 4, 000 scanned images, written by 400 writers including 200 males and 200 females. These 4, 000 images are further segmented into 36, 082 text-line images. Each text-line image is annotated/ transcribed with UTF-8 codecs. The dataset can be used for many deep learning-based applications including, text recognition, skew detection, gender classification and age-groups classification.

**INDEX TERMS** Artificial intelligence, document image analysis, handwritten text, natural language processing, optical character recognition, speech recognition, Pashto, standard dataset.

## I. INTRODUCTION

The abundance of document images is due to the advancement of recent technologies including acquisition via cameras and scanners. Also, the use of smart phones has increased the creation of document images by taking snapshots of documents and then sharing them in different social media's applications. Such document images are in pixel form and cannot be read or analyzed directly on a computer. In short, these images need conversion from pixel data to a representation that can be easily understood and analyzed by computer. Thus, the Document Image Analysis (DIA) helps to convert document images into machine-readable codes (i.e., Unicode and UTF-8 Codec) [1], [2], [3].

The goal of a DIA system is to extract text and graphics from document images. As images require large storage and their exchange is also very expensive. Therefore, the conversion of document images enables us to exchange, analyze, and store data on the cost of less recourses i.e., time and space. In a typical DIA, it is the Optical Character Recognition (OCR) component that plays a key role in digitizing image documents. Though there exist various effective OCR systems for other cursive scripts, the Pashto language lacks such a system. The major reasons include insufficient real training

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

data (datasets) and less significant research addressing the language specific challenges [4], [5], [6], [7], [8], [9].

Datasets have a significant role in recognizing handwritten scripts. Moreover, these datasets are essential for developing, evaluating, and comparing various text recognition methods. Standard datasets provide a foundation for comparing and evaluating different techniques. However, creating a dataset is a laborious job and the process involves collecting maximum extent of samples from different targeted population [10], [11]. In addition to that, the development of datasets becomes more challenging while it deals with a language whose research is just in its introductory phase [7]. Further, the researchers always try to find out an appropriate dataset that covers all the possible levels of diversity in the target language [12], [13]. Handwritten text recognition has become an important area in the DIA and finding a good dataset for recognition is the main issue. Very little attention has been paid to develop a dataset for the Pashto handwritten text recognition.

Therefore, the major objective of this study is to introduce a dataset that covers different genres of Pashto language and to the best of our knowledge; there is no such dataset that contains handwritten text images regarding the Pashto language. The new dataset is named as Pashto Handwritten Text Image-base (PHTI). The focus is more on covering the different genre of the Pashto language rather than creating text-line images. The major contents of the PHTI are taken from poetry, short stories, news, religion, culture, jokes, and sports. The newly dataset contains images of text-lines with fully annotated ground-truth. The main contributions of this research work are following.

- The dataset comprising of $4,000$ pages of handwritten materials in the Pashto language.
- The $4,000$ images are further segmented into $36,082$ text-line images.
- The dataset is annotated/ transcribed for the supervised learning paradigm with UTF-8 codecs.
- The textual analyses; that is total words, unique words, unique characters and their frequencies are given.
- The dataset also provides a platform for the DIA community to benchmark deep leaning-based recognition systems.

The rest of the paper is organized as follows. Section II reviews the existing datasets that are addressing the Pashto language in the domain of DIA. Section III describes the basic features of the Pashto language including information regarding character set. Section IV explains the overall process of how we have created the PHTI dataset. Section V shares some potential applications that can be benchmarked via PHTI dataset using the supervised learning domain. Finally, section VI concludes the overall work.

## II. RELATED DATASETS

There have been abundant datasets available that have contributed to the domain of DIA for cursive script languages such as Arabic [14], [15], [16], [17], [18], Urdu [19], [20], [21], and Persian [22], [23], [24]. However, this section reviews the datasets that only address the Pashto language. Following are the datasets available for the Pashto language in the DIA research.

Wahab et al. [25] created a synthetic dataset, comprises of 1000 unique ligatures in four different font sizes (4 images of every ligature). The main contribution was the creation of a dataset rather than an OCR system for the Pashto language. In 2016, Ahmad et al. [26] introduced a dataset named KPTI. The dataset was developed for the Pashto cursive script and contains $1,026$ images acquired from the scribed books; the images are further segmented into images of $17,015$ text-line using method reported in [6] and [27]. The scribed text is considered harder than printed material, and comparatively easier than handwritten text.

Khan et al. [28] in 2018 presented a medium-sized dataset consisting of $4,488$ images from 102 different samples of 44 Pashto characters. An A4-sized paper was divided into six columns to collect from the participants. Thus, one person wrote six times the 44 characters of the Pashto language.

Khan et al. [29] presented the Pashto Handwritten Numerals Database (PHND) in 2019. The dataset consists of $50,000$ scanned images of the Pashto Numerals and is publicly available for the research community. The data was collected from three universities: University of Malakand, University of Azad Jammu and Kashmir, and University of Peshawar. The total participants were $1,250$; with a ratio of 65% male and 35% female. Each person wrote four times the digits from 0 to 9.

Khan et al. [30] developed a medium-sized dataset for the recognition of handwritten Pashto characters. The dataset includes $8,800$ images of the Pashto characters, with 200 different representations of each letter.

Amin et al. [12] developed another dataset named Poha in 2020. The poha dataset contains $26,400$ images of 44 isolated Pashto characters and 10 numerals. The data was collected from the Department of Pashto in the University of Peshawar, Pakistan and Tongmyoung University Busan, South Korea. The total participants were 350 while 300 were from University of Peshawar, Pakistan among them.

Similarly, Din et al. [31] developed a benchmark for Pashto handwritten character dataset of Pashto characters; the dataset consists of $43,000$ images. The data was collected on A4-sized paper from 350 university students aged $19-24$. The participants were native speakers of the Pashto language and studied Pashto at their primary school level.

Another dataset presented by Khan et al. [32] named Handwritten Pashto Character Image Dataset (HPCID). The HPCID dataset consists of $14,784$ samples; (336 variants of 44 characters) of the Pashto. The data was collected on A4 sized paper from 360 participants including 169 males and 163 females of different ages, and educational background.

Similarly, Huang et al. [33] presented another medium-sized dataset in 2021. Their dataset consists of $11,352$ images of characters (258 samples of each Pashto character). The data was collected from the Department of Pashto, University

**TABLE 1. Pashto datasets and their important statistics.**

| Dataset/Author Name | Year | Writers | Contents | Statistics |
|---|---|---|---|---|
| FAST-NU [25] | 2009 | Synthetic | Ligatures | 1000 Unique Ligatures (4 font sizes) |
| KPTI [26] | 2016 | Scribed | Sentences | 17, 015 Text-line images of Pashto |
| Khan et al. [28] | 2018 | 17 | Characters | 4, 488 Images of Pashto characters |
| PHND [29] | 2019 | 1250 | Numerals | 50, 000 Images of Pashto numerals |
| Khan et al. [30] | 2020 | 30+ | Character | 88, 00 Images of Pashto characters |
| Poha [12] | 2020 | 350 | Char/Digits | 26, 400 Images of characters & numerals |
| PHCD [31] | 2020 | 350 | Characters | 43, 000 Images of Pashto characters |
| HPCID [32] | 2021 | 336 | Characters | 14, 784 Images of Pashto characters |
| Huang et al. [33] | 2021 | 258 | Characters | 11, 352 Images of Pashto characters |
| Rehman et al. [34] | 2021 | 550 | Digits | 60, 000 Images of Pashto digits |
| Gold Standard Pashto Dataset [35] | 2021 | Printed Books | Sentences | 300 Text-line images of Pashto |

of Swabi, Pakistan with different gender, age, and educational background.

In 2021, Rehman et al. [34] developed a dataset for Pashto digits, the dataset consists of 60, 000 images of Pashto digits. Likewise, KPTI, Han et al. [35] presented a dataset named Gold-Standard Pashto Dataset, consists of 300 text-line images, selected from three Pashto books. Table 1 provides a comparative overview of Pashto script datasets.

The literature suggests that none of the datasets, except KPTI [26], collected continuous text-lines of the Pashto language. The majority of the datasets are developed for isolated characters and numerals. The KPTI dataset is based on scribed material, which means calligraphers write the samples, and such calligraphers are famous for their beautiful/ uniform writing. Though the KPTI dataset provides a clear aspect for being used in generalization, it lacks the real handwritten samples of the Pashto language. Therefore, the literature itself reveals the gap and we need a comprehensive dataset that will not only show the diverse materials of the Pashto language but also exploit the natural variations of the native hand writers.

## III. PASHTO LANGUAGE

The Pashto language is originated from Eastern Iranian language [7], which is derived from the Indo-European family. Pashto is the second largest language in Pakistan and one of the two official languages of Afghanistan. Pashto's native speakers are ranging from 55-60 million across the world [7], [30]. Pashto language writing is cursive by nature and its writing has close resemblance with Arabic script as it uses several Arabic characters. However, they may have differences in their structure, syntax, pronunciation and lexicon etc. [31], [36]. It is written from right to left, but numerals are written from left to right. According to a prominent historian, Abdul Hai Habibi [37], the first-ever book of Pashto was written in the 8th century. It reveals the fact that the Pashto language has a long association with novel, history, poetry and religious themes. Poetry is very famous in the Pashto language.

### A. PASHTO LANGUAGE IN THE GLOBAL CONTEXT

Apart from Afghanistan and Pakistan, Pashto is taught in the United Kingdom (UK) and United States (US) universities and Jamia Millia Islamia in New Delhi, India [38]. The United Nations Organization (UNO) officially recognizes 82 languages around the globe. Pashto is ranked 43rd among them [38]. Pashto is also spoken by Pashto speakers around the world, with a significant population in the regions of Germany, the United Arab Emirates, the US, UK, Iran, Tajikistan, Canada, India, Malaysia, and Singapore [38], [39]. It has also a special role in international media houses such as BBC,[1] Voice of America, Afghanistan National Television and Pakistan Television, AVT Khyber etc.

### B. PASHTO CHARACTERS SET

An interesting aspect of the Pashto language is its character set, as it is the superset of Arabic, Urdu and Persian character sets. The generic character set of Pashto provides a solid hypothesis for transfer learning and domain adaptation due to similar visual cues [40], [41]. FIGURE 1 illustrates the mutual relationship between the characters of Pashto, Urdu, Arabic, and Persian.

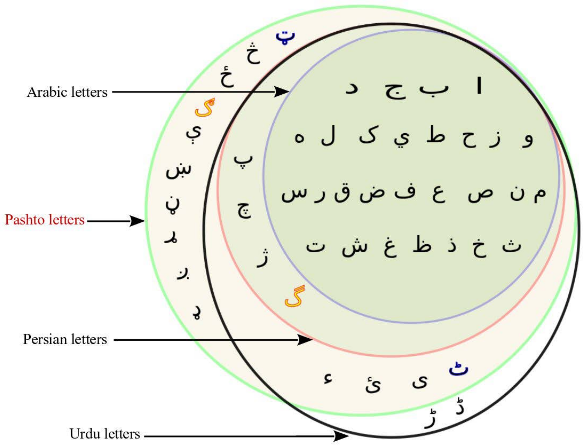


**FIGURE 1. Pashto character set and overlapping similar alphabets with Urdu, Persian and Arabic [7].**

## IV. DESCRIPTION OF THE PHTI DATASET

According to the researchers in [29] and [42], the need for a sufficient amount of training data is a considerable

[1] https://www.bbc.com/pashto

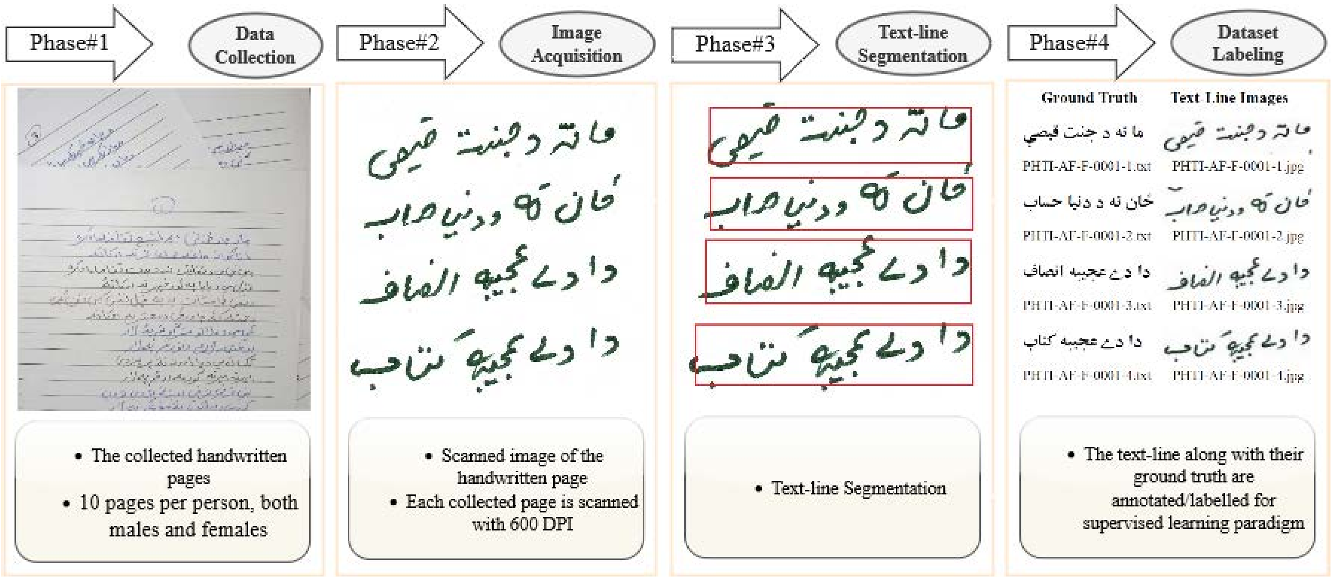


**FIGURE 2.** **PHTI dataset and its development process.**

requirement for deep learning-based methods. The more data from all genres we have, the more chances will be there for generalization of deep learning-based methods. Thus, PHTI is a real dataset developed for the research community keeping in view the generalization of the recognition systems. Further, this section explains the process of how we created the PHTI dataset. The overall process includes; data collection process, image acquisition process, text-line segmentation process, annotation process and detailed statistics of the PHTI dataset. FIGURE 2 shows the graphical representation of the process with which the PHTI dataset is created. The dataset is freely available via GitHub link.[2]

### A. DATA COLLECTION PROCESS

The PHTI dataset is developed with an intention to cover many Pashto language genres such as prose, poetry, short story, history, sports, BBC Pashto, and religion. In this context, data were collected from a diverse background of learners having different levels of educational background, including University, College, School, and Deeni Madaras.[3] Students of both sexes participated in the process, and we provided each student with ten pages of sample text to write in their natural writing style. Also, the participants were advised to use any paper of their own choice. Table 2 shows education background, age, gender, and data collection statistics. The data were taken from 17 different sources, which cover a variety of genres in the Pashto language. Table 3 shows each source and number of text-lines extracted from a specific source.

### B. IMAGE ACQUISITION

Image acquisition is the first step in dataset creation. Scanned-based images are considered suitable for digitization, while camera-captured images mostly contain skewness, blurriness, perspective distortion, wrapping skew, shadow, and light reflection. On the other hand, scanned-based images often include skewness [7]. We have tried our best to avoid skewness while scanning the collected data. Each collected handwritten sample is scanned with 600 DPI using an HP Scanjet scanner. FIGURE 3 shows the scanned pages/images of the dataset along with their ground-truth. Each acquired image is named PHTI-XX-Y-ZZZZ.jpg; PHTI is a constant prefix that refers to our new dataset. The "XX" annotates the source name, and "Y" represents the gender, while "ZZZZ" describes the page number in the respective source.

**TABLE 2.** **Education/profession, age, and gender wise data collection statistics.**

| Education/ Profession | Age | Male | Female | Total |
|---|---|---|---|---|
| $10^{th}$ Grade Students | $15-20$ Years | 50 | 50 | 100 |
| Bachelor Student | $20-30$ Years | 50 | 50 | 100 |
| MPhil/Ph.D Scholars | $30-40$ Years | 50 | 50 | 100 |
| Faculty & Staff | Above 40 Years | 50 | 50 | 100 |

### C. TEXT-LINE SEGMENTATION

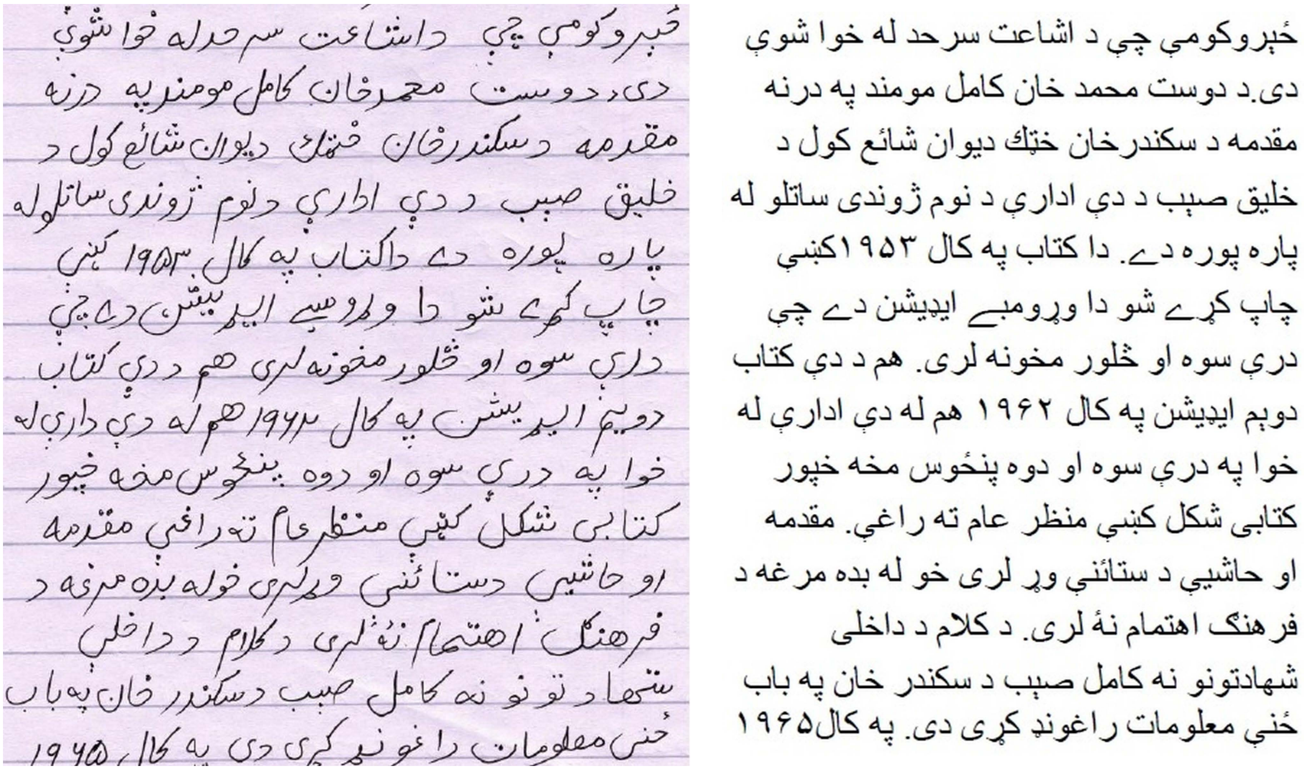
A text-line image is an image of one text-line from a scanned page. In offline datasets, text line segmentation is an essential pre-processing step. It is crucial because the performance of OCR systems depends mainly on the quality and cleanness of text-line image fed as input. One of the complex challenges in developing a reliable dataset is the text-line segmentation of handwritten documents because of the natural variation in handwriting [41], [42], [43]. The PHTI images are further segmented into 36,082 text-line images using the method reported in [6] and [27]. After segmentation into text-line images, the notion used for image file names is extended with another number. For example; a name for a text-line

[2] https://github.com/adqecsbbu/PHTI

[3] Schools/Institutes that teaches law and religion.

**TABLE 3.** Source wise statistics.

| Source | | Text-line Extracted | Test 15% | Validation 15% | Training 70% | Contents |
|---|---|---|---|---|---|---|
| Books | Alwat | 2,012 | 302 | 302 | 1,408 | Poetry |
| | Da Madiny Yaduna | 1,641 | 246 | 246 | 1,149 | Traveling story |
| | Pashto Literature | 3,134 | 470 | 470 | 2,194 | Literary terms |
| | Sandareez Ilham | 1,886 | 283 | 283 | 1,320 | Religion |
| | Atharzan | 1,053 | 158 | 158 | 737 | |
| | Salam Pakistan | 1,672 | 251 | 251 | 1,170 | |
| | Rangona Sangsar | 1,250 | 188 | 188 | 875 | |
| Dissertation | 20th Century Sociology in Pashto | 2,593 | 389 | 389 | 1,815 | Sociology |
| | History of Pashto Academy | 5,487 | 823 | 823 | 3,841 | Culture |
| | Prof. M. Nawaz Tair | 3,279 | 492 | 492 | 2,295 | History |
| | Aesthetics | 1,439 | 216 | 216 | 1,007 | Education |
| | History & Geography of Malakand | 2,466 | 370 | 370 | 1,726 | |
| | Khushal Khan Khattak Poetry | 2,468 | 370 | 370 | 1,728 | |
| General | Sports | 1,220 | 183 | 183 | 854 | Sports |
| | Afsana | 1,338 | 201 | 201 | 937 | Short Story |
| | News & Health | 1,496 | 224 | 224 | 1,047 | JokesNews |
| | Pashto Jokes & Tappay | 1,648 | 247 | 247 | 1,154 | |
| **Sum** | | **36,082** | **5,412** | **5,412** | **25,25** | |



(a) Acquired Image

ځېروكومې چې د اشاعت سرحد له خوا شوې
دى.د دوست محمد خان كامل مومند په درنه
مقدمه د سكندرخان ختټك ديوان شائع كول د
خليق صېب د دې ادارې د نوم ژوندى ساتلو له
پاره پوره دے. دا كتاب په كال ۱۹۵۳كښې
چاپ كړے شو دا وړومبے ايډيشن دے چې
درې سوه او خلور مخونه لرى. هم د دې كتاب
دوېم ايډيشن په كال ۱۹۶۲ هم له دې ادارې له
خوا په درې سوه او دوه پنځوس مخه خپور
كتابى شكل كښې منظر عام ته راغې. مقدمه
او حاشيې د ستائنې وړ لرى خو له بده مرغه د
فرهنګ اهتمام نۀ لرى. د كلام د داخلى
شهادتونو نه كامل صېب د سكندر خان په باب
څنې معلومات راغونډ كړى دى. په كال۱۹۶۵

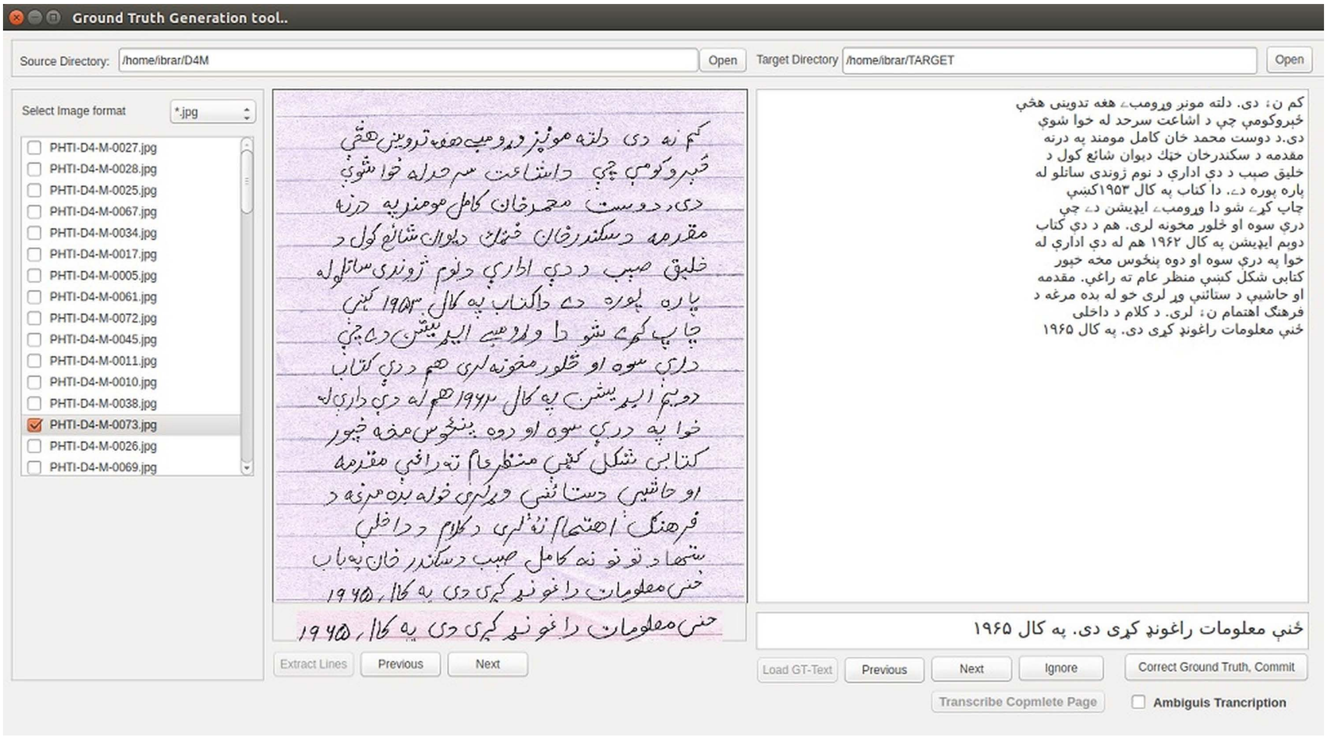
(b) Ground truth

**FIGURE 3.** An instance from PHTI dataset of acquired page along with ground truth.

image file is now "PHTI-XX-Y-ZZZZ-N.jpg". Here, the last "N" in the file name refers to a certain text-line number in a document image. FIGURE 4 shows the graphical user interface developed by [7] for text-line segmentation and dataset labeling. Table 4 shows the naming conventions for a text-line image and its corresponding ground-truth.

### D. DATASET LABELING

To exploit the supervised learning paradigm, data must be in input-output pairs. It means, the scanned images need proper and clean annotation/ labeling to meet the requirement of recognition systems based on supervised learning domain. Usually, labeling is considered to be the most laborious and costly process. However, for the PHTI dataset, we have given the material to the participants that are mostly in transcribed form. For example; books, dissertations and news were already in digitized form. However, each scanned image as well as image of a segmented text-line was manually checked and validated regarding its ground-truth, and if there was a mismatch or an error, then the image as well as its



Source Directory: /home/ibrar/D4M

Target Directory /home/ibrar/TARGET

Select Image format *.jpg

- □ PHTI-D4-M-0027.jpg
- □ PHTI-D4-M-0028.jpg
- □ PHTI-D4-M-0025.jpg
- □ PHTI-D4-M-0067.jpg
- □ PHTI-D4-M-0034.jpg
- □ PHTI-D4-M-0017.jpg
- □ PHTI-D4-M-0005.jpg
- □ PHTI-D4-M-0061.jpg
- □ PHTI-D4-M-0072.jpg
- □ PHTI-D4-M-0045.jpg
- □ PHTI-D4-M-0011.jpg
- □ PHTI-D4-M-0010.jpg
- □ PHTI-D4-M-0038.jpg
- ☑ PHTI-D4-M-0073.jpg
- □ PHTI-D4-M-0026.jpg
- □ PHTI-D4-M-0069.jpg

کم ن؛ دی. دلته مونږ ورومبې هغه تدوینی هڅې
څېروکومې چې د اشاعت سرحد له خوا شوې
دی.د دوست محمد خان کامل مومند په درنه
مقدمه د سکندرخان خټک دیوان شائع کول د
خلیق صېب د دې اداري د نوم ژوندی ساتلو له
پاره پوره دے. دا کتاب په کال ۱۹۵۳کښې
چاپ کړے شو دا ورومبے ایډیشن دے چې
درې سوه او څلور مخونه لری. هم د دې کتاب
دویم ایډیشن په کال ۱۹۶۲ هم له دې اداري له
خوا په درې سوه او دوه پنځوس مخه خپور
کتابی شکل کښې منظر عام ته راغې. مقدمه
او حاشیې د ستائنې وړ لری خو له بده مرغه د
فرهنگ اهتمام ن؛ لری. د کلام د داخلی
څنې معلومات راغونډ کړی دی. په کال ۱۹۶۵

څنې معلومات راغونډ کړی دی. په کال ۱۹۶۵

**FIGURE 4.** Text-line segmentation and dataset labeling.

**TABLE 4.** Naming Convention of the PHTI dataset.

| Image/ Ground-truth | Naming Convention | Legends |
|---|---|---|
| Image | PHTI-AF-F-0001.jpg | PHTI: Pashto Handwritten Text Imagebase |
| Ground-truth | PHTI-AF-F-0001.txt | AF= Source, F=Gender(F: Female, M: Male) |
| Text-line | PHTI-AF-F-0001-1.jpg | 0001 as Page No from a Particular Source) |
| Ground-truth | PHTI-AF-F-0001-1.txt | 1 as Text-line No in a Particular Page) |

annotation was corrected. However, there might be ±4 % error in the validation process. The ground-truth is stored in a text file with UTF-8 codec. The file name is the same as the image file name, except the extension.

### E. STATISTICS OF PHTI

This section provides the complete statistics of the PHTI dataset, including the number of writers, total pages, total extracted text-lines, the number of unique characters and words, minimum and maximum characters per line, and the most frequently occurring words. A total of 400 persons including 200 males and 200 females participated, and a total of 3, 970/4, 000 pages of handwriting were scanned. However, only 30 pages had errors, omissions, and cutting; therefore, they were ignored. The remaining 3, 970 images are further segmented into overall images of 36, 082 text-lines. The overall text-line images are split into 70% training, 15% test, and 15% validation sets. Table 5 shows the statistics of the PHTI dataset while Table 6 shows the top most occurring words and their frequencies in the PHTI dataset.

**TABLE 5.** Statistics of PHTI.

| Description | Number |
|---|---|
| Total writer | 400 |
| Female writer | 200 |
| Male writer | 200 |
| Total pages | 3, 970 |
| Pages per writer | 10 |
| Average line per page | 10 |
| Total text-lines images | 36, 082 |
| Text-line written by females | 18, 069 |
| Text-line written by males | 18, 013 |
| Total words | 4, 20, 961 |
| Unique words | 33, 330 |
| Total character | 169 |
| Minimum height of image | 28 |
| Maximum height of image | 237 |
| Minimum width of image | 78 |
| Maximum width of image | 1, 263 |

## V. APPLICATIONS OF THE PHTI DATASET

The PHTI dataset is designed in a manner to facilitate the following research areas in several aspects. The applications are given in the following sections.

**TABLE 6.** Top 50 most occurring Pashto words in PHTI.

| S.No | Word | Frequency | S.No | Word | Frequency |
|---|---|---|---|---|---|
| 1 | د | 24,706 | 26 | پښتو | 1,291 |
| 2 | په | 13,385 | 27 | خان | 1,267 |
| 3 | او | 9,425 | 28 | زما | 1,096 |
| 4 | چې | 6,446 | 29 | مې | 1,086 |
| 5 | کبنې | 5,794 | 30 | که | 1,066 |
| 6 | نه | 3,673 | 31 | يې | 1,059 |
| 7 | دے | 3,420 | 32 | نو | 1,009 |
| 8 | ته | 2,952 | 33 | هر | 974 |
| 9 | دَ | 2,920 | 34 | داسې | 916 |
| 10 | دي | 2,792 | 35 | شي | 880 |
| 11 | دا | 2,679 | 36 | کبنې | 857 |
| 12 | هم | 2,652 | 37 | تر | 802 |
| 13 | نې | 2,584 | 38 | شو | 800 |
| 14 | به | 2,326 | 39 | وي | 799 |
| 15 | يو | 2,235 | 40 | ډېر | 799 |
| 16 | هغه | 1,928 | 41 | وو | 758 |
| 17 | سره | 1,927 | 42 | بيا | 707 |
| 18 | دغه | 1,717 | 43 | څۀ | 701 |
| 19 | پۀ | 1,643 | 44 | وه | 690 |
| 20 | له | 1,606 | 45 | ستا | 690 |
| 21 | ده | 1,602 | 46 | څه | 630 |
| 22 | نۀ | 1,564 | 47 | بې | 601 |
| 23 | خو | 1,563 | 48 | دے | 603 |
| 24 | خپل | 1,430 | 49 | وخت | 593 |
| 25 | دي | 1,399 | 50 | ما | 583 |

### A. NATURAL LANGUAGE PROCESSING (NLP)

The NLP research community regarding the Pashto language can utilize all the annotated data of the PHTI dataset, as we have collected all the ground-truth data in one text file. The natural flow of each resource is kept intact. Therefore, the dataset is a suitable candidate for the exploration of research regarding NLP for the Pashto language. The same data can also be used for text-to-speech conversion, keyword spotting and for the analysis of words embedding.

### B. GENDER AND AGE CLASSIFICATION

The PHTI dataset is created by collecting handwritten materials equally from opposite genders. Thus, the PHTI dataset can be used to differentiate among the distinctive features exhibiting different writing styles associated with different genders while writing the Pashto language. Similarly, the PHTI dataset can also be used for age classification, as the materials were equally collected from various age groups of both males and females.

### C. SKEW AND LINE-SEGMENTATION

Although the instances in the PHTI that have skew are very limited, one can use this dataset for evaluating skew-detection and correction approaches. The maximum skew angle present in the images is in between $-5°$ to $+5°$. Another, the PHTI dataset in terms of scanned documents has $3,970$ images. Due to these $3,970$ images, the PHTI dataset is a suitable benchmark for evaluating the available line-segmentation approaches. Similarly, new segmentation approaches may also be proposed with the help of this dataset.

### D. OCR APPLICATION

The major contribution that we expect from the PHTI dataset is in the area of OCR application. We believe that PHTI is so far the most comprehensive and largest dataset that contains sufficient material regarding the Pashto language. Also, the contents are rich enough that almost all the possible styles of handwriting are present there. Collecting data from different persons having different qualification backgrounds ensures that the models learned with this data will have high chances for generalization. The deep-learning models can be easily trained and evaluated for the Character Error Rate (CER) and Word Error Rate (WER) using the PHTI dataset.

## VI. CONCLUSION

In this work, we addressed the unavailability of handwritten text images for the Pashto language to facilitate the research community in the domain of DIA. Thus, we have created the most comprehensive and the largest dataset so far in the Pashto language. The dataset is written by 400 individuals of different genders, ages, education, qualification, and profession. The name of the dataset is PHTI. The dataset consists of 3, 970 scanned pages. The scanned pages are further segmented into 36, 082 text-line images; for each text-line image, a corresponding ground-truth exists in the UTF-8 codecs. The data collection was made from 17 different sources, including poetry, short stories, history, culture, news, health, and sports, to cover the elementary classes of the target language for maximum generalization. The dataset is freely available for all research communities especially interested in Pashto cursive script recognition, gender and age classification, NLP, skew, and text-line segmentation.

## VII. FUTURE WORK

To our knowledge, the PHTI dataset is the first and largest dataset that presents handwritten Pashto text-lines written by 400 different writers. The future work includes adding more Pashto pages to the dataset, including left and right-handed writer samples, age identification, writer identification, and baseline for recognition.

## ACKNOWLEDGMENT

The authors would like to thank Dr. Ali Khail Daryab, Chairperson, Department of Pashto Language, University of Malakand, Dr. Nasrallah Jan Wazir, the Director Pashto Academy, University of Peshawar, Dr. Jan Alam, and Dr. Khalil-Ur Rahman for their help and support regarding Pashto text collection and for providing them with books, dissertation, and other valuable materials.

## REFERENCES

[1] G. Nagy, "Twenty years of document image analysis in PAMI," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 38–62, Jan. 2000.

[2] H. S. Baird, H. Bunke, and K. Yamamoto, *Structured Document Image Analysis*. Berlin, Germany: Springer, 2012.

[3] J. Memon, M. Sami, R. A. Khan, and M. Uddin, "Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR)," *IEEE Access*, vol. 8, pp. 142642–142668, 2020.

[4] R. Ahmad, S. Naz, M. Afzal, S. Rashid, M. Liwicki, and A. Dengel, "A deep learning based Arabic script recognition system: Benchmark on KHAT," *Int. Arab J. Inf. Technol.*, vol. 17, no. 3, pp. 299–305, May 2020.

[5] R. Ahmad, S. H. Amin, and M. A. U. Khan, "Scale and rotation invariant recognition of cursive pashto script using SIFT features," in *Proc. 6th Int. Conf. Emerg. Technol. (ICET)*, Oct. 2010, pp. 299–303.

[6] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, and A. Dengel, "Text-line segmentation of large titles and headings in Arabic like script," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Apr. 2017, pp. 168–172.

[7] R. Ahmad, "An end-to-end OCR system for Pashto cursive script," Ph.D. dissertation, Dept. Comput. Sci., Technische Universität Kaiserslautern, Kaiserslautern, Germany, 2018.

[8] I. Bar-Yosef, *Computer Vision Analysis of Historical Documents*. Beersheba, Israel: Ben Gurion Univ., 2009.

[9] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, A. Dengel, and T. Breuel, "Recognizable units in Pashto language for OCR," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1246–1250.

[10] U.-V. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition," in *Proc. 5th Int. Conf. Document Anal. Recognit. (ICDAR)*, 1999, pp. 705–708.

[11] R. Hussain, A. Raza, I. Siddiqi, K. Khurshid, and C. Djeddi, "A comprehensive survey of handwritten document benchmarks: Structure, usage and evaluation," *EURASIP J. Image Video Process.*, vol. 2015, no. 1, pp. 1–24, Dec. 2015.

[12] M. S. Amin, S. M. Yasir, and H. Ahn, "Recognition of Pashto handwritten characters based on deep learning," *Sensors*, vol. 20, no. 20, p. 5884, Oct. 2020.

[13] M. W. Sagheer, C. L. He, N. Nobile, and C. Y. Suen, "A new large Urdu database for off-line handwriting recognition," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2009, pp. 538–546.

[14] I. Saleh Al-Sheikh, M. Mohd, and L. Warlina, "A review of Arabic text recognition dataset," *Asia–Pacific J. Inf. Technol. Multimedia*, vol. 9, no. 1, pp. 69–81, Jun. 2020.

[15] S. Yousfi, S.-A. Berrani, and C. Garcia, "ALIF: A dataset for Arabic embedded text recognition in TV broadcast," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1221–1225.

[16] M. Tounsi, I. Moalla, and A. M. Alimi, "ARASTI: A database for Arabic scene text recognition," in *Proc. 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, Apr. 2017, pp. 140–144.

[17] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Märgner, and H. E. Abed, "KHATT: Arabic offline handwritten text database," in *Proc. Int. Conf. Frontiers Handwriting Recognit.*, Sep. 2012, pp. 449–454.

[18] A. Lawgali, M. Angelova, and A. Bouridane, "HACDB: Handwritten Arabic characters database for automatic character recognition," in *Proc. Eur. Workshop Vis. Inf. Process. (EUVIP)*, 2013, pp. 255–259.

[19] S. B. Ahmed, S. Naz, S. Swati, I. Razzak, A. I. Umar, and A. A. Khan, "UCOM offline dataset—An Urdu handwritten dataset generation," *Int. Arab J. Inf. Technol.*, vol. 14, no. 2, pp. 1–7, 2017.

[20] A. A. Chandio, M. Asikuzzaman, M. Pickering, and M. Leghari, "Cursive-text: A comprehensive dataset for end-to-end Urdu text recognition in natural scene images," *Data Brief*, vol. 31, Aug. 2020, Art. no. 105749.

[21] H. Ali, A. Ullah, T. Iqbal, and S. Khattak, "Pioneer dataset and automatic recognition of Urdu handwritten characters using a deep autoencoder and convolutional neural network," *Social Netw. Appl. Sci.*, vol. 2, no. 2, pp. 1–12, Feb. 2020.

[22] A. Soleimani, K. Fouladi, and B. N. Araabi, "UTSig: A Persian offline signature dataset," *IET Biometrics*, vol. 6, no. 1, pp. 1–8, Jan. 2017.

[23] A. E. Ghahnavieh, M. Enayati, and A. A. Raie, "Introducing a large dataset of Persian license plate characters," *J. Electron. Imag.*, vol. 23, no. 2, Apr. 2014, Art. no. 023015.

[24] A. Alaei, P. Nagabhushan, and U. Pal, "A new dataset of Persian handwritten documents and its segmentation," in *Proc. 7th Iranian Conf. Mach. Vis. Image Process.*, Nov. 2011, pp. 1–5.

[25] M. Wahab, H. Amin, and F. Ahmed, "Shape analysis of Pashto script and creation of image database for OCR," in *Proc. Int. Conf. Emerg. Technol.*, Oct. 2009, pp. 287–290.

[26] 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 *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Oct. 2016, pp. 453–458.

[27] R. Ahmad, S. Naz, and I. Razzak, "Efficient skew detection and correction in scanned document images through clustering of probabilistic Hough transforms," *Pattern Recognit. Lett.*, vol. 152, pp. 93–99, Dec. 2021.

[28] 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," 2019, *arXiv:1904.03391*.

[29] K. Khan, B.-H. Roh, J. Ali, R. U. Khan, I. Uddin, S. Hassan, R. Riaz, and N. Ahmad, "PHND: Pashtu handwritten numerals database and deep learning benchmark," *PLoS ONE*, vol. 15, no. 9, Sep. 2020, Art. no. e0238423.

[30] S. Khan, S. Nazir, H. U. Khan, and A. Hussain, "Pashto characters recognition using multi-class enabled support vector machine," *Comput., Mater. Continua*, vol. 67, no. 3, pp. 2831–2844, 2021.

[31] I. Uddin, D. A. Ramli, A. Khan, J. I. Bangash, N. Fayyaz, A. Khan, and M. Kundi, "Benchmark Pashto handwritten character dataset and Pashto object character recognition (OCR) using deep neural network with rule activation function," *Complexity*, vol. 2021, pp. 1–16, Mar. 2021.

[32] S. Khan, H. U. Khan, and S. Nazir, "Offline Pashto characters dataset for OCR systems," *Secur. Commun. Netw.*, vol. 2021, pp. 1–7, Jul. 2021.

[33] 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, pp. 1–8, Mar. 2021.

[34] M. Z. Rehman, N. M. Nawi, M. Arshad, and A. Khan, "Recognition of cursive Pashto optical digits and characters with trio deep learning neural network models," *Electronics*, vol. 10, no. 20, p. 2508, Oct. 2021.

[35] Y. Han and M. Rychlik, "Development of a gold-standard Pashto dataset and a segmentation app," *Inf. Technol. Libraries*, vol. 40, no. 1, pp. 1–15, Mar. 2021.

[36] H. Tegey and B. Robson, "A reference grammar of Pashto," Center Appl. Linguistics, Washington, DC, USA, Tech. Rep. ED 399825, 1996.

[37] P. A. Habibi, "The cultural, social and intellectual state of the people of Afghanistan in the era just before the advent of Islam," *Afghanistan*, vol. 2, no. 3, pp. 1–7, 1967.

[38] J. Iqbal, A. Zaman, and A. Ghafar, "Inclusion of Pashto in 'O'level Cambridge education," *VFAST Trans. Educ. Social Sci.*, vol. 1, no. 1, pp. 35–38, 2013.

[39] A. Birch, B. Haddow, A. V. M. Barone, J. Helcl, J. Waldendorf, F. Sánchez-Martínez, M. L. Forcada, V. M. Sánchez-Cartagena, J. A. Pérez-Ortiz, M. Esplà-Gomis, W. Aziz, L. Murady, S. Sariisik, P. van der Kreeft, and K. Macquarrie, "Surprise language challenge: Developing a neural machine translation system between Pashto and English in two months," in *Proc. 18th Biennial Mach. Transl. Summit*, vol. 1, 2021, pp. 92–102.

[40] M. Husnain, M. M. S. Missen, S. Mumtaz, M. Z. Jhanidr, M. Coustaty, M. M. Luqman, J.-M. Ogier, and G. S. Choi, "Recognition of Urdu handwritten characters using convolutional neural network," *Appl. Sci.*, vol. 9, no. 13, p. 2758, Jul. 2019.

[41] R. Ahmad, S. Naz, M. Z. Afzal, S. F. Rashid, M. Liwicki, and A. Dengel, "The impact of visual similarities of Arabic-like scripts regarding learning in an OCR system," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, Nov. 2017, pp. 15–19.

[42] K. Khan, "Pashtu numerals recognition through convolutional neural networks," *J. Appl. Emerg. Sci.*, vol. 9, no. 2, p. 91, 2019.

[43] J. H. AlKhateeb, "Word based off-line handwritten Arabic classification and recognition. Design of automatic recognition system for large vocabulary offline handwritten Arabic words using machine learning approaches," Ph.D. dissertation, Univ. Bradford, Bradford, U.K., 2010.

**IBRAR HUSSAIN** received the B.C.S. degree (Hons.) from the Department of Computer Science, University of Peshawar, Pakistan, and the M.S. degree in computer science from the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Dir Upper. He is currently a Ph.D. Scholar at the Department of Computer Science and Information Technology, University of Malakand, Chakdara, Dir Lower, Khyber Puktunkhwa, Pakistan.

**RIAZ AHMAD** received the M.S. degree (Hons.) in computer science from NUCES (FAST) University, Pakistan, in 2010, and the Ph.D. degree from the Technical University of Kaiserslautern, Germany, in 2018. He also worked as a member of the German Research Center for Artificial Intelligence (DFKI), Multimedia Analysis and Data Mining (MADM) Research Group, Kaiserslautern, Germany. Currently, he is heading the Computer Science Department, Shaheed Benazir Bhutto University, Sheringal, Pakistan. His research interests include document image analysis, image processing, and optical character recognition. More specifically, his work examines the challenges posed by cursive script languages in the field of OCR systems. In addition to that, he is studying the behavior of deep learning architectures in the field of OCR in terms of invariant approaches against scale and rotation variation in Pashto cursive text.

**SIRAJ MUHAMMAD** received the M.Phil. degree from Quaid-i-Azam University, Islamabad, in 2010, and the Ph.D. degree from the Asian Institute of Technology (AIT), Thailand, in 2020. He was a Software Engineer at Elixir Technologies of Pakistan, Islamabad, from 2010 to 2011. Currently, he is an Assistant Professor with the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Pakistan. His research interests include reverse engineering, computer vision, image processing, deep learning, and natural language processing.

**KHALIL ULLAH** received the Graduate degree in computer systems engineering from the University of Engineering and Technology Peshawar, Pakistan, in 2006, the Master of Science (M.S.) degree in electronics and communications engineering from Myongji University, South Korea, in 2009, and the Ph.D. degree in biomedical engineering from the LISiN Politecnico di Torino under Erasmus Mundus Expert II Fellowship, in 2016. Currently, he is acting as an Assistant Professor and the Head of the Software Engineering Department, University of Malakand. His research interests include extracting muscle anatomical and physiological information from high-density electromyography, computer vision, digital signal and image processing, and deep learning with applications to medical healthcare.

**HABIB SHAH** received the Ph.D. degree from the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, in 2013. He is currently an Assistant Professor and the Head of the Research Unit, College of Computer Science, King Khalid University, Saudi Arabia. His research interests include artificial intelligence, learning algorithms, data mining techniques, the IoT, time series analysis, and optimization. He has successfully published more than 50 articles in various international SCI and Scopus journals and conference proceedings. He is an editorial board, a guest editor, and act as a reviewer for various journals and conferences as well. He has also served as a program committee member and a co-organizer for numerous international conferences/workshops. Currently, he is working on three research projects of KKU.

**ABDALLAH NAMOUN** (Member, IEEE) received the bachelor's degree in computer science and the Ph.D. degree in informatics from the University of Manchester, U.K., in 2004 and 2009, respectively. He is currently an Associate Professor in intelligent interactive systems and the Head of the Information Systems Department, Faculty of Computer and Information Systems, Islamic University of Madinah. He has authored more than 50 publications in research areas spanning intelligent systems, human–computer interaction, software engineering, and technology acceptance and adoption. He has extensive experience in leading complex research projects (worth more than 21 million Euros) with several distinguished SMEs, such as SAP, BT, and ATOS. He has investigated user needs and interaction with modern interactive technologies, the design of composite software services, and methods for testing the usability and acceptance of human interfaces. His research interests include integrating state-of-the-art artificial intelligence approaches in designing and developing interactive systems.