

MULTI-ASPECT URDU HANDWRITING DATA COLLECTION

MUHAMMAD FAHAD

Government Graduate College Civil Lines Multan, Pakistan.

MALIK MUHAMMAD SAAD MISSEN

Faculty of Computing, The Islamia University of Bahawalpur, Pakistan.

MUJTABA HUSNAIN*

Faculty of Computing, The Islamia University of Bahawalpur, Pakistan.
Email: mujtaba.husnain@iub.edu.pk

ALISAMAD

Faculty of Computing, The Islamia University of Bahawalpur, Pakistan.

DALER ALI

Faculty of Computing, The Islamia University of Bahawalpur, Pakistan.

ASAD ALI

Department of Computer Science and Information Technology, National College of Business Administration and Economics Bahawalpur Campus, Pakistan.

Abstract

Urdu script is categorized as one of the cursive and bidirectional script derived from Arabic and Persian script; this is the reason Urdu script shares almost similar challenges and issues but with higher complexity. There is a lack of freely available public datasets for the research in the field of Urdu handwriting recognition. In this paper, we propose a multi-aspect Urdu handwriting data collection by inviting a number of native Urdu speakers from different social groups. To make the dataset more comprehensive, both the isolated characters and the ligatures are included in the dataset. Furthermore, the persons having physical disability are also invited for data collection to make the corpus more comprehensive. We also give a review of existing data collections for Urdu handwriting recognition and give a comparison of the proposed data collection with existing ones.

Keywords: Urdu Handwritten Text, Intelligent Character Recognition, Multi-Aspect Data Generation

1. INTRODUCTION

Handwriting recognition is an active area of research in the field of pattern recognition and has various application in industrial and professional applications. Some of these applications include forms processing in government, administrative, health and academic institutes; postal address recognition, processing of bank cheques etc. Handwriting recognition concerns with automatic transforming a source language into its symbolic representation. The source language can be represented either in its spatial (offline) or temporal (online) [1] form in graphical marks. In depth analysis of handwritten text give rise to a number of useful applications such as author profiling, named entity recognition, recognition of overlapped characters etc.

In late 1950's the first Optical Character Recognition (OCR) system was developed for the recognition of Latin text [2] [3] which deals with recognition of numerals only. With the advancements in OCR, the systems available nowadays is expanded to

recognize Latin script, and characters of a variety of other languages like Chinese, Japanese, Arabic, Persian etc. Optical character recognition (OCR) of Urdu script was started in late 2000 and the first work on Urdu OCR is published in 2004. The literature review identified the fact that there has been lack of research efforts in Urdu handwritten text recognition as compared to recognition of other language scripts [6] [7]. There are few Urdu OCR systems for printed text that are commercially available [8] [9] but to-date there is no system available for Urdu handwritten text recognizing.

In verbal communication, Urdu language adopted many dialects across the regions but in formal writing Urdu script uses standard way. It is also observed that Urdu written script shares similarities with other languages like Arabic and Persian [10]. Therefore, automatic interpretation of Urdu handwritten text would have prevalent and ubiquitous benefits.

Urdu handwriting is used for official assignments and communications in all the organizations in Pakistan. Most of the time, all this data (provided in the form of handwritten applications or forms) is typed into computers for further processing that requires a huge man power, processing equipment, time, money and other resources. For example, data entry in the National Database and Registration Authority (NADRA) offices for processing requests of National Identity Cards (NIC), student's applications in government institutes, signature on banker cheques etc. require an automatic Urdu handwritten text recognition tool to recognize the text and process the documentation in real time environment. Furthermore, the system should provide facility to save the information in some appropriate database. This will reduce a significant number of resources in daily official matters considering the nature of this task.

The development of Urdu handwritten recognition system can assist in reading historic Urdu manuscripts to make the content of these manuscripts available. The content of manuscripts is written in clear and readable way as compared to handwritten text which makes the task of recognition of contents of manuscript much simpler. On the other hand, some issues associated with handwritten text make the task of developing the system for recognition of handwritten text more challenging and complicated. Some of these issues are differences in writing style (even from the same author), image degradation due to cursive nature of the script, poor quality or illegible handwriting etc. Urdu script is considered as more complicated as compared to Arabic and Persian script.

One of the main reasons for the Urdu script to attain less attention is developing a state-of-art OCRs system is the lack of a standard database for handwritten Urdu script. Although there are some available databases for printed Urdu script however the availability for Urdu handwritten data collection is much less as compared to other Urdu-based scripts like Persian and Arabic. Furthermore, it is also a challenging task to use Urdu handwritten script in automation process due to having varying writing styles (even from a single author) and also a single character

entry needs two to three keystroke combinations. In order to avoid this complexity, there is a need to develop a system that help in automatic conversion of handwritten Urdu characters into their original counterpart. Therefore, a large corpus is needed to develop such intelligent systems that train the system for recognizing handwritten Urdu characters.

These issues motivated us to develop a multi-aspect Urdu handwritten text database that includes both the characters and words of Urdu handwritten script written in different styles from a number of native Urdu speakers from different domains.

This paper is organized as follows: A brief introduction of Urdu script is given in Section 2. In Section 3, we describe basic steps in ICR by describing the usual process of text recognition. Then in Section 4, state-of-the art discussion on the existing databases of Urdu script is given in detail. Section 5 provides detailed complete information about our proposed corpus on Urdu handwritten text. Open issues, future directions and detailed analysis on our proposed work in comparison with existing databases is presented in Section 6. Section 7 concludes the article.

2. URDU SCRIPT

Urdu is the national language of Pakistan and also one of two official languages of Pakistan [11] (with the other being English). It is widely spoken and understood as a second language by a majority of people of Pakistan [12] [13] and also being adopted increasingly as a first language by the people living in urban areas of Pakistan.

Urdu script is written from right to left while digits are written from left to right, this is the reason Urdu can be considered as one of the bidirectional languages. Urdu script consists of 38 basic letters shown in Figure 1.

This alphabet set is also considered as super set of all Urdu script-based languages' alphabets, i.e., Arabic contains 28 and Persian contains 32 [14]. Furthermore, Urdu script also contains some additional alphabets to express the Hindi phonemes. Both Hindi and Urdu languages [14] have same phonology with only difference in written script.

All Urdu script-based languages such as Arabic and Persian have some unique characteristics i.e. (i) the script of these languages is written from right to left in cursive style and (ii) the script of these languages is context sensitive i.e., written in the form of ligatures which is a combination of a single or many alphabets. Due to this context sensitivity, most of the alphabets have different shapes depending on their position and their adjoining character in the word [5]. This connectivity of alphabets [15] has enriched the Urdu vocabulary with almost 24,000 ligatures.

In Urdu script, alphabets are classified in two groups: joiner and non-joiner.

Figure 1: Basic alphabets and numerals of Urdu script

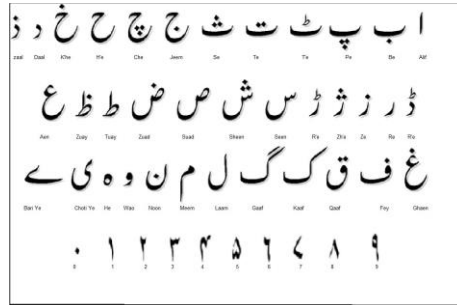
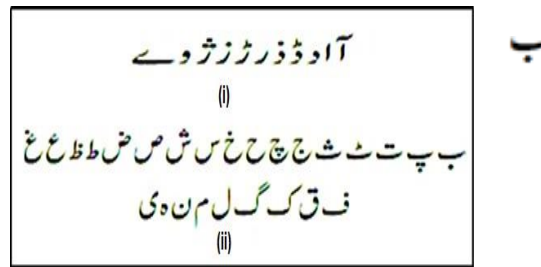


Figure 2: (a) Non joiner and (b) Joiner alphabets in Urdu script

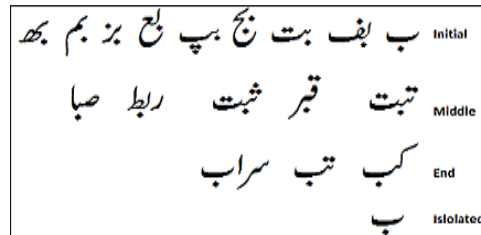


The joiner characters join to other characters on initial, middle and end position in the ligature. While non-joiner appears in isolated form. Figure 2 shows list of joiner and non-joiner alphabets of Urdu script. The joining capability of some Urdu alphabets allow the writers to write in careless way in which letters overlap each other. This overlapping makes the segmentation of words to characters more complex and challenging while performing text recognition using segmentation- based approach.

Naskh and Nastaliq are two widely used fonts in which Urdu script is written. Nastaliq is more complex than Naskh [14] [6] [16] [17] as Nastaliq has variety of variations of shape of an alphabet depending on its position in the word as compared to Naskh e.g., the second alphabet (bay) has different shapes while appearing in any position in a word shown in Figure 3.

Above mentioned issues make the recognition of Urdu handwritten text more challenging as compared to any other script. There are also some inherent characteristic features of Urdu script that may help in recognition process of Urdu handwritten text. These features include accent and diacritical marks that help in differentiating one character from the other; different shapes of same letter based on its position in the word; and presence of the virtual line (baseline). It is also mentioned by a number of researchers that reason behind the assumption that recognition of Urdu handwritten text is more difficult is absence of the appropriate resources and required techniques.

Figure 3: Different shapes of Urdu alphabet at different position in a word



3. INTELLIGENT CHARACTER RECOGNITION (ICR)

In the field of computer vision and pattern recognition, intelligent character recognition (ICR) is the process of recognition of handwritten text that is given as input to a computer. The process of handwriting recognition primarily depends on optical character recognition (OCR). ICR is different from OCR in a sense that ICR is associated with handwritten text recognition while OCR works on recognition of printed text. However, a complete handwriting recognition system also includes correct segmentation of words into characters, formatting of the extracted segment and finding the most reasonable words. Forrest of paper the term ICR will be used for handwritten text recognition. There are some issues in ICR such as change in font, slope of line, different writing style even from single writer, overlapping joining letters, missing placement of dots and diacritics aka secondary strokes etc. that make process of ICR more challenging than the recognition of printed text. These issues become more complicated in Urdu ICR because of its cursive script in which character changes its shape based on its position in a word [18]. These issues are discussed in detail in subsequent subsections below.

3.1 Urdu Based ICR Systems

Urdu based ICR systems can be divided into two types i.e., Online and Offline. In Online ICR, real time recognition of characters is performed using sensors to detect and analyze the pen tip movement, stroke position, baseline detection etc. While in Offline, character recognition implicates the automatic conversion of handwritten text from scanned image of paper. In practice, the offline character recognition is a more complex process than online character recognition [19] [20] [21]. In practice, the text data is given as input in the form of scanned image that is analyzed and recognized as machine readable characters. The text data may have different fonts and handwriting styles that needs to be preprocessed to produce an ideal and clear view of the input data. A typical ICR system comprised of three phases: (i) preprocessing, (ii) segmentation and (iii) recognition. In first phase, a set of operations are performed in order to reduce the ink-noise ratio because the input set of handwritten documents may include inconsistent text due to having different writing styles. These operations include skew correction and/or smoothing, chain coding and baseline removal etc. [14] [6] [4] [22]. In segmentation phase, the scanned image is segmented at three levels as mentioned in [19] [23]. In order to get the finer results

from segmentation, one must have to move through all the three levels. In last phase, recognition is performed in which the segmented text data is scanned and matched with the stored training set data. The application of ICR has increased its efficacy towards automatic recognition of real world handwritten documents to make them useful for various business and academic applications.

3.2 Contribution in this paper

In this paper, our contribution revolves around following objectives:

To provide a multi-aspectual dataset with ground-truth to the researchers for Urdu ICR development, this will save a lot of time and effort of the research teams working on Urdu ICR. To collect data by inviting the native Urdu speakers of both the genders from different fields, age. Furthermore, in order to make the dataset more comprehensive, the physically disabled persons are also invited. To include handwriting of handicaps, partially blind, primary level students, calligraphers and individuals while traveling. To provide an annotated data collection with ground-truth for researchers working for Urdu processing technologies.

4. REVIEW OF EXISTING URDU DATA COLLECTIONS

Some Offline Urdu handwritten databases available are CALAM [24], UPTI [10] [25], UCOM [26] or UNHD and CENPARMI [27]. Only UNHD is available freely while access to other data collections is not free. The concept of creating Urdu handwritten dataset was first conceived in 2013 at computer science department of COMSATS Institute of Information Technology (CIIT) Abbottabad, Pakistan. Firstly, data was collected from 100 students on 6 papers having 8 text lines on each paper. Base lines were removed and text line segmentation was performed on collected data [26]. This Urdu dataset was collected at CIIT. With this affiliation we named it as UCOM dataset. UCOM was published in 2015. The extension of UCOM dataset is UNHD which is an abbreviation of 'Urdu Nastaliq Handwritten Dataset'. In these Urdu handwritten samples gathered from school students, college graduates, and office going individuals. In this text for dataset broadened from 48 lines to 700 unique text lines. These lines included Urdu numeric and Urdu constraint handwritten samples.

CENPARMI [27] is considered to be the first handwritten-Corpus for Urdu language. This consists of digits and characters of Urdu script. It is a huge database for Urdu-handwritten texts. It contains 57 Urdu-words, 44 isolated characters, dates in different formats, isolated-digits, numeric strings with and without-decimal points, and 5 special symbols. It formed the world's first database for offline Urdu handwritten recognition; which was designed at Centre for Pattern-Recognition and Machine-Intelligence (CENPARMI). Furthermore, CENPARMI focuses only on isolated characters and digits and some selected words. Urdu Printed-Text Image (UPTI) [10] [25] is a dataset that was developed for research-community as an analogy to APTI dataset. This consists of different adaptation to-measure accuracy of the recognition-system for Urdu Language. These adaptations consist of ligatures or sub words, degraded-text-lines and degraded-ligatures. The text lines included were selected-

from Jang newspapers. These lines include different topics of politics, religion and social-issues. This dataset consists of 10000 images with having Urdu text lines and 970649 characters [25]. CALAM provides us a way for healing and fetching of information in a scientific and systematic aspect. This uses design and development of an annotated- corpus of handwritten text image [24]. This is a suitable mechanism to-annotate linguist handwritten-image dataset. Away from pure text glossary, CALAM contribute some additional linguistics features-about the nature of language such as-document of corpus to-other language. CALAM provides us a terrace for linguist Corpus convenient for all types of linguistic linked to research, where a large-scale of fine grade systematic-data crosswise language is maintained in both handwritten and machine-readable format. The details of available Urdu datasets are summarized in Table 1.

5. MULTI-ASPECT URDU HANDWRITTEN DATA COLLECTION

The proposed data collection is unique in its approach towards building an annotated data collection for Urdu handwriting recognition. We propose the data collection with focus on those writers that are likely to have very unique writing styles because of their professions or because of some physical disabilities. Currently, there exists no data collection which involves such variety of writers. Another uniqueness of proposed data collection lies in the use of same text for all the writers. This activity opens the ways for research in author- identification, author attribution and graphology tasks in Urdu handwriting. We make this data collection freely available for research purposes which is another positive feature of proposed data collection.

Table 1: Details of Urdu datasets

Dataset	Statistics	Price
	Total number of writers 500	Public
	Text lines per page 8	
UNHD [28]	Total number of text lines 10,000	
	Total Number of words 312, 000	
	Total number of characters 187, 200	
	Total number of writers 250	US\$250
UPTI [10]	Total number of text images 10,000	
	Text lines per page 6	
	Total number of characters 9, 70, 650	
	Total number of writers 343	US\$500
CENPARMI [27]	Total number of text images 19,432	
	Total samples of Urdu handwritten digits 180	
	Total number of writers 725	US\$400
	Total number of images 1,200	
CALAM [29]	Total number of text lines 3,043	
	Total number of words 46, 664	
	Total number of ligatures 101,181	

Preparing this data collection involves following steps:

Selecting participants for writing,

Selecting content for Writing,

Preparing annotation forms and annotation guidelines

Annotating the dataset

5.1. Selection of Participants

This step makes our data collection unique because we have chosen a variety of writers not only with different demographics but with the likelihood of producing different styles of Urdu handwriting. In this subsection, we provide details of all categories of selected participants along with images of their handwritten samples.

5.1.1 Students

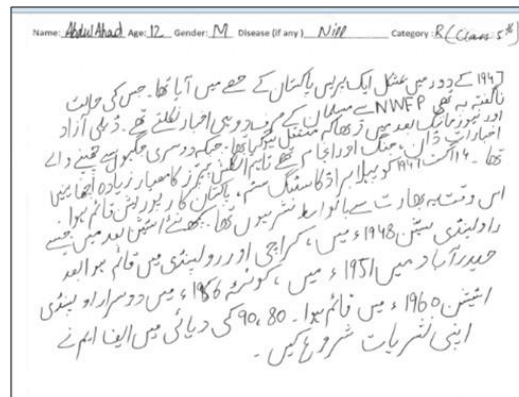
Firstly, we have included the student's category, which is further divided into three sub-categories. First of them is Primary level (grade 5) students. In which students of both gender male/female are included with age group of between 10 – 12 years. Among these, we selected 5 male and 5 female students. We also want to include left hand writer but unfortunately, we did not have any one among these 10.

Second sub-category of students is Matric level (grade 10) students. We selected the students from both the genders with age group of 13 – 16 years. We have 5 students from either gender, 2 of them use left hand while writing.

Third sub-category of students is Graduate level students of both the genders having age between 20-30 years. We have 7 male and 8 female students. Unfortunately, we could find any student who could write with left-hand.

Please see Figure 4 and Figure 5 for writings of students' category.

(a) Primary Level Male Writing Style



(b) Primary Level Female Writing Style

Name: NAWFPA, Age: 10, Gender: F, Disease (if any): Nil, Category: R (Class: 5th)

1947 کے دور میں نکل ایک ایس پاکستان کے حصے میں آیا تھا جس کی حالت ناگتہ تھی۔
NWFPA سے مسلمانوں کے صرف دو ہی اہلکار تھے ڈپٹی آرڈر اور نوز مارگٹ بعد میں ڈپٹی
کرنل بن گیا۔ چونکہ مدرسے کی تعلیموں سے بچھینے والے اخبارات ڈپٹی آرڈر کے دل کو براہ راست
تاہم انگلش پبلسز کا اخبار زیادہ اہلکار تھے۔ 11 اگست 1947ء کو یہ لہور آئی لائٹنگ سسٹم پاکستان
کا رویش نام ہوا۔ اس وقت ہزاروں سے لاکھوں نوجوان تھے جنھیں انڈین ایجوکیشن
بیس رویش نام تھی۔ 1948ء میں پاکستانی اور رویش نام کے نام سے ایک ایس جیو ایس
1946ء کو یہ لہور آئی 1956ء میں دوسرا رویش نام انڈین ایجوکیشن نام ہوا۔ 80-90
کی دہائی میں لائٹنگ سسٹم نے اپنی شہرت شروع کی۔

c. Matric Level Male Writing Style

Name: M. J. Q. al, Age: 15, Gender: Male, Disease (if any): Nil, Category: R (10th)

1947 کے دور میں مشکل ایک ہریس پاکستان کے حصے میں آیا تھا۔
جس کی حالت ناگتہ تھی NWFPA سے مسلمانوں کے صرف دو ہی
اخبار تھے۔ ڈپٹی آرڈر اور نوز مارگٹ بعد میں ڈپٹی کرنل ہو گیا
تھا۔ جبکہ دوسری جگہوں سے بچھینے والے اخبارات ڈپٹی کرنل اور
انجیل کے نام انگلش پبلسز کا اخبار زیادہ اہلکار نہیں تھے۔ 11 اگست
1947ء کو یہ لہور آئی لائٹنگ سسٹم پاکستان کا رویش نام ہوا۔
اس وقت ہزاروں سے لاکھوں نوجوان تھے جنھیں انڈین ایجوکیشن
بیس رویش نام تھی۔ 1948ء میں پاکستانی اور رویش نام کے نام سے ایک ایس
جیو ایس 1946ء کو یہ لہور آئی 1956ء میں دوسرا رویش نام انڈین ایجوکیشن
نام ہوا۔ 80-90 کی دہائی میں لائٹنگ سسٹم نے اپنی شہرت شروع کی۔

Figure 4: Handwriting samples of Maric level students of both the genders

ہی کا نام سننے ہی ایک نوعوت سے باخبر کا نام لہو ہے۔ یہ
دنیا میں سب سے زیادہ پائے جانے والا جانور ہے۔ ہی کا سائنسی نام
(Domesticus) ہے۔ اس کا خاندان (Felidae) کہلاتا ہے۔
پہر گھنٹہ فرور ہے۔ ایڈ انوائس کے مطابق ہی 9500 سال
سے انسانوں کے درمیان موجود ہے۔ ہی کی اوسط اونچائی 25
سے 35 سینٹی میٹر کی مہائی 30 سینٹی میٹر تک ہوتی ہے اس کے
جسم کا درجہ حرارت 38.6 سینٹی گریڈ ہوتا ہے۔ جب کہ ایک منٹ
ہی ان کا دل 120 سے 140 مرتبہ دھڑکتا ہے۔ گرمی کی حالت میں
سائنسی لینے کی رفتار گھنٹے سے 40 مرتبہ فی منٹ ہوتی ہے۔ دنیا کی
جھولی ترین ہی کا وزن 108 کلوگرام ہوتا ہے۔

(a) Graduate Level Male Writing Style

(b) Graduate Level Female Writing Style

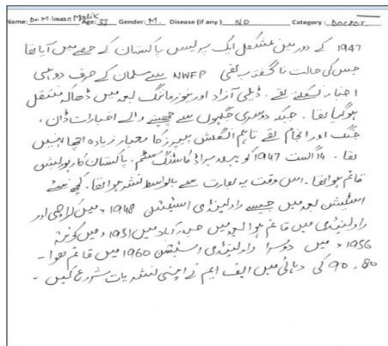


Figure 5: Handwriting samples of Graduate level students of both the genders

5.1.2. Doctors

Second category of the invited participants is the doctor's category. It includes doctors of both gender male/female. We selected 5 doctors each from the both genders. Unfortunately, we could not find any left-hand writer in doctor category. The writing samples from the doctor category are shown in Figure 6.

(a) Male Doctor Writing Style



(b) Female Doctor Writing Style

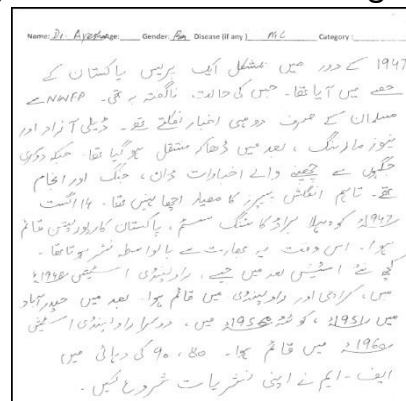
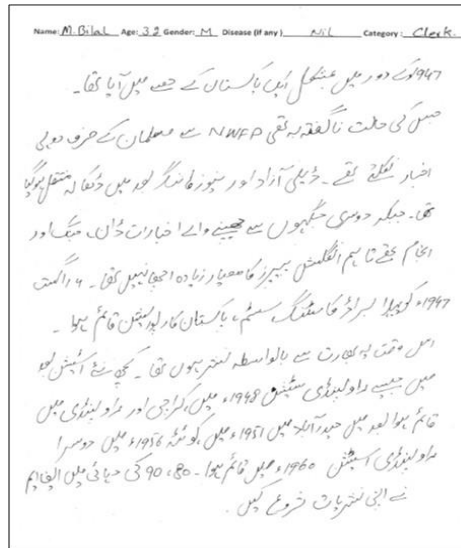


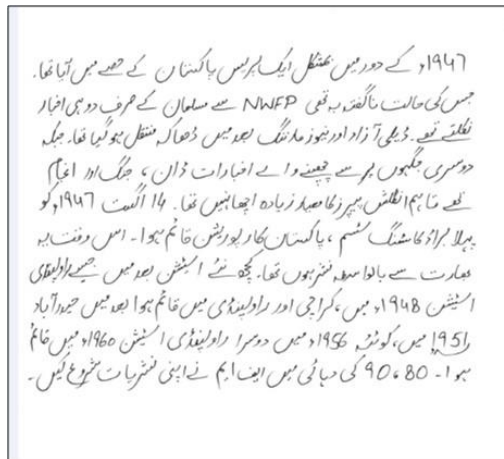
Figure 6: Handwriting samples of doctors of both the genders

5.1.3. Office Clerks

The third category is of clerks working in government and administrative offices. We selected 5 individuals each from male and female category. All 10 clerks are right-handed. Figures 7 depict some of the handwritten samples collected from clerk category.



(a) Male Office Clerk Writing Style



(b) Female Office Clerk Writing Style

Figure 7: Handwriting samples of clerks of either gender

(b) Female Random Traveler Writing Style

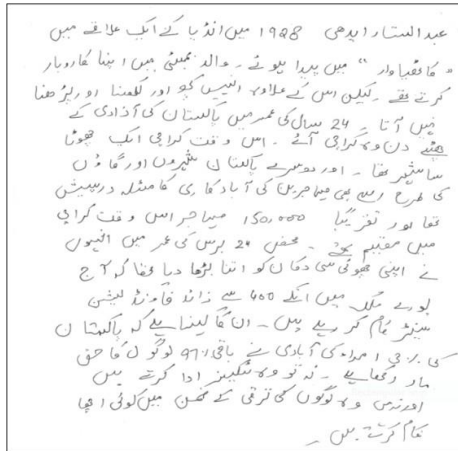


Figure 9: Handwriting samples of random travelers of either gender

5.1.6 Stamp Papers Seller

The stamp paper sellers are the persons who used to sell and write official stamp papers often used for some legal activities. This category is most interesting among all since all the stamp sellers adopt a similar writing style and use black ink for writing. This category includes only males and all are right-handed. Unfortunately, we couldn't find any left-handed seller. Please see Figure 10 for writings of stamp seller category.

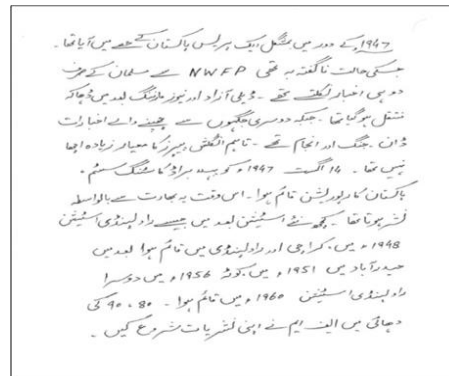


Figure 10: Stamp Paper Sellers Writing Style

5.1.7 Old Educated

The seventh category is of old educated people. It includes both gender male/female. Most of the persons in this category are either retired from their jobs or running their own business. There are 6 males and 4 female writers. All writers are right-handed. Figure 11 show sample of handwritten text writings of old age educators' category.

بلبل کا نام سننے میں نکلنا حکومت سے جانوروں کا جانور مہرنا ہے۔ یہ
 دنیا میں سب سے زیادہ مالچے جانے والا جانور ہے۔ بلبل کا سائنسی
 نام (Domesticus Felis) ہے۔ اس کا فائدہ
 (Felines) لگانا ہے۔ یہ گوشت خور جانور ہے۔ ایک انداز سے
 9500 سال سے انسانوں کے درمیان موجود
 ہے۔ بلبل کی اوسط اونچائی 73 سے 75 سینٹی میٹر تک ہوتی ہے۔ اکثر
 بلبلوں کی لمبائی 30 سینٹی میٹر تک ہوتی ہے ان کے جسم کا درجہ
 حرارت 38.6 سینٹی گریڈ تک ہوتا ہے۔ جب کہ ایک منٹ میں ان کا
 دل 120 سے 140 مرتبہ دھڑکتا ہے۔ آرام کی حالت میں سانس
 میں سانس لینے کی رفتار 16 سے 20 مرتبہ فی منٹ ہوتی ہے۔

(a) Old Educated Male Writing Style

Name: Shehzaad Age: 45 Gender: Fe Disease (if any): NA Category: Expert
 1947 کے دور میں بمشکل ایک پریس پاکستان کے حصے میں
 جس کی حالت ناگفتہ بہ تھی NWFP سے مسلمان کے فرقہ دو
 ہی اخبار نکلتے تھے۔ ڈبلیو آزاد اور نیوز مانتگ بعد میں
 ڈھاکہ منتقل ہو گیا تھا۔ جبکہ دوسری جگہوں سے چھپنے
 والے اخبارات ڈان، جنگ اور انجام تھے تاہم انگلش
 پیپرز کا معیار کا معیار زیادہ اچھا نہیں تھا۔ 11 اگست 1947ء کو
 پہلا براڈ کاسٹنگ سسٹم پاکستان کا رپورٹیشن قائم ہوا۔
 اس وقت یہ بھارت سے بالواسطہ نشر ہوتی تھی۔ کچھ نئے اسٹیشن
 بعد میں جیسے راولپنڈی اسٹیشن 1948ء میں کراچی اور راولپنڈی
 میں قائم ہوا بعد میں حیدرآباد میں 1957ء میں،
 1956ء میں دوسرا راولپنڈی اسٹیشن 1960ء قائم ہوا۔
 90-80 کی دہائی میں الفہم نے اپنی نشریات
 شروع کیں۔

(b) Old Educated Female Writing Style

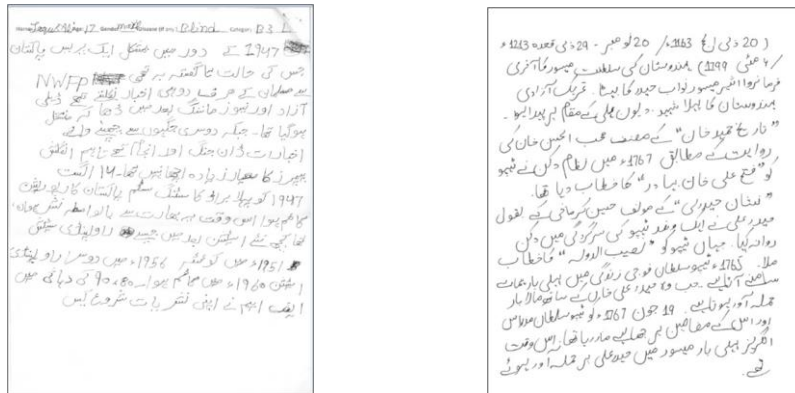
Figure 11: Handwriting samples of old educated persons

5.1.8 Special Persons

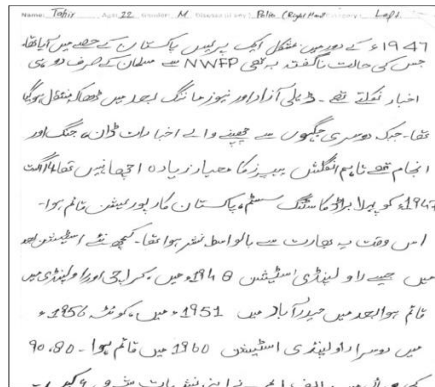
This category includes the persons who are physically disabled such as partially blind, deaf, blind etc. The handwritten samples from this category make our data set more comprehensive and different from the datasets of same kind. The detailed discussion about the handwritten samples collected from the persons of this category is given below.

Partially Blind. The last one is special person's category, which is further divided into three sub-categories. First of them is partially blind. We selected the partially blind people of both the gender with age between 15 –20 years and invited 10 persons from either gender. Among these 20, one is found left-handed and remaining are right-handed.

Physically Disabled: Second sub-category includes the native Urdu speakers that have some physical disability issues like vision impairment, dwarfism, missing or crossed fingers etc. We selected 10 male writers. Two of them are left-handed writers and others are right-handed. A very different type of handwriting style was observed from the writings of physically disabled persons as shown in Figure 12.



(a) Partially Blind (B3) Writing Style (b) Writing style of one-eyed person (other is dead by birth)



(c) Writing style of person with right hand effected from polio disease



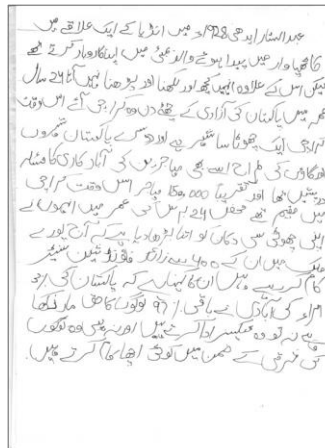
(d) Writing style of person having cut on right hand index figure

Figure 12: Handwriting samples of some physically disabled persons

Ambidextrous: The last one is Ambidextrous. It describes the individuals who can use either hand while writing. In this sub-category, we found two females of age between 18 to 27. The writing samples of this sub-category are shown in Figure 13.



(a) Right hand writing style of ambidextrous



(b) Writing style of same person with left hand

Figure 13: Handwriting samples of ambidextrous person

5.2 Selecting Content for Writing

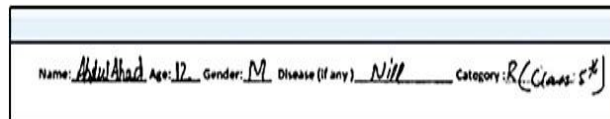
In order to maintain the symmetry among the text, all the authors are given to write the same script in their own handwriting. Furthermore, the corpus also consists of an aligned transcription for image, line by line, phrase by phrase, or word by word. The corpus is completely marked up for content information to support content detection and evaluation of systems like linguistic handwriting recognition, signature verification, and author identification.

5.3 Preparing Annotation Forms and Annotation Guidelines

Forth collection of data set, we have invited a total of 84 individuals from different categories discussed above. Each individual was asked to write the set of 10 A4 size

pages. Furthermore, each individual was directed to write the printed text in his/her pages in black ink to reduce ink-noise ratio. They were also given directions to how to write on the pages. Furthermore, they were also asked to provide some personal information like name, age, gender, disease (if any), category (right/left hand writer, profession like student, doctor, etc.). A sample of this information is depicted in the figure 14 below (it is not made public with data collection; it is only an example).

Figure 14: Header of the Annotation Form



Name: Abdul Ahad, Age: 17, Gender: M, Disease (if any): Nil, Category: R (Class: 5th)

All the writers are directed to follow the given instructions. Some of the individuals also draw baseline by themselves using lead pencil for their convenience while writing. Also, some of the writers wrote in blue because writing in blue is more common than in black.

5.4 Annotating Data Set

Aletheia [30] is a leading system for scientific use and still cost effective. It is used for recognition and annotation of scanned documents. Its main function is to create and to view page partition and OCR ground truth. Storage format of page is XML. Third party software supported page is also viewed by this. We are using Aletheia version 3(pro), which is latest version in market. Listed below are some important features of Aletheia:

Segment the page elements on four levels; regions, text lines, words and glyphs
Remove noise from selected element by smearing option
Add annotated text of selected element
Change color document into black and white image.

Border the document

Provides the metadata

Page collections

6. DATA ANALYSIS

This section will cover the analysis of our data set with the existing datasets at different level.

6.1 Data Set Details

Table 2 describes a summarized review of our data collection in comparison with existing Urdu data collections. The proposed data collection stands unique as far as variety of different styles is concerned contributing significant number of words.

Table 2: Comparison of Existing Datasets with Our Dataset

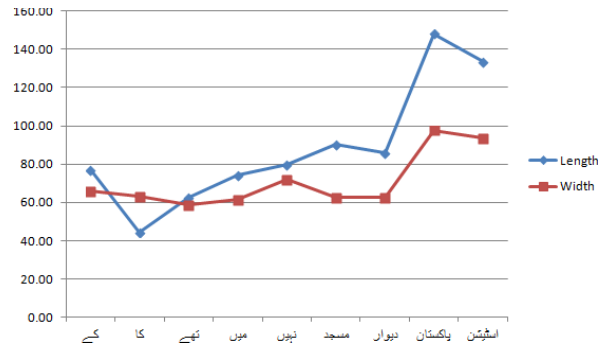
Details	UCOM	UNHD	CENPARMI	UPTI	CALAM	Our Data Set
No of words per-page	104	-	-	-	-	130
No. of text lines per-page	8	5 to 8	-	-	3 to 5	8 to 12
No. of pages per writer	6	6	2	-	-	10
Approx. no. of words per writer	620	624	-	-	-	1381
Total number of words	62000	312000	20 digits, 38 numeral stings, 43 characters, 57 words, 5 special symbols	-	46664	114623
Total number of lines	6400	10,000	-	10063	3043	9379
Total number of pages	600	3000	686	-	1200	830
Number of writers	100	500	343	-	725	83
Categories	students only	Students and Professionals	3 (lefthanded/ righthanded)	Text consisted of different social issues	6 (subject based)	8
Special persons writing	-	-	-	-	-	10
Publicly available	Yes	yes	No	No	No	Yes

In the following subsections, we analyze writing patterns of different set of authors for different aspects. Handwriting analysis is described as a scientific study and analysis of handwriting at different granularity levels. It is a way of interpreting behavior from peculiarities in handwriting. The scientific name for handwriting analysis is Graphology [31]. It helps in determining the psychological behavior and profession of the authors. Unfortunately, we lack psychological theories for Urdu handwriting styles; however, as starting point in this direction, we plan to compare lengths and width of different words because size of the word does matter in graphology.

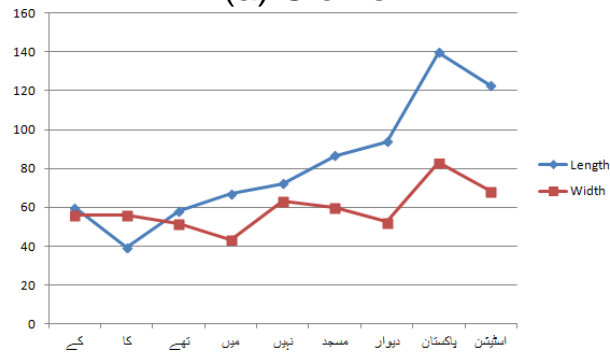
6.2 Comparing Length and Width of Words over Categories

In this subsection, we compare length and width of a selected set of words over different categories. Words are selected according to their length i.e., words are chosen from a category of words of length two, three, four, five, six and seven. The idea behind this interesting task is to seek if a particular category of authors can be found using a particular style for a set of words. Figure 17 represent graphs for various categories. All graphs do not reveal something very big but there are some interesting observations that can be made. For example, it is evident from the graphs that as the width of a word increases (with a greater number of alphabets), its height also increases. Normally, width and height of a word are not co-related but in this particular scenario it is understood because as the width of a word increases it

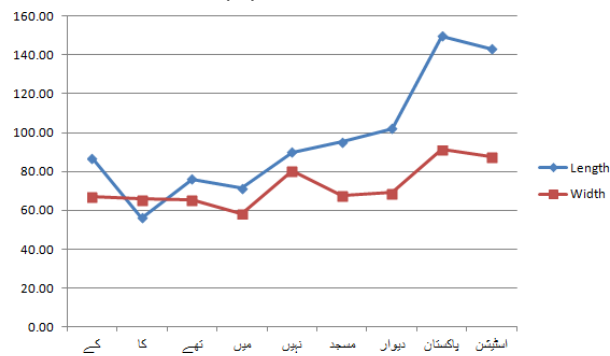
involves various types of alphabets involving accents which cause an increase in height too. The first four words mentioned on x-axis start around 60 pixels in height and width for all categories. Average of 5th, 6th and 7th words remains around 80 while last two words get peaks because of their increased length and height. There is nothing extra ordinary that can be mentioned as far as cross-category examination is concerned.



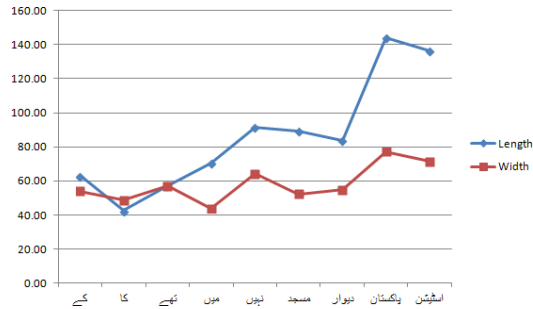
(a) Clerks



(b) Doctors

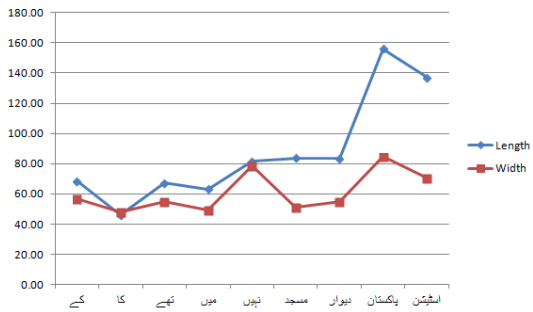


(c) Land record officials

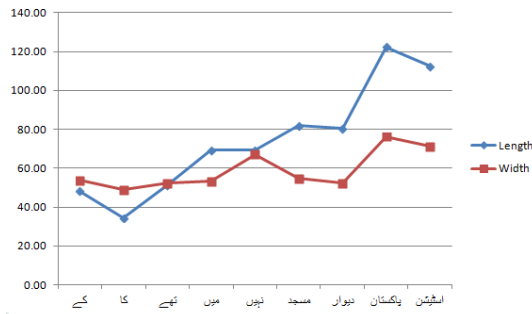


(d) Old educated persons

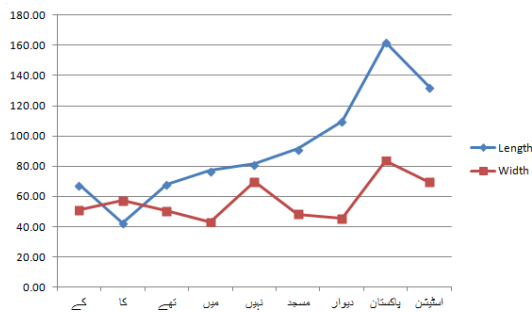
Figure 15: Comparing Length and Width of words over clerk, doctor, land record officials and old educated Categories



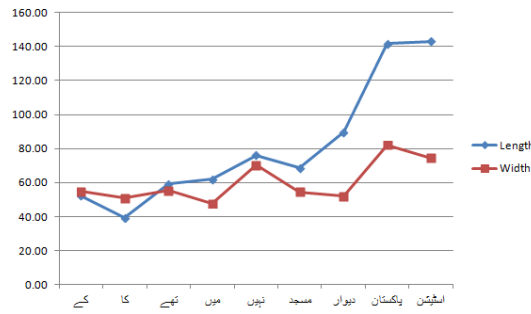
(a) Special persons



(b) Students

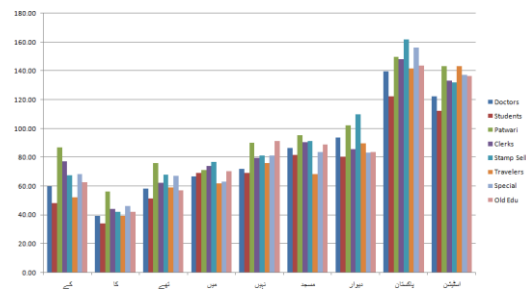


(c) Stamp sellers

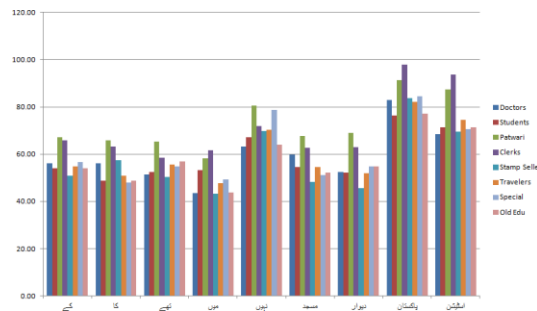


(d) Random travelers

Figure 16: Comparing Length and Width of words over special persons, students, stamp sellers and random travelers Categories



(a) Special persons

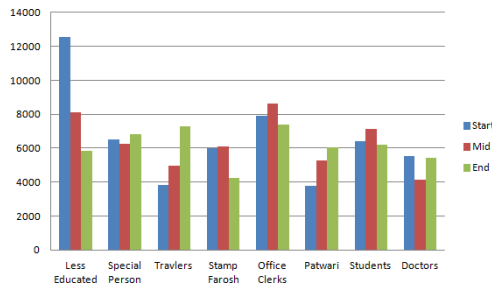


(b) Students

Figure 17: Graph length and width of words for all categories

6.3 Comparison for the size of words according to its position in script

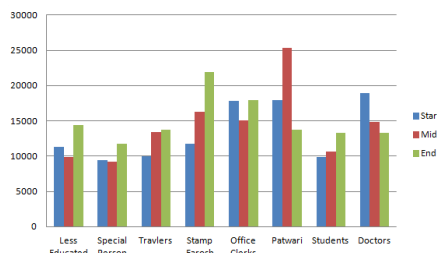
In this subsection, we analyze if position of a word affects its width while writing. This analysis is performed over all categories of authors by selecting a set of words of different lengths. Figure 18 show the width of different words with respect to their positions written by different authors of different categories. The idea behind this analysis is to see if authors squeeze their words when they have a limited space problem. A detailed observation of these graphs shows that most of the writers write freely when in start of a line while they squeeze their words when approaching end of the line. It can be important feature to be considered when working on handwriting recognition.



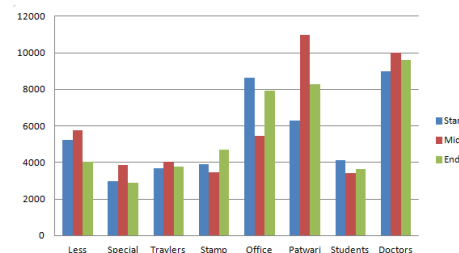
(a) Graph for word NAHI



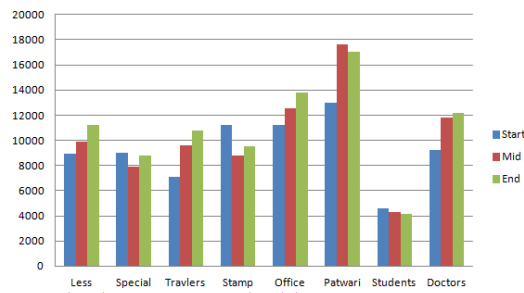
(b) Graph for word DEWAR



(c) Graph for word PAKISTAN



(d) Graph for word MASJID



(e) Graph for word STATION

Figure 18: Comparing Length and Width of different words over clerk, doctor, land record officials and old educated Categories

6.4 Analyzing Different Patterns of a Letter According to Category of Authors

Tables 3, 4, 5, 6, 7, 8 and 9 highlight different styles used by a particular category of authors for a specific Urdu alphabet. Looking at various patterns present in different tables for all alphabets, it can be concluded that highest variance in writing styles can be observed in Patwaris (see table 6) and Stamp Sellers (see table 8) category while others do differ in their style but more or less show similar patterns.

Letters	Used Styles	Letters	Letters Family	Used Styles
ا Series	ا آ آ	ا series	ا آ	ا ا ا
ب Series	ب ت ت	ب series	ب پ ...	ب ب ت
ج Series	ج ح ح	ج Series	ج چ ج ...	ج ج
د Series	د > د	د Series	د د	د د
ر series	ر ر ر	ر Series	ر ز ...	ر ر
س series	س س س	س Series	س ش س	س س
ص series	ص ض ض	ص Series	ص ض ص	ص ض
ع series	ع ع ع	ع Series	ع غ ع	ع ع
ه series	ه ه ه	ط Series	ط ظ ط	ط ظ ط
ف Series	ف ف ف	ف series	ف ف	ف ف
ق Series	ق ق	ق Series	ق ق	ق ق
ک Series	ک گ گ	ک Series	ک گ ک	ک ک
ل series	ل ل ل	ل Series	ل ل	ل ل
م series	م م م	م Series	م م	م م م
ن Series	ن ن ن	ن Series	ن ن	ن ن ن
و series	و و و	و series	و و و	و و و
ی series	ی ی ی	ی Series	ی ی ی	ی ی ی

Table 3: Letter styles used by less educated persons

Table 4: Letter styles used by doctors

Table 5: Letter styles used by Travelers

Letters	Used Styles
ا Series	ا آ آ
ب Series	ب ب ت
ج Series	ج ح ح
د series	د >
ر Series	ر ر
س series	س س س
ص Series	ص ض ض
ع series	ع ع ع
ط series	ط ظ ط
ف Series	ف ف ف
ق Series	ق ق ق
ک Series	ک گ ک
ل Series	ل ل ل
م series	م م م
ن series	ن ن ن
و Series	و و و
ی series	ی ی ی

Table 6: Letter styles used by Patwaris (Land Record Holders)

Letters	Used Styles
آ Series	آ آ آ
Series	آ آ آ
س series	س س س
ر series	ر ر ر
س series	س س س
س series	س س س
ع series	
د series	د د د
ف Series	ف ف ف
series	
ق Series	ق ق ق
ک Series	ک ک ک
ل series	ل ل ل
ز Series	ز ز ز
ن Series	ن ن ن
د series	د د د
ی series	ی ی ی

7. CONCLUSION

Urdu script is categorized as one of the cursive and bidirectional script derived from Arabic and Persian this is the reason it shares almost similar challenges and issues but with higher complexity. In this paper, we propose a multi aspectual Urdu handwriting data collection. We also give a review of existing data collections for Urdu handwriting recognition and give a comparison of the proposed data collection with existing ones. The uniqueness of the proposed data collection lies in the variety of authors known especially for their typical handwriting. Making same text written by all authors also make this data collection suitable for research tasks like author attribution, graphology, gender prediction, age prediction, profession prediction, education-level prediction etc. The proposed data collection will be made freely available for research purposes. Our future direction of research is to take forward this work by training machine learning models on this data and to propose algorithms that employ these trained models for different tasks that have been listed above.

Table 7: Letter styles used by Special (Physically Disabled) Persons

Letters	Used Styles
ا Series	ا ا
ب series	ب ب
ج Series	ج ج
د Series	د د
ر series	ر
س series	س
ص Series	ص ص
ع Series	ع ع
ط series	ط
ف Series	ف ف
ق Series	ق
ك series	ك ك
ل series	ل ل
م Series	م م
ن Series	ن ن
و series	و و و
ي Series	ي ي ي

Table 8: Letter styles used by Stamp Sellers category

Letters	Used Styles
ا Series	ا
ب Series	ب ب ب
ج Series	ج ج ج
د Series	د
س series	س س
series	
ص Series	ص ص
ع series	ع ع
ط Series	ط ط
ف Series	ف ف
ق Series	ق
ل series	ل
م Series series	م ل
ن Series	ن ن
و series	و و
ي series	ي ي

We defined an XML-based handwritten text image corpus and the annotation methodology that has the prospective budding to provide researchers all the facilities for document image processing research problems. These problems include author identification; text verification; recognition of text pages at line, word, and ligature levels; and separation of handwritten and printed texts. Furthermore, our database would be helpful in the design of an automatic intelligent system for direct processing of massive handwritten forms collected for census data.

Table 9: Letter styles used by Students category

Letters	Used Styles
ا Series	ا
آ Series	آ آ آ
ح Series	ح ح ح
د Series	د
ر series	ر
س Series	س س
ص Series	ص ص
ع Series	ع ع ع
ط Series	ط ط
ف Series	ف ف
ق Series	ق
ک Series	ک ک
ل Series	ل ل
ز series	ز ز ز
و series	و و و
ہ Series	ہ ہ ہ
ی Series	ی ی ی ی

Also, it can be very widely used for language transcription and transliteration applications acting as an information exchange center. To date, only four datasets are available for handwritten Urdu script. The aim of this work is to build a resource that would provide ground-truth annotation for handwritten text images. We propose floating the dataset as an open source on cloud storage free for academic use, where permissions for usage would be given on request.

References

- 1) C. Bahlmann, Directional features in online handwriting recognition, Pattern Recognition 39 (1) (2006) 115–125.
- 2) S. Mori, H. Nishida, H. Yamada, Optical character recognition, John Wiley & Sons, Inc., 1999.
- 3) H. F. Schantz, The history of OCR, optical character recognition, Recognition Technologies Users Association Manchester, VT, 1982.
- 4) K. U. Khan, et al., Online urdu handwritten character recognition: Initial half form single stroke characters, in: 2014 12th International Conference on Frontiers of Information Technology (FIT), IEEE, 2014, pp. 292–297.

- 5) M. I. Razzak, F. Anwar, S. A. Husain, A. Belaid, M. Sher, Hmm and fuzzylogic: a hybrid approach for online urdu script-based languages character recognition, Knowledge-Based Systems 23 (8) (2010) 914–923.
- 6) N. H. Khan, A. Adnan, S. Basar, An analysis of off-line and on-line approaches in urdu character recognition.
- 7) M. W. Sagheer, C. L. He, N. Nobile, C. Y. Suen, Holistic urdu handwritten word recognition using support vector machine, in: Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE, 2010, pp. 1900–1903.
- 8) A. Wali, S. Hussain, Context sensitive shape-substitution in nastaliq writing system: Analysis and formulation, in: Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, Springer, 2007, pp. 53–58.
- 9) Q. U. A. Akram, S. Hussain, Ligature-based font size independent ocr for noori nastalique writing style, in: Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop on, IEEE, 2017, pp. 129–133.
- 11) N. Sabbour, F. Shafait, A segmentation-free approach to arabic and urdu ocr, in: Document Recognition and Retrieval XX, Vol. 8658, International Society for Optics and Photonics, 2013, p. 86580N.
- 12) G. F. Simons, Ethnologue: Languages of the world, sil International, 2017. [12] T. Rahman, Language and politics in pakistan, Language 133 (1998) 9.
- 13) A. Mahboob, R. Jain, Bilingual education in india and pakistan, Bilingual and Multilingual Education (2016) 1–14.
- 14) M. I. Razzak, Online urdu character recognition in unconstrained environment, Ph.D. thesis, INTERNATIONAL ISLAMIC UNIVERSITY, ISLAMABAD (2011).
- 15) G. S. Lehal, Choice of recognizable units for urdu ocr, in: Proceeding of the workshop on document analysis and recognition, ACM, 2012, pp. 79–85.
- 16) U. Pal, A. Sarkar, Recognition of printed urdu script, in: null, IEEE, 2003, p. 1183.
- 17) U. Pal, R. Jayadevan, N. Sharma, Handwriting recognition in indian regional scripts: a survey of offline techniques, ACM Transactions on Asian Language Information Processing (TALIP) 11 (1) (2012) 1.
- 18) C. Jawahar, A. Kumar, A. Phaneendra, K. Jinesh, V. Govindaraju, S. R. Setlur, Guide to ocr for indic scripts: Document recognition and retrieval.
- 19) R. P. dos Santos, G. S. Clemente, T. I. Ren, G. D. Cavalcanti, and Text line segmentation based on morphology and histogram projection, in: Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, IEEE, 2009, pp. 651–655.
- 20) S. Marinai, P. Nesi, Projection based segmentation of musical sheets, in: Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on, IEEE, 1999, pp. 515–518.
- 21) A. Ali, M. Ahmad, N. Rafiq, J. Akber, U. Ahmad, S. Akmal, Language independent optical character recognition for hand written text, in: Multitopic Conference, 2004. Proceedings of INMIC 2004. 8th International, IEEE, 2004, pp. 79–84.

- 22) UI-Hasan, S. B. Ahmed, F. Rashid, F. Shafait, T. M. Breuel, and Offline printed urdu nastaleeq script recognition with bidirectional lstm networks, in: Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, IEEE, 2013, pp. 1061–1065.
- 23) S. Kumar, A. Karim, A review on recognition of handwritten urdu characters using neural networks. International Journal of Advanced Research in Computer Science 8 (9).
- 24) P. Choudhary, N. Nain, Calam: Linguistic structure to annotate hand- written text image corpus, in: Computational Intelligence in Data Mining-Volume 2, Springer, 2015, pp. 449–460.
- 25) S. Naz, A. I. Umar, R. Ahmed, M. I. Razzak, S. F. Rashid, F. Shafait, Urdu nastaliq text recognition using implicit segmentation based on multi- dimensional long short term memory neural networks, SpringerPlus 5 (1) (2016) 2010.
- 26) S. B. Ahmed, S. Naz, S. Swati, M. I. Razzak, A. I. Umar, A. A. Khan, Ucom offline dataset- an urdu handwritten dataset generation., Int. ArabJ. Inf. Technol. 14 (2) (2017) 239–245.
- 27) M. W. Sagheer, C. L. He, N. Nobile, C. Y. Suen, A new large urdu database for off-line handwriting recognition, in: International Conference on Image Analysis and Processing, Springer, 2009, pp. 538–546.
- 28) S. B. Ahmed, S. Naz, S. Swati, M. I. Razzak, Handwritten urdu character recognition using one-dimensional blstm classifier, Neural Computing and Applications (2017) 1–9.
- 29) P. Choudhary, 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 (TALLIP) 15 (4) (2016) 26.
- 30) S. Kunz, F. Brecht, B. Fabian, M. Aleksy, M. Wauer, Aletheia—improving industrial service lifecycle management by semantic data federations, in: 2010 24th IEEE International Conference on Advanced Information Net- working and Applications, IEEE, 2010, pp. 1308–1314.
- 31) P. K. Grewal, D. Prashar, Behavior prediction through handwriting anal- ysis, IJCST 3 (2) (2012) 520–523.