A Comprehensive Image Dataset of Urdu Nastalique Document Images

*Qurat-ul-Ain Akram Aneeta Niazi Farah Adeeba Saba Urooj

Sarmad Hussain Sana Shams

Center for Language Engineering, Al-Khawarizmi Institute of Computer Science
University of Engineering and Technology
Lahore, Pakistan
**ainie.akram@kics.edu.pk*

Abstract

Reporting the standard image dataset along with ground truth information has become important in pattern recognition and Optical Character Recognition (OCR) research. Nastalique writing style is mostly used to write Urdu books, magazines and newspapers. In this paper, a large image dataset of Urdu document images written using Nastalique writing style has been reported. This data has been collected to cover the range of font sizes from 14 to 40. The ground truth typed corpus has been developed along image corpus. A total of 2,912 document images are scanned from 413 books, among them 593, 595, 150, 149, 151 ,461, 202, 186, 226 and 199 images are scanned for 14, 16, 18, 20, 22, 24, 28, 32, 36 and 40 font sizes respectively.

Keywords— *Urdu Image Dataset, Urdu Document Images, OCR, Noori Nastalique, Ligature, Main body, Diacritic*

Introduction

The research on the development of Optical Character Recognition (OCR) has long history. OCRs for printed and handwritten text of different languages including English, Russian, Chinese, Devanagari, Urdu and Arabic have been developed [1-4]. The accuracy of the OCRs depends on thorough analysis of the variations of the document images and effectiveness of the developed techniques on large dataset. In addition, a large amount of standard data is also required covering all real life varieties of document images, to evaluate and compare different techniques. Recently, using standard datasets of different languages, different competitions have been organized to compare and evaluate different techniques of OCR including document analysis and classification and recognition [5,6]. These annual competitions play an important role for the development and maturation of the algorithms which result in overall performance improvement of OCR systems.

Based on the above discussion, standard corpus is required for pattern recognition and OCR research. The accessibility of the benchmark corpus not only facilitates the researchers to do research but also provides platform to evaluate different techniques.

English is written in Latin script. The development of OCR for printed English text is quite easy as compared to the cursive scripts. It has 52 letters in characters set, each character has single shape. Normally projection profile methods are used to segment English document image into lines, words and characters. The recognition of handwritten English text is a challenging task. Martin and Bunke [7] report the English handwritten dataset of Lancaster-Oslo/Bergen (LOB) text corpus. The text lines extracted from different domains are printed in a proper format on a form. The form is designed properly so that domain of the printed text, text number, printed text, handwritten text and name of the writer can be extracted easily. The writers are asked to write the printed text in the specified area of the form. Filled forms are scanned at 300 DPI at gray scale resolution of 8-bit using HP-Scanjet 6100. The scanned images are saved in tiff format with LZW compression. A total of 556 forms are filled by 250 writers. After pre-processing, total of 4,881 handwritten line images are extracted. These lines have 43,751 words instances and 6,625 words vocabulary. Each line has 8 to 9 words on average. A separate ASCII file is maintained containing the information of each printed and handwritten text line.

Marti and Bunke [8] report English corpus twice as large as corpus [7]. The reported handwritten data set is extracted from 1,066 filled forms, written by approximately 400 different writers. The dataset is comprised of 92,85 handwritten lines and 82,227 word instances covering 10,841 vocabulary words.

This dataset has on average 8.59 lines per form. The average number of words per text line is 8.98.

Arabic language belongs to cursive script. In cursive languages, one or more characters form a ligature. The main stroke of the ligature is called RASM (or main body) and secondary stroke(s) is called IJAM(s) or diacritic(s) [9]. Normally, Naskh writing style is used to write Arabic text. In Naskh, the characters in ligature are written along the baseline. Each character has at most four shapes based on the position in a ligature, such as initial, medial, final and isolated shapes, shown in Figure1.

Due to the cursive nature of the Arabic text, the development of the OCR for Arabic is a challenging task. The main reason of limited research on Arabic OCR is the unavailability of dataset along with character level ground truth information for Arabic language.

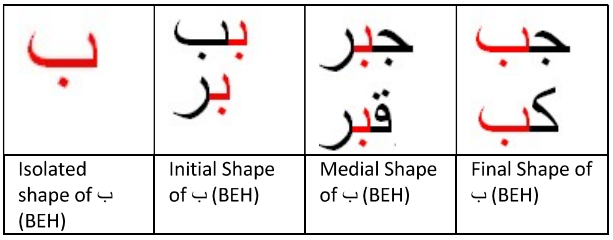


Figure 1. Consistent isolated, initial, medial and final shapes of ب (BEH) in Naskh

Margner and Pechwitz [10] report a process of genrating the synthetic Arabic dataset. The ARABTex is used to generate the Arabic documents from ASCII text. Ground Truth (GT) information consists of character code, font style, size and positional information of character. They also develop statistical HMM based OCR using 946 Tunisian town names dataset, and report reasonable accuracy.

Al-Ma'adeed et al. [11] develop Arabic handwritten database (AHDB). They first design the form to automatically extract information of respective handwritten text. Twenty nine high frequent words and sixty seven words used to write a cheque are printed on the form. In addition, the writers have to write three sentences representing numbers and quantities of cheque. Total of 104 forms written by 104 writers are scanned at 600 DPI using Hewlett Packard 6350 scanner. Mozaffari et al. [12] present handwritten dataset of 52,380 isolated characters and 17,740 numbers which are extracted from filled exam forms of schools. The filled forms are scanned at 300 DPI in gray scale format. Each character image is stored as a 77x95 BMP image. The presented IFHCDB database consists of 52,380 isolated characters and 17,740 numbers, and is divided into training (70%) and testing (30%) data. Kharma et al. [13] develop Arabic handwritten database which has Arabic words, numbers, signatures and complete sentences. Five hundred students are selected to develop this dataset. Each student is instructed to copy a predefined numbers, digits and sentences. The filled forms are scanned in both gray scale, and Black and White (BW) formats using HP scanner. The handwritten digits, words, sentences and signatures are cropped from original grayscale images and saved in BMP file formats. Digits, words, sentences and signatures are also extracted from BW images. This database includes 37,000 Arabic words, 10,000 digits, 2,500 signatures and 500 Arabic sentences. Pechwitz et al. [14] report another publically available IFN/ENIT-database of Arabic. This handwritten dataset contains 946 Tunisian town/villages names along with postcode. A total of 411 writers filled 2,265 forms. The filled forms are scanned at 300 DPI with BW format. During scanning, the page numbers and other information is maintained manually. This dataset contains 26,459 Tunisian town/village names and 212,211 handwritten characters. The GT information including postcode, global word, character shape sequence, baseline (y1,y2), baseline quality, number of words, number of PAW (Part of Arabic Word), number of characters and writing quality is stored against each town/village name image.

Urdu is cursive language which belongs to Arabic script. Normally, Nastalique writing style is used to publish Urdu content such as books, magazines and newspapers in Pakistan. Nastalique is written diagonally and has complex rules to place marks and diacritics. Nastalique has context sensitive character shaping [15]. Based on the shapes similarity, the Urdu character RASMs are divided into 21 classes. Unlike Naskh, Nastalique has character as well as ligature overlapping (Figure 2). The shape of a character depends on the context in which it appears, illustrated in Figure . The detailed analysis of Nastalique is discussed in [15, 16, 17].

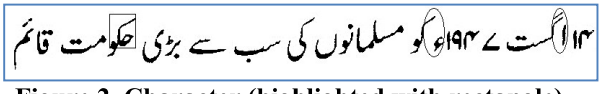
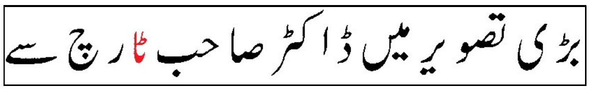


Figure 2. Character (highlighted with rectangle) and Ligature (highlighted with circles) Overlapping

Figure 3. Contextual Character Shaping of Character ب (BEH) at initial position in Nastalique

Nastalique writing style is a compact writing style and due to paper, ink and printing qualities, sometimes the diacritics and RASM are attached as shown in Figure 4. Nastalique has thick-thin-thick transitions of stroke while writing the character/ligature. Due to printing qualities, sometimes the Urdu ligatures are broken at the thin stroke (Figure 5).

(a) Example of RASMs attachment



(b) Example of RASM and diacritic attachment

Figure 4. Diacritics and RASM attachment, the attached connected components are highlighted with red color

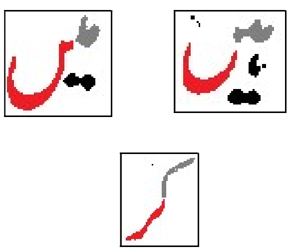


Figure 5. Broken connected components of ligatures (broken connected components with different colors)

The development of OCR for the Nastalique is a challenging task due to the aforementioned characteristics of Nastalique and behavior of paper and printing qualities. Limited research exists in the literature for recognition of Urdu document images written using Nastalique writing style. Real image corpus of published Urdu Nastalique documents is not available for the development of Urdu document processing algorithms and also for the evaluation of different research approaches. Usually researchers use their own corpus of Nastalique writing style. Most of them have been developed manually for large font sizes. [4, 18-25].

In this paper, a comprehensive dataset of Urdu Nastalique document images scanned from Urdu books is presented. This dataset is publically available for the researchers to do research in Urdu document analysis, pattern recognition and OCR.

Methodology

A survey has been conducted to analyze font style and font sizes of Urdu books and magazines for the development of Urdu image corpus. According to the survey, most of the Urdu books and magazines are written using Nastalique writing style having 14 to 40 font sizes. The normal text of the Urdu books is written using 14 and 16 font sizes. In children books, the normal text is written in larger font sizes range from 18-22 font sizes. The remaining font sizes are normally used to write headings. Therefore, based on this analysis, three categories of the document images are defined (1) normal text, (2) normal text for children and poetry books and (3) headings text. The complete process for books collection, corpus acquisition, corpus labeling, and ground truth data generation has been designed for the development of this image corpus. Each of the sub-processes is detailed in subsequent sections.

1.1 Corpus Collection

Corpus collection process is divided into two main phases i.e. corpus design and corpus development. Corpus design deals with selection of books from which the selected document pages will be scanned. Books for each font size category are selected on the basis of defined criterion to ensure variety of domains, paper quality, print quality, paper transparency, and publisher and publication date. Corpus development involves scanning, organization and GT generation of the scanned images. Details are presented in subsequent sections.

Image corpus for 14 to 40 font sizes has been developed for the coverage of different available font sizes. To generate presented image corpus, first step is the selection and purchasing of Urdu books. The selection criterion is detailed below.

1. **Character Set and Symbols**: The image corpus should cover the following:
    a. Urdu alphabet given in Fig.6 below
    b. Latin digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
    c. English characters (A-Z, a-z)
    d. Urdu digits (۰, ۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹)
    e. Urdu aerab (ً ُ ٗ ٍ ٰ ِ َ ْ)
    f. Other symbols of Urdu, as follows: (ؔ ؕ ؑ ؐ ؒ ؓ ٘ ؟ () " ' ۔ ، : ؛)

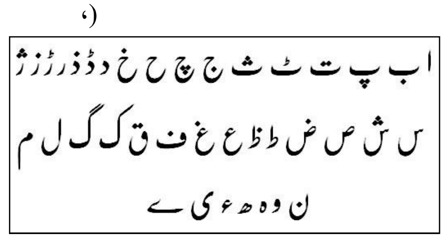


Figure 6. Urdu character set

2. **Font Size and Style:** Based on the survey findings books written using Noori Nastalique should picked up. The selected font sizes are 14-40.
3. **Multiple Domains:** Books are selected from multiple domains for each font size category to address the coverage of a balanced corpus.
4. **Publishers and Publication Date Variety:** Books published from multiple publishers of different cities are selected. In addition, variety of publishers within a city is also considered. Other than publisher, publication date also affects printing as well as paper quality of books. Therefore, while selecting the books, this parameter is also considered and books having variety of publication dates are selected.
5. **Page/Printing Quality:** Paper and printing qualities also affect image quality. All these varieties are included in the Urdu text image corpus to have the standard dataset for Urdu images.

1.2 Corpus Development from Books

Based on the availability of books according to the above mentioned criteria, the number of books and pages are selected according to font size category. To estimate font size of the printed text, Urdu character set and two characters high frequent Urdu ligatures are typed at multiple font sizes range from 14 to 40. These are then printed on transparencies which are placed on the printed text of books to find font size. For normal font size i.e. 14 and 16 font sizes, at least 100 books are selected and five pages from each book are scanned to generate image. In addition, table of content (TOC) page and page with or image/figure are also scanned for the researchers who want to do research on document layout analysis. For the second category of font size i.e. children books which are available less in frequency as compared to the first category, at least 30 books for each of the selected font size i.e. 18, 20 and 22 font sizes, are selected and from each book at least 5 pages are selected to scan. To generate image corpus of third category i.e. heading text, at least 20 books for each of 24, 28, 32, 36 and 40 font sizes are selected and at least 10 headings from each book are marked to scan. The number of books, number of scanned images, and domains coverage for each font sizes are provided in Section 3.

The selected pages of each book are scanned at 300 DPI using HP Scanjet G3110 scanner. During scanning, both BW and gray scale versions are generated for each font size except for 16. For each version, two types of images are scanned; (1) image without cropping the region of interest and (2) image with cropping the region of interest, both samples of gray scale and BW are shown in Figure 7. The images without cropping the region of interest is developed for the researchers who want to do research on page frame detection of Urdu document images. To generate image corpus for headings, the heading textual area is extracted and saved during scanning. All the images are saved in JPG or BMP file format.

1.3 Corpus Organization

The intelligent labeling of the image corpus is essential for the research and development. This is normally done manually and is time consuming task to ensure error free data labeling. The data labeling helps to extract the desired data automatically. Image corpus for each font size is maintained separately to maintain Urdu image corpus in an orderly manner. Moreover, gray scale and BW images are placed separately. Each version of cropped (edited) and un-cropped (unedited) of BW and gray scale are also maintained.

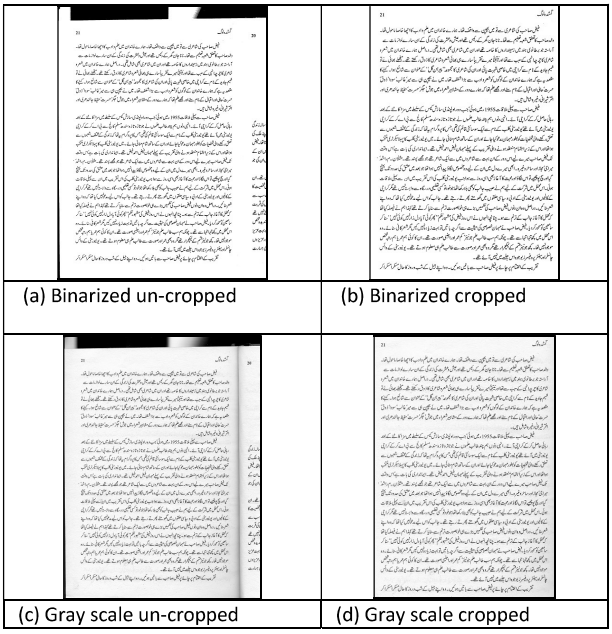


Figure 7. Sample of binarized and gray scale cropped and un-cropped region of interest

The naming convention of images is defined to automatically extract information related to the book, font size, editing and color versioning etc. Each image name for normal text (for 14 to 22 font sized text images) has following tags.

A_B_*C_D_E_F_G.jpg

e.g. BW_UE_B13_R_P26_F14.jpg where

1. ***A*** represents the image format information i.e. gray scale represented by **G** or Black and White represented by **BW**.
2. ***B*** tag represents whether the scanned image is cropped (edited) represented by **E**, or un-cropped (unedited) represented by **UE**.
3. ****C*** When ***B*** tag is E then **C** tag is used to indicate editing type which is cropped for this corpus. Image name does not have C tag when ***B*** tag is UE.
4. ***D*** tag defines the book number (assigned manually) of the image from which it is scanned. The book number has **B** as prefix letter indicating book. This book number can be used to get further information about book including book name, author, publisher, publications date and domain which is maintained in separate file.
5. ***E*** indicates content type of the scanned image. The image can have normal (or regular) text represented as **R**, figure represented as **I**, table of contents represented as **T.**
6. ***F*** is correspond to the page number of book which is scanned to generate the image. The page number is defined with letter **P** as prefix.
7. ***G*** is last tag used for the font size of image. Depending on the font sizes appeared in the text of the image, there can be multiple entries of font size, each is defined with prefix F.

The image corpus for each font size of headings is also maintained separately. As heading images are actually cropped from the document image during scanning. Therefore, cropped and un-cropped versions are not maintained explicitly. The naming convention for the heading image is defined as follows:

A_D_H_F_H#_G.jpg

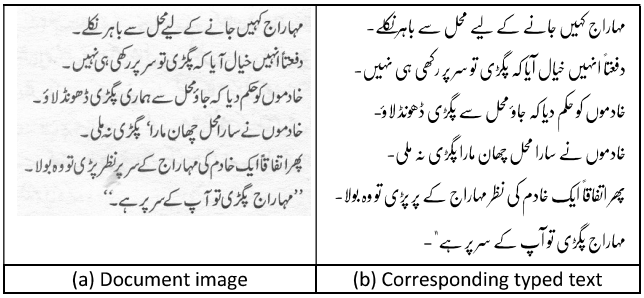
e.g. G_B149_H_P34_H1_F32.jpg where

A, ***D***, ***F*** and ***G*** tags are same as mentioned above. The ***H*** is used to define the type of the image i.e. H indicating the image is of heading. There can be more than one heading on same page. The **H#** is used to define the sequence number of heading in the document image from which the heading image is extracted. The heading number is defined with prefix letter **H** as can be seen in above example.

The complete information of each font size corpus is also maintained manually in separate file during scanning of images. This file provides information related to the book ID, book name, author name, publisher, year of publication, city, total number of pages, domain, image name, available font sizes in image, and columns (either 1or 2) of each scanned image. This information is cross verified to generate the error free details of an image.

1.4 Text Corpus of Images

Parallel typed version of each image is also generated as ground truth data, to process and recognize document images of the reported image corpus. This GT data will assist the researchers to extract training and testing data for classification and recognition by developing segmentation of lines and ligatures systems. Furthermore, this parallel text corpus of the reported image is also helpful for the researchers to develop language models using contextual information for post-processing of OCR system to improve the accuracy. Each scanned document image is typed by two typists. They are given instruction to type text as is and enter carriage return where required to have exact mirror of the image. This means number of lines in text files must be same as number of lines in document image (Figure 8). A total of 2,843 images are typed. Both versions of typed data are manually verified and mistakes are removed. During verification of the text pages, it has been observed that in some pages, typist typed correctly but in document image there were typo mistakes. Therefore, for the training and recognition of Urdu OCR it has been ensured that text corpus should be the mirror of image and those typo errors are remained in the text version. The detailed statistics of text corpus are given in Section 3.



(a) Document image	(b) Corresponding typed text

Figure 9. Sample of image and corresponding typed text

Corpus Statistics

The image corpus has been developed covering variety of domains for each font size. During development, complete information about the page is maintained. The summarized information of number of books, domains and authors is given in Table 1.

Table 1: Statistics of Urdu image corpus

Font size	Book/Magazine count	Number of document images	Domains	Authors
14	101	593	18	76
16	116	595	19	100
18	30	150	10	23
20	45	149	2	24
22	56	151	2	21
24	21	461	18	24
28	21	202	6	21
32	23	186	9	21
36	31	226	7	22
40	26	199	7	22

The scanned document images contain normal text, TOC and figures for the Urdu documents layout analysis research and development. The presented corpora contain total of 29 document images which have figures and 84 document images of TOC. The document images having normal text also have variation of paper, printing, headings, headers and footers etc. The sample layouts of figures, normal text and TOC are shown in Figure 9 and Figure 10.

Table 2. Additional characters' Unicode of Urdu

Identical Form	Decomposed form
ئ (U+0626)	◌ٔ+ی (U+06CC + U+0654)
ؤ (U+0624)	◌ٔ+و (U+0648+ U+0654)
أ (U+0623)	◌ٔ+ا (U+ 0627+ U+0654)
ۂ (U+06C2)	◌ٔ+ہ (U+06C1+ U+0654)
ۃ (U+06C3)	

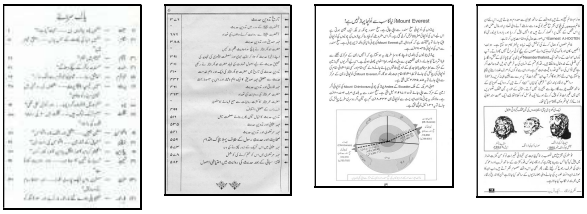


Figure 10. Samples of document images having figures and TOC

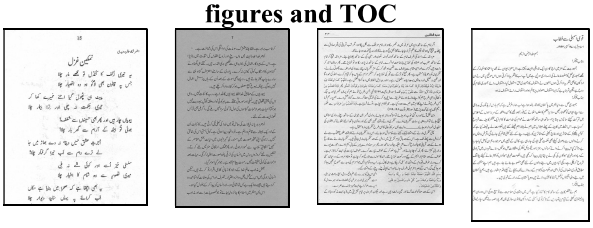


Figure 11.Sample layouts of normal text having variation of paper and printing qualities

Table 3. Font wise Urdu Letters, Urdu digits, English letters and digits, Urdu Aerab and symbols statistics

Font Size	Total Characters	Unique Urdu Characters	Unique Urdu Digits	Unique English Characters	Unique Latin Digits	Unique Urdu Aerab	Unique Symbols
14	726,385	45	10	52	10	12	32
16	579,730	45	10	51	10	13	31
18	111,178	44	10	34	10	10	22
20	101,559	44	10	20	10	9	15
22	81,718	43	8	3	10	10	18
24	7,807	43	3	10	6	7	18
28	2,730	42	3	16	4	6	12
32	5,519	40	0	4	10	9	11
36	3,462	42	4	11	3	7	13
40	2,949	42	0	0	0	9	12

Table 4. Font wise lines and ligature statistics of corpus

Font Size	Total document images	Lines	Total Ligatures	Unique Ligature	Average Lines per image	Average Ligatures per Line
14	591	13,712	386,648	6,452	23	28
16	528	11,080	306,080	5,938	20	27
18	150	2,622	60,056	2,872	18	23
20	149	2,318	54,657	2,204	16	24
22	151	1,857	43,121	1,865	12	23
24	461	463	3,961	883	1	9
28	202	203	1,424	502	1	7

32	186	274	2,874	616	2	11
36	226	260	1,776	537	1	7
40	199	222	1,510	498	1	7

Conclusion

In this paper, a comprehensive image corpus of Nastalique writing style is presented. The complete process to select books according to the define criteria, scan and organize the images in orderly manner is defined. In addition, ground truth typed data is also developed. A total of 2, 912 images are selected from 413 books. Among these, 593, 595, 150, 149 and 151 images are scanned for 14, 16, 18, 20 and 22 font sizes. The image corpus for headings contains 461, 202, 186, 226 and 199 heading images for 24, 28, 32, 36 and 40 font sizes respectively. The subset of the reported document image corpus for 14, 16, 18, 20, 22, 24, 28, 32, 36 and 40 are publically available for researchers at [26-35]. Moreover, the typed corpus of each font size is prepared as ground truth information which is also publically available at [36-45].

Acknowledgements

This work has been supported by Urdu Nastalique OCR research project grant by ICTR&D Fund, Ministry of IT, Govt. of Pakistan. See www.UrduOCR.net for details.

References

[1]. R. Smith, "An Overview of the Tesseract OCR Engine," in Ninth Int. Conference on Document Analysis and Recognition (ICDAR), 2007.

[2]. R. Smith, D. Antonova and D.-S. Lee, "Adapting the Tesseract open source OCR engine for multilingual OCR," in International Workshop on Multilingual OCR, Barcelona, Spain, 2009.

[3]. N. Sankaran and C. V. Jawahar, "Recognition of printed Devanagari text using BLSTM Neural Network," in 21st International Conference on Pattern Recognition (ICPR), 2012.

[4]. N. Sabbour and F. Shafait, "A Segmentation Free Approach to Arabic and Urdu OCR," in SPIE, Volume 8658, 2013.

[5]. Antonacopoulos, S. Pletschacher, C. Clausner and C. Papadopoulos, "Competition on Historical Newspaper Layout Analysis (HNLA2013)," in 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.

[6]. D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. Gomez-i-Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan and L. P. Heras, "ICDAR 2013 Robust Reading Competition," in 12th International Conference on Document Analysis and Recognition (ICDAR), 2013.

[7]. U. -V. Marti and H. Bunke, "A full English sentence database for off-line handwriting recognition," in Fifth International Conference on Document Analysis and Recognition, 1999.

[8]. U. -V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," International Journal on Document Analysis and Recognition, vol. 5, pp. 39-46, 2002.

[9]. M. Davis, "Unicode Text Segmentation," Addison-Wesley Professional, 2013.

[10]. V. Margner and M. Pechwitz, "Synthetic data for Arabic OCR system development," in Sixth International Conference on Document Analysis and Recognition, 2001.

[11]. S. Al-Ma'adeed, D. Elliman and C. A. Higgins, "A data base for Arabic handwritten text recognition research," in Eighth International Workshop on Frontiers in Handwriting Recognition, 2002.

[12]. S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban and S. M. Golzan, "A Comprehensive Isolated Farsi/Arabic Character Database for Handwritten OCR Research," in Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule (France), 2006.

[13]. N. Kharma, M. Ahmed and R. Ward, "A new comprehensive database of handwritten Arabic words, numbers, and signatures used for OCR testing," in 1999 IEEE Canadian Conference on Electrical and Computer Engineering, 1999.

[14]. M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze and H. Amiri, "IFN/ENIT - database of handwritten Arabic words," in CIFED 2002, 2002.

[15]. Wali and S. Hussain, "Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation," in International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (CISSE), 2006.

[16]. S. Hussain, "www.LICT4D.asia/Fonts/Nafees_Nastalique," in 12th AMIC Annual Conference on E-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore, 2003.

[17]. S. Hussain, S. Rahman, A. Wali, A. Gulzar and S. J. Rahman, "Grammatical Analysis of Nastalique Writing Style of Urdu," Center for Research in Urdu Language Processing, FAST-NU, Lahore, 2002.

[18]. Q. Akram, S. Hussain, A. Niazi, U. Anjum and F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique," in 11th IAPR Workshop on Document Analysis Systems, Tours, France, 2014.

[19]. Hasan, S. B. Ahmed, S. F. Rashid, F. Shafait and T. M. Breuel, "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," in International Conference on Document Analysis and Recognition, 2013.

[20]. Muaz, "Urdu Optical Character Recognition System," Unpublished, MS Thesis Report, National University of Computer and Emerging Sciences , Lahore, 2010.

[21]. S. A. Sattar, "A Technique For The Design And Implementation Of An OCR For Printed Nastalique Text," Unpublished, Degree of Doctor of Philosophy

Thesis Report, N.E.D University of Engineering and Technology, Karachi, Pakistan, 2009.
[22].D. A. Satti, "Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach".
[23].F. Shafait, D. Keysers and T. M. Breuel, "Layout Analysis of Urdu Document Images," in INMIC'06. IEEE, 2006.
[24].S. Tariq and S. Hussain, "Segmentation Based Urdu Nastalique OCR," in 18th Iberoamerican Congress on Pattern Recognition (CIARP 2013), Havana, Cuba, 2013.
[25].M. Naz, Q. Akram and S. Hussain, "Binarization and its Evaluation for Urdu Nastalique Document Images," in INMIC, Lahore, 2013.
[26]."CLE Urdu Image Corpus 14 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurduimage corpus14pt.htm. [Accessed 30 9 2016].
[27]."CLE Urdu Image Corpus 16 Point Size," Center for Language Engineering (CLE), 30 06 2012. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus16pt.htm. [Accessed 30 09 2016].
[28]."CLE Urdu Image Corpus 18 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus18pt.htm. [Accessed 30 09 2016].
[29]."CLE Urdu Image Corpus 20 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus20pt.htm. [Accessed 30 09 2016].
[30]."CLE Urdu Image Corpus 22 Point Size," Center for Language Engineering (CLE), 31 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus22pt.htm. [Accessed 30 09 2016].
[31]."CLE Urdu Image Corpus 24 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus24pt.htm. [Accessed 30 09 2016].
[32]."CLE Urdu Image Corpus 28 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus28pt.htm. [Accessed 30 09 2016].
[33]."CLE Urdu Image Corpus 32 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus32pt.htm. [Accessed 30 09 2016].
[34]."CLE Urdu Image Corpus 36 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus36pt.htm. [Accessed 30 09 2016].
[35]."CLE Urdu Image Corpus 40 Point Size," Center for Language Engineering (CLE), 06 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurduimagecorpus40pt.htm. [Accessed 30 09 2016].
[36]."CLE Urdu Text Corpus 14 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus14pt.htm. [Accessed 30 09 2016].
[37]."CLE Urdu Text Corpus 16 Point Size," Center for Language Engineering (CLE), 30 06 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus16pt.htm. [Accessed 30 09 2016].
[38]."CLE Urdu Text Corpus 18 Point Size," Center for Language Engineering (CLE), 30 10 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus18pt.htm. [Accessed 30 09 2016].
[39]."CLE Urdu Text Corpus 20 Point Size," Center for Language Engineering (CLE), 12 07 2012. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus20pt.htm. [Accessed 30 09 2016].
[40]."CLE Urdu Text Corpus 22 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus22pt.htm. [Accessed 30 09 2016].
[41]."CLE Urdu Text Corpus 24 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus24pt.htm. [Accessed 30 09 2016].
[42]."CLE Urdu Text Corpus 28 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus28pt.htm. [Accessed 30 09 2016].
[43]."CLE Urdu Text Corpus 32 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus32pt.htm. [Accessed 30 09 2016].
[44]."CLE Urdu Text Corpus 36 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus36pt.htm. [Accessed 30 09 2016].
[45]."CLE Urdu Text Corpus 40 Point Size," Center for Language Engineering (CLE), 07 11 2014. [Online]. Available: http://cle.org.pk/clestore/cleurdutextcorpus40pt.htm. [Accessed 30 09 2016].