2012 International Conference on Frontiers in Handwriting Recognition

An Unconstrained Benchmark Urdu Handwritten Sentence Database with Automatic Line Segmentation

Ahsen Raza
National University
of Sciences & Technology
Islamabad, Pakistan

Imran Siddiqi
Bahria University
Islamabad,
Pakistan

Ali Abidi
National University
of Sciences & Technology
Islamabad, Pakistan

Fahim Arif
National University
of Sciences & Technology
Islamabad, Pakistan

ashen.raza@mcs.edu.pk, imran.siddiqi@bahria.edu.pk, abidi@mcs.edu.pk, fahim@mcs.edu.pk

***Abstract*— In this paper we present and announce a novel off-line sentence database of Urdu handwritten documents along with a few preprocessing and text line segmentation procedures. Despite an increased research interest in Urdu handwritten document analysis over the recent years, a standard benchmark dataset, which could be used in Urdu handwriting recognition tasks, has been missing. Based on our own developed and updated corpus named CENIP-UCCP (Center for Image Processing- Urdu Corpus Construction Project), we have developed an Urdu handwritten database. The corpus is a collection of a variety of Urdu texts that were used to generate forms. These forms were subsequently filled by native writers in their natural handwritings. Six categories of text were used to generate these forms with each category using approximately 66 forms. Up till now, the database comprises 400 digitized forms produced by 200 different writers. The database is completely labeled for content information as well as content detection and supports the evaluation of systems like Urdu handwriting recognition, line segmentation and writer identification. The database was also experimented with the proposed Urdu text line segmentation scheme rendering promising segmentation results.**

***Keywords*---- Handwriting recognition, database, corpus, Urdu.**

I. Introduction

The development of standard databases is important in any research domain. This provides a platform for the comparison and evaluation of different algorithms and techniques on the same grid, without any bias [2, 11]. Recent years have witnessed an increasing demand of such benchmarking datasets in different research areas. Such databases not only save the researcher from the tedious task of compiling and labeling the data but also provide the possibility of objectively comparing different systems on the same data set. A step ahead to this is the organization of various evaluation campaigns which allow the comparison of different techniques under the same experimental conditions [16, 22]. The document analysis and recognition (DAR) community has also developed a number of standard databases allowing researchers to quantify their techniques on common benchmarks. The most popular of these are the databases for handwriting recognition. These include the databases like CEDAR [12], MNIST [13], CENPARMI [14], IAM [11] and RIMES [16] for offline, while IAM-OnDB [18, 19], UNIPEN [23] and IRONOFF [24] for online document recognition. These databases have been used for a variety of recognition tasks like character and word recognition, line/word segmentation, word spotting, document layout analysis, document segmentation and, writer identification and verification.

Urdu is one of the major languages of the Indian sub-continent with speakers all over the world. Urdu document analysis and recognition has gained research interest during the last decade [2-9, 15, 17]. Current areas of interest include Urdu OCR, handwriting recognition and word spotting based retrieval systems. With the advancement in these areas, there will naturally be a need to evaluate and compare the proposed techniques on standard data sets. To the best of authors' knowledge, there is no unconstrained database of Urdu handwritten documents that could be employed for Urdu handwriting recognition and related tasks. Only one constrained collection of Urdu handwritings can be found in the literature [7]. This data set, however, comprises a limited vocabulary and does not capture the semantic and syntactic variations of the script so conclusive experiments cannot be performed.

In this paper, we present the first version of a database comprising complete Urdu sentences. At present, the database consists of 400 handwritten forms, written by 200 different writers and contains a total of 23833 printed Urdu words in 2051 lines of text. To capture the maximum syntactic variations, forms were filled by a variety of writers having diverse backgrounds and coming from geographically distributed locations in Pakistan. The database is labeled by finding the coordinates of each line of text as well as its transcription and hence can be used to evaluate handwriting recognition and related systems.

In the next section, we discuss corpus acquisition followed by generation and filling of forms. We then discuss some characteristics and statistics of the collected data and present the naming conventions and ground truth labeling. Next we discuss the proposed text line segmentation procedure and

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.177

its results. Finally, we give the concluding remarks and discuss some future enhancements to the present database.

II. Corpus Acquisition

In the first phase of database development, the corpus was extracted from authentic sources and arranged as plain text called plain corpora [11]. Electronically available resources are the most suitable for collection of text but unfortunately there are very few repositories of Urdu text in UTF-8 format. Most of the sources present Urdu text as graphics which cannot be used in any text based application [3, 4, 30]. In our case, text was extracted from 6 broad domains with content not older than 2010. Text was extracted from two reliable and updated electronic sources which include BBC (British Broadcasting Corporation) Urdu [34] and Jang Pakistan [35]. In some cases, in addition to Urdu content, the text also contained occurrences of numerals. All such occurrences were separately identified and recorded as well.

In the next section, we discuss in detail the design and structure of forms and their filling.

III. Forms & Writers

To generate the forms, the corpus was split into 4 to 6 lines of text containing approximately 5 to 6 sentences with at least 50 to 65 words. This fragmented text was then copied to forms which were printed and distributed to writers for copying the text in their natural handwritings. Each writer was given two forms with different texts on each form.

The structure of the forms is inspired from that of the IAM database [11] and is indicated in Figure 1. Each form consists of four parts. The first part comprises the title "Urdu Text Database" and a unique identification number. To generate the form numbers, each category is assigned a letter code, each form within a category is assigned a two-digit code and each writer is assigned a three-digit code. These codes are then used to name the forms according to the following convention:

[CategoryCode]-[FormCode]- [WriterID]

For example "S-01-001" specifies that the text of the form belongs to the text category "Sports", 01 represents form count in this category while 001 identifies the writer of the form.

Second part of the form consists of 4 to 6 lines of printed text from the corpus while the third part of the form was left blank where the writers were asked to copy the printed text. In the last part, the writers could optionally provide their names.

For collecting writing samples, we selected individuals belonging to different backgrounds and geographically distributed across Pakistan to fill the forms. Writers were asked to copy the printed text in their natural handwriting without any constraints on the writing style or the writing instrument. In most of the cases, individuals used blue or black ball points as the writing instrument. Twenty data collection sessions were organized with 10 individuals per session (on the average).

The filled forms were scanned with HP-Scan jet 6800 using Redires 9.0 as the scanning software. The resolution was set to 300dpi in true color. The scanned forms were saved in a TIFF format with LZW compression. Each form was completely scanned including printed as well as handwritten part. This allows the utilization of these forms in an application to distinguish between handwriting and printed texts [11] as well.

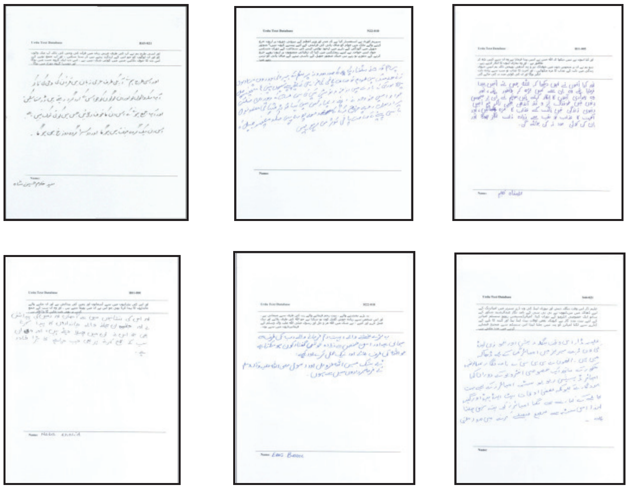


Figure 1. Sample printed and filled forms

IV. Characteristics & Statistics

The database comprises a total of 400 filled forms contributed by 200 writers. The corpus used to generate the forms was extracted from two huge and authentic electronic sources namely BBC Urdu [34] and Jang news [35] in six different categories as listed in Table 1. These categories are chosen to be science fiction, entertainment, sports, blogs, religion and editorials. The number of forms in each category is summarized in Figure 2 while more detailed distribution of sentences and words in each of the categories can be found in Figure 3 and Figure 4 respectively. The distribution of words per line of text is indicated in Figure 5. There are a total of 23833 printed Urdu words in the database. Corresponding handwritten words are 23812. The database contains a total of 2068 Urdu sentences with 2051 Urdu lines. The database also contains 783 numerals. More detailed statistics of the database can be found in Table 1.

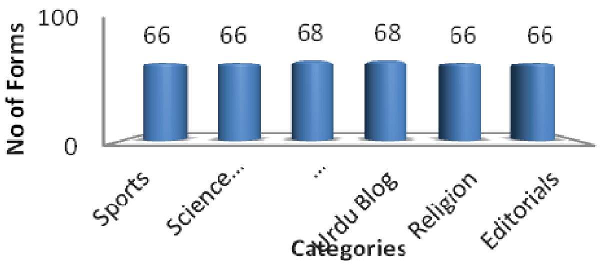


Figure 2. Distribution of forms in categories

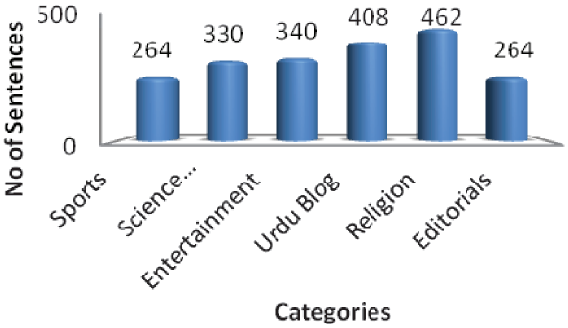


Figure 3. Distribution of sentences in categories

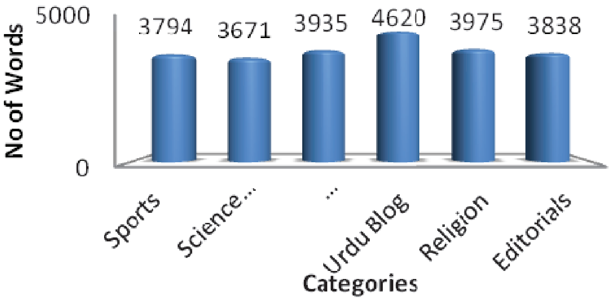


Figure 4. Distribution of words in categories

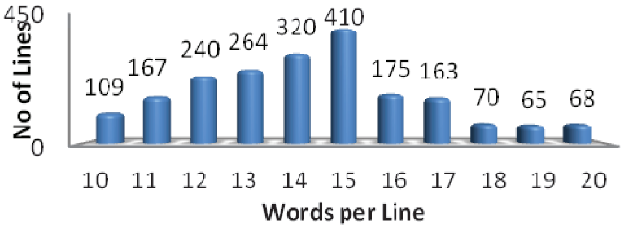


Figure 5. Distribution of words in lines

V. GROUND TRUTH DATA LABELING

The labeling of ground truth data is imperative to any database. This, naturally, is a time consuming, expensive and error prone task [1, 11]. For our collection, we have carried out two types of labeling, one for content information and the other for content detection. Each text line in the handwritten text is identified manually using simple software (Figure 6) that allows opening an image and drawing rectangles over the textual regions. The x and y coordinates and the width and height of each line of text are stored in a data file. This data can be used to evaluate line segmentation algorithms. The next level, word segmentation, has not been considered in the present version of the database and will be included in the next version allowing evaluation of word segmentation schemes as well.

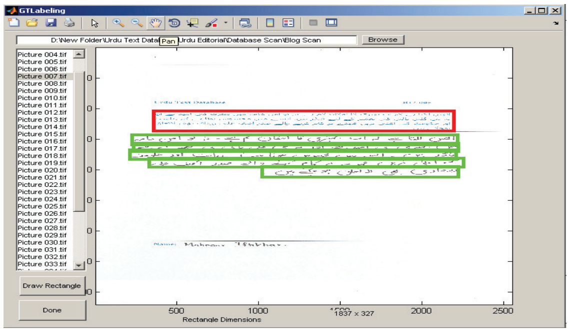
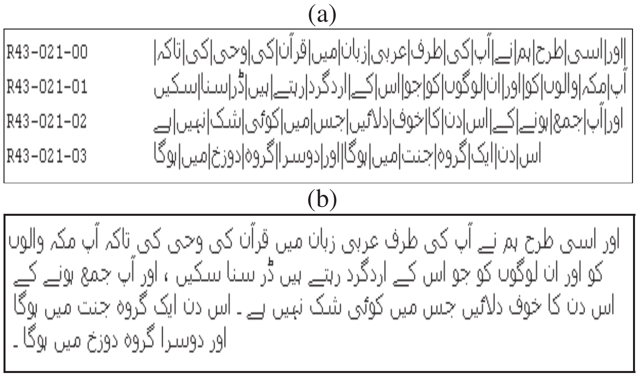


Figure 6. A snap shot of ground truth labeling software for content detection

For evaluation of handwriting recognition systems, the ground truth data corresponds to the actual transcription of each line of text. For this, we generated two data files for each scanned document, one containing the transcription of the printed text in a paragraph while the other containing the ground truth text corresponding to each line of handwritten text. An illustration of these files can be seen in Figure 7.



(a)

R43-021-00	|اور|اسی|طرح|ہم|نے|آپ|کی|طرف|عربی|زبان|میں|قرآن|کی|وحی|کی|تاکہ|
R43-021-01	آپ|مکہ|والوں|کو|اور|ان|لوگوں|کو|جو|اس|کے|اردگرد|رہتے|ہیں|ڈر|سنا|سکیں
R43-021-02	اور|آپ|جمع|ہونے|کے|اس|دن|کا|خوف|دلائیں|جس|میں|کوئی|شک|نہیں|ہے
R43-021-03	اس|دن|ایک|گروہ|جنت|میں|ہوگا|اور|دوسرا|گروہ|دوزخ|میں|ہوگا

(b)

اور اسی طرح ہم نے آپ کی طرف عربی زبان میں قرآن کی وحی کی تاکہ آپ مکہ والوں
کو اور ان لوگوں کو جو اس کے اردگرد رہتے ہیں ڈر سنا سکیں ، اور آپ جمع ہونے کے
اس دن کا خوف دلائیں جس میں کوئی شک نہیں ہے ۔ اس دن ایک گروہ جنت میں ہوگا
اور دوسرا گروہ دوزخ میں ہوگا ۔

Figure 7. An example of the date files corresponding to Figure 1 **(a):** ground truth of the handwritten text. **(b):** ground truth of the machine printed text.

Table 1 Some statistics of the database

Category	Category Symbol	Number of Forms	Average Urdu words/ Form	Average Urdu lines/ Form	Average Urdu sentences/ Form	Average numeral(s) /Form	No of Writers	Total Urdu Sentences	Total Urdu words	Total Urdu Lines	Total numerals
Sports	S	66	57	5	4	2	33	264	3694	334	132
Science Fiction	F	66	54	6	5	1	33	330	3635	337	130
Entertainment	E	68	60	6	5	1	34	340	3935	356	134
Urdu Blog	B	68	71	6	6	2	34	408	4620	348	137
Religion	R	66	57	7	7	1	33	462	3970	336	126
Editorials	N	66	57	5	4	1	33	264	3838	340	124
Overall	**-**	**400**	**59**	**6**	**5**	**1**	**200**	**2068**	**23833**	**2051**	**783**

VI. TEXT LINE SEGMENTATION

In this section we present the proposed text line segmentation procedure and compare the segmentation results with the manual segmentation. The procedure is primarily based on a set of image processing operations. A general flow of the scheme can be seen in Figure 8. As a first step we apply cubic interpolation to the input (gray scale) image as this gives a smooth estimate of the gray level at any desired point in the image [25]. This is followed by the application of a median filter to remove any salt and pepper noise that may be introduced during image acquisition.

We then detect and extract handwritten text from the form. Since the handwritten part of the form is separated by long black horizontal lines, one may detect these lines and extract the handwritten part. This is done using the well known Hough transform for line detection [32, 33]. A horizontal line separating two parts is characterized by a high peak in Hough transform. These separating lines were also used to detect and correct the skew in the scanned forms.

Once the handwritten part is extracted, we binarize the extracted region of interest. After evaluating a number of binarization techniques [26-29], we used the classical Otsu's global thresholding algorithm. To split the binarized text into individual lines, we employed the well-known horizontal projection profiles [20, 21, 36] which are scanned for local minima. If the value at a local minimum is zero, it corresponds to an ideal separation between two text lines. If the value is greater than zero, it means there is some intersection between the words of upper and lowers lines.

To handle these intersection issues, we first find the connected components in the image and filter out all the components with area below a predefined threshold. These components are assumed to be dots and diacritic marks and are not considered for line segmentation. For each of the remaining components, we compute its centroid. If the centroid lies above the cutting line of the horizontal projection profile, the component is assigned to the upper line otherwise it is assumed to belong to the lower line. The line segmentation procedure is illustrated on an image in Figure 9.

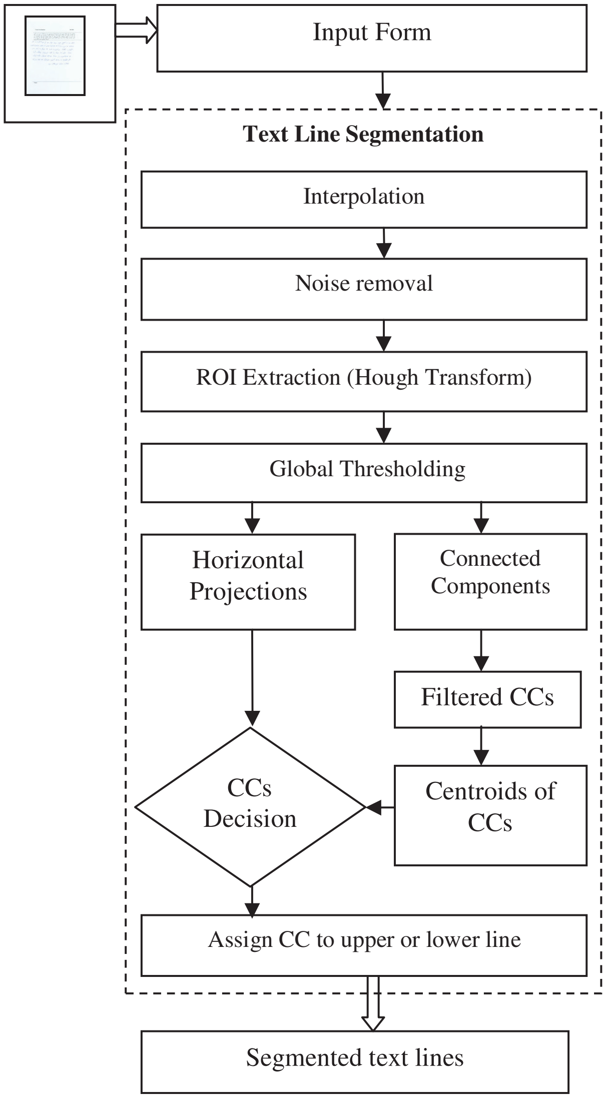
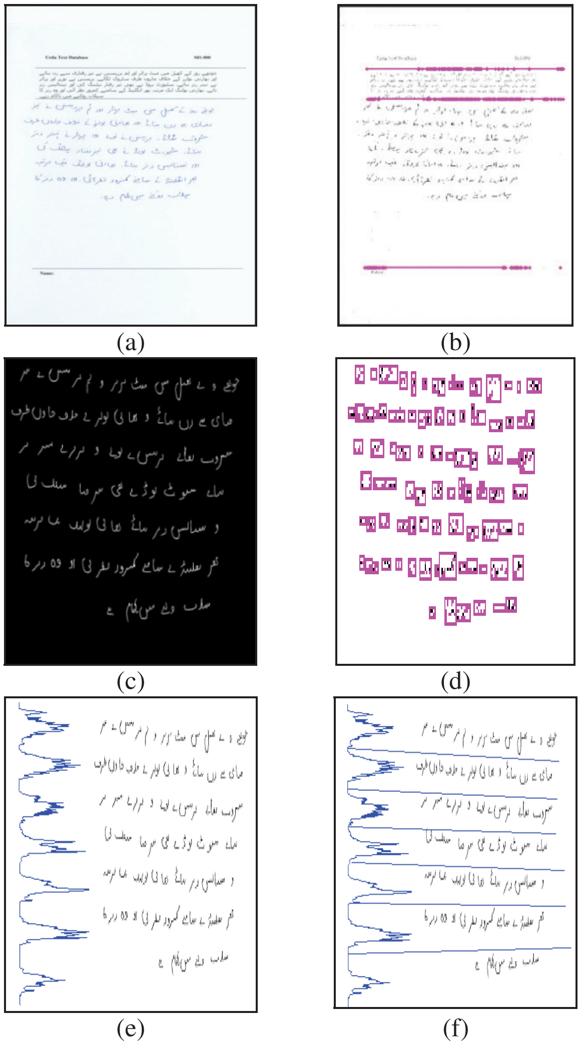


Figure 8. A general flow of Urdu text line segmentation procedure.



(a) (b) (c) (d) (e) (f)

Figure 9. Various steps of text line segmentation procedure; (a): origional form; (b): hough transform for line detection; (c):ROI extracted; (d):filtered connected components and centroids of connected components. (e):Horizontal projection profile; (f): HPP based line segmentation

VII. Experimental Results

Experiments are carried out to compare the performance of the automatic text/line segmentation scheme with that of manual segmentation. The handwritten text is extracted from the forms with no errors. For line segmentation, out of 2051 text lines, 2001 were correctly segmented and only 50 (2.4%) could not be properly segmented. The detailed segmentation results can be found in Table 2.This category wise result distribution in Table 2 signifies the challenge of text line segmentation in each category which is naturally dependent on writings produced by the writers of that particular category. We also compared the line segmentation results with and without skew correction the former, naturally, comes out to be more effective. The segmentation results with and without skew correction for each of the categories are summarized in Figure 10.

Table 2. Urdu text line segmentation results for each category of text.

S/No	Category Symbol	No of Correctly Segmented Lines	Lines with error	Error %
1	S	322	12	0.6%
2	F	322	15	0.7%
3	E	346	10	0.5%
4	B	344	04	0.2%
5	R	331	05	0.2%
6	N	336	04	0.2%
Overall	-	**2001**	**50**	**2.4%**

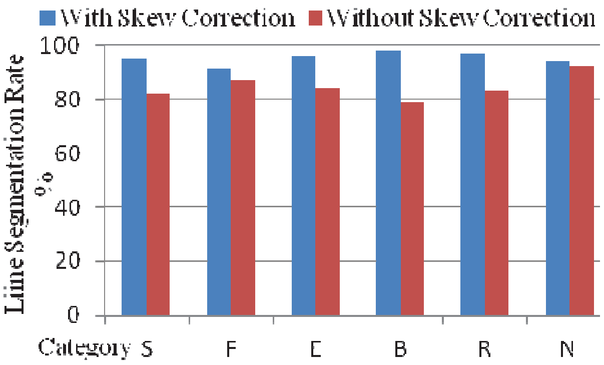


Figure 10. Line segmentation rate comparison with skew correction and non-skew correction parameters of forms.

VIII. Future Perspectives

In this paper, we have presented a novel sentence database of Urdu handwriting. The database is based on our own developed corpus named as CENIP-UCCP. The database is labeled with coordinates of each line of text as well as the transcription of the printed and handwritten texts. The main aim of the present study was to provide a benchmark sentence database in Urdu language. At present, the database mainly supports tasks like Urdu text/handwriting recognition, line segmentation and printed/handwritten text segmentation. Since the writer ID is also stored on each form, the same data set can be used to evaluate writer identification/verification systems as well. The database was also experimented with an automatic line segmentation scheme reporting high segmentation rates.

In future, we intend to enhance the size of the current version of database. We hope that the size of database would be doubled encompassing diversity of more natural writing styles in its future version. We also plan to provide ground truth data at word level allowing the evaluation of word segmentation algorithms. The database would be made publically available soon. The authors expect that this database would be helpful for researchers working on Urdu handwriting recognition and related tasks.

REFERENCES

[1] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong and R. Young, "ICDAR 2003 Robust Reading Competitions". Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003.

[2] H. Malik and M. Fahiem, "Segmentation of Printed Urdu Scripts Using Structural Features", In Proceedings of Second International Conference in Visulaization, pp.191-195.2009.

[3] M. Ijaz and S. Hussain,. "Corpus Based Urdu. Lexicon Development", In Proceedings of the Conference on. Language and Technology, Peshwar Pakistan, pp. 1-12, 2007.

[4] A. Ali, S. Siddiq and K. Malik, "Development of Parallel Corpus and English to Urdu Statistical Machine Translation", International Journal of Engineering & Technology IJET-IJENS Vol: 10 No: 05, pp. 31-33, 2010.

[5] A. damek, M. Humayoun, H. Hammarström and A. Ranta, "Urdu Morphology, Orthography and Lexicon Extraction", MSc Thesis, Department of Computing Science, Chalmers University of Technology, 2006.

[6] N. Durrani and S. Hussain, "Urdu Word Segmentation", In Proceedings of 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, 2010.

[7] M. Waqas, C. Lei, N. Nobile, and C. Y. Suen "A New Large Urdu Database for Off-Line Handwriting Recognition", In Proceedings of ICIAP, LNCS 5716, pp. 538–546, 2009.

[8] D. Becker and K. Riaz , "A Study in Urdu Corpus Construction", In Proceedings of 3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics. Taipei, Taiwan. 2002.

[9] M. Ijaz and S. Hussain, "Corpus Based Urdu Lexicon Development", In Proceedings of the Conference on Language and Technology. University of Peshawar, Pakistan, 2007.

[10] R. Wilkinson, J. Geist, S. Janet, P. Grother, C. Burges, R. Creecy, B. Hammond, J. Hull, N. Larsen, T. Vogl and C. Wilson, The first census optical character recognition systems conf. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, MD, 1992.

[11] U.V. Marti and H. Bunke. "A full english sentence database for off-line handwriting recognition", In Proceedings of International Conference on Document Analysis and Recognition (ICDAR)", pp. 705–708, 1999.

[12] J. Hull. "A database for handwritten text recognition research",In Proceedings of IEEE Trans. on Pattern Analysis and Machine Intelligence, pp.550–554, 1994.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition", In Proceedings of IEEE. pp. 2278-2324, 1998.

[14] C. Suen, C. Nadal, R. Legault, T. Mai, and L. Lam. "Computer recognition of unconstrained handwritten numerals", In Proceedings of the IEEE, pp.1162–1180, 1992.

[15] M. Sagheer, M. Waqas, N. Nobile, C. Y. Suen, C. Lei "A Novel Handwritten Urdu Word Spotting Based on Connected Components Analysis", In Proceedings of 20th International Conference on Pattern Recognition(ICPR), 2010.

[16] E. Augustin, M. Carré, E. Grosicki, J.M. Brodin, E. Geoffrois and F. Preteux. "Rimes evaluation campaign for handwritten mail processing". In Proceedings of the Workshop on Frontiers in Handwriting Recognition. pp. 231–235,2006.

[17] U. Pal and A. Sarkar "Recognition of Printed Urdu Script", In Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR),2003.

[18] M. Liwicki and H. Bunke, "IAM-OnDB - An On-Line English Sentence Database Acquired from Handwritten Text on a Whiteboard". In Proceedings of the Eighth International Conference on Document Analysis and Recognition(ICDAR), 2005.

[19] E. Indermühle, M. Liwicki and H. Bunke, "IAMonDo database: an Online Handwritten Document Database with Non-uniform Contents", In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 2010.

[20] U.V. Marti and H. Bunke, "Text Line Segmentation and Word Recognition in a System for General Writer Independent Handwriting Recognition", Institut for Informatik und angewandte Mathematik Universit¨at Bern, Neubr¨uckstrasse 10, CH-3012 Bern, Switzerland.

[21] L. Likforman, A. Zahour and B. Taconet, "Text Line Segmentation of Historical Documents: a Survey", International Journal on Document Analysis and Recognition(IJDAR), Springer, 2006.

[22] D. Pallett, "A Look at NIST's Benchmark ASR Tests: Past, Present, and Future", In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding,2003.

[23] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman and S. Janet, "Unipen project of on-line data exchange and recognizer benchmarks". In Proceedings of the 12th International Conference on Pattern Recognition(ICPR), 1994.

[24] C. Gaudin, P.M. Lallican, P. Binter and S. Knerr, "The ireste on/off (ironoff) dual handwriting database". In Proceedings of International Conference of document analysis and recognition(ICDAR),1999.

[25] R. C. Gonzalez and R. E. Woods. "Digital Image Processing". 2nd edition, 2001.

[26] W. Niblack, "An introduction to digital image processing", Prentice-Hall, Englewood Cliffs (NJ). pp. 115-116, 1986.

[27] J. Sauvola, T. Seppanen, S. Haapakoski and M.Pietikainen, "Adaptive Document Binarization", In Proceedings of 4th Int. Conf. On Document Analysis and Recognition(ICDAR), Ulm, Germany, pp.147-152, 1997.

[28] C. Wolf, J.M. Jolion and F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Documents". In Proceedings of the 16th International Conference on Pattern Recognition ICPR'02, Quebec, Canada, pp. 1037-1040, 2002.

[29] N. Otsu, "A threshold selection method from gray-level histograms", In Proceedings of IEEE Transactions on Systems, Man and Cybernetics, 1979.

[30] A. Raza , S. Hussain, H. Sarfraz, I. Ullah and Z. Sarfraz, "Design an development of phonetically rich Urdu speech corpus". National University of Computer & Emerging Sciences (NUCES), Pakistan.

[31] M. Stephanin and Strassel, "Linguistic Resources for Arabic Handwriting Recognition Data Consortium, Philadelphia, USA.

[32] L. Likforman, A. Hanimyan and C. Faure, "A Houghbased algorithm for extracting text lines in handwrittendocuments", In Proceedings of third International Conference on Document Analysis and Recognition(ICDAR), pp. 774-777, 1995.

[33] G. Louloudis, B. Gatos, I. Pratikakis and K. Halatsis,"A Block-Based Hough Transform Mapping for Text Line Detection in Handwritten Documents", In Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition(IWFHR), La Baule, Oct. 2006.

[34] http://www.bbc.co.uk/urdu/

[35] www.jang.com.pk/

[36] A. Nicolaou1 and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines", In Proceedings of 10th International Conference on Document Analysis and Recognition(ICDAR), pp-626-630. 2009.