UCOM Offline Dataset-An Urdu Handwritten Dataset Generation

Saad Bin Ahmed[1], Saeeda Naz[2,3], Salahuddin Swati[4], Imran Razzak[1], Arif Iqbal Umar[2], and Akbar Ali Khan[4]

[1]College of Public Health and Health Informatics, King Saud Bin Abdul Aziz University for Health Sciences, Saudi Arabia

[2]Department of Information Technology, Hazara University, Pakistan

[3]GGPGC No.1, Abbottabad, Higher Education Department, Pakistan

[4]COMSATS Institute of Information Technology, Pakistan

Abstract: *A benchmark database for character recognition is an essential part for efficient and robust development. Unfortunately, there is no comprehensive handwritten dataset for Urdu language that would be used to compare the state of the art techniques in the field of optical character recognition. In this paper, we present a new and publically available dataset comprising 600 pages of handwritten Urdu text written in Nasta'liq style in conjunction with detailed ground truth for the evaluation of handwritten Urdu character recognition. This dataset contains text lines written in Nasta'liq style by limited individuals on A4 size paper. The acquired data on page was scanned and text lines were segmented. UCOM database covers all Urdu characters and ligatures with different variation in addition to Urdu numeric data. We have considered that ligature consists of up to five characters in this dataset. The UCOM dataset can be used for handwritten character recogntition as well as writer identification. We proposed and evaluated the strength of Recurrent Neural Networks (RNN) on UCOM offline database sample text line.*

Keywords: *Recurrent neural networks, optical character recognition, cursive, offline handwriting.*

Received January 15, 2015; accepted June 21, 2016

1. Introduction

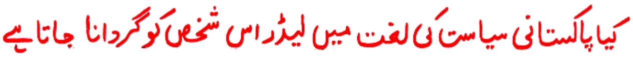
Document image analysis is intended to interpret digital documents. Such documents can be a synthetic data or in scanned form. The Optical Character Recognition (OCR) is a specialized technology which reads document images and translates them into searchable text. The OCR systems capable of recognizing characters and thereby words and sentences have been routinely developed and improved. Due to extensive research in the field of OCR since decade, there is utmost urge to generate standard and reliable dataset for evaluation of such techniques.

In the literature, several works have been done on the OCR systems for cursive [1, 2, 3, 9, 15, 21, 22, 23, 24] and non-cursive [4, 7, 8, 10] scripts. The dataset for cursive scripts (e.g., Arabic, Urdu, Jawi, Persian, Sindhi and Pashtu) are limited and not been thoroughly addressed in contrast to non-cursive script [4, 7, 17, 18, 19, 20]. The non-cursive script is easy to recognize as compare to cursive script due to its independent occurrences of letters. Mostly, researchers develop their own dataset for implementation and testing that is not available publically for comparison of the state of the arts techniques.

Essoukri *et al*. [6] developed an Arabic relational database for Arabic OCR systems named ARABASE for Arabic Script and Evaluated their two systems. One system evaluates the recognition accuracy of printed Arabic writing using the generalized hough transform while the other system deals with handwritten Arabic script using planar hidden markov model.Al-Ma'adeed *et al*. [1] presented Database For off-line Arabic Handwriting (AHDB) and collected samples from 100 writers. CENPARMI database developed [2] that have the most popular words used in writing bank legal amounts. It consists of 3,000 handwritten cheque images. It consists of labelled 2,499 legal amount images, 29,498 sub-word images, and 15,175 digit images.

There are some struggle have been done for printed Urdu character recognition dataset [5, 22] and very limited work presented for handwritten isolated digits and letters, numeral strings, bank related words, special symbols, and dates [21]. Whereas up to our knowledge, no work has been done for the development of handwritten Urdu dataset having Urdu text in Nasta'liq font. Therefore, we proposed a standard cursive dataset for handwritten Urdu language that would be made available online freely for researchers as a benchmark dataset. Figure 1 shows

Urdu handwritten sentence which contains Urdu words.



کیا پاکستانی سیاست کی لغت میں لیڈر اس شخص کو گردانا جاتا ہے

Figure 1. Urdu handwritten sentence.

This paper is organized in various sections. Section 2 summarizes different challenges that exist in Urdu language. These challenges envision research ideas especially in Urdu language. The detailed description about acquisition of dataset has been explained in section 3. The acquired data needs pre-processing which has been thoroughly discussed in section 4. The dataset labels have been discussed in section 5. The applicability of UCOM dataset has been explored in section 6. Section 7 represents different scenarios of possible applications. The conclusions have been depicted in section 8.

2. Challenges in Urdu Language

Urdu is a complete language with its own writing script, that is a mixture of Arabic and Persian script. Urdu script contains 58 characters and more than 10 commonly used fonts i.e., Nasta'liq, Naskh, Noori Nasakh, Noori Nasta'liq, Koofi, etc. Nasta'liq is a special calligraphic way of writing and is the most popular fonts. Unlike Arabic, Nasta'liq is written diagonally from right-to-left and top-to-bottom as depicted in Figure 2.

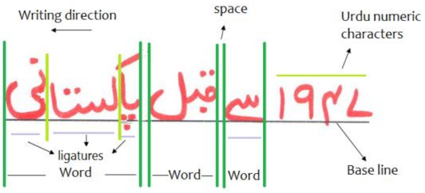


Figure 2. Representation of Urdu text with ligatures and Urdu digits.

An interesting fact about Urdu is that its number system is written from left to right. Thus, it has both the properties of left to right and right to left writing systems. It does not have a baseline rather the text is centre justified [11, 14]. It is context sensitive language and are written in the form of ligatures that may comprise a single or many different characters to form a word [12, 13]. Most of the characters have different shapes depending on their position in the ligature e.g. a letter may appear differently depending on its position as an isolated, middle, centre, or ending character. Whenever, non-joiner character appears at final position in a word or isolated form and it will always terminate the word as end character of word and maintain its full shape as shown in Figure 3. The joiner character occur at final, isolated, initial position or at middle position and it may completely change its shape at middle position and initial position as depicted in Figure 4. Thus, a character in Naskh has four basic shapes depending upon character position whereas in case of Nasta'liq, it is observed that these different shapes vary with both side neighbouring characters as well as position of that character in a ligature as shown in Figure 5-b.

Urdu also uses punctuation marks to separate sentences and leave white space between ligatures and words for separation. Furthermore, characters may overlap with each other and are very rich in diacritical marks; i.e., Urdu contains 22 diacritical marks and these additional diacritical marks associated with ligature represent short vowels or other sounds. Some diacritical marks are compulsory whereas some diacritical marks are optional and only added to help in pronunciation. Based on the above complexity analysis of Nasta'liq script, machine based recognition of Nasta'liq is much more complex as compare to Naskh writing style.

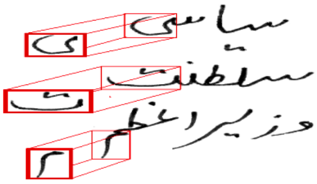
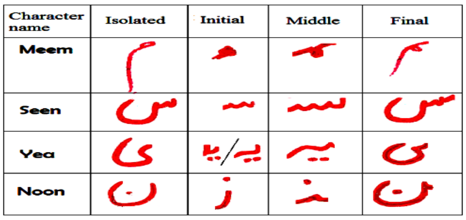
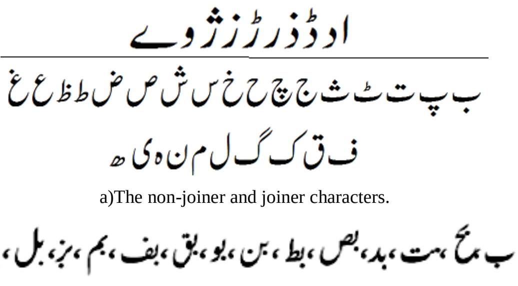


Figure 3. Urdu words with last character appeared in actual shape.



Character name	Isolated	Initial	Middle	Final
Meem	م	مـ	ـمـ	ـم
Seen	س	سـ	ـسـ	ـس
Yea	ی	یـ/یـ	ـیـ	ـی
Noon	ن	نـ	ـنـ	ـن

Figure 4. Urdu characters with its isolated, initial, middle and final shapes that may occur in an Urdu word.


ا د ڈ ذ ر ڑ ز ژ و ے
ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ
ف ق ک گ ل م ن ہ ی ھ

a)The non-joiner and joiner characters.

بمئے ، بت ، بد ، بص ، بط ، بن ، بو ، بق ، بف ، بم ، بز ، بل ، با ، بي

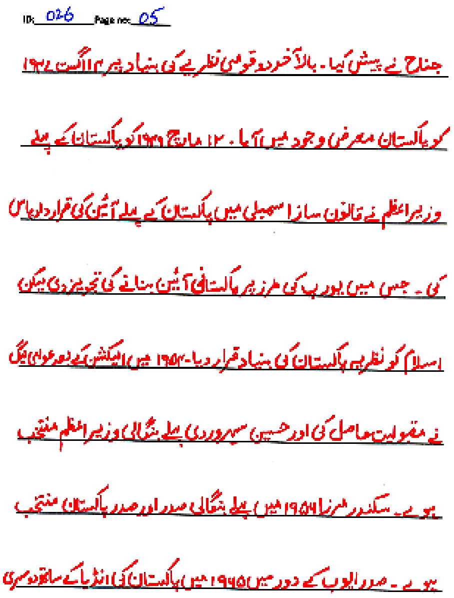
b) Shape depending on the position of occurrence of character in a word/sub-word.

Figure 5. Characters and shapes of characters in Urdu.

3. Database Generation

In order to preform evaluation and development of Urdu handwritten OCR, we present comprehensive handwritten Urdu dataset for Nasta'liq writing style. The dataset is publically available and can be obtained freely by sending email to corresponding author. The data was taken in red colour to ensure integrity of taken samples. The integrity of data refers to maintain its pixel value during noise removal. The noise is usually in black color and it is easy to determine it if text has other than black color. The dataset was acquired at COMSATS Institute of Information Technology (CIIT), Abbotabad, Pakistan. Each individual was trained before dataset gathering and are asked to write the provided text in natural way. The initial data were gathered from 100 native writers (both male and female). Each individual are asked to write 6 pages whereas each page contain 8 text lines.Same content pages were provided to each individual as represented in Figure 6.

Thus, each individual wrote 48 text lines. Table 1 shows the complete depiction of gathered data. We provided 6 blank pages to each individual with their identification and page numbers and asked them to write given Urdu text as represented in Figure 6. Some text are taken on given baselines while other are collected without it. The skew is corrected, noise and baselines were removed by GNU image manipulation tool as represented in Figure 7-a and 7-b. Each individual is asked to write given text. The UCOM dataset set is obtained with identification of each individual so that it can also be used for writer identification in addition to handwritten character recognition. After dataset collection, the text pages were scanned on a flatbed scanner with 300 dpi and tagged with the writer number. Furthermore, we performed skew correction and segmented the pages into the text lines. The dataset consist of 62,000 words written by 100 individuals with total of 6,400 lines. There are approximately 1240 words written by a single author and the total number of words written by all 100 authors is 62,000. There are 53,248 number of characters exist in UCOM dataset to date. There is a plan to increase number of authors to 300 later as a second phase of data collection with a bit different variations given to individuals.



ID: 026 Page no: 05

جناح نے پیش کیا۔ بالآخر دو قومی نظریے کی بنیاد پر ۱۴ اگست ۱۹۴۷

کو پاکستان معرض وجود میں آیا۔ ۱۲ مارچ ۱۹۴۹ کو پاکستان کے پہلے

وزیر اعظم نے قانون ساز اسمبلی میں پاکستان کے پہلے آئین کی قرارداد پاس

کی۔ جس میں یورپ کی طرز پر پاکستانی آئین بنانے کی تجویز دی لیکن

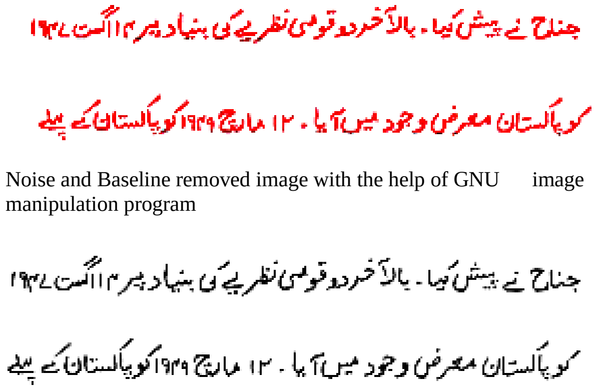
اسلام کو نظریہ پاکستان کی بنیاد قرار دیا۔ ۱۹۵۴ میں الیکشن کے بعد عوامی لیگ

نے مقبولیت حاصل کی اور حسین سہروردی پہلے بنگالی وزیر اعظم منتخب

ہوئے۔ سکندر مرزا ۱۹۵۶ میں پہلے بنگالی صدر اور صدر پاکستان منتخب

ہوئے۔ صدر ایوب کے دور میں ۱۹۶۵ میں پاکستان کی انڈیا کے ساتھ دوسری

Figure 6. Handwritten Urdu text sample on A4 Size paper.



جناح نے پیش کیا۔ بالآخر دو قومی نظریے کی بنیاد پر ۱۴ اگست ۱۹۴۷

کو پاکستان معرض وجود میں آیا۔ ۱۲ مارچ ۱۹۴۹ کو پاکستان کے پہلے

a) Noise and Baseline removed image with the help of GNU image manipulation program

جناح نے پیش کیا۔ بالآخر دو قومی نظریے کی بنیاد پر ۱۴ اگست ۱۹۴۷

کو پاکستان معرض وجود میں آیا۔ ۱۲ مارچ ۱۹۴۹ کو پاکستان کے پہلے

b) Grayscaled segmented textlines.

Figure 7. Noise and baselines were removed by GNU image manipulation tool.

Table 1. Statistics of UCOM dataset.

UCOM Details	Statistics
Number. of text lines per page	8 Text Lines
Number. of pages written by a writer	6 Pages
Approx. number of words per page	104
Approx. written number of words by a writer	620
Approx.number of characters per page (assume 4 characters per word)	416
Approx. number of characters by a writer	6656
Total number of characters	53248
Total number of words	62000
Total text lines	6400
Total number of pages	600
Number of writers	100

In machine learning, ground truth provides standard to learn patterns in supervised mode. It is considered as backbone for supervised learning tasks. To evaluate the performance of recognition system, ground truth needs to be determined accurately and labelled correctly. In Figure 8, the Urdu handwritten word is represented by its pixel values and these pixels are highlighted in red box to depict the feature values which may be provided to the classifier with its corresponding ground truth.

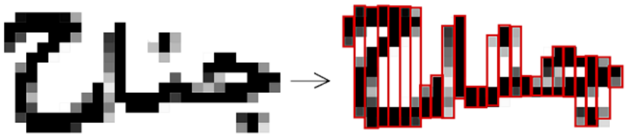


Figure 8. Groundtruth identification.

The utf8 encoding of every non-Latin character is determined. Like other non-Latin languages, Urdu has also its utf8 encoding characters regardless to its position. In reading Urdu, character positions are vital in determination of word for reader. As specified in Figure 4, suppose the character meem occurred at different positions. There is no need to declare different utf8 codes with respect to each position of meem. So, we have only one code of meem for its every position, graphically it may have different shapes but its utf8 code would be same. We labelled every character with its four possibilities of occurrences in Urdu word. The character appeared in isolation is labelled as meem_iso, at initial position as meem_i, at middle position meem_m and at final position as meem_f respectively in ground truth information file for each line.

4. Pre-Processing of UCOM Database

The pre-processing steps applied on UCOM dataset includes removal of baseline and noise, grey scale conversion and skew correction, and text lines segmentation as shown in Figure 6.

4.1. Removal of Baseline and Noise

After taking data from different individuals, the text page was scanned. Noise may occur during the scanning process and need to be removed. Since, the text is in red colour, therefore all the other colours considered as noise, are removed. So, we have only our text lines image. There is built-in noise associated to scanned documents in GIMP. We removed baselines and noise by GNU Image manipulation program with respect to colour information. For suppression of noisy data, we also used median filter.

4.2. Greyscale Conversion and Skew Correction

As in Urdu, it is cumbersome to learn each shape of every character in presence of variations with respect to one character. Therefore, we are dealing with pixels values to accommodate different shapes information. The original image was converted into grey scale. The baselines provided on a page a help to an author to write in straight line. For skew correction, we used variance of horizontal projection method. In this method the image is project at different angles and calculates the variance of horizontal projection. Horizontal projection is the sum of each row of the image. The horizontal projection of un-skewed image will likely have the maximum value.

4.3. Textline Segmentation

In our pre-processing step, the text lines were segmented by projection profile in horizontal direction. The detail for different methods for Arabic text segmentation has been presented in [16]. The steps of text line segmentation are given below.

1. The Greyscale image G_i is converted to Binary Image (B_i).
2. Horizontal projection Hp_i of B_i is calculated, which is the sum of each row.
3. Hp_i is scanned if a non-zero number is found that position is marked as X (line upper row no) and then continue search for zero value if a zero value is found that position is marked as Y (line lower row no). Here, X and Y indicate the text line positions.
4. The text line positioned at X and Y is cropped from Grayscale image G_i.
5. The steps 3 and 4 is repeated for the whole image.

5. Dataset Evaluation

The UCOM dataset set is obtained with identification of each individual so that it can also be used for writer identification in addition to handwritten character recognition. The acquired dataset is completely applicable on any type of classifier. We have used Recurrent Neural Networks (RNNs) for dataset evaluation. As learned from literature [7, 10] the basic constraint of Multi-Layer Perceptrons (MLPs) is to map input into output vector without considering the previous computations at output unit while RNNs has flexibility of tracing back previous computations. In this way history also takes a part in computations at hidden layer. The internal state of the network is retained by looped connection that makes an influence at output level. The RNNs are meant to retain the previous sequence information [8]. Figure 9 shows complete depiction of proposed system.

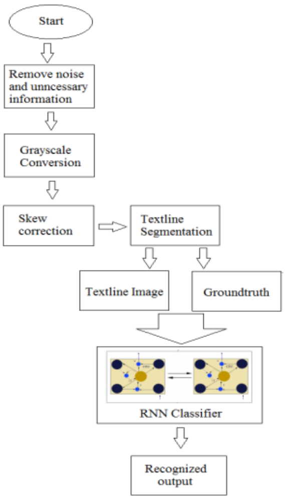


Figure 9. Proposed Recognition System

The features of an image play a crucial role in text-line recognition. The image is segmented into candidate regions and features are extracted from each candidate region as marked in Figure 10. After pre-processing, we performed segmentation of text lines. The window size of 30×1 (x-height) traverses over the given text line image while maintaining the aspect ratio to get corresponding pixel values as feature values which is given to classifier for learning as shown in Figure 10. To test the applicability of UCOM handwritten dataset, we took 50 text line images as train dataset and 20 text-lines as test dataset. The error was calculated by edit distance measurement and 0.04~0.06% error was reported on little subset of UCOM offline dataset to assess the applicability of proposed dataset with respect to given classifier. There is a plan to apply Recurrent Neural

Network classifier on extended version of UCOM dataset. The produced results are motivated enough that encourages in conducting detail experiment of given dataset.

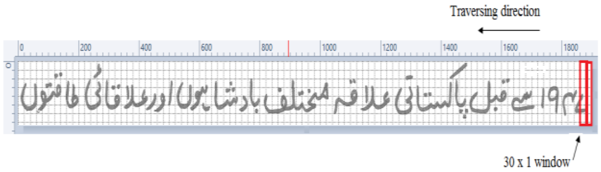


Figure 10. Normalized to x-height 30×1 window size in red colour.

6. Usage Recommendations

The UCOM offline Urdu dataset is a combination of all characters that appears at isolated, initial, middle or final position. The UCOM dataset set is stored with identification number of each individual. Initially, 100 writers are involved in preparation of UCOM offline database. The obtained text pages are segmented into text lines and tagged with respect to individual identity. The researcher will be able to access the tagged line as well as tagged pages and raw dataset. The dataset is publically available and can be obtained freely by sending email to corresponding author. Depending upon the use of database, we recommend partitioning the database into train set, validation set and test set.

Based on UCOM offline dataset, there are some typical scenarios where document analysis tasks can be performed. Our recommendations are mentioned as follows.

7. Research Scenarios

Based on UCOM offline dataset there are some typical scenarios where document analysis tasks can be performed. Our recommendations are mentioned as follows.

- The UCOM offline handwritten database will be used for construction of ligatures/sub words' database by using applying different techniques to segment lines into words and words into ligatures/sub words.
- Another variation of UCOM dataset usage is to take geometrical information of every character and maintain the information in separate table against every character. The captured information in this way can be trained and used for character recognition.
- To determine ligatures from Urdu words is another research area which can be performed using UCOM database.
- The dataset is stored and tagged with identification of user, thus can be used to perform writer identification on data sample. For, user identification can be performed dataset of 600 pages form 100 individual.
- To evaluate the potential of state of the art techniques on cursive scripts like Urdu, the proposed dataset can be used.

8. Conclusions and Future Work

In this paper, we developed a new dataset for Urdu language written in Nasta'liq writing style. Being a cursive nature, Urdu has no standard dataset available publicly. The basic motive of preparing UCOM offline database is to compile Urdu text and make it available to research community free of cost.In order to evaluate the UCOM dataset, we took 50 text line images as train dataset and 20 text-lines as test dataset. The 0.04~0.06% error was reported on subset of UCOM offline dataset. It is planned to extend this database up to 300 writers. Currently, UCOM database covers almost all characters with different variations in addition to Urdu numeric data. At first phase, data was gathered from 100 individuals. As the dataset is tagged with user identity, thus it can also be used for writer identification. Other future task includes writer identification, apply different feature extraction approaches, and apply different classifiers to recognize the text and word recognition with the help of dictionary and language modelling.

Acknowledgment

The UCOM text lines were taken from telecommunication students at bachelor level of 4th semester (Fall13) at COMSATS Institute of CIIT, Abbottabad, Pakistan.

References

[1] Al-Ma'adeed S., Elliman D., and Higgins C., "A Database for Arabic Handwritten Text Recognition Research," *in Proceeding of the 8th International Workshop on Frontiers in Handwriting Recognition*, Niagara on the Lake, pp. 485-489, 2002.

[2] Al-Ohali Y., Cheriet M., and Suen C., "Databases for Recognition of Handwritten Arabic Cheques," *The Journal of Pattern Recognition*, vol. 36, no. 1, pp. 111-121, 2003.

[3] Biadsy F., El-Sana J., and Habash N., "Online Arabic Handwriting Recognition using Hidden Markov Models," *in Proceeding of the 10th International Workshop on Frontiers of Handwriting and Recognition*, pp. 1-6, 2006.

[4] Breuel T., "The OCRopus Open Source OCR System," *in Proceeding of Document Recognition and Retrieval*, San Jose, pp. 1-15, 2008.

[5] Center for Language Engineering, www.cle.org.pk/clestore/imagecorpora.htm, Last Visited 2015.

[6] Essoukri N., Amara B., Mazhoud O., Bouzrara N., and Ellouze N., "ARABASE: A Relational Database for Arabic OCR Systems," *The International Arab Journal of Information Technology*, vol. 2, no. 4, pp.259-266, 2005.

[7] Gosselin B., "Multilayer Perceptrons Combination Applied to Handwritten Character Recognition," *in Neural Processing Letters*, vol. 3, no. 1, pp. 3-10, 1996.

[8] Graves A., "Supervised Sequence Labeling with Recurrent Neural Networks," *Studies in Computational Intelligence*, vol. 385, pp. 3-124, 2012.

[9] Javed S. and Hussain S., "Segmentation Based Urdu Nastalique OCR," *Progress in Pattern Recognition*, vol. 8259, pp. 41-49, 2013.

[10] Marti U. and Bunke H., "Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 65-90, 2002.

[11] Naz S., Hayat K., Razzak M., Anwar M., and Akbar H., "Arabic Script Based Language Character Recognition: Nasta'liq vs Naskh analysis," *in Proceeding of Computer and Information Technology World Congress*, Tunisia, pp. 1-7, 2013.

[12] Naz S., Hayat K., Razzak M., Anwar M., Madani S., and Khan S., "The Optical Character Recognition of Urdu-Like Cursive Scripts," *Progress in Pattern Recognition*, vol. 47, no. 3, pp. 1229-1248, 2014.

[13] Naz S., Hayat K., Razzak M., Anwar M., and Zar S., "Challenges in Baseline Detection of Arabic Script Based Languages," *Springer International Publishing in Intelligent Systems for Science and Information*, vol. 542, pp. 181-196, 2014.

[14] Razzak M. and Hussain S.,"Locally Baseline Detection for Online Arabic Script Based Languages Character Recognition," *International Journal of the Physical Sciences*, vol. 5, no. 7, pp. 955-959, 2010.

[15] Sabbour N. and Shafait F., "A Segmentation-Free Approach to Arabic and Urdu OCR," *in Proceeding of Document Recognition and Retrieval*, pp. 1-12,2013.

[16] Sari T. and Sellami M., "Overview of Some Algorithms of Off-Line Arabic Handwriting Segmentation," *The International Arab Journal of Information Technology*, vol. 4, no. 4, pp. 289-300, 2007.

[17] Seiler R., Schenkel M., and Eggimannn F., "Off-Line Cursive Handwriting Recognition Compared with On-Line Recognition," *in Proceeding of 13th International Conference on Pattern Recognition*, Vienna, pp. 505-509, 1996.

[18] Smith R., "An Overview of the Tesseract OCR Engine," *in Proceeding of 9th International Conference on Document Analysis and Recognition*, Parana, pp. 629-633, 2007.

[19] Rashid S., Safait F., and Breuel T., "Scanning Neural Network for Text Line," *in Proceeding of 10th International Workshop on recognition Document Analysis Systems*, Gold Coast, pp. 105-109, 2012.

[20] Tonouchi Y., "Path Evaluation and Character Classifier Training on Integrated Segmentation and Recognition of Online Handwritten Japanese Character String," *in Proceeding of 12th International Conference on Frontiers in Handwriting Recognition*, Kolkata, pp. 513-517, 2010.

[21] Sagheer M., He C., Nobile N., and Suen C., *A New Large Urdu Database for Off-Line Handwriting Recognition*, springer, 2009.

[22] Ul-Hasan A., Ahmed S., Rashid S., Shafait F., and Breuel T., "Offline Printed Urdu Nastaleeq Script Recognition with Bidirectional LSTM Networks," *in Proceeding of 12th International Conference on Document Analysis and Recognition*, Washington, pp. 1061-1065, 2013.

[23] Wang Y., Ding X., and Liu C., "MQDF Discriminative Learning Based Offline Handwritten Chinese Character Recognition," *in Proceeding of International Conference on Document Analysis and Recognition*, Beijing, pp. 1100-1104, 2011.

[24] Zafar M., Mohamad D., and Othman R., "On-line Handwritten Character Recognition: an Implementation of Counter Propagation Neural Net," *in Proceeding of IEEE International Conference on Engineering of Intelligent Systems*, pp. 232-237, 2005.

Saad Bin Ahmed is serving as Lecturer at King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia (KSAU-HS). He is completed his Master of computer sciences in intelligent systems from University of Technology, Kaiserslautern, Germany and has been served as research assistant at Image Understanding and Pattern Recognition (IUPR) research group at University of Technology, Kasierslautern, Germany. Mr. Saad had served as Lecturer at COMSATS institute of information technology, Abbottabad, Pakistan and Iqra University, Islamabad, Pakistan. He has also performed his duties as project supervisor at Allama Iqbal Open University (AIOU), Islamabad, Pakistan. His area of interests is document image analysis, medical image processing and optical character recognition. Mr. Saad is in field of image analysis since 10 years and has been involved in various pioneer research like handwritten Urdu character recognition.

Saeeda Naz an Assistant Professor by designation and Head of Computer Science Department at GGPGC No.1, Abbottabad, Higher Education Department of Government of Khyber-Pakhtunkhwa, Pakistan, since 2008. She received her BS degree from the University of Peshawar (UoP), Peshawar, Pakistan in 2006 and MS (Computer Science) degree from the COMSATS Institute of Information Technology (CIIT), Pakistan in 2012. Currently, she is doing her PhD in Computer Science from Hazara University, Department of Information Technology, Mansehra, Pakistan. She has published two book chapters and more than 20 papers in peer reviewed national and international conferences and journals. Her areas of interest are Optical Character Recognition, Pattern Recognition, Machine Learning, Medical Imaging and Natural Language Processing.

Salahuddin Swati is currently working as a lecturer in department of computer science, COMSATS institute of information Technology Abbottabad. He did his MS (Computer Science) from CIIT Abbottabad in 2013. His research interest include Image processing, Computer Vision, Machine learning, document image analysis, watermarking, video compression and pattern recognition.

Imran Razzak currently working as an assistant professor at King Saud bin Abdulaziz University for Health Sciences, has completed his Masters (2007) and doctoral studies (2011) in the field of computer sciences form International Islamic University, Islamabad (IIUI). Before joining KSAU-HS, he had been working at University of Technology (Malaysia), King Saud University (Saudi Arabia) and Air University (Pakistan). Dr. Razzak, till now is an author of one patient, three book chapters and more than forty five publications / Research findings, published in International Journals and Conferences. His area of expertise include, image processing and intelligent systems for medical and document imaging.

Arif Umar was born at district Haripur Pakistan. He obtained his MSc (Computer Science) degree from University of Peshawar, Peshawar, Pakistan and PhD (Computer Science) degree from BeiHang University (BUAA), Beijing P.R. China. His research interests include Data Mining, Machine Learning, Information Retrieval, Digital Image Processing, Computer Networks Security and Sensor Networks. He has at his credit 22 years' experience of teaching, research, planning and academic management. Currently he is working as Assistant Professor (Computer Science) at Hazara University Mansehra Pakistan.

Akbar Khan was born in Khyber Pakhtunkhwa (KP), Pakistan. He received the M.Sc. degree in Electronics from the University of Peshawar (UoP), Peshawar, Pakistan and MS in Computer Engineering from the COMSATS Institute of Information Technology (CIIT), Abbottabad, Pakistan. He was as lecturer in Electronics at Higher Education Department of Government of KPK, Pakistan from 1999 to 2008. Since 2008, he has been with the CIIT as Assistant Professor of Electrical Engineering. He is currently a Ph.D. student at Quaid-i-Azam University (QAU), Islamabad, Pakistan. His research interests include image processing and computer vision; and more recently, the pattern and character recognition.