State of the art in Nastaleeq Script Recognition

Harmohan Sharma[#1], Dharam Veer Sharma[#2]
[#1]Department of Computer Science, Multani Mal Modi College, Patiala, Punjab, India
[#2]Department of Computer Science, Punjabi University, Patiala, Punjab, India

***Abstract* —** *OCR of Nastaleeq script has gained a lot of importance during recent past owing to the requirements of preserving historic manuscripts and making such manuscripts searchable besides other applications of OCR. Nastaleeq, being a complex script, has largely remained untouched for automation till now. Whatever little work has been done so far, it has proved insufficient to fulfil the needs. Developing OCR for Urdu script based languages becomes even more complex than other languages like Latin and Chinese due to complexities of Urdu scripts, i.e. cursive nature of writing Urdu, context sensitive shapes, overlapping between ligatures, use of joiners, formation of ligatures within the words and space between the ligatures. Moreover, this paper analyzes understanding of Urdu language, characteristics of Nastaleeq script and the complexities involved in developing the Urdu OCR.*

Keywords — *Optical Character Recognition, Nastaleeq, Ligature recognition.*

I. INTRODUCTION

The procedure of converting scanned images of machine printed or handwritten text (numerals, letters, and symbols) into a computer processable format is termed as Optical Character Recognition (OCR). Although OCR system can be developed for different purposes, for different languages, an OCR system contains some necessary basic steps. Image Acquisition, Pre-processing, Segmentation, Feature extraction, Classification and Post-processing are different phases of an OCR. Image acquisition or Digitization is the process of converting printed documents into equivalent electronic form. The process of Imaging helps in the conversion of a scanned document into a corresponding electronic bitmap image. The preliminary step for the purpose of data collection is the process of digitization. However a digital camera may serve the purpose of capturing images from the printed documents, the usage of scanners is also found to be more apt and popular for the purpose. The two types of recognition systems are feasible depending on the acquisition type of the input data; they are On–line system and Off–line systems. The pre-processing phase is a set of operations that apply successive transformations on an image. It takes in a raw image, reduces noise and distortion, removes skewness and performs skeletonizing of the image thereby simplifying the processing of the rest of the stages. The operations that help to achieve this goal include thresholding, binarization, thinning, filtering, smoothening, edge detection and skew detection. These are applied to the text image, so as to get a noise-free, blur-free OCR ready image of the text region. For recognizing the cursive script, Segmentation based and Segmentation free approaches are there to deal with connected characters in a word. Segmentation of the Urdu script is a daunting task as the script has a strong cursive mode. The feature extraction stage analyzes a text segment and selects a set of features that can be used to uniquely identify the text segment. The selection of a stable and representative set of features is the heart of pattern recognition system design. Features can be classified into two categories as Local features, which are usually *geometric* (e.g. concave/convex parts, number of endpoints, branches, joints etc) and Global features, which are usually *topological* (connectivity, projection profiles, number of holes, etc) or *statistical* (invariant moments etc.). The most important decision making stage of an OCR system is the classification stage which uses the features extracted in the previous stage to classify the text segment according to preset rules. The final stage, the post-processing stage, improves recognition by refining the conclusions taken by the previous stage and recognizes words by using context. This stage plays a crucial role in producing the best possible solution and involves techniques that strongly rely on character frequencies, lexicons, and other context dependent information.

II. URDU SCRIPT : AN OVERVIEW

Urdu is a Central Indo-Aryan language of the Indo-Iranian branch, belonging to the Indo-European family of languages spoken by nearly 250 million people in India, Pakistan and other neighbouring countries. Urdu is the national language of Pakistan, and one of the 23 scheduled languages of India and also enjoys the status of being one of the official languages of five Indian states [1]. The word ‘Urdu’ comes from the Turkish word ‘ordu’ means ‘camp’ or ‘army’. It emerged in northern India about one thousand years ago out of a mixture of Hindi and other local languages with the Persian language spoken by armies and merchants from Persia. It also

included many words from Arabic and Turkish languages. Urdu borrows words from many languages which enhances its appeal as poetic language. Probably because of complex amalgam of various races and their rich cultural background, the sweetness and decency in Urdu language is unmatched and unparalleled.

III. URDU WRITING STYLE

The two well-known approaches of writing Urdu are Naskh and Nastaleeq. Of the two popular styles of writing, Nastaleeq is used by the Urdu script, whereas Naskh Style is used by both Persian and Arabic scripts. The Nastaliq script, developed by Mir Ali Tabrizi way back in the 14th century, has its roots from two styles Nash and Taliq [2] and is admired as for its calligraphic style and artistic beauty. Naskh is usually written with equal vertical depth above and below the base line. The curves are fully deep and the words generally well spaced and words are spread horizontally along the base line taking more space for writing a ligature. The Naskh style is easier for character recognition compared to Nastaleeq due to its linearity in writing, having a single base line and non-overlapping of characters in adjacent ligatures. Nastaleeq style of writing is highly cursive in nature with multiple base lines, overlapping of characters in adjacent ligatures and vertical stacking of characters within a single ligature. All this makes character recognition in Nastaleeq more challenging. The time of Mughal Dynasty in South Asia saw a surge in the use of Nastaleeq. A subtle version of this script is used for printing Urdu documents. Nastaleeq typesetting was a challenging job at the initial stage. This became easier with the computerization of the Nastaleeq script that by Mirza Ahmed Jameel in the year 1980. Mirza managed to create 20,000 Mono-type Nastaleeq ligatures used in computers [3] and named it Noori Nastaleeq. This brought about a drastic change in the printing of Urdu script. Noori Nastaleeq was used in the LaserComp Machine in early eighties by Mono-Type. The Pakistani Urdu Daily called “Daily Jung” purchase this machine at a whopping cost of 10 million. Later on different PC interfaces Inpage, OpenType Nastaleeq like Jameel Noori Nastaleeq, Alvi Nastaleeq and Faiz Lahori Nastaleeq were created and are still widely used. The basic attributes of Nastaleeq writing style is followed by all Nastaleeq originated fonts. Urdu, as mentioned, is written in the Nastaleeq style of Persian calligraphy, has 40 characters and 10 diacritical marks. The other characters/ symbols used in Urdu are 20 digits (both Roman and Urdu), 7 Punctuation Marks, 5 Honorifics and 2 Poetic Marks. Many other sources give slightly different character set.

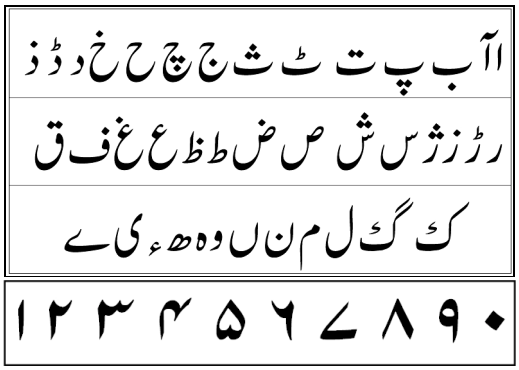
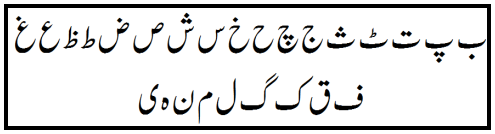


Fig. 1: Urdu character set and Numerals

(a)



(b)

Fig. 2: The isolated form of
(a) Non-joiners and (b) Joiners in Urdu.

Fig. 3: An example of the Nastaleeq script

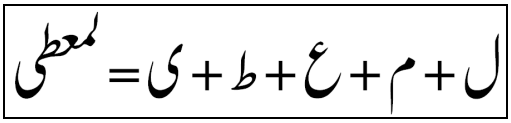


Fig. 4: Classification of ligature instead of individual characters

It is also interesting to note that the 40 main characters in Urdu script can be classified into 21 classes based on visual similarity (see Table 1). As can be seen from the table, the similar looking characters in same class can be differentiated by presence/absence of dots, diacritic symbols or a small line. For example, for class 2 the basic shape of all the five characters is identical, the only distinction is in the dots/diacritic marks present above or below the basic shape. The above property can be used for reducing the ligature classes for identification. Instead of taking 40 basic characters, we consider the basic shapes of the characters and classify the dots and diacritic marks separately. This reduces the basic classes at character level from 40 to 21.

TABLE 1: Classification of Urdu Characters

Sr. No.	Basic Shape	Shape Set	No. of Elements
1	ا	ا آ	2
2	ب	ب پ ت ٹ ث	5
3	ح	ج چ ح خ	4
4	د	د ڈ ذ	3
5	ر	ر ڑ ز ژ	4
6	س	س ش	2
7	ص	ص ض	2
8	ط	ط ظ	2
9	ع	ع غ	2
10	ف	ف	1
11	ق	ق	1
12	ک	ک گ	2
13	ل	ل	1
14	م	م	1
15	ں	ن ں	2
16	و	و	1
17	ہ	ہ	1
18	ھ	ھ	1
19	ء	ء	1
20	ی	ی	1
21	ے	ے	1

IV. Characteristics of Nastaleeq Script

This section provides a comprehensive list of characteristics of the Nastaleeq characters. The script of Urdu language is typically written in the Nastaleeq style, which has its roots in the Persian calligraphy; On the other hand, the Arabic is generally represented in the Naskh style. The Nastaleeq style emerged from a subtle merger of the 2 writing systems namely Naskh and Taleeq. It seems to be a complex script as the calligraphic touch makes it look, both cursive and context sensitive in nature. The character recognition point of view puts forth the following observations -

- Text is written right to left in both printed and handwritten forms.
- Numbers are written from left to right.
- Urdu Nastaleeq script is inherently cursive in nature.
- It is written diagonally from top right to bottom left with stacking of characters.
- Urdu characters do not discriminate between lower-case and upper-case forms.
- Urdu characters relatively change their shapes according to the contextual use, often depending upon the characters that precede or follow. In general, they tend to acquire one of the four shapes namely isolated or standalone, initial, medial and final, in the word. A character may have up to 45 different shapes. For example *bay*, which is the second letter in the Urdu alphabet has 19 different shapes for its initial form.
- All Urdu characters can be categorized into 2 sub-categories aptly named non-joiners and joiners. The joiners can be in either the isolated, initial, medial or final shape and can also merge with the succeeding character, whereas on a contrast the non-joiners acquire only the isolated and final shape and as expected do not join with the next character.
- A group of joiners and/or non-joiners joined together give rise to a ligature. A ligature can be identified as a connected component of characters. A word in Urdu is a collection of one or more ligatures. A ligature ends either with a space or with a non-joining character. The ligatures are tilted at a certain angle towards the right side. Due to this diagonal nature,the Nastaleeq consumes less horizontal space as compared to Naskh.
- There is vertical overlapping, both within ligatures, and among the ligatures.
- Though there is a strong likeness in shape amongst many letters of the Urdu alphabet, yet they clearly stand apart from each other by the presence and the positioning of the dots. The placements of the dots may either be above or below the letters.
- *Nuktas* (dots) positions may be replaced during joining process. The *nuktas* present in Urdu characters may change their positions when joined with other characters.

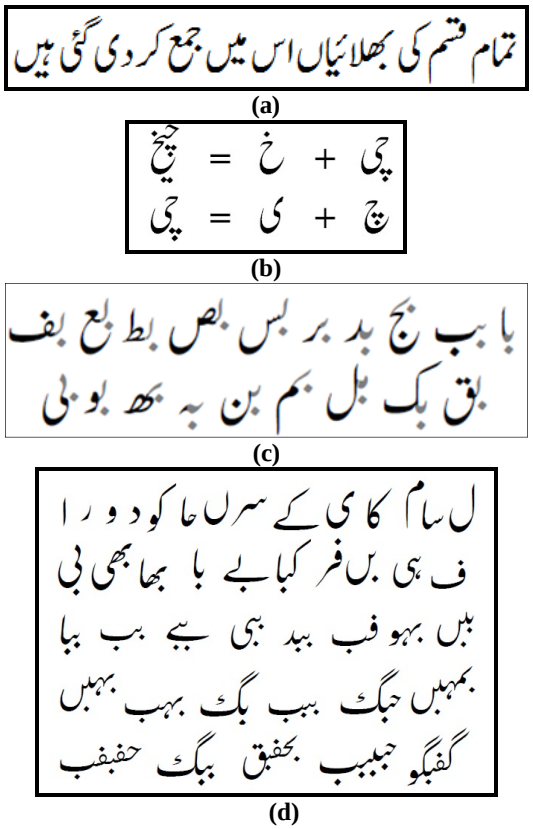


Fig. 5: **(a)** Vertical Overlap among & between ligatures, **(b)** Nukta replacement during character joining **(c)** Different shapes of character bey in initial position [4] **(d)** Ligatures (row wise) without any secondary component and with one, two, three and four secondary components.

V. Review of Literature

One of the major challenges in Urdu OCR is of finding a good database for offline (both machine printed and hand written) and online recognition. No such database is available. Several methods for recognition of other scripts like Latin, Chinese, and Arabic scripts have been proposed [5,6,7]. Among Indian scripts, some pioneering work has been done on Oriya [8], Bangla [9,10] and Devnagari [11,12] scripts. Some studies have also been reported on Gurmukhi [13], Telugu [14] and Tamil [15] scripts. However, to the best of our awareness, no major work has been done on Urdu OCR.

Naz et al. [16] endeavour to provide survey on OCR work on Urdu-like cursive scripts with special emphasis on segmentation-based and segmentation-free Nastaleeq and Naskh scripts. The survey encompasses work related to the recognition of printed, handwritten, online character and on the various techniques being applied for preprocessing, segmentation, feature extraction, classification, and recognition. The study of the current recognition methods for Urdu OCR acknowledges that most of the research focused on isolated characters or ligature based recognition.

TABLE 2: Domain-wise corpus size distribution

Domains Sub domains	Raw Corpora	
	Size	Distinct words
C1. Sports / Games		
C1.1. Sports (special events)	1666304	23118
C2. News		
C2.1. Local and international affairs C2.2. Editorials and opinions	8957259	67365
C3. Finance		
C3.1. Business, domestic and foreign market	1162019	17024
C4. Culture/Entertainment		
C4.1. Music, theatre, exhibitions, review articles on literature C4.2. Travel / tourism	3845117	59214
C5. Consumer Information		
C5.1. Health C5.2. Popular science C5.3. Consumer technology	1980723	34151
C6. Personal communications		
C6.1. Emails, online discussions, editorials, e-zines	1685424	30469
Total	**19296846**	**104341**

Farah Adeeba [17] extracted a wordlist from 19.3 million corpus gathered from a wide range of domains as mentioned in the table 2, and also mentioned a list of 2430 most frequently used Urdu Ligatures with their respective frequency.

In [18], at CENPARMI in Montreal, Canada, authors have created an Urdu handwriting database of 109,588 images from 343 various writers around various regions of the world for the Urdu off-line handwriting recognition which includes 14,890 samples of 44 isolated characters, 60,329 samples of isolated digits, 1705 samples of five special symbols,19,432 samples of 57 financial related Urdu words, 12,914 samples of numeral strings with/without decimal points and 318 samples of Urdu dates in different patterns. Experiments are conducted on Urdu digits recognition with an accuracy of 98.61%. In image pre-processing, noise removal, normalization (grayscale and size) and binarization, gradient feature extraction based on Robert's operator on each normalized image and SVM using a Radial RBF kernel function is applied for classification and followed by an in-depth error analysis is given for the recognition results.

Razzak et al. [19] presented the issues of pre-processing steps for handwritten online character recognition from both online and offline side to reduce the variation and proposed various processing techniques like de-hooking, smoothing and interpolation, slant estimation, stroke mapping, baseline estimation and skew correction in the preprocessing steps to improve and normalize the Urdu handwritten stroke for recognition system.

Fuzzy logic rules are practiced to essence the strokes and then unite the strokes to figure ligatures using some linguistic rules. 89.2% recognition rate is claimed in the presented approach.

Pal and Sarkar [20] have suggested a system for recognition of text printed in Urdu, a relatively difficult script because of large set of characters and many similar shaped characters, in which the document image is captured using a flatbed scanner and passed through skew correction using Hough transform, line segmentation and character segmentation modules. Segmentation of character is completed by a combination of component labeling and vertical projection profile methods. A combination of techniques has been used for the recognition of isolated or individual characters. Topological features, the conceptual features acquired from a water reservoir, contour based features, are the techniques, to name a few. A prototype of the system has been tested on printed Urdu characters of Naskh and Nastaleeq types. The system recognizes basic characters only and does not deal with recognition of compound characters. The system's identification rate for individual text lines is 98.3% and character segmentation accuracy is 96.6%. The segmentation errors largely exist because of touching characters. The recognition rate for the basic characters and numerals is 97.8%. The technique proves to be a failure when the target data is composed compound characters of variating sizes and fonts, whereas it has proved to be a boon when dealing with isolated characters and numerals.

Aamir Wali et al. [21] presented a paper in which they have discussed the variety of shapes of Urdu characters. The sole objective of this paper is to spot and recognize characters on the basis of their shape and on the contextual placement of these characters in ligatures. Since it is not possible to study all possible ligatures, they have limited their study up to 4 character ligatures. The study mentions 474 shapes of different ligatures based on the classification of Urdu characters.

All the earlier OCR's created for Noori Nastaleeq were designed and implemented for a fixed font size. But this restriction had to be surpassed as all Urdu magazines, newspapers, Urdu dailies and books written in the Noori Nastaleeq script are font size independent with a varying range of font size used. In [22], Akram et al. proposed a technique to overcome this hurdle, so as to make the OCR font-size independent. The working of the technique is based on extracting the outline frame of the ligatures. In stage one, splines are used to extract the outline of the input ligatures. This is followed by the application of the scaling factor in accordance with the font size so as to set to the size for which the OCR is trained. Once this is achieved, the scaled outline is converted into the image. This technique has been tried and tested on manually generated Urdu single character ligatures and has resulted in an accuracy of 98% while on the other hand, 96% accuracy has been achieved for the data scanned from printed data such as magazines and books.

A template matching approach [23,24], which relies on the cross-correlation is reported which maintains a file of the character shapes of the Nastaleeq font. The concept of cross-correlation is used to match each of the character shape in the font file, line-by-line, with the shapes identified in the text image. Concurrently, the system writes the character codes into a text file in the sequence in which the characters are encountered. The same authors constructed a finite state Nastaleeq text recognizer [25] for reading and recognizing each of the character shapes of Nastaleeq in the segmented text image.

A hybrid approach was adopted for breaking up the text by the line segmentation implemented by the top-down technique, while the bottom-up segmentation furnished the contributing ligatures. While presenting this approach, Lehal [26] faced quite a few hindrances like the horizontal overlapping of ligatures, diacritic issues, merged ligatures or even broken lines. The count of associated primary and secondary components is also analyzed. 42,441 connected components from 45 Urdu images are considered and it is found that 55.5% of the components are primary and remaining are secondary components. The system is tested to classify the primary and secondary components correctly with 99.02% accuracy.

The recognition set of ligatures is componentized into different classes namely primary and secondary connected components using connected component analysis procedure. Though the individual count of the ligatures covered stands at 4657 (90% coverage), which is substantially reduced to 2212 post segmentation, thereby increasing the percentile of data coverage to about 99%. The split-up of this count is 2190 Primary Connected Components and 22 Secondary Connected Components. Urdu script also has 41 isolated characters that are further classified as 21 primary and secondary connected components. It is concluded thus that classes for recognition of Urdu Text stands at 2233 inclusive of 2211 primary and 22 secondary components [27].

Further, Lehal [28] proposed a system to recognize 9262 ligatures formed from 2190 primary and 17 secondary components. An assortment of combinations of DCT, Gabor filters and Zoning based features along with kNN, HMM and SVM (polynomial) classifiers have been experimented and attain a recognition accuracy of 98% on pre-segmented ligatures. The Unicode string of the ligature is constructed once the primary and secondary components are identified. For the same, a Binary Search Tree from the code book file is made and all the nodes in the BST represent the primary component codes. All the associated secondary components are managed in a linked list.

Initially the search starts in the nodes of the primary component in the BST followed by the search in the secondary level linked list attached to that node, to derive the Unicode String.

Using ligatures as the basic units of recognition, Shabbir and Siddiqi [29] proposed a recognition technique that was segmentation free and size invariant. The ligatures were extracted from the Nastaleeq font used on Urdu words. A connected component labeling method was implemented on binarized document images of Urdu text to extract the ligatures amounting to about 250 ligature clusters. The Ligatures thus extracted are further categorized into primary ligatures and diacritics, being recognized using right-to-left HMMs. The information helps the Diacritics to be grouped with specific ligatures whilst the detailed dictionary validation helps in the ligature recognition. Ultimately, the Unicode string of the word is then compiled on to a text file.

Husain [30] presented a paper for the off-line recognition of cursive Urdu text written in Noori Nastaleeq script. Here, ligature based identification has been adopted instead of character based identification. The system is trained for a small number of ligatures about 200 selected ligatures. The performance of the system on images containing only the trained ligatures was 100%.

In [31], Husain et al. presented a method for recognition of online Cursive Urdu hand written Nastaleeq Script. The system is currently trained for 250 ligatures. This system minimizes the errors by using segmentation free approach. By using multiple features, authors improved number of ligatures that can be identified. 250 base ligatures and 6 secondary strokes are recognized successfully. These when combined form 864 single character, 2 character and 3 character ligatures and can recognize 50000 common words from Urdu dictionary successfully. The Recognition rate of base ligatures is 93% and of the secondary strokes is 98%.

Haider and Khan [32] presented a system for online recognition of isolated single-stroke handwritten Urdu characters. In this system the character set was initially split into four groups on the basis of number of strokes i.e. single, two, three and four-stroke characters. But the projected method work for single-stroke handwritten Urdu characters only and mainly composed of four steps - data acquisition, preprocessing, features extraction and classifier design. Some novel features were extracted and used for classification. Distinct classifiers are figured on 85 instances of character set taken from 35 individuals of unlike age groups and acquired a recognition rate of 95%, 92%, 89% and 87% for Probabilistic Neural Network (PNN), Correlation Classifier, Back Propagation Neural Networks (BPNN) using template vectors as targets and BPNN based classifier using scalar targets respectively.

Khattak et al. [33] presented a holistic approach and scale-invariant technique for recognition of frequently occurring Urdu ligatures in Nastaleeq font. The proposed methodology relies on separating the primary and secondary components of ligatures. The primary components correspond to the main body of the ligature which may have zero or more dots or diacritics as secondary components. A combination of Projection features, Concavity features and Curvature features of ligatures are extracted using right-to-left sliding windows on each ligature image and are fed to the models for training. The system trained and evaluated on a total 2,028 high frequency Urdu ligatures from a standard database achieved a recognition rate of 97.93%.

Javed et al. [34] proposed a segmentation free approach for Nastaleeq Urdu OCR in which the ligature as a whole is used instead of segmenting it into smaller units. Global transformational features originating from a non-segmented ligature was extracted out by the authors before being fed into Hidden Markov Model (HMM) recognizer. For recognizing a ligature, first, they identify its shape and then recognize it by seeing the class of feature vector which it belongs to. A total of 3655 ligatures (1282 unique ligatures) from the 5000 high frequency words in a corpus-based dictionary are tested and 3375 ligatures are accurately identified, giving an accuracy of 92%.

VI. Conclusion and Challenges Ahead

This paper revolves around the evolutionary progress of Nastaleeq script keeping in view of OCR. The paper also presents an analytical briefing on the recognition techniques including segmentation on the Nastaleeq. In spite of the diverse OCR based researches on the Nastaleeq script, there are still disconcerted issues in the recognition methodologies. The primary setback is the lack of a standard database populated with a vast range of letters, digits and words that form samples for both off-line and on-line recognition of the Nastaleeq script. As of now, the built-up recognition methods have been tested effectively on the limited databases that have been collected privately by individual researches and such databases have very limited access to the outsiders. Enriching the database with an abundance of data for sampling would add to the effective enhancement of the on going research in this field. The highly calligraphic and cursive nature of the script increases the complexity of segmentation of the words as well as numerical arrays. Most of the research till date is around individual ligatures or even ligatures with up to 4 secondary components. The research outcomes have not been able to give a generalized approach to

the cursive script. The segmentation aspect has not been delved upon in a major way. We hope the results of the efforts in Nastaleeq script recognition will be unified and will be made available globally in order to yield high performance recognition systems for this beautiful calligraphic script.

REFERENCES

[1] Gurpreet Singh Lehal, "A Word Segmentation System for Handling Space Omission Problem in Urdu Script" in the Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), the 23rd International Conference on Computational Linguistics (COLING), Beijing, pp 43–50, August 2010.

[2] M. Asad, A. S. Butt, S. Chaudhry and S. Hussain, "Rule-based Expert System for Urdu Nastaleeq justification", in the Proceedings of 8th International Multitopic Conference (INMIC 2004), pp 591–596, 2004.

[3] Prof (Dr) Syed M. Abdul Khair Kashfi, "Noori Nastaliq Revolution in Urdu Composing", Elite Publishers Limited, D -118, SITE, Karachi, Pakistan, 2008.

[4] M. G. A. Malik, C. Boitet and P. Bhattacharyya, "Analysis of Noori Nasta'leeq for Major Pakistani Languages", in the Proceedings of the 2nd Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'2010), Penang, Malaysia, pp 95-103, 2010.

[5] S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR Research and Development", in Proceedings of the IEEE, vol 80, issue 7, pp 1029-1058, 1992.

[6] G. Nagy, "Chinese Character Recognition - A twenty five years retrospective", in Proceedings of the ICPR, pp 109 - 114, 1988.

[7] Atallah Mahmoud AL-Shatnawi, Safwan AL-Salaimeh, Farah Hanna AL-Zawaideh and Khairuddin Omar, "Offline Arabic Text Recognition – An Overview", in World of Computer Science and Information Technology Journal (WCSIT), vol 1(5), pp 184-192, 2011.

[8] B. B. Chaudhuri, U. Pal and M. Mitra, "Automatic Recognition of Printed Oriya Script", Sadhana, vol 27, part 1, pp 23-34, 2002.

[9] B. B. Chaudhuri and U. Pal, "A Complete Printed Bangla OCR System", in Pattern Recognition, vol 31, pp 531-549, 1998.

[10] Md. Mahbub Alam and Dr. M. Abul Kashem, "A Complete Bangla OCR System for Printed Characters", in JCIT, vol 1, issue 01, pp 30-35, 2010.

[11] U. Pal and B. B. Chaudhuri, "Printed Devnagari Script OCR System", Vivek, vol 10, pp 12-24, 1997.

[12] Vikas J. Dongre and Vijay H. Mankar, "A Review of Research on Devnagari Character Recognition", in the International Journal of Computer Applications, vol 12(2), pp 8 -15, 2010.

[13] G S Lehal and Chandan Singh, "A Gurmukhi Script Recognition System", in Proceedings of the 15th International Conference on Pattern Recognition, vol 2, pp 557- 560, 2000.

[14] A. Negi, C. Bhagvati and B. Krishna, "An OCR System for Telugu", in the Proceedings of 6th ICDAR, pp 1110 - 1114, 2001.

[15] G. Sirmony, R Chandrasekaran and M. Chandrasekaran, "Computer Recognition of Printed Tamil Charcters", in Pattern Recognition, vol 10, issue 4, pp 243-247, 1978.

[16] Saeeda Naz, Khizar Hayat, Muhammad Imran Razzak, Muhammad Waqas Anwar, Sajjad A. Madani and Samee U. Khan, "The Optical Character Recognition of Urdu-like Cursive Scripts", in Pattern Recognition, vol. 47, Issue 3, pp 1229–1248, 2014.

[17] Farah Adeeba, "Urdu 2430 Most Frequently Used Ligatures" Center for Language Engineering Al-Khwarizmi Institute of Computer Science University of Engineering and Technology Lahore, Pakistan

http://www.cle.org.pk/software/ling_resources/UrduHighFreqLigature.htm).

[18] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile and Ching Y. Suen, "A New Large Urdu Database for Off-Line Handwriting Recognition", in Image Analysis and Processing (ICIAP 2009) vol 5716, pp 538–546, 2009.

[19] Muhammad Imran Razzak, Syed Afaq Husain, Abdulrahman A. Mirza and Abdel Belaïd, "Fuzzy Based Preprocessing using Fusion of Online and Offline trait for Online Urdu Script based languages Character Recognition", in International Journal of Innovative Computing, Information and Control, vol 8, number (5(A)), pp 3149–3161, 2012.

[20] U. Pal and A. Sarkar, "Recognition of Printed Urdu Script", in Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR), pp 1183-1187, 2003.

[21] Aamir Wali, Atif Gulzar, Ayesha Zia, Muhammad Ahmad Ghazali, Muhammad Irfan Rafiq, Muhammad Saqib Niaz, Sara Hussain, and Sheraz Bashir "Contextual Shape Analysis of Nastaleeq", CRULP Annual Student Report, pp 288-302, 2001-2002.

[22] Qurat ul Ain Akram, Sarmad Hussain and Zulfiqar Habib, "Font Size Independent OCR for Noori Nastaleeq" in the Proceedings of Graduate Colloquium on Computer Sciences, Department of Computer Science, FAST-NU Lahore, vol 1, 2010

[23] Sohail A. Sattar,Shamsul Haque, Mahmod K. Pathan and Quintin Gee, "Implementation Challenges for Nastaliq Character Recognition", in Wireless Networks, Information Processing and Systems, ser. Communications in Computer and Information Science, vol 20, Springer, Berlin, Heidelberg, pp 279-285, 2009.

[24] S. A. Sattar, "A Technique for the Design and Implementation of an OCR for Printed Nastalique Text" (Ph.D. dissertation), NED University of Engineering & Technology, Karachi, Pakistan, 2009.

[25] Sohail Abdul Sattar, Shams-ul Haque and Mahmood Khan Pathan, "A Finite State Model for Urdu Nastalique Optical Character Recognition", in International Journal of Computer Science and Network Security (IJCSNS) vol 9(9), 2009.

[26] Gurpreet Singh Lehal, "Ligature Segmentation for Urdu OCR," in 12th International Conference on Document Analysis and Recognition (ICDAR), pp 1130 -1134, 2013.

[27] Gurpreet Singh Lehal, "Choice of Recognizable Units for URDU OCR," in Proceeding of the workshop on Document Analysis and Recognition (DAR), pp 79-85, 2012.

[28] Gurpreet Singh Lehal and Ankur Rana, "Recognition of Nastalique Urdu Ligatures", in Proceedings of the 4th International Workshop on Multilingual OCR, USA, 2013.

[29] Safia Shabbir and Imran Siddiqi, "Optical Character Recognition System for Urdu Words in Nastaliq Font", in International Journal of Advanced Computer Science and Applications (IJACSA), vol 7, No. 5, pp 567-576, 2016.

[30] S. A. Husain, "A Multi-tier Holistic approach for Urdu Nastaliq Recognition", International Multitopic Conference INMIC, Karachi, 2002,

[31] S. A. Husain, Asma Sajjad and Fareeha Anwar, "Online Urdu Character Recognition System", in the IAPR Conference on Machine Vision Applications, Tokyo, Japan, pp 98-102, 2007.

[32] Ihtesham Haider and Kamran Ullah Khan, "Online Recognition of Single Stroke Handwritten Urdu Characters", in Proceedings of the 13th International Multi topic IEEE Conference (INMIC'09) , pp 1–6, 2009.

[33] Israr Uddin Khattak, Imran Siddiqi, Shehzad Khalid and Chawki Djeddi, "Recognition of Urdu Ligatures - A Holistic Approach", in 13th International Conference on Document Analysis and Recognition (ICDAR), pp 71-75, 2015.

[34] Sobia T. Javed, Sarmad Hussain, Ameera Maqbool, Samia Asloob, Sehrish Jamil and Huma Moin, "Segmentation Free Nastalique Urdu OCR", World Academy of Science, Engineering and Technology, issue 70, pp 457-462, 2010.