The Nucleus **51**, No. 3 (2014) 361-367

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

Diacritics Recognition Based Urdu Nastalique OCR System

S. Nazir* and A. Javed

Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan

(Received January 31, 2014 and accepted in revised form September 15, 2014)

Improvements and new developments in the field of Artificial Intelligence have opened new horizons in the advancement of machines that originally have limited intelligence. As compared to human brain, machines have already better computational speed and storage however there is still much room to improve the capability to acquire and process data and draw conclusions from it on its own. Optical Character Recognition (OCR) deals exclusively with printed designs and hand written text in nature. Plenty of developments have been made in OCR so far in recognition of Latin, Asian, Arabic and Western texts. As far as Urdu is concerned the work is almost non-existent when compared with the languages cited above. One of its main reasons is the use of extremely complex characters of Nastalique style in Urdu. A methodology for the recognition and processing of the diacritics of Nastalique script is presented in this research work. The proposed technique is effective in recognizing cursive texts with invariant font size of 48. A dataset of 6728 main Urdu Nastalique ligatures is used for the testing purposes which shows that this new technique has the capacity to recognize Nastalique ligatures by having an accuracy of 97.40%. The proposed research work also focuses to improve the existing base mark association process of the Urdu OCR system.

Keywords: Image Processing, Image Segmentation, Nastalique Font, Optical Character Recognition Software, Text Recognition, Urdu OCR

1. Introduction

Urdu is one of the popular languages of South Asia and national language of Pakistan [1]. Urdu is spoken by 250 million people across the globe including 100-130 million native speakers which are generally living in Pakistan and India [2]. Urdu is recognized as a literature wealthy language of sub-continent and most of the literature in sub-continent was written in Urdu in the past [3]. Urdu uses particular calligraphic model for publishing known as Nastalique while common type of Arabic and Persian is Nasakh [4].

Urdu is written in Arabic script from right to left unlike Hindi which can be written in Devanagari script. Digits in Urdu are written from left to right. Urdu has 37 standard characters with few extensions to standard characters which can make total of 41 characters when written in isolated form [5]. These extensions include alif-madaa (آ) to alif (ا), Nun-gunna (ں) to nun (ن) and choti-hey (ۃ) and dochashmi-hey (ھ) to hey (ہ). Urdu inherits the property of cursiveness from Arabic script [6]. Cursiveness means different shapes of joining characters depending on the sequence of joining [7]. It makes text recognition quite a challenging task to accomplish for Urdu Language. This property makes Urdu text recognition more difficult. Several of those characters will also be used as diacritic marks [8]. Separate diacritics will also be used in Urdu such as for example zer, zaber, pesh, madd. The formation of ligature can include characters joining on both sides of neighboring characters as well as those joining at only one side [9]. Nine out of 36 standard characters in Urdu remains separate and don't join with any character to form single ligature, thus breaks the word in to multiple ligatures [7]. They are also referred to as terminators. The remaining 26 may join from both sides [10]. Dots are extremely popular in Urdu language script and combination of one and many dots above, under or in the midst of characters are often used to differentiate between them [11]. Out of 38, 17characters have dots associated with them and occasionally diacritics such as toy (ٹ) and hamza (ء) are used to recognize characters [12]. There exist quite large variations in different characters styles as well as their measurement. The possibility of characters overlapping in horizontal direction also happens in Urdu Script [13].

2. Literature Review

Urdu is one of the most widely spoken languages and has more than 100 million speakers across the world [14]. Urdu can be written in more than one style. Nastalique script is most commonly used font style for writing Urdu documents [15]. It is highly cursive and context sensitive in nature [16]. Some of its characteristics are:

- It starts from the top right and ends at bottom left corner.
- The joins are formed by cups like shape.
- Characters in different ligature are overlapping.
- Characters shapes are context dependent.

* Corresponding author : saima_nazir_91@yahoo.com

The shape of the letters depends on the neighboring letter and its position [17]. A letter in Urdu language has different shapes depending on its position. There are four positions of a letter i.e. Initial, Medial, Final and Isolated [3]. The mark placement and alignment justification mechanism are quite intricate in Urdu script. Characters can be classified on the bases of similarity in their base shape. They differ from each other on the diacritical's marks associated with them [18].

Javed et al. [19] proposed the idea of using either Segmentation based approach or Segmentation free approach for Urdu Nastalique script recognition. Segmentation free approach is considered better for Nastalique script recognition due to highly cursive nature. Their methodology entails extraction of global transformation features followed by Hidden Markov Model (HMM) recognizer. The process of recognition is divided into Training and Recognition.

Vivek et al. [20] characterize the 36 Urdu characters based on the different features such as number of dots, dots position and placement, height width ratio etc. Pal et al. [21] have implemented the combination of vertical projection method and the component labeling technique for character segmentation. Zarah et al. [22] used template matching for the recognition of Urdu script. Diacritics were recognized first and then the main body was recognized using the visible features recognition strategy. Frinken et al. [23] used the vertical projection method for the character segmentation and these segmented characters are refined until they attained satisfactory performance.

Durrani et al. [24] used a segmentation free approach to recognize the ligatures. They extracted the statistical features of main body. These statistical features have been fed into multi-tier neural network model for recognizing the main body of characters. Their Algorithm provides quite encouraging results on test samples. Hussain et al. [25] presented 6 marks and 10 digits for Urdu publications which is a part of Urdu Zabta Takhti 1.01, the declared national standard for Urdu by Government of Pakistan. Ahmad et al. [26] proposed a novel and robust recognition approach for printed scripts of Urdu Nastalique font without considering lexicon. Shamsher et al. [27] used forward feed Neural Network to recognize Urdu OCR script.

Some researchers proposed solutions for isolated characters only like in [28-31]. Iqbal et al. [30] proposed a system for Urdu OCR by processing Urdu Nastalique font and generates the recognized output in Roman Urdu.

3. Methodology

The image used for further processing is a scanned image of printed Urdu text. It is scanned with 50 DPI. The overall work flow of the system is given in the Figure 1.

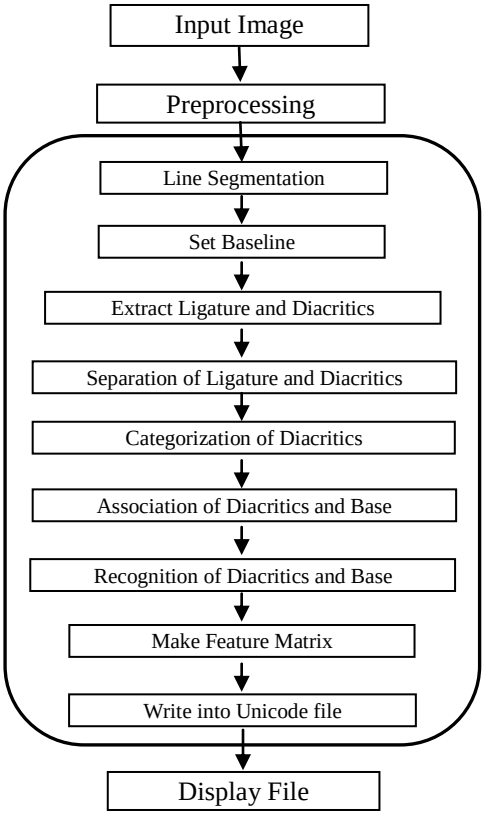


Figure 1. Flow chart of proposed methodology.

The scanned image has been fed into our proposed algorithm. It enhances the image and resolves the issue of breaking components. Line segmentation is done using horizontal projection method followed by baseline identification, which is used to segment the characters into ligature and diacritics. Ligature and diacritics are associated via new proposed technique which utilizes vertical projection method. Co-relation method is used for character recognition and feature matrix is generated as output.

3.1 Image Acquisition

Image acquisition has been attained by using Scanner which can be used to scan Urdu document page at a resolution of 200 dpi as shown in Figure 2. In our case document pages have fixed size text and can be proved that all the letter and diacritics are found in the variety of context. These test images composed of both single and multiple lines. Image acquisition is an arduous task since facets like noise, orientation and other factors may influence the acquisition process.

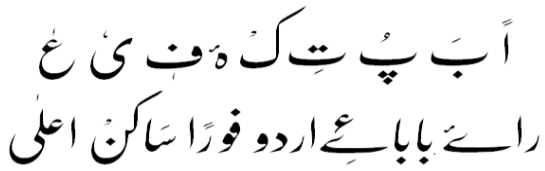


Figure 2. Acquired image.

3.2 Preprocessing

The acquired image is often corrupted by various degradations so some sort of preprocessing is required to enhance image to such an extent so that actual processing can be performed easily [32]. Preprocessing involves number of practices which can be applied depending on the recognition needs and nature of resource images. A few of the steps involved in preprocessing are however frequent in a lot of the recognition processes.

Document image binarization is performed in the preprocessing stage of different document image processing related applications such as optical character recognition (OCR) and document image retrieval. It converts a gray-scale document image into a binary document image by calculating the threshold level of gray scale image. RGB image is transformed into binary image on the basis of fixed threshold value. The negative of binary image is produced such that background pixels are represented as black and object pixels are represented as white. The degradation of breaking components of characters has been resolved via applying Dilation operator on this negative image as shown in Figure 3.

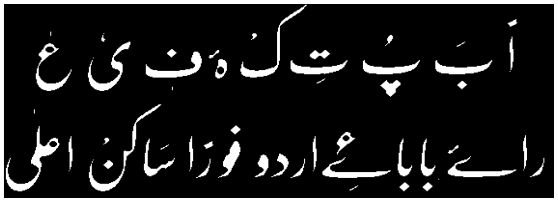


Figure 3. Image after preprocessing.

3.3 Line Segmentation

Vertical histogram is computed for the segmentation of image into lines. Lines are separated by the white spaces in the image.

Consider the image $\mathrm{Im}(i,j)$ of size $R \times C$, the horizontal projection of all the foreground pixels can be given as

$$\mathrm{Ph}(i) = \sum_{j=1}^{n} \mathrm{Im}(i,j)\ 1 \le i \le m \qquad (1)$$

3.4 Horizontal Histogram

Horizontal histogram show here a zero band of black pixel value and text line is separated from there. This is shown in Figure 4.

Threshold value is also determined between the two lines by depending on their heights. If the number of rows between two lines is less than this threshold value, they are not considered as a separate line.

Figure 4. Histogram of horizontal projection.

3.5 Baseline Identification

As stated above, Nastalique is written diagonally, it starts from the top right and ends at bottom left corner. The diagonal writing style causes the first and last characters to lie on the top left corner and the baseline respectively. The row containing maximum pixels is marked as baseline after line segmentation phases shown in Figure 5. According to the Nastalique style writing rules, the baseline should lie below the middle row of the line. Any candidate for baseline in the upper half should be ignored. Diacritic can lie above or below the baseline.

Figure 5. Baseline identification using horizontal projection information.

3.6 Line Segmentation in Ligature and Diacritics

As in Nastalique style, the ligature overlaps in both vertical and horizontal projection. Therefore line segmentation based on vertical histogram is not conceivable. Connected component method is used for ligature separation. A component is formed by checking the 8-neighbors of each pixel and then adding into it. The connected components are shown in Figure 6.

Figure 6. Connected components.

Ligature normally lies on the baseline. So all the components that lies on the baseline are considered as ligature. Threshold is set on the basis of the size of ligature versus diacritics to isolate the ligature from their diacritics. In Figure 7 diacritics and ligatures are shown in white grey colors respectively. Baseline identification is used to categories the diacritics into two categories. They are:

Diacritics above baseline (e.g. Arabic Fatha, Arabic Superscript Alef etc)

Diacritics below baseline (e.g. Arabic Kasra, Arabic Subscript Alef etc)

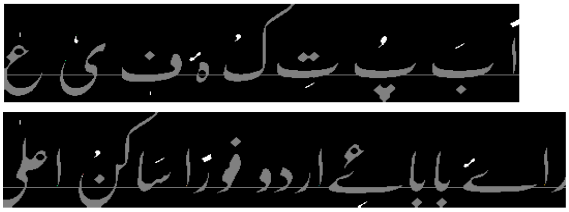


Figure 7. Line segmentation into ligatures and diacritics.

Segmentation is achieved via thinning process which is quite efficient to locate the structural relationship of the characters followed by segmenting them on the basis of the junction points. Thinned body of a sample images is displayed in Figure 8. Segmentation is based on the method of single stroke cutting on the junction point. The thinned image is traversed form left to right. End point of last letter is taken as initial node. The traversal of characters after thinning operation is accomplished by starting from the initial point of character. In Figure 9 start, end and break point is encircled in green, red and pink color respectively. Whenever a black pixel is encountered with more than one neighboring pixels, the character is segmented into two parts.

Figure 8. Sample Image and its Thinned body.

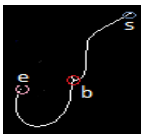


Figure 9. Start, End and Break Points.

3.7 Diacritics and Ligature Association

According to a technique implemented in Hussainet.al. [33] for base mark association, centroid is calculated of each ligature and mark, and then this calculated center-to-center distance is used to associate the main body of the ligature with the mark present above and below that ligature. We can't use this technique for ligature diacritics association because of the position of the diacritics placement on the letters. For example when we place "Arabic Damma" on Alef, It is positioned such that the main Body "Alef" does not lie under the center of "Arabic Damma". It is presented in Figure 10.In this proposed methodology we have implemented a new technique. According to the proposed technique the start and end point of the diacritics is calculated followed by diacritic association with the base ligature that lies between this start and end point using the vertical projection method as shown in Figure 11.This proposed method works reasonably well for all the diacritics base association.

Figure 10. Centroid to Centroid relationship between diacritic and base.

Figure 11. Vertical Projection of Starting and Ending point of diacritics to the base.

3.8 Recognition

Character recognition is achieved via using correlation method. It is the typical method used in pattern matching. It deals with the two dimensional information of the image to identify the character image.

3.9 Feature Matrix

The feature vector contains the code of the base, the mark and the diacritics associated with them. We use the following code for diacritics in feature matrix.

4. Experimental Results

A total of 4560 ligatures are used as the data set and it is also confirmed that all the letters and diacritics are used in the variety of context. These test images are comprised of both single line and multiple lines.

OCR accuracy can be measured as the ratio of correctly recognized characters to the total number of character samples tested as shown in equation 2.

$$\text{OCR Accuracy} = \frac{\text{No. of correctly recognized characters}}{\text{Total No. of characters tested}} \times 100 \quad (2)$$

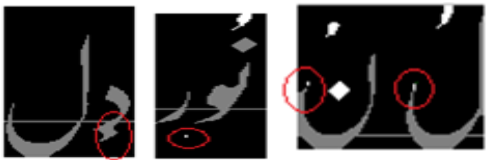
To illustrate the accuracy of URDU OCR algorithm, the performance is tested using many pages scanned at 50 dpi. These pages are published using Noori

Nastalique font with the point size 48. The line segmentation process is tested by running the algorithm on whole data set. All the lines are correctly segmented with 100% accuracy. These segmented lines are used for the baseline identification. The baseline is also identified with 100% accuracy. The process of Ligature and diacritics extraction is also tested on the sample data set. This process accomplishes 99% accuracy. It is also confirmed that all the Urdu letters and diacritics are tested in the variety of context. The diacritics are correctly categorized in two categories i.e. diacritics above baseline and diacritics below baseline. The technique implemented for diacritics and ligature association provides 100% accurate results and the recognition method gives 98% accurate result. All the experiments are performed using MATLAB as software tool.

5. Analysis and Discussion

Testing and future analysis shows the following type of errors in OCR preprocessing which needs further improvement.

- Due to noise or placement, the diacritic connects with the base and is thus not separable. The example in Figure 12a depicts the associated diacritical mark 'Arabic Kasra' which is colored as base ligature when attached.
- Sometimes extra noise is introduced which is miss-classified as a diacritic. This is shown in Figure 12b.
- Poor printing quality can also cause main body to break into fine junction and introduce false diacritics as shown 'Laam' and 'Noon'. The starting portion is identified as 'Laam' and 'Noon' and the last point is identified as 'Arabic Superscript Alef'. This is shown in Figure 12c.



(a) **(b)** **(c)**

Figure 12. Error Scenarios

5.1. Class-wise Diacritics Analysis

Urdu diacritic marks are divided into four groups.

1. Arabic Fathatan, Arabic Kasratan, Arabic Fatha and Arabic Kasra.
2. Arabic Superscript Alef and Arabic Subscript Alef.
3. Arabic Dammatan, Arabic Damma, Arabic Sakun, Arabic Noon Gunna and Arabic Inverted Damma.
4. Arabic Madda above, Arabic Hamza above and Arabic Shadd.

Table 1. Diacritics codes.

Main Body	Description	Unicode
ً	Arabic Fathatan (Do Zabar)	U+064B
ٌ	Arabic Dammatan (Do Pesh)	U+064C
ٍ	Arabic Kasratan (Do Zair)	U+064D
َ	Arabic Fatha (Zabar)	U+064E
ُ	Arabic Damma (Pesh)	U+064F
ِ	Arabic Kasra (Zair)	U+0650
ّ	Arabic Shadd (Thasdid)	U+0651
ْ	Arabic Sukun (Jazm)	U+0652
ٓ	Arabic Madda Above (Maada)	U+0653
ٔ	Arabic Hamza Above (Hamza)	U+0654
ٖ	Arabic Subscript Alef (Khari Zair)	U+0656
ٗ	Arabic Inverted Damma (Ulti Pesh)	U+0657
٘	Arabic Mark Noon Gunna (Noon Gunna)	U+0658
ٰ	Arabic Superscript Alef (Khari Zabar)	U+0670

Table 2. Class wise diacritics analysis

Class No.	Occurrence	Recognized	Accuracy
1	1086	1020	94
2	1937	1937	100
3	2728	2690	98
4	977	977	100

The class wise analysis is shown in Table 2 and the comparative analyses with the existing implemented Urdu OCR techniques have been conducted and results are represented in Table 3.

6. Conclusion and Future Work

The Nastalique style of Arabic to write Urdu is highly cursive in nature. It is complex script because of its diagonal nature and context sensitivity. The diacritics placement, in addition to other factors makes the preprocessing very challenging. This research work addresses the various factors in processing and recognition of the diacritics. This system achieves 98%

Table 3. Comparative analysis of proposed and existing techniques.

Technique	Preprocessing	Line Segmentation	Baseline Identification	Ligature and Diacritic segmentation	Diacritics and Ligature association	Recognition	Post-processing
Wali [3]	No	Yes	No	No	No	Yes	No
Razzak [10]	No	Yes	No	No	No	No	No
Pathan [14]	Yes	Yes	No	No	No	Yes	No
Tariq [29]	Yes	Yes	No	No	No	Yes	Yes
Iqbal [30]	Yes	Yes	No	No	No	Yes	Yes
Nawaz [31]	No	Yes	No	No	No	Yes	No
Javed [33]	Yes	Yes	Yes	No	No	Yes	No
Proposed Technique	Yes	Yes	Yes	Yes	Yes	Yes	Yes

accuracy. The overall accuracy can be increased by providing the larger training data to the system, adding language specific rules and applying more sophisticated statistical techniques.

Optical character recognition is an open and challenging field for researchers to implement new ideas. The basic challenge is font dependency. There exist lots of rooms for researchers to propose effective and efficient solutions related to these problems.

As a future direction pre-processor engine can be made to aid the recognition method, with the help of some advance techniques of image processing that can increase the recognition accuracy. Size independent module can be added to enable this application for variety of font sizes.

References

[1] A.Ul-Hasan, S.B. Ahmed, F.Rashid, F.Shafait and T.M.Breuel, 12th International Conference on IEEE 7 (2013) pp. 1061-1065.

[2] N. Fareen, M.A.Khan and A.Durrani, Survey of Urdu OCR: an Offline Approach, Proceedings of the Conference on Language & Technology Lahore, Pakistan, 9-10 November (2012) pp. 67-72.

[3] A. Wali and S. Hussain, Context Sensitive Shape-Substitution in Nastaliq Writing System: Analysis and Formulation, Proceedings of 2nd International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering (2006), Bridgeport, USA, 4-14 December (2006) pp. 53-58.

[4] S. Naz, K. Hayat, M.I. Razzak, M.W. Anwar and H. Akbar, Arabic Script Based Character Segmentation: A Review, Computer and Information Technology (WCCIT), World Congress on. IEEE, 22-24 June (2013).

[5] S.A. Husain and S.H. Amin, A Multi-tier Holistic approach for Urdu Nastaliq Recognition, Proceedings of International Multi Topic Conference, Karachi, Pakistan (2002) pp. 84-84.

[6] M. Kherallah, N. Tagougui, A. Alimi, H.E. Abed and V. Margner, Online Arabic Handwriting Recognition Competition, Document Analysis and Recognition (ICDAR), International Conference on IEEE, Beijing, China, 18-21 September (2011) pp.1454-1458.

[7] A. Das and U. Bhattacharya, ISI graphy: A Tool for Online Handwriting Sample Database Generation, Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Fourth National Conference, Jodhpur, India (2013) pp. 1 - 4

[8] S. Naz, K. Hayat, M.I. Razzak, M.V.Anwar and H. Akbar, Challenges in Baseline Detection of Cursive Script Languages, Science and Information Conference, London, UK, 7-9 October (2013) pp. 551-556.

[9] D.A. Satti and K. Saleem, Complexities and Implementation Challenges in Offline Urdu Nastaliq OCR, Proceedings of the 4th Conference on Language & Technology, Lahore, Pakistan, 9-10 November (2012) pp. 85-91.

[10] M.I. Razzak, S.A. Hussain, M. Sher and Z.S. Khan, Combining offline and Online Preprocessing for Online Urdu Character Recognition, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong, China,18-20 March, **1 (**2009) 18.

[11] R.J. Rodrigues and A.C.G. Thome, Cursive Character Recognition – A Character Segmentation Method using Projection Profile-based Technique, Proceedings of SCI /ISAS **V** (2000).

[12] S. Naz, K. Hayat, M.I. Razzak, M.V. Anwar and H. Akbar, Arabic Script Based Language Character Recognition: Nasta'liq vs Naskh analysis, IEEE World Congress on Computer and Information Technology, Sousse, Tunisia, 22-24 June (2013) pp. 1-7.

[13] N. Sankaran and C. V. Jawahar, Recognition of Printed Devanagari Text Using BLSTM Neural Network, ICPR. IEEE, Tsukuba, Japan (2012) pp. 322-325.

[14] I.K. Pathan, A.A. Ali and R.J. Ramteke, International Journal on Advances in Computational Research **4** (2012) 117.

[15] A. Jain, A. Dubey, G. Rachit, N. Jain and T. Pooja, International Journal of Innovative Research and Studies **2** (2013) 86.

[16] G.S. Lehal, Choice of Recognizable Units for Urdu OCR, Proceeding of the Workshop on Document Analysis and Recognition, Mumbai, India, 16 December (2012) pp. 79-85.

[17] F.O. Deborah, O.E. Olusayo and F.A. Alade, Innovative Systems Design and Engineering **3** (2012) 10.

[18] D. Megherbi, S.M. Lodhi and J.A. Boulenouar, A Fuzzy Logic-Based Technique for Urdu Character Representation and Recognition, Proceedings of SPIE **3962**, San Jose, California (2000) pp. 13-24.

[19] S.T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil and H. Moin, Proceedings of World Academy of Science, Engineering and Technology **46** (2010) 456.

[20] A. Vivek and A.M. Meggiolaro, Sign language Recognition using Competitive Learning in the HAVENET Neural Network, Proceedings of SPIE **3962**, San Jose, California (2000).

[21] U. Pal and A. Sarkar, Recognition for Printed Urdu Script, Proceedings of the Seventh International Conference on Document Analysis and Recognition **2**, IEEE Computer Society, Edinburgh, Scotland, 3-6 August (2003) pp. 1183-1187.

[22] Z. Shah and F. Saleem, Ligature Based Optical Character Recognition of Urdu, Nastaleeq Font, 6th Multi Topic International Conference, Karachi, Pakistan, 27-28 December (2002) pp. 25, 27-28.

[23] V. Frinken, A. Fischer, R. Manmatha and H. Bunke, IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 211.

[24] N. Durrani and S. Hussain. Urdu Word Segmentation, Human Language Technologies, The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 2 June (2010) pp. 528-536.

[25] S. Hussain and M. Afzal, Urdu Computing Standards: UZT 1.01, Proceedings of the IEEE International Multi-Topic Conference, Lahore, Pakistan (2001) pp. 223-228.

[26] Z. Ahmad, J.K. Orakzai, I. Shamsher and A. Adnan, Urdu Nastaleeq Optical Character Recognition, Proceedings of World Academy of Science, Engineering and Technology **26** (2007) pp. 2380-2383.

[27] I. Shamsher, Z. Ahmad, J.K. Orakzai and A. Adnan, Int. J. Comp. Info. Sys. & Control Engg. **1** (2007) 2978.

[28] A. Hussain, F. Anwar and A. Sajjad, Online Urdu Character Recognition System, MVA2007 IAPR Conference on Machine Vision Applications, 16-18 May, Tokyo, Japan (2007)pp. 98-101.

[29] J. Tariq, U. Nauman and M.U. Naru, Soft Converter: A Novel Approach to Construct OCR for Printed Urdu Isolated Characters, IEEE 2nd International Conference on Computer in Engineering and Technology, Chengdu, China **3** (2010) pp. V3-495.

[30] F. Iqbal, A. Latif, N. Kanwal and T. Altaf, Conversion of Urdu Nastaliq to Roman Urdu using OCR, IEEE 4th International Conference on Interaction Sciences, Busan, Korea (2011) pp. 19-22.

[31] T. Nawaz, S.A.H.S. Naqvi, H. Rehman and A. Faiz, Int. J. of Image Processing **3** (2009) 92.

[32] S. Belongie, J. Malik, and J. Puzicha, IEEE Trans. Pattern Anal. Mach. Intell. **24** (2002) 509

[33] S.T. Javed and S. Hussain, Improving Nastalique Specific Pre-Recognition Process for Urdu OCR, Proceedings of 13th IEEE International Multitopic Conference, Islamabad, Pakistan, 14-15 December (2009) pp. 1-6.