University of Sindh Journal of Information and Communication Technology
(USJICT)
Volume 3, Issue 1, January 2019

ISSN-E: 2523-1235, ISSN-P: 2521-5582
© Published by University of Sindh, Jamshoro
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Issues & Challenges in Urdu OCR

[1]Urooba Zaki, [1]Dil Nawaz Hakro,[1]Khalil-ur-Rehman Khoumbati, [2]M. Ahmed Zaki, [1]Maryam Hameed

[1]Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
[2]Department of Telecommunication, Mehran University of Engineering & Technology Jamshoro, Pakistan
uroobazaki524@gmail.com, dill.nawaz@gmail.com, khalil.khoumbati@usindh.edu.pk, ahmed_zaki64532@yahoo.com, maryamGhori133@gmail.com

Abstract: Optical character recognition is a technique that is used to recognized printed and handwritten text into editable text format. There has been a lot of work done through this technology in identifying characters of different languages with variety of scripts. In which Latin scripts with isolated characters (non-cursive) like English are easy to recognize and significant advances have been made in the recognition; whereas, Arabic and its related cursive languages like Urdu have more complicated and intermingled scripts, are not much worked. This paper discusses a detail of various scripts of Urdu language also discuss issues and challenges regarding Urdu OCR. due to its cursive nature which include cursiveness, more characters dots, large set of characters for recognition, more base shape group characters, placement of dots, ambiguity between the characters and ligatures with very slight difference, context sensitive shapes, ligatures, noise, skew and fonts in Urdu OCR. This paper provides a better understanding toward all the possible engendering dilemmas related to Urdu character recognition.

Keywords: Urdu Recognition Challenges; Urdu OCR; Optical Character Recognition; Character Dots; Urdu Character shapes;

I. Introduction

Optical character recognition (OCR) is considered to be a way to convert pictorial data into text and enable the fastest mode of data input [1]. Optical Character Recognition (OCR) as the name implies a technology through which the image form of data of typed, written or hard prints can be converted into editable text formats. It is widely used as a form of printed information input of paper documents, identity documents, invoices, bank statements, receipts computer, business cards, email, static data printing or proper documentation [2]. In addition to character recognition, single characters (Latin Script) are easier to recognize than Cursive characters. [3]. Since the OCR development for like script is still in its initial phase compared to the OCR's development for Latin posts, it still needs more attention to Arabic script and related language script [4],[5]. Urdu Language inherits some of its feature from Arabic like script with more characters and more diverse placement and orientation of dots. Therefore, the development of Urdu OCR is expected to be a major problem.

Urdu is the official national language of Pakistan derived from Farsi script [2] starts from right side to left side as are in Arabic and Persian [6]. Urdu script follows bi-directional writing style in which characters and words are written from right whereas numbers starts from left.. So basically, there is no difference between the basic form of the characters and the only difference is in placement and orientation of the number of dots.

II. Related work

OCR follows two techniques for recognition namely segmentation based and segmentation free. Segmentation based involves individual characters recognition where as segmentation free or ligature based recognition involves whole word recognition [4]. This section is further divided into segmentation free and segmentation based approach for presenting the related work of Urdu OCR.

A. Segmentation Free (Ligature based) Recognition

Din et al. [7], proposed a hybrid technique based on vertical and horizontal projection for segmentation of line and ligature of offline Urdu documents. The system composed of two stages of segmentation in which line segmentation is the first stage after image binarization through global threshold technique. Line segmentation is done by using the horizontal projection technique. The system dilates the whole image text with the square structure. At the next level, segmentation of the ligature is done in terms of segmentation of the primary and secondary ligatures using the vertical projection technique. The system was tested for 30 documents consisting of 310 lines of text in these 306 have been segmented successfully with an accuracy of 98.7%. These lines comprise 7364 ligatures in these 6,811 ligatures recognize with an accuracy of 92.5%.

Rana & Lehal [2], proposed segmentation free OCR technique for Urdu nastalique script. The system prepared sample Urdu data by taking data from various books. The system consists of 10082 ligatures that are classified into 16 classes as primary and secondary component. The next step is the feature extraction which involves different technique which are discrete cosine transformation (DCT) for feature extraction, Gabor feature for edge detection, directional feature for calculating distance between white and black pixel and Gradient feature for calculating magnitude and direction. Recognition is the last step, done by using different techniques including support vector machine (SVM) with linear and polynomial kernel and K nearest neighbor (KNN) classifier technique. The system was tested on 110 pages of Urdu printed text. Support vector machine and k nearest neighbor technique recognize 3746 primary component just in 559 and 170 second. The accuracy of the system is 90.29% reported.

Lehal [6], presents a system that recognized the Urdu ligatures in nastalique writing style. The system comprises of two steps named as offline ligature recognition and online ligature recognition. Offline ligature recognition done by identifying the ligatures with 2190 primary and 17 secondary component, train data with 15 samples for each ligature and lastly create the code book of primary, secondary and Unicode for correspond ligature. Whereas online ligature recognition done by segmenting the lines, splitting the ligature using connected component technique, recognize the primary and secondary components by using Gabor, DCT and zoning feature extraction techniques and SVM, KNN and HMM as classifier and then compare it with Unicode string. The system able to recognized 9262 ligature with 98% accuracy.

Javed et al. [8], used Hidden Markov model and rule based post-processor technique for the development of Urdu OCR in nastalique script. The system consists of two phases training and recognition phase. Training phase comprise of several steps which are separation of lines from text page, identify baseline, extraction of main body and Thinning. In the recognition phase, first step is the segmentation of ligatures then each segment is sent to Hidden Markov Model (HMM) for recognition. The system tested 1692 ligatures from 18600 words in which 1569 ligatures was correctly identified with accuracy 92.73%.

Javed & Hussain [9], proposed the preprocessing stage for Urdu character recognition in nastalique writing style. The system composed of two stages. The first stage is the segmentation of page into lines using horizontal and vertical projection technique. The next stage is the segmentation of lines into ligature done by follow the steps namely division of line into ligature, baseline detection using horizontal projection of pixel and association of base and marks is the last step done by calculating the center for each shape. The system was tested with varied font size from 34 to 38 of twenty pages (ten lines per page) from three different books. An accuracy of 100% was reported for line segmentation. The lines comprises of 3655 ligature out of which 3436 ligature accurately separated with accuracy of 94%.

Sattar et al. [10], proposed a novel technique of segmentation free approach for Urdu character recognition in nastalique script based on cross correlation. The system consist of two level of segmentation in which line segmentation is the first stage and ligature and isolated character segmentation is next stage. For testing purpose, MatLab is used. The system tested for small subset of Urdu words in nastalique script with same font size and shows an encouraging result.

Husain et al. [12], proposed a recognition system for online Urdu handwriting. The segmentation free system consists of steps namely stroke acquisition, preprocessing involves smoothing to remove hooks, feature extraction, association and classification of special stroke with base stroke by using BNN technique, identification of ligature , valid ligature form word, identify the word from word dictionary then the last step is to written a word into text file. The system was trained for 240 ligatures with addition of 6 diacritics. The system was successful to recognize 864 ligatures, 50000 words of these ligatures. An accuracy of 93% and 98% was reported for recognition of base and secondary stroke.

B. Segmentation Based (Isolated character) Recognition

Naz et al. [13], proposed a segmentation based recognition system for Urdu nastalique script. The system consists of several steps; conversion of image into grey scale, feature extraction used zoning feature technique. Zoning feature are calculated by calculating the density of the pixel based on zone. The next step is the classification done by using two techniques namely two dimensional long short term memory (2DLSTM) and CTC output layer. The system used Urdu printed text. Dataset consist of 10,000 text line, 771,339 character and 44 labels. The dataset breaks into 68% (6800), 16% (1800) and 16% (1800) for training, validation and testing set. The system performed four experiment based on zone size (30, 50, 70 and 90). The recognition rate 82.62%, 87.34 %, 92.96% and 93.38% for 30, 50, 70 and 90 zone size was reported.

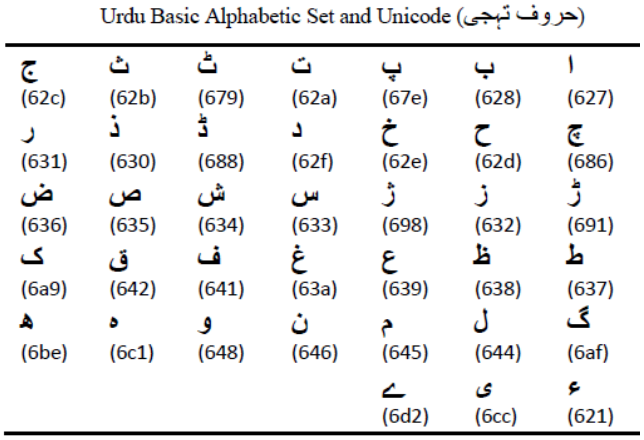
Khan et al. [14], proposed an Urdu OCR system. The system consists of three steps namely preprocessing, feature extraction and classification. Preprocessing consists of noise removal by filtering technique, binarization based on global threshold, normalization of the image, skew correction and edges detection. Next step is feature extraction done by using three different methods. Last step is classification the done by using decision tree algorithm. The system used MatLab for preprocessing and feature extraction and used J-48 for classification. The system has its own database, tested 441 character with accuracy of 92.06%, 54.32% and 32.13% for the Hu Moment, Zernike Moment and PCA technique.

Wahab & Haque [15], proposed a segmentation free approach for online as well as offline Urdu Character Recognition. The framework composed of stages in which preprocessing is the first one involves noise detection, smoothing and binarization of image. The next stage is extraction of text lines that involves text area extraction, extraction of primary and secondary stroke, base line detection, stroke identification and feature extraction done by using sliding window technique and Hu Moment algorithm. The last step is recognition done by using K-Nearest Neighbors algorithm for matching features based on Euclidean distance for 10 nearest neighbors. The system was developed on MatLab and Microsoft C #. Net. The system was tested for both online and offline document of varies font and script. An accuracy of 97.09%, 98.86% and 97.12% for extracting text lines, extraction of primary and secondary stroke and recognition was reported.

Malik and khan [16] proposed a system for online Urdu handwritten. For the recognition of online Urdu handwritten the system consists of six steps. Data acquisition performed by using x and y coordinates, online stroke detection, preprocessing by removing the repeating point and noise by using Time domain filtering. They used analytical approach of segmentation for feature extraction. Removed slant through rule base slant analysis. The last step of the system is classification done by matching the character from database. The technique involve in this phase was tree based dictionary search which reduce the searching space up to 96.2%. The system was trained to test 39 isolated character, 10 Urdu numbers, 200 two character Urdu ligature with one hat feature. An accuracy of 93.3%, 93%, 93% and 78% for recognition of hat feature, isolated character, numbers and two character ligatures are reported. The overall result of recognition of the system is 90.25%.

III. URDU SCRIPT

Urdu language is most spoken in five Indian states and regarded as fifth fluently spoken language with approx. 4.7% of world population [17]. Urdu alphabet comprises of 58 characters set [14] shown in Figure 2 having 38 basic characters shown in Figure 1 called "horof-e-thahji" and most of these letters are Arabic and small from Persian [6] whereas Arabic has 28 characters and Persian has 32 characters. Urdu is one of the languages which includes features, scripts, and writing as in Persian and Arabic languages.[18]. There are many different font styles available in Urdu i-e Nastaliq, Naskh, Noori Naskh, Noori Nastaliq, Kofi, etc. Out of them, two are common: Nastaliq and Naskh [19]. Many challenges are encountered when developing an OCR system for Urdu because of the nature of Urdu script which will be discussed in more detail in this section.



Urdu Basic Alphabetic Set and Unicode (حروف تہجی)

ج (62c)	ث (62b)	ٹ (679)	ت (62a)	پ (67e)	ب (628)	ا (627)
ر (631)	ذ (630)	ڈ (688)	د (62f)	خ (62e)	ح (62d)	چ (686)
ض (636)	ص (635)	ش (634)	س (633)	ژ (698)	ز (632)	ڑ (691)
ک (6a9)	ق (642)	ف (641)	غ (63a)	ع (639)	ظ (638)	ط (637)
ھ (6be)	ہ (6c1)	و (648)	ن (646)	م (645)	ل (644)	گ (6af)
				ے (6d2)	ی (6cc)	ء (621)

Figure 1: Basic Character of Urdu

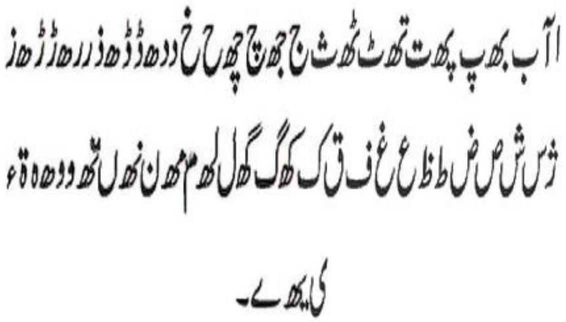


Figure 2: 58 Character Set of Urdu language [14]

A. Bi-Directional Data

Bi-directional means script is written in two dimension i-e x and –x-axis. Urdu follows bidirectional rules because numbers starts from left whereas characters starts from right [12] shown in Figure 3.

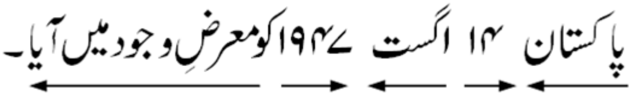
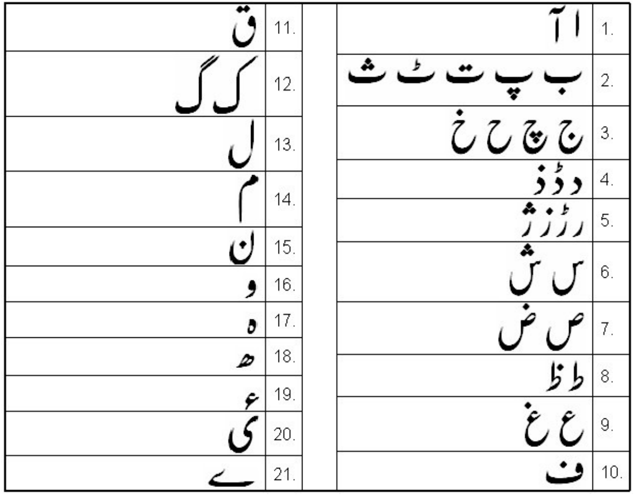


Figure 3: Bi-direction Data [12]

B. Joiner & Non-Joiner

Urdu has 38 basic characters. These basic characters are divided into joiner and non-joiner. Joiner characters are those that have the ability to combine with other characters and form a new one whereas the non-joiner are isolated character that are not able to join with other. All alphabetical 38 letters can be split into 21 classes as shown in Figure 4 with respect to their basic shapes to minimize the diversity [20].



ق	11.	آ ا	1.
ک گ	12.	ب پ ت ٹ ث	2.
ل	13.	ج چ ح خ	3.
م	14.	د ڈ ذ	4.
ن	15.	ر ڑ ز ژ	5.
و	16.	س ش	6.
ہ	17.	ص ض	7.
ھ	18.	ط ظ	8.
ء	19.	ع غ	9.
ی	20.	ف	10.
ے	21.		

Figure 4: 21 Classes of Urdu Alphabets

C. Importance of Diacritics

Urdu characters have some marks, called diacritics. Diacritics are the sign which when written below and above the character change the pronunciation as well as the character orientation. Diacritics include dots, tuay, etc.

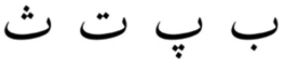


Figure 5: Position of Dot

Another type of diacritics called "Aerab" helps in word pronunciation. Figure 6 represent the position of Aerab [20].

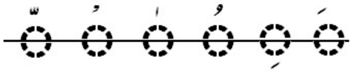


Figure 6: Aerab (Right to left: Zabr, Zair, Paish, Khari Zabr, Juzm, Shud)

D. Dots

The letters of Urdu shown in Table 1 possess 1, 2, or 3 dots. The dot position may be up, down, or inside the base shape.

Table 1: Characters Arrangement According To Number of Dots

NUMBER OF DOTS	=	CHARACTERS
With single dot	10	– غ – ظ – ض – ز – ذ – خ – ج – ب ن – ف
With two dots	2	ق – ت
With three dots	5	ش – ژ – چ – ث – پ
Without dots	18	– ع – ط – ص – س – ر – د – ح – ا ے – ی – ء – ھ – ہ – و – م – ل – گ – ک
With small (ط)	3	ڑ – ڈ – ٹ
Total numbers of characters	38	

E. Placement of Dots

Dots are arranged in different places. The number of dot range is up to 1 to 3, and are set in different places. It is inside, at the top, bottom of the base shape of the character. For example, placement of single dot shown in Table 2, Table 3 and 4 represents the placement of two and three dots.

Table 2: Single Dot Placement

PLACEMENT	CHARACTERS	TOTAL (10)
Below	ب	1
Above	– ظ – ض – ز – ذ – خ ن – ف – غ	8
Inside	ج	1

Table 3: Two Dots Placement

PLACEMENT	CHARACTERS	TOTAL (2)
Below	No character	0
Above	ق – ت	2
Inside	No character	0

Table 4: Three Dots Placement

PLACEMENT	CHARACTERS	TOTAL (5)
Below	پ	1
Above	ش – ژ – ث	3
Inside	چ	1

Table 5 shows the overall summary of placements of dots in Urdu. We can take the result that characters have dots above the base shape are increase in number.

Table 5: Summary of Dots Placement

Placement of Dots	Characters	No:
Dots above the characters	ز – ذ – خ – ش – ژ – ث ت – ن – ف – غ – ظ – ض – ق –	13
Dots below the characters	ب – پ	2

Dots inside the characters	ج – چ	2
Without dots	– س – ر – د – ح – ا – ل – گ – ک – ع – ط – ص ے – ی – ء – ھ – ہ – و – م	18
With small (ط)	ژ – ڈ – ٹ	3
Total characters	***38***	

F. Shape Sensitivity

Urdu script is shape sensitive that means that most of the characters of Urdu scrip have up to four different shapes depending on the position i-e isolated, middle, left and right.

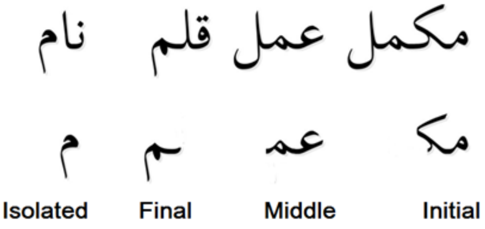


Figure 7: Shape of "Meem (م)"

G. Broken Ligature

Sometime due to some technical problem in Urdu printed text the strokes are absent in between the character and hence the world is not recognized.

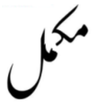


Figure 8: Broken Ligature

H. Space between ligature and word

Urdu script is written in such a way that there is space between the ligatures and a whole word. So it becomes a challenging task for segmentation and recognition.

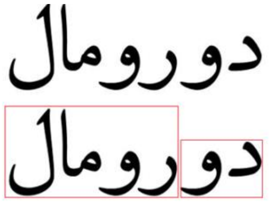


Figure 9: Space between Words

I. Line Segmentation

Segmentation is the process of OCR and line segmentation in Urdu may face two problems that are ligature touching and overlapping.

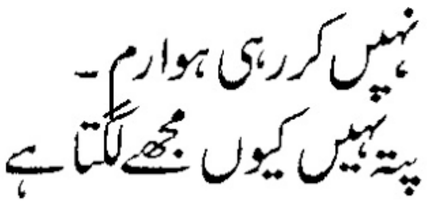


Figure 10: Ligature Touching & Overlapping

J. Base-line:

Base line as the name implies a horizontal line place at the base cutting all the characters and secondary characters at some points.

IV. URDU LIGATURE

Ligature means connected character. To create ligature or words, letters are connected to its succeeding or preceding character. This connection is called the cursive nature of Urdu language. Urdu scripts have minimum one ligature and maximum four as shown in Figure 11. According to the OCR, a ligature consists of primary component and secondary components as shown in Figure 12. The primary component indicates the basic form, while the secondary component corresponds to the marks of points and diacritics and to the special symbols associated with the ligature [6].

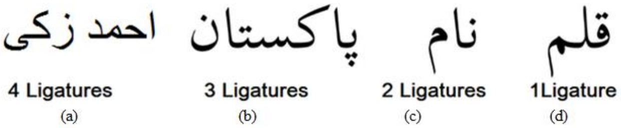


Figure 11: Various Ligatures (a) 4 ligatures (b) 3 ligatures (c) 2 ligatures (d) 1 ligature

Figure 12: Ligature

Table 6: Primary &Secondary Component

Primary Component	
Secondary Components	

The words can be created from isolated, single and multiple ligatures. Table 7 shows some examples to create words in Urdu.

Table 7: Formation of word by ligatures

1	**Word by isolated character**	روزه – دروازه – اور
2	**Word by one ligature**	نغمہ – خط – لفظ
3	**Word by multiple ligature**	– دونوں – حروف یزدستاو

V. Font Style

Urdu script have diagonally writing style from right to left and have varsity in font style like Nastalique, Naskh,, Kofi, sals, devani and Raka [12] [27] and many more some of them shown in Figure 13.

VI. Noise

In an OCR system, the noise is an important factor that causes by different factors such as scanning, ink print during print and old documents. Searching between the characters dots and the noise is a difficult task. In Urdu script, due to the presence of dots it becomes more challenging task to find out the different between noise and dots of a character. Although some researchers use noise removal techniques, [21] chose to use high quality paper image that free of noise.

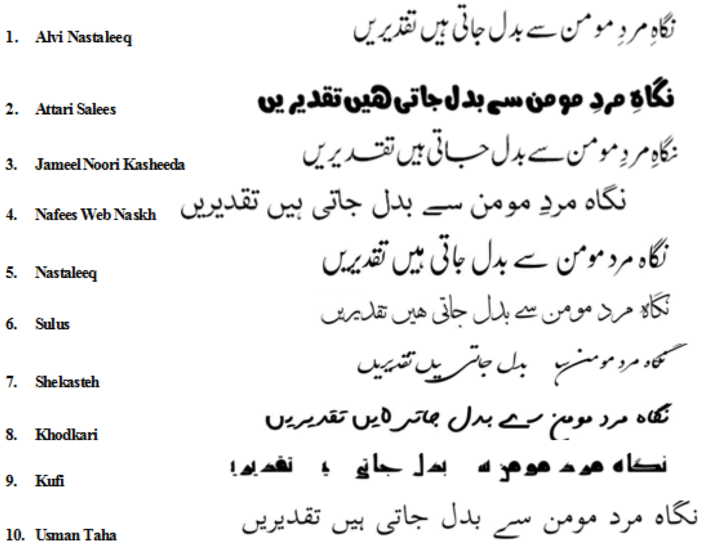


Figure 13: Fonts

VII. Skew

Skew is basically the boundary of scanned text image. Therefore, we must pay attention to the scan. Identifying the correct boundary of text and remove extra region called skew detection and its correction. Al-shatnawi & Omar [23], used polygon method for Arabic text images whereas for Indian script Chaudhuri & Pal [24], proposed an algorithm and Desai [26], work on Gujarati handwritten number to detect and correct skew.

VIII. Result and Discussion

Urdu OCR includes some additional challenges as compared to Arabic and Persian script because it possesses more basic characters as in Arabic and Persian. Cursive nature of Urdu language can make recognition more challenging task. Urdu word has minimum one ligature and maximum four. According to the OCR, a ligature classified as primary and secondary components. A word can be made from single, multiple ligatures and only with isolated characters. Hence there is a huge collection of words with the versatility of different combination of characters with ligatures. Dots are arranged in different places. The number of dot range is up to 1 to 3, and are set in different places. It is inside, at the top, the bottom of the base shape of the character. The orientation of the dot may be different according to base shape of the character. Urdu script is shaped sensitive a single character has up to four different shapes depending on the position i-e isolated, middle, left and right. Urdu is difficult to recognize because most of the characters have the same basic formats, but placement orientation and numbers of dots make a basic shape different. A basic form like " ٮ" has five different letters in Urdu whereas in Arabic have only three characters for the same basic form. Sometimes due to some technical problem in Urdu printed text, the strokes are absent in between the character and hence the world is not recognized. Due to the isolated character, it is sometimes a challenging task to find the space between the ligatures and a whole word. So it becomes a challenging task for segmentation and recognition. Face problems that are ligature touching, overlapping and due to baseline sometimes it cuts the characters and secondary characters at some points. Noise and Skew can make the recognition even more challenging. Furthermore, Urdu script has diagonally writing style from right to left and have varsity in font style like Nastalique, Naskh, Kofi, Noori Naskh, Noori Nastaliq, Kofi, etc. Out of them, two are common: Nastaliq and Naskh. Table 8 shows the overall Summary of issues and challenges.

Table 8: Summary of issues and challenges

Issues & Challenges	***Explanation***
Bi-directional writing	Characters written from right to left whereas number written from left to right
More basic characters	Urdu possess 38 basic characters which are more from Arabic and Persian script
Word range	A single word may have minimum one ligature and maximum four
Versatility of words	A word can be made from a single, multiple ligatures and only with isolated characters
Range of dots	A character have minimum single dot and maximum three
Placement of dots	Dots may be places top, bottom and inside the character
Orientation of dots	Above, below and inside the character
Shape sensitive	Single character has up to four different shapes depending on the position i-e isolated, middle, left and right.
Same base shape	Most of the characters have same base shape wth different placement orientation and numbers of dots
Missing word	Strokes are absent in between the character and hence the world is not recognized
Space between word	Challenging task for segmentation and recognition of some words due to Space between the ligatures and a whole word
Ligature touching, overlapping	Sometimes it is difficult to recognize a character because of overlapping and character touching problem
Noise & Skew	Noise and Skew can make the recognition even more challenging.
Fonts	Urdu script have diagonally writing style from right to left and have varsity in font style

IX. CONCLUSION

Urdu derived from Farsi hence adopts writing style from Arabic and Persian. It has more ligatures hence words are formed by more characters in Urdu script which is the major problem in recognition. Depending on the placement, orientation and number of dots a base shape may have up to 5 multiple characters. So, that an individual letters make multiple characters. A lot of Fonts can be used in the Urdu script. To build the whole OCR for Urdu, a lot of cautions are needed.

X. FUTURE WORK

For future research, the technology used by the researcher will be investigated for Urdu OCR. Font family and style are taken into account for the enhancement of Urdu OCR.

REFERENCES

[1] U. Pal and B. B. Chaudhuri, "Indian script character recognition : a survey," vol. 37, pp. 1887–1899, 2004.

[2] A. Rana and G. S. Lehal, "Smart Computing Prototype for Industry 4.0 Revolution with IOT and Bigdata Implementation Model," *Indian J. Sci. Technol.*, vol. 8, no. December, pp. 1–7, 2015.

[3] D. N. Hakro, "Enhanced Segmentation and Feature Extraction for Sindhi Optical Character Recognition," 2015.

[4] V. Narang, S. Roy, O. V. R. Murthy, and M. Hanmandlu, "Devanagari character recognition in scene images," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 902–906, 2013.

[5] M. Tanvir and S. A. Mahmoud, "Arabic handwriting recognition using structural and syntactic pattern attributes," *Pattern Recognit.*, vol. 46, no. 1, pp. 141–154, 2013.

[6] G. S. Lehal, "Choice of Recognizable Units for Urdu OCR," *Proceeding Work. Doc. Anal. Recognit.*, pp. 79–85, 2012.

[7] I. S. and S. K. Israr Ud Din, Zumra Malik, "Line and Ligature Segmentation in Printed Urdu Document Images Line and Ligature Segmentation in Printed Urdu Document Images," *J. Appl. Environ. Biol. Sci.*, vol. 6, no. March, pp. 114–120, 2016.

[8] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, "Segmentation Free Nastalique Urdu OCR," 2010.

[9] S. T. Javed and S. Hussain, "Improving Nastalique-Specific Pre-Recognition Process for Urdu OCR," 2009.

[10] S. A. Sattar, S. and Haque, and M. K. Pathan, "Nastaliq Optical Character Recognition," pp. 329–331, 2008.

[11] S. A. Husain, A. Sajjad, and F. Anwar, "Online Urdu Character Recognition System," *MVA2007 IAPR Conf. Mach. Vis. Appl.*, pp. 1–7, 2007.

[12] S. T. Javed and S. Hussain, "Segmentation Based Urdu Nastalique OCR," pp. 2–8.

[13] S. Naz, K. Hayat, M. Imran, M. Waqas, S. A. Madani, and S. U. Khan, "The optical character recognition of Urdu-like cursive scripts," *Pattern Recognit.*, pp. 1–20, 2013.

[14] K. Khan, R. U. Khan, A. Alkhalifah, and N. Ahmad, "Urdu Text Classification Using Decision Trees," *2015 12th Int. Conf. High-*

capacity Opt. Networks Enabling/Emerging Technol., pp. 1–4, 2015.

[15] A. Wahab and S. Haque, "Optical Character Recognition System for Urdu .," vol. 8, no. 2, 2010.

[16] S. Malik and S. A. Khan, "Urdu online handwriting recognition," *Emerg. Technol. 2005. Proc. IEEE Symp.*, pp. 27–31, 2005.

[17] S. Shabbir and I. Siddiqi, "Optical Character Recognition System for Urdu Words in Nastaliq Font," *Inf. Emerg. Technol. (ICIET), 2010 Int. Conf.*, vol. 7, 2016.

[18] S. Sardar and A. Wahab, "Optical Character Recognition System for Urdu .," *2010 Int. Conf. Inf. Emerg. Technol.*, pp. 1–5, 2010.

[19] M. I. Razzak, S. A. Hussain, M. Sher, and Z. S. Khan, "Combining Offline and Online Preprocessing for Online Urdu Character Recognition," vol. I, pp. 18–21, 2009.

[20] S. Wali, A and Rehman, "Implementation of Reverse Chaining Mechanism in Pango for Rendering Nastaliq Script," in *Second Workshop of Computational Approaches to Arabic Scriptbased Languages, Stanford University, USA*, 2007.

[21] S. Ajward, N. Jayasundara, S. Madushika, and R. Ragel, "Converting Printed Sinhala Documents to Formatted Editable Text," *2010 Fifth Int. Conf. Inf. Autom. Sustain.*, no. 1, pp. 138–143, 2010.

[22] A. M. Al-shatnawi and K. Omar, "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity Images Based on Centre of Gravity," no. September, 2015.

[23] A. M. Al-Shatnawi, "A skew detection and correction technique for Arabic script text-line based on subwords bounding," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, pp. 324–328, 2015.

[24] B. B. Chaudhuri and U. Pal, "Skew Angle Detection of Digitized Indian Script Documents," vol. 19, no. 2, pp. 182–186, 1997.

[25] A. A. D. Ã, "Gujarati handwritten numeral optical character reorganization through neural network," *Pattern Recognit.*, vol. 43, no. 7, pp. 2582–2589, 2010.

[26] J. M. Patel and A. A. Desai, "Gujarati Text Localization , Extraction and Binarization from Images International Journal of Computer Sciences and Engineering Open Access," no. September, 2018.

[27] Soomro, W. J., Ismaili, I. A., & Shoro, G. M. (2018). Optical Character Recognition System for Sindhi Text: A Survey. University of Sindh Journal of Information and Communication Technology, 2(2), 1-7.