International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:8, 2007

Urdu Nastaleeq Optical Character Recognition

Zaheer Ahmad, Jehanzeb Khan Orakzai, Inam Shamsher, and Awais Adnan

Abstract**—**This paper discusses the Urdu script characteristics, Urdu Nastaleeq and a simple but a novel and robust technique to recognize the printed Urdu script without a lexicon. Urdu being a family of Arabic script is cursive and complex script in its nature, the main complexity of Urdu compound/connected text is not its connections but the forms/shapes the characters change when it is placed at initial, middle or at the end of a word. The characters recognition technique presented here is using the inherited complexity of Urdu script to solve the problem. A word is scanned and analyzed for the level of its complexity, the point where the level of complexity changes is marked for a character, segmented and feeded to Neural Networks. A prototype of the system has been tested on Urdu text and currently achieves 93.4% accuracy on the average.

Keywords**—**Cursive Script, OCR, Urdu.

I. Introduction

URDU is the national language of Pakistan, is spoken by more than 60 million speakers in over 20 countries [2].It is a cursive script, written from right to left, like Arabic and Farsi but with some additional alphabets, therefore OCRs used for Arabic or Farsi will not suit the needs for Urdu script.

In this paper a character is segmented using a three steps approach, firstly, lines of text are identified, secondly words are identified and thirdly each character is segmented and extracted from a word/sub-word using its complexity level to be feeded to neural network for final recognition/classification.

The main focus of the paper is character segmentation and extraction from a word or sub-word, text lines, words identification and Neural Networks used for character segmentation and identification has not been described in detail.

II. Urdu Script

Urdu is one of the popular Indian script in the Indian subcontinent and national language of Pakistan evolved in the subcontinent from the mixture of Arabic, Turkish, Farsi and Hindi Languages with 58 character set defined by National Language Authority Pakistan as shown in Fig. 1.But only 40 basic and one *do-chashmi-hey* is used to form all composite alphabets; so a total of 41 alphabets Urdu shares a common script and many characteristics of Arabic script with additional set of alphabets.

Authors are with the Center for Computing, Institute of Management Sciences, Peshawar, Pakistan.

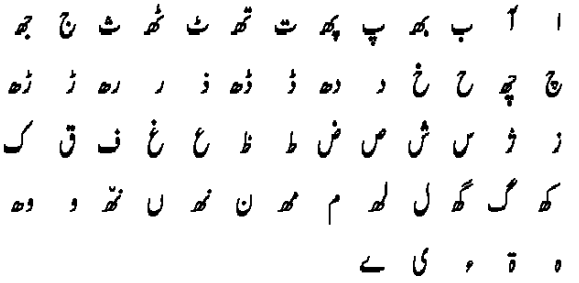

Fig. 1 Character Set (58 alphabets) of Urdu Script

Most of Urdu characters when combined form a degree of about 45 to the horizontal line because of which Urdu script reading is faster than roman script but on the other hand it makes it harder for the novice readers and the machines to recognize the word or segment one character from the rest.

Unlike the English script there is no capital or small characters in Urdu, but the last character of a word can be considered as a capital character as in many cases it presents the full form of the character and the characters at initial and middle positions are considered as small. Every character has a stand alone shape besides different joining forms, but some of the alphabet like the characters making the word Urdu (ودرا) or of the similar category are not joinable or cannot be connected. Urdu alphabet utilizes consonant letters, vowels, diacritic marks, numerals, punctuations and a few superscripts signs.

The graphic representation of each alphabet has more than one form depending on its position and context in the word. In general each letter has four forms that is beginning, middle, final and standalone as shown in Table I.

TABLE I
Characters and its Different Forms

#	Char حرف	Forms لااکشا	Name اِسم	Name
0	ء		ہمزہ	hamzah
1	ا	ا	الف	alif
1a	آ	آ	الف مدّ	alif madd
2	ب	بببب	بے	bē
2h		ھب ھب	بے ھ	bhē
3	پ	پپپپ	پے	pē
3h		ھپ ھپ	پے ھ	phē
4	ت	تتتت	تے	tē
4h		ھت ھت	تے ھ	thē

5	ٹ	ٹٹٹ	ےٹ	t.ē
5a	ٹ	ٹٹٹ	ےٹ	t.ē
5h		ٹھ ٹھ	ےٹھ	t.hē
6	ث	ثثث	ےث	s̲ē
7	ج	ججج	ےیج	jīm
7h		ھج ھج	ےھج	jhē
8	چ	چچچ	ےچ	čē = cē
8h		ھچ ھچ	ےھچ	čhē = chē
9	ح	ححح	ےحےڑب	bar.ī Hē
10	خ	خخخ	ےخ	xē = kh̲ē
11	د	د	لاد	dāl
11h		ھد ھد	ےھد	dhē
12	ڈ	ڈ	لاڈ	d.āl
12a	ڈ	ڈ	لاڈ	d.āl
12h		ھڈ ھڈ	ےھڈ	d.hē
13	ذ	ذ	لاذ	z̲āl
14	ر	ر	ےر	rē
15	ڑ	ڑ	ےڑ	r.ē
15a	ڑ	ڑ	ےڑ	r.ē
15h		ھڑ ھڑ	ےھڑ	r.hē
16	ز	ز	ےز	zē
17	ژ	ژ	ےژ	žē = zhē
18	س	سسس	نیس	sīn
19	ش	ششش	نیش	šīn = shīn
20	ص	صصص	داص	Sād, Suād
21	ض	ضضض	داض	Żād, Żuād
22	ط	ططط	ےوط	Tōē
23	ظ	ظظظ	ےوظ	Zōē
24	ع	ععع	نیع	‘ain
25	غ	غغغ	نیغ	ğain
26	ف	فففف	ےف	fē
27	ق	قققق	فاق	qāf
28	ک	ککک	فاک	kāf
28h		ھک ھک	ےھک	khē
29	گ	گگگ	فاگ	gāf
29h		ھگ ھگ	ےھگ	ghē
30	ل	لللل	مال	lām
31	م	ممم	میم	mīm
32	ن	ننن	نون	nūn
32a	ں	ں	نون غنّہ	nūn-e ğunnah
33	و	و	واو	vāō
34	ہ	ہہہ	یٹوھچ ےہ	čhōt.ī hē
34a	ہ	ہہہ	یٹوھچ ےہ	čhōt.ī hē
34b	ھ	ھھھ	یمشچود ےھ	dō-čašmī hē
35	ی	ییی	یٹوھچ ےی	čhōt.ī yē
35a	ئ	ئئئ	ہزمہ	hamzah
35b	ے	ے	ےیڑب	bar.ī yē

III. Nastaleeq

Urdu is written in Arabic script. Arabic script has many traditional writing styles, including Naskh (mostly used for Arabic language), Taleeq, Kufi, Divani, Sulus, Riqa, etc. Naskh and Taleeq styles of writing were combined into the very spatially concise Nastaleeq writing style. Nastaleeq writing system for Urdu is character based, bidirectional (mainly R to L), diagonal, non-monotonic, cursive, context sensitive writing system with a significant number of marks (dots and other diacritics). This makes Nastaleeq one of the most complex writing styles and challenging to develop an OCR for it. Nastaleeq is a complex cursive style of writing Arabic script based languages e.g. Urdu and Persian. Each letter has precise writing rules, relative to the width of the flat nib of the pen, called *qat*. The measurement of some letters in terms of *qat* is given in Fig. 2.

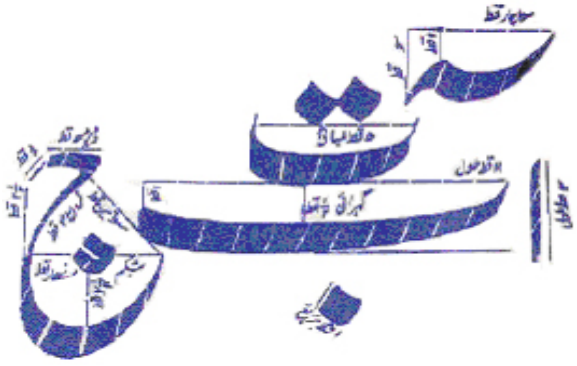

Fig. 2 Characters measurements used in Nastaleeq

As Nastaleeq is a writing style for Arabic script, it inherits its bidirectional nature, where the characters are written from R to L but numbers are written from L to R. Hand written Nastaleeq has been developed as art in the Muslim world where it replaced all other forms of art like painting etc.

IV. Feature Extraction

A. Preprocessing Algorithms

The main problem arising in segmentation is the possibility of overlapping of characters in a word or sub-word which occurs quite often especially in cursive languages. Eliminating the possibility of overlapping by stretching the words horizontally to make space between two connected characters is shown in Fig. 3. Text lines, words boundary identification and overlapping and other preprocessing techniques have been adopted from [1] [3], [4], [5] and [6].

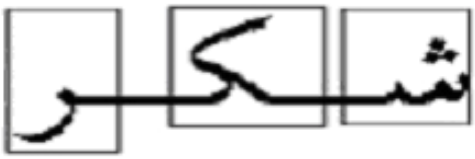

Fig. 3 Horizontally Stretched Word

B. Segmentation

The segmentation phase is based on the level of complexity offered by a character during scanning. The characters are grouped into three levels of complexity, simple, semi complex and complex as shown in Table II.

TABLE II
COMPLEXITY WISE GROUPING OF CHARACTERS

Char	Forms	Complexity Level
فرح	ا ش ک ا ل	
ا	ا	Simple
ب	ب ب ب	Simple
ٹ	ٹ ٹ ٹ	Simple
ش	ش ش ش	Simple
ء		Semi Complex
ج	ج ج ج	Semi Complex
د	د	Semi Complex
ڈ	ڈ	Semi Complex
ک	ک ک ک	Semi Complex
گ	گ گ گ	Semi Complex
ل	ل ل ل	Semi Complex
ی	ی ی ی	Semi Complex
ے	ے	Semi Complex
ن	ن ن ن	Semi Complex
ض	ض ض ض	Complex
ظ	ظ ظ ظ	Complex
غ	غ غ غ	Complex
ف	ف ف ف	Complex
ق	ق ق ق	Complex
م	م م م	Complex
و	و	Complex
ہ	ہ ہ ہ	Complex
ھ	ھ ھ ھ	Complex

The complexity of a character is measured by analyzing the topological features, the number of holes, the width, height of holes and the direction of these holes but the decisive part is played by the lines which are encountered by the scanner during the scanning process. The character dō-čašmī hē (ھ) is made of two holes by three lines (connecting each other) therefore is considered a complex character similary a single hole or closed / loop character like mīm (م) is also considered as a complex character. A character with semi opened shape or with two line from one side like bar.ī Hē (ح) is considered semi complex shape, all the remaining characters are considered simple ones. Deciding a complex or semi complex shape it has taken care that the lines should be connected at some point and the distance of any two lines should not exceed a specified limit keeping in view the size of the fonts under consideration like the distance between the two lines of bar.ī Hē (ح) .

To avoid complex calculation and improve the efficiency the scanning is carried out both vertically and horizontally. During which an isolated word is scanned vertically from right to left, double and triple lines characters are looked firstly from the upper side of the image, if the character is closed with two or more lines such that it makes a hole it is further scanned horizontally from top to bottom to define the hole as a vertical or horizontal hole and calculate the distance between the lines of holes. If the character is closed from three sides but open or semi open from one side these characters are semi complex characters. All other characters are considered simple one.

The character is scanned and the level of complexity is stored during the scanning as complex, semi complex or simple, when the level is getting changed (the loop is starting / ending) the change is verified from the right side scan if it also confirms the change the beginning /end for complexity is marked, reaching on the other side of the character the same process is performed, now the character is marked on two ends as a single / isolated character. It is extracted from the word and the search for a new character continues.

C. Character Recognition

Character Recognition has been performed using Neural Networks; the technique used here is described in detail in our paper [6]. All the procedure is the same except that this time the training was done using different forms of a character.

V. CORPORA

For experiments, we collected a corpora consisting of two sets of images (and associated transcriptions): computer generated, i.e. synthetic, images and real-world images consisting of scans of commonly available hardcopy Urdu documents which do not contain any other language.

VI. Assumptions Made

During the whole process of segmentation and character recognition the input script is assumed to be diacritic (Erabs) free. The font's size has been kept fixed or the image has been resized to make the fonts suitable for segmentation and recognition.

VII. Results

Old and newly written scripts were used to evaluate results on good and bad quality paper which produced results on the average as 93.4%, which can further be improved using a lexicon and focusing more onto frequently used characters in the script.

VIII. Future Directions

This Urdu character recognition system is developed on diacritic (Erabs) free Urdu text. Further research is needed to develop a system that recognize diacritic of Urdu, Arabic and other languages having the same properties, with an integrated lexicon to further improve the results.

IX. Conclusion

This paper describes a system for OCR of printed Urdu script. The recognition accuracy of our prototype is promising, but more work is needed. Our character segmentation method should include handling a larger variety of characters including roman script that occurs often in images obtained from Urdu documents. We also need to recognize characters having diacritic to make it a complete OCR system. In general, the system needs to be tested and fine-tuned on a wider variety of images containing characters in diverse fonts and size.

References

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

[2] Raymond G. Gordon, "Ethnologue: Languages of the World Fifteenth Edition" SIL International, 2005.

[3] Khalid Saeed, "New Approaches for Cursive Languages Recognition: Machine and Hand Written Script and Texts".

[4] K. Saeed, Three-Agent System for Cursive Script Recognition, " Proc. CVPRIP '2000 Computer Vsion, Pattern Recognition and Image Processing-5th Joint Conf. on Information Sciences, JCIS'200, Vol.2, PP.244-247, Feb 27-March 3, N.Jersry 2000.

[5] K. Saeed, R Niedzielski, "Experiments on Thinning of Cursive-Style Alphabets, "Inter Conf. on information Technologies ITESB '99, June 24-25, Minsk 1999.

[6] Inam shamsheer, Zaheer Ahmad, Jehanzeb Khan Orakzai, Awais Adnan, "OCR For Printed Urdu Script Using Feed Forward Neural Network," MLPR 2007 :International Conference on Machine Learning and Pattern Recognition", 2007.