Advances in Computational Research
ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, 2012, pp.-117-121
Available online at http://www.bioinfo.in/contents.php?id=33

RECOGNITION OF OFFLINE HANDWRITTEN ISOLATED URDU CHARACTER

IMRAN KHAN PATHAN*, ABDULBARI AHMED ALI AND RAMTEKE R.J.

Department of Computer Science, North Maharashtra University, Jalgaon, MS, India
*Corresponding Author: Email- imk.pathan@yahoo.co.in, abdulbariaa@yahoo.com, rakeshramteke@yahoo.co.in

Received: February 21, 2012; Accepted: March 06, 2012

Abstract- This paper presents an approach for recognition of offline handwritten isolated Urdu character based on Invariant Moments. Handwritten Urdu character recognition is lagging behind due to segmentation dilemma and complexity of Urdu letter writing. An attempt is made to apply Moment Invariant technique followed by Primary and secondary component separation. The Urdu letters were grouped into single component and multi-component characters. If letter is multi-component then Secondary component were separated from primary component. SVM is adopted for classification and position of secondary component (Above, Below and middle) is considered for recognition. For each of 46 characters 200 image samples were used for training and 600 for testing respectively. In this manner overall 36800 handwritten characters were used to apply the technique. Overall performance rate is found to be 93.59% for all offline handwritten isolated Urdu characters. It is possible to enhance the accuracy of system by combining more structural and statistical features.
Keywords- Primary Component, secondary Component, offline Handwritten Urdu isolated characters, feature extraction

Citation: Imran Khan Pathan, Abdulbari Ahmed Ali and Ramteke R.J. (2012) Recognition of Offline Handwritten Isolated Urdu Character. Advances in Computational Research, ISSN: 0975-3273 & E-ISSN: 0975-9085, Volume 4, Issue 1, pp.-117-121.

Introduction

In the Race course of consistent development and evolution human has upgraded knowledge associated with numerous aspects of life. To spread the acquired knowledge to next generations related with different fields, it is frequently stored in handwritten and printed form. With advancement in printing technology and its eventual introduction to the world, the volume of printed material has skyrocketed. With the progress of optical character recognition technology, now it is possible to scan documents as an image and to make it editable and searchable for further information processing. However, for most of the languages there is no efficient way to search through this printed or handwritten material quickly and efficiently due to unavailability of robust character recognition system. Urdu is one these languages which need robust Character recognition system to convert huge handwritten as well as printed data into editable form. Development of robust Urdu OCR system falls behind due to word and character segmentation dilemma.

As compared to other languages very limited research work is done in Urdu character recognition, due to segmentation dilemma and varies in character shape with respect to changing in its position most of researchers focuses on machine-printed Urdu text. Some degree of work is done in printed isolated character recognition for specific font and size, no complete work exists in Urdu word recognition. Work on printed Urdu script OCR is continue in Phase II of [1], in which segmentation free approach will be followed where the ligature as a whole is used instead of segmenting it into smaller units because character segmentation is consider one of the very difficult task in Urdu. Center for Research in Urdu Language Processing (CRULP) is also doing research and development in linguistic and computational aspects of Urdu, has also launched corpus and associated tools for Urdu text processing.

Pal and Sarkar [2] present an approach for recognizing printed Urdu characters and numerals using binary-tree classifier followed

by water-reservoir features to classify similar looking characters into subsets. They achieved an accuracy of 97.8% on 3050 characters and numerals [3]. S. Hoque [4] proposed an offline OCR using chain code quantization approach, [5] Proposed a Nastaliq character recognition system by using artificial neural network. Nain [6] present an edge recognition approach using first order difference chain coding that actually represents the curvature of a planner curve. Syed [7] introduce a recognizer for Urdu NooriNastaliq script using ligature based approach instead of character based. In this work a multi-tier approach has been utilized. Shah [8] also presented an OCR for Urdu Nastaliq font. It was a ligature based recognizer but use template matching technique for classification. Both Arabic and Urdu text recognition systems have to address similar challenges, thus review of Arabic text recognition research evaluation may would be helpful in case of Urdu character recognition system. In [9] a fruitful review of various methodologies for Arabic handwriting recognition can be found. Maged [10] used geometrical features to describe the complete skeleton of character for an automatic recognition system of hand written Arabic characters. Amin [11] focus on thinning process and feature extraction and also discuss the problems come across in producing satisfactory thinned form of text and describe solutions to those complexity.
Limited research work is done for Urdu handwritten Character recognition.

Overview of Urdu:
Presently, Urdu is the official language of Pakistan and one of the sixteen major languages constitutionally recognized in India [12]. Urdu derived 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; as shown in figure 1 (b) total of 41 alphabets Urdu shares a common script and many characteristics of Arabic script with additional set of alphabets [13].

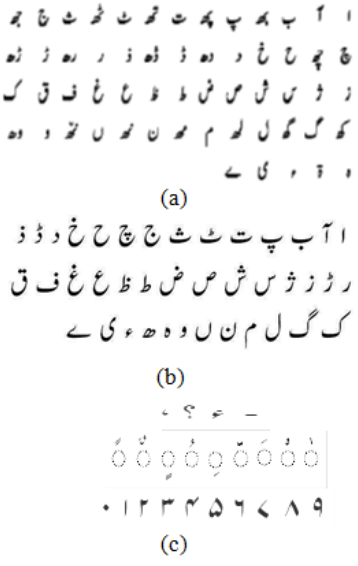

Fig. 1- (a) 58 character set defined by N.L.A.P (b) Basic Urdu Character set (c) punctuation, diacritic marks, and numerals

Each letter has multiple forms depending on its position in the word. Each letter is drawn in an isolated form when it is written alone, and is drawn in up to three other forms when it is written connected to other letters in word. For example letter Khah has four forms: isolated (خ), initial (خ), medial (خ) and final(خ). Some of the letters in Urdu has two forms. Possible shapes of all characters are given in Table-1
Another unique feature of Urdu is that the Urdu words are usually written without short vowels or diacritic symbols. Any machine transliteration or text to speech synthesis system has to automatically guess and insert these missing symbols. This is a non-trivial problem and requires an in-depth statistical analysis [14].

Table-1 Different forms of characters depending on its position in the word

Sr. No	Isolated	Initial	Middle	Ender
1	آ			
2	ا			ا
3	ب	بـ	ـبـ	ـب
4	پ	پـ	ـپـ	ـپ
5	ت	تـ	ـتـ	ـت
6	ٹ	ٹـ	ـٹـ	ـٹ
7	ث	ثـ	ـثـ	ـث
8	ج	جـ	ـجـ	ـج
9	چ	چـ	ـچـ	ـچ
10	ح	حـ	ـحـ	ـح
11	خ	خـ	ـخـ	ـخ
12	د			ـد
13	ڈ			ـڈ
14	ذ			ـذ
15	ر			ـر
16	ڑ			ـڑ
17	ز			ـز
18	ژ			
19	س	سـ	ـسـ	ـس
20	ش	شـ	ـشـ	ـش
21	ص	صـ	ـصـ	ـص
22	ض	ضـ	ـضـ	ـض
23	ط	طـ	ـطـ	ـط
24	ظ	ظـ	ـظـ	ـظ
25	ع	عـ	ـعـ	ـع
26	غ	غـ	ـغـ	ـغ
27	ف	فـ	ـفـ	ـف
28	ق	قـ	ـقـ	ـق
29	ک	کـ	ـکـ	ـک
30	گ	گـ	ـگـ	ـگ
31	ل	لـ	ـلـ	ـل
32	م	مـ	ـمـ	ـم
33	ن	نـ	ـنـ	ـن
34	ں			
35	و			ـو
36	ہ	ہـ		ـہ
37	ھ			
38	ء			
39	ی	یـ	ـیـ	ـی
40	ے	یـ	ـیـ	ـے
41	بھ			
42	پھ			
43	تھ			
44	ٹھ			
45	جھ			
46	چھ			
47	کھ			
48	گھ			

But generally characters acquire one of these four shapes, namely isolated, initial, medial and final. Urdu characters can be divided into two groups, separators and non-separators. These are also known and non-joiners and joiners respectively. The separators or non-joiners can acquire only isolated and final shape. On contrary non-separators or joiners can acquire all the four shapes. The isolated form of each of these alphabet as shown in Fig 2(a) are not joinable or cannot be connected with other Fig 2 (b) shows alphabets which are joiner and may join with other characters.

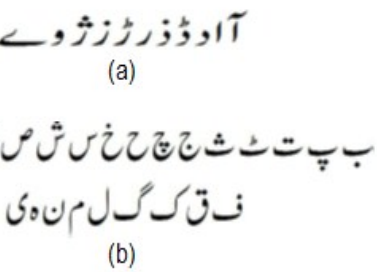

Fig. 2- (a) Non joiner characters (b) joiner Characters of Urdu

Characteristics of handwritten Urdu Letters

Urdu characters generally comprise of minimum one and maximum four components. The letters can be grouped into single component letters and multi- component letters, further multi-components may be grouped into primary components and secondary components. Position and type of secondary components play vital role for identifying the characters. For example characters Say (ش) and Pay (پ) consist same primary component (i.e. س), but type and positions of secondary components are different, above and below respectively. Type and positions of secondary components can be used as an important feature of Urdu Characters. While considering secondary components as features it is considered that number secondary components may be 1, 2 or 3 and position of secondary components may be above, below or middle. Even some letters can only be distinguished by their secondary components for example, Jeem (ج) and (چ) differ only by the number of dots one or three. In case of printed letters these secondary components will be disconnected from primary components and also secondary components are disconnected from other remaining components of the letters. But in case of handwritten letters it may not work because there are important variations in drawing the secondary components as shown in Table-3.

Table 2- Component based Urdu letter

Without Secondary components		ا ح د ر س ص ط ع ل م ں و ه ھ ء ی ے
With Secondary components	**Secondary components Above**	آ ت ٹ ث خ ڈ ذ ڑ ز ژ ش ض ظ غ ف ق ن گ
	Secondary components Within	ج چ
	Secondary components Below	ب پ

There are important variations in writing Urdu Characters specially drawing secondary components; mostly in drawing two dots and three dots [15]. As shown in Table 2 Samples A1, B1, and F4, three dots come in three variations Broken characters arises challenges in character recognition, as shown in E5 and F6 ,'Do-Chasmi-Hey' are written with two loops and with one loop respectively. Also in A2 and A3 characters 'Wow' is written with loop and without loop respectively which changes the structural features and feature extraction become tricky

Table 3- Samples showing variations in handwritten letters

	A	B	C	D	E	F
1	ج	چ				
2	و	و		گ	گ	گ
3	س	س	س		پ	پ
4				پ	ب	ب
5	س	س			ں	ں

System Methodology:

In this paper Moment Invariants (MI) are used to evaluate seven distributed parameters for handwritten isolated Urdu characters. In any character recognition system the characters are proposed to extract features that uniquely represent properties of the character [16]. The MIs are well known to be invariant under tralsation, rotation, scaling and reflection. They are measures of the pixel distribution around the center of gravity of the character and allow capturing the global charatershpe information. In the present work, the moment invariatnsar are evaluated s using centeralmoment of the image function f(x,y) up to third order[17]. Urdu Characterscanbe grouped into single component and multi component characters as shown in Table 2.

Initially it is verified whether character consists of single component or more than one component. If character is single component then image is normalized into 60 X 60 and divided into three horizontal zones for features extraction as shown in figure 3 (b). From each zone 7 Moment Invariant features and from whole image 7 Moment Invariant features were computed, in this manner28 Moment Invariant features were determined.

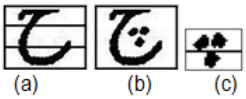

Fig. 3- (a) Character Cheem (b) Primary component of Cheem (c) Secondary component of Character cheem.

Afterwards SVM is used for classification and character is put into appropriate class and finally recognized the single stork component character. If Character belongs to multi component character groups shown in Table 3 - Samples A1, D4 and F3, first primary and secondary component are separated and stored on different locations. The image of primary stroke is normalized into 60 X 60 by maintain height and width ratio [18] and further divided into 3 zones. The secondary component is normalized into 22 X 22 and divided into 2 horizontal zones as shown in figure After separating character into primary and secondary two alternative were considered first is counting number of secondary strokes and their types like (ب)with one dot below, pey(پ) with three dots below, say (ش) three dot above and Tte (ٹ) one toe above. But it may give incorrect results in case of handwritten character recognition due to variation in writing secondary components. For example characters Pay (پ) should consists of a primary component (i.e. س) and secondary components (i.e. three dots below). But in case of handwritten characters the structure of secondary components varies from writer to writer. The variation in writing Character Pay (پ) is shown in Table 3 - D4, E4 and F4 are which should have three dots below primary component (س) but E4 consist one dot below to a zigzag, and F4 have (ں) as secondary component.

Counting number of secondary strokes and their types may give unsatisfactory results in case of handwritten characters. That why Moment Invariant of secondary component are considered as a feature instead of number of secondary components and types. The segmented secondary components were normalized into 22 X 22. Then normalized image is divided into two horizontal zones for

features extraction as shown in figure3 (c). From each of two zone 7 Moment Invariant features and from whole image 7 Moment Invariant features were computed, in this manner 21 Moment Invariant features were determined. Later using SVM the type of secondary component is recognized which could be one of the component shown in figure 5.

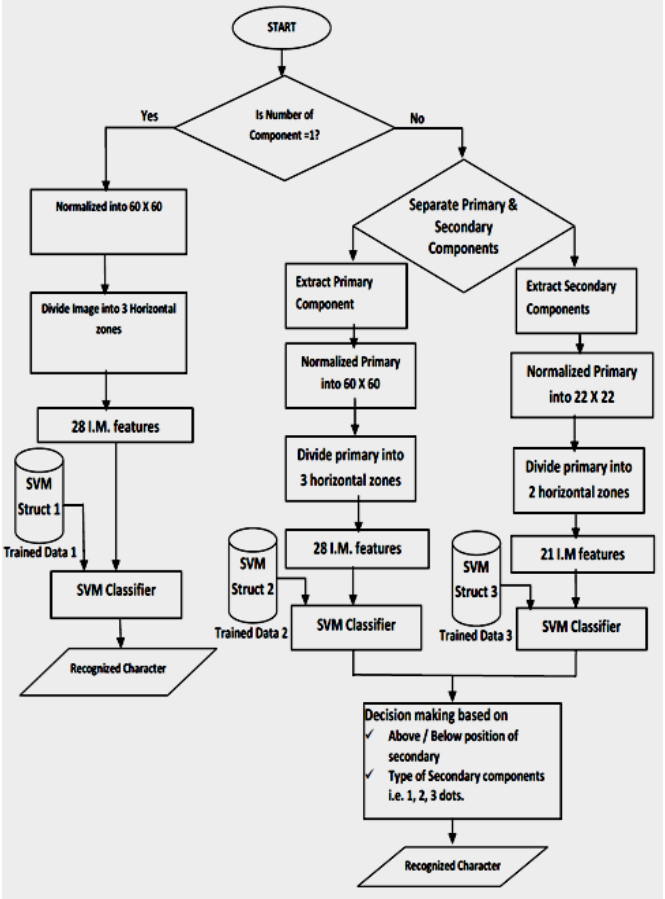

Fig. 4- System Architecture for Feature Extraction of Handwritten Isolated Urdu Character

Table 4 - Accuracy of Secondary components

Secondary Sahpe	Name	Recognition rate
	One dot	98.97
	Two dot	94.42
	Three dot	96.15
	Toe	94.88

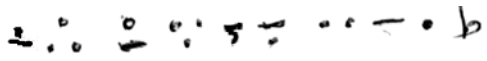

Fig. 5- Possible Primary Components

The position of secondary component is also one of the important features in Urdu character recognition as shown in Table 2. Due to its importance with Invariant Moment of Secondary component we have also considered its position whether Above, Below or middle. Based on 28 features of primary component and 21 features of secondary component further decision making is conceded with the help position of secondary components and character is recognized as shown in Figure 4.

Result and Conclusion

The proposed method is applied on 36800 handwritten Urdu characters. For each of 46 characters 200 image samples were used for training and 600 for testing respectively. The characters were grouped into single component and multi-component. Further multi-component is segmented into primary and secondary component. Finally invariant moments of primary as shown in Table-4 and secondary component were calculated and classified using SVM. For decision making position of secondary component i.e. Above, Below or Middle is used with combination of SVM result and character improvingly recognized as shown in table 5.

Table 5 - Recognition Rate of all Urdu Characters

Sr. No	Letters	Recog-nition	Sr. No	Letters	Recog-nition
1	Alif	99.35	24	Ghain	91.79
2	Bey	95.64	25	Fay	91.74
3	Pay	98.38	26	Quaf	88.70
4	Tey	95.90	27	Kaaf	89.95
5	Tay	96.36	28	Gaaf	92.59
6	Say	90.71	29	Laam	98.44
7	Jeem	91.42	30	Meem	90.29
8	Chey	86.47	31	Noon	91.14
9	Hay	89.95	32	Noon_Gunna	96.20
10	Khay	94.85	33	Waow	83.73
11	Daal	98.88	34	Hey	88.74
12	Dhal	89.15	35	Do Cha. Hey	95.14
13	Zaal	90.15	36	Hamza	94.59
14	Ray	93.54	37	Choti Ye	99.14
15	Dhy	84.86	38	Badi ye	94.99
16	Zay	88.26	39	Bha	96.87
17	Seen	94.39	40	Pha	94.88
18	Sheen	94.98	41	Tha	99.14
19	Suad	98.49	42	Ttha	94.99
20	Zuad	96.68	43	Jha	89.66
21	Toe	97.16	44	Cha	94.88
22	Zoe	97.51	45	Kha	91.14
23	Ain	99.78	46	Gha	93.35
Average Recognition Rate					93.59

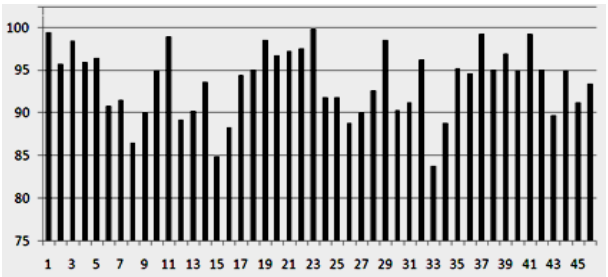

Chart of Character Recognition Rate shown in Table 4

Conclusion

In this paper, an attempt is made to apply Moment Invariant based feature extraction technique followed by Primary and secondary component separation. No standard data set of isolated Urdu Character is available therefore dataset of 36800 isolated charac-

ters is created. The Urdu letters were grouped into single component and multi-component characters and passed through system as mentioned in figure 4. We achieved overall accuracy up to 93.59% for all offline handwritten isolated Urdu characters. Enhance in recognition rate could possible by combining more statistical and structural features with.

References

[1] Project proposal (2011) "*Software Requirement, Design, & Testing documents, Development of Robust Document Image Understanding System for Documents in Indian Scripts Phase II" ,Sponsored By, Ministry of Communication & Information Technology, Govt. of India*.

[2] Pal U. and Anirban Sarkar (2003) 7*th International Conference on Document Analysis and Recognition,* 1183-1187.

[3] Lorigo L.M., Govindaraju V. (2006) *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5).

[4] Hoque S., Sirlantzis K., Fairhurst M.C. (2003) 7*th International Conference on Document Analysis and Recognition*, 2, 834-838.

[5] Husain S.A. and Amin S.H. (2002) *International Multi Topic IEEE INMIC*.

[6] Nain N., Laxmi V., Bhadviya B. (2007) 3*rd International IEEE Conference,* 821- 825

[7] Syed Afaq Hussain and Syed Hassan Amin (2002) *International Multi Topic IEEE INMIC.*

[8] Zahra A. Shah and Farah Saleem (2002) *International Multi Topic IEEE INMIC.*

[9] Liana M. Lorigo, Venugopal Govindaraju (2006) *IEEE Trans. Pattern Anal. Mach. Intell*. 28(5), 712-724.

[10] Maged Mohamed Fahmy and Maged Mohamed (2001) *Studies in Informatics and Control*, 10(2).

[11] Mandana Kairanifar and Adnan Amin (1999) 5*th International Conference on Document Analysis and Recognition,* 213.

[12] C.M. Naim (1999) *Introductory Urdu - Volume I, Book Published by National Council for Promotion of Urdu Language*.

[13] Zaheer Ahmad, Jehanzeb Khan Orakzai, InamShamsher and Awais Adnan (2007) *World Academy of Science Engineering and Technology*, 26.

[14] Gurpreet Singh Lehal (2010) 23*rd International Conference on Computational Linguistics*.

[15] Gheith A. Abandah and Mohammed Z. Khedher (2009) International Journal of Computer Processing of Languages, 22(1), 1-25

[16] Ramteke R.J., Mehrotra S.C. (2008) *International Journal of Computer Processing of Oriental Languages*.

[17] Ramteke R.J., Mehrotra S.C. (2006) *IEEE Conference on Cybernetics and Intelligent Systems,* 1-6.

[18] Abdulbari Ahmed Ali, Ramteke R.J. (2011) *International Journal of Machine Intelligence*, 3(3), 116-120.