*Sindh Univ. Res. Jour. (Sci. Ser.) Vol.46 (2):143-152 (2014)*

**SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)**

# Issues and Challenges in Sindhi OCR

D. N. HAKRO[++], I. A. ISMAILI*, A. Z. TALIB, Z. BHATTI*, G. N. MOJAI**

School of Computer Sciences, University Sains Malaysia, Malaysia, PO box 11800, USM Penang, Malaysia

Received 3rd April 2014 and Revised 13th May 2014

**Abstract:** Optical Character Recognition (OCR) is the reading (recognition) of a written or printed document. Many of the languages are enriched with the OCR but OCR is lacking in Sindhi Language which has a golden 5000 year history. OCRs for some of the languages including Latin script and some other languages with isolated characters (non-cursive) are easy to develop whereas developing an OCR for a cursive language and a language possessing a large set of characters such as Sindhi is a challenging job. Sindhi Language has 52 characters as compared to 28 in Arabic, 32 in Persian and 39 in Urdu. This paper presents the various scripts of Sindhi Language including very old scripts, and issues and challenges in Sindhi OCR posed by cursive nature and other features of the current standard script. The main challenges include cursiveness, more characters dots, and variation of the placement and orientation of dots, four dotted characters, a large set of characters for recognition, Unicode representation, more base shape group characters, same base shape with variation in number and placement and orientation of dots, ambiguity between the characters with very slight difference, more characters with dots, context sensitive shapes, ligatures, noise, skew and fonts in Sindhi OCR. We also provide a summary of issues and challenges for the development of Sindhi OCR. This summary is useful for the researchers of OCR as well on Sindhi computing.

**Keywords**: Optical Character Recognition, Character Dots, Sindhi, Script, Character shapes.

## 1. INTRODUCTION

OCR is considered as a way in which data in the form of text image can be transformed into digital format in order to allow a faster method for data entry (Pal and Sarkar, 2003). Character recognition is much easier for non-cursive scripts (such as Latin), compared to cursive (connected) ones. Significant research solutions have been proposed for recognitions of both types of script; cursive and non-cursive (Jelodar *et al.,* 2005). Arabic script and script of languages adopting it still need much attention (Parvez and Mahmoud, 2013; Slimane *et al.,* 2013) because development of OCRs for such scripts are still in infancy compared to development of OCRs for the Latin script. Sindhi Language is one of the languages adopting the Arabic script and inherits some of its characteristics but with bigger number of characters; more differentiating characters, more characters with dots and more variety of placement as well orientations of dots. Thus more and greater challenges are expected in the development of Sindhi OCR.

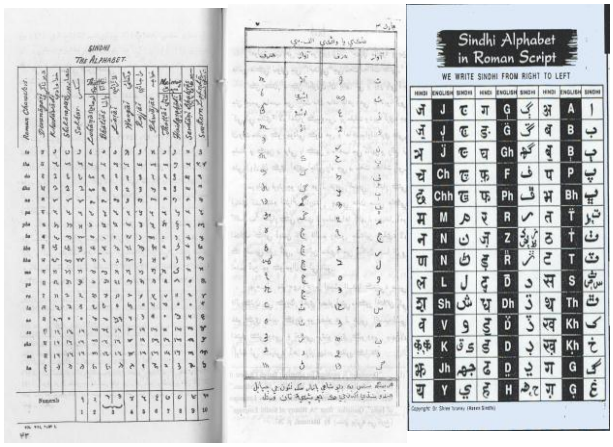
Sindhi Language is considered as one of the seven languages of heaven (Alana, 2004) and has a 5000 year old history (About Indus Civilization, 2013). Sindhi script follows the same script and rules of bi-directional writing of the Arabic script like other Perso-Arabic scripts. These scripts are written from right to left and for writing numbers the writing is from left to right. For writing, the writing pen is also moved backward to form a cursive character shape. The shapes of an individual character also change according to their position and connection with the preceding or succeeding characters. The main challenge is in cases where there is no difference in the base shape of the character and the difference is only in the number of dots and placement of the dots diacritics in forming a ligature.

## 2. MATERIAL AND METHODS

### A. Sindhi Script

Sindhi was written in many forms but over time three types of script have been used commonly. As Sindhi Language possesses a very rich history, various types of script were used in different regions; some of them are as shown in **(Fig.1)** and the most commonly used scripts are shown in Fig. 1(c). Sindhi Language script has a sum total of 52 letters in its alphabet as compared to its parent language Arabic with 28 letters, and its neighboring languages such as Persian (32), Urdu (39) and Pashto (44). In this regard, Sindhi constitutes the largest extension of the original Arabic script.

[++]Corresponding Author E-mail: dill.nawaz@gmail.com, iai_a@yahoo.com, azht@cs.usm.my, zeeshan.bhatti@live.iium.edu.my ghulam_itprof@yahoo.com, Corresponding author: Dil Nawaz, email:dill.nawaz@gmail.com
*Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
**Institute of Mathematics and Computer Science (Bioinformatics),University of Sindh, Jamshoro, Pakistan




(a) (b) (c)

**Fig. 1: Previous scripts for Sindhi language**

**(a) khudabadi, khowjki etc,**
**(b) Initial script (Wanki Alif-bay ( واڻڪي ا ب )),**
**(c) Currently used script (copyright: Shree Israney)**
**copyright: (a) & (b) Sindhi Sooratkhati (Allana, 1993).**
**Copyright: (c) Dr.Shree Israney**

## B. Writing System In Sindhi Language

The writing system for Sindhi starts from right to left for the characters, whereas for the numbers, the system is from left to right as depicted in **(Fig. 2)**. This poses a challenge for the development of the OCR for the language because at the time of recognition, if a number comes between the characters then the output writing mechanism must be reversed.

دلنواز جي پيدائش جي تاريخ ١٩٧٩ / ٠٢ / ٢٦ آهي

Dil Nawaz je paidaish jee tarikhh 26 / 02 / 1979 aahe (Roman)

**Fig. 2: Writing direction of Sindhi script**

## C. Script Complexitites In Sindhi Ocr

Very little work has been done on Sindhi Language OCR and Sindhi OCR development has just started as the Sindhi script poses more challenges because of the large character set and its complexities. The cursiveness and context sensitivity are the two major problems faced by a number of Arabic script OCR developers around the world. These problems are explained here with regard to the Sindhi script together with some suggestions.

## D. Cursiveness

Sindhi script, like Arabic script is a script with most of the characters, have more than one shapes when forming a word, ligature or a compound word. **(Fig. 3)** shows examples of words that may be formed from one or more ligatures and isolated characters.

سنڌ يو نيو ر سنّي دلنواز

**Fig. 3: Ligatures and isolated characters**

In **(Fig. 3)** the word "يونيورسنّي " (university) is a combination of three ligatures "يو " , "نيو " and "سنّي " as well as an isolated character "ر ". This word can be decomposed into its basic components as depicted in **(Fig. 4)**.

يو نيو ر سنّي = يو + نيو + ر + سنّي
يو = ي + و A Ligature
نيو = ن + ي + و A Ligature
ر = ر An Isolated Character
سنّي = س + ڻ + ي A Ligature

**Fig. 4: Word formation: from ligatures and isolated characters**

Since Sindhi script has more characters, more words can be formed and more combinations of words can result in more variety of compound words. With regard to Sindhi OCR implementation, more combinations of words pose more problems in recognition.

## E. Importance Of Dots In Sindhi Script

In Sindhi script, the dots are so important in extending the original Arabic script. In Sindhi changing the number of dots, and placement and orientation of dots can make another character. Sindhi script made use of two accents; one is dot and another is small " ط " as in " ڻ " . The character, "ڻ" is written in front of an imperative to make a verb in Sindhi Language. The details on dots and accents in Sindhi characters are given in Table 2. There are more characters with dots than those without dots.

## F. Number of Dots

The characters in Sindhi script possess one, two, three or four dots (most of the Arabic extended scripts only have up to three dots) as given in Table 1. The position of the dot may be above, below or inside the character's base shape. The dots may be in vertical order or horizontal order. The number and location of the dots pose main difficulty in optical character recognition where one dotted character with a noisy dot may look like another character because the characters possess the same shape but with different number of dots. These noisy dots may change one character to another character. Only one dot may differentiate one character from the other. No other language adopting the Arabic script possesses more four dotted characters than Sindhi. Sindhi script also possesses more characters with dots (one to four) than any other scripts. A sum total of 35 characters with dots ranging from one to four dots (one with diacritic) and with different placement and orientation pose many challenges not faced by other OCRs researchers around the world. Characters recognized with missing dot or additional

dot will change the meaning of the sentence, as they have entirely different sounds.

**Table 1: Dots in Sindhi characters**

| Number of Dots | = | Characters |
|---|---|---|
| With single dot | 12 | ب,ج,جھ,خ,ڊ,ذ,ز,ض,ظ,غ,ف,ن |
| With Two dots | 11 | ت,ٹ,ج,چ,ڈ,ڊ,ق,ڪ,گ,ي |
| With Three dots | 06 | پ,ٽ,ٿ,ڇ,ڏ,ش |
| With Four dots | 05 | ٻ,ٺ,ڃ,ڙ,ڦ |
| With small (ط) | 01 | ٺ |
| Without dot | 17 | س,ص,ط,ع,ک,گ,گھ,ل,م,و,ه,ء,ا,ح,د,ر, |
| **Total number characters** | **52** | |

## G. Placement and Orientation of Dots

The dots are placed at various locations and in a variety of combinations. The number of dots ranges from one to four and the dots are placed at different locations; above, below and inside the base shape of a character. For example in character " **ب** " the dot is placed below the base shape, whereas in "**ض**" the dot is placed above the base shape and for the character "**ج**" the dot is placed inside in. **(Table 2)** depicts the placement of a single dot.

**Table 2: Placement of a single dot**

| S.No: | Placement | Characters |
|---|---|---|
| 1. | Below | ب, جھ, ڊ |
| 2. | Above | خ, ذ, ز, غ |
| 3. | Inside | ج, ن |

In characters with two dots the dots are placed in one of two orientations: vertical or horizontal, and at below, above or inside the base shape of the character. **(Table 3)** shows the placement and orientation of character with two dots. Sindhi is the only script among scripts that adopt the Arabic script including Arabic itself that possesses characters with two dots oriented vertically. Some characters may possess diagonal dots and these dots even change their orientation angle when specific type of font is used as given in Table 3.

**Table 3: Placement and orientation of two dots**

| S.No: | Placement | Direction | |
|---|---|---|---|
| | | Horizontal | Vertical |
| 1. | Above | ت, گ | ٹ |
| 2. | Below | ي | ب, ڪ |
| 3. | Inside | چ | ڃ |

The set of three dots in Sindhi script are placed either in upward or downward direction in a triangle shape as shown in Table 4. A different orientation of three dots also exists in Sindhi script such as the second entry of Table 4 where a three dot orientation pointing downward above a character besides the normal upward orientation. The former orientation is not used in the scripts of neighboring languages and Arabic.

**Table 4: Placement and orientation of three dots**

| S.No: | Placement | Orientation | Character |
|---|---|---|---|
| 1. | Above | Upward | ث, ش |
| 2. | Above | Downward | ٽ |
| | | | |
| 3. | Below | Upward | No Character |
| 4. | Below | Downward | پ |
| | | | |
| 5. | Inside | Upward | No Character |
| 6. | Inside | Downward | ڃ |

Four dots are used specifically in Sindhi script to represent a character and are written inside, below or above the character base shape, as shown in Table 5. This poses another challenge since these dots are very near to each other and it is even more challenging when these dots are inside a character as in the third entry of the **(Table 5)**.

**Table 5: Placement of four dots**

| S.No: | Placement | Characters |
|---|---|---|
| 1. | Above Character | ڻ,ڙ,ڦ |
| 2. | Below Character | ٻ |
| 3. | Inside Character | ڃ |

A summary of various positions and orientations of dots in Sindhi Script is given in Table 6. We can conclude that there are more characters with dots placed above the character base shape.

**Table 6: Placement of dots**

| Placement of the Dot | No: | Characters |
|---|---|---|
| Dots above characters | 19 | خ,ذ,ڈ,ذ,ڙ,ز,ش,ض,ظ,غ,ف,ڦ,ق,گ,ت,ٹ,ٹ,ٹ,ٹ,ٹ |
| Dots below characters | 09 | ب, ٻ, پ, جھ, ڊ,ڊ,ڳ,ي |
| Dots in between characters | 06 | ج,ڃ,ڃ,ڃ,چ,ن |
| With diacritic (ط) | 01 | ٺ |
| Without dot | 17 | ص,ط,ع,ے,ک,گ,گھ,ل,م,و,ه,ء,ا,ح,د,ر,س |
| **Total Characters** | **52** | |

## H. Character Number

Character is regarded as the smallest unit of a writing system. So, while digitizing characters, every character is represented by one corresponding code. For example, character "D" is represented with 0044 and "د " with 062F in the Unicode table. But, for Sindhi two characters " جھ" and " گھ", are represented by two Unicode characters compared to traditional one code per character. The character "جھ" is represented by two Unicode characters " ج " and " ھ ", whereas " گھ " is represented with two Unicode characters " گ " and " ھ ". This poses another challenge in recognizing Sindhi characters since a classifier may produce errors in recognizing these characters because it is very difficult to decide whether these characters are going to be recognized as a single character or as two separate characters. For example, The character "گھ " can be recognized as either a single character or two separate characters "گ " and "ھ".

## I. Unicode Placement

Arabic letters are placed in consecutive order whereas the extended characters for scripts adopting the Arabic script are placed at various different locations. As such, letters which are extended in Sindhi script and not available in the Arabic script are placed at various locations in the Unicode table. Designers have given two ranges: 1) Arabic-Indic digits (U+0660 - U+0669) designated for Arabic use and 2) (U+06F0 - U+06F9) Eastern-Arabic Indic digits, employed by a number of different languages such as Sindhi and Urdu. This poses another problem in searching and sorting of the Sindhi characters which may be necessary after recognition when implementing grammar checking and other facilities in the OCR.

## J. Character Shape Groups

Sindhi scripts, as well as its mother script Arabic use a combination of base shapes and dots. These shapes are combined into shape groups that may or may not contain the dots as well as the diacritic marks to produce different sound. There are a total of 52 characters and they are divided into different shape groups as shown in **(Table 7)**.

**Table 7: Base shape groups in Sindhi**

| S.No: | Base Shape | Shape Set | No: of Elements | | S.No: | Base Shape | Shape Set | No: of Elements |
|---|---|---|---|---|---|---|---|---|
| 1 | ا | ا | 1 | | 12 | ق | ق | 1 |
| 2 | ب | ب، ٻ، پ ،ٺ،ٽ،ٿ، ٹ ،پ | 9 | | 13 | ڪ | ڪ | 1 |
| 3 | ح | ج , چ , ڃ , چ , ڇ , ح , خ | 7 | | 14 | ک | ک،گ،ڳ،گ | 4 |
| 4 | حه | جھ | 1 | | 15 | گھ | گھ | 1 |
| 5 | د | د,ڈ,ڈ, ڊ , ڍ,ذ | 6 | | 16 | ل | ل | 1 |
| 6 | ر | ر ، ڑ ، ز | 3 | | 17 | م | م | 1 |
| 7 | س | س ، ش | 2 | | 18 | ں | ن , ڻ | 2 |
| 8 | ص | ص ، ض | 2 | | 19 | و | و | 1 |
| 9 | ط | ط ، ظ | 2 | | 20 | ه | ھ | 1 |
| 10 | ع | ع ، غ | 2 | | 21 | ء | ء | 1 |
| 11 | ف | ف ، ڦ | 2 | | 22 | ى | ي | 1 |
| **Subtotal** | | | **37** | | **Subtotal** | | | **15** |
| **Total number of characters** | | | | | | | | **52** |
| **Total number of group shapes** | | | | | | | | **22** |

With a particular base shape, the number, placement and orientation of dots make different characters in Sindhi script. shows characters formed from the second base shape of **(Table 8)** where characters may be easy to discern by human beings but it can be very challenging for the OCR. Nawaz *et al.* (2009) used morphology techniques (Soille, 2003) to remove dots and diacritics by using pepper noise and thinning process (Gonzalez *et al*,. 2nd ed) respectively on binary images.

ب ٻ پ ٺ ٽ ٿ ٹ پ

**Fig. 5: Same base shape with different number of dots, and their positions and orientations**

It is a very difficult and challenging task to decide whether a dot is actually a dot or a noise. It is

very difficult for the recognition engine to recognize the character such as "ب". The second letter in Sindhi alphabet will become a third letter "پ" and the fifth letter "ت" will become seventh letter "ٹ" in Sindhi alphabet with just a dot added to the letter. A single dot can change the character and thus the corresponding word and sentence. The Sindhi language authorities can therefore apply new combinations of dots (i.e. number, placement and orientation) to form more letters, and thus the alphabet can be extended further and more sounds can be produced if needed.

## K. Context Sensitive Shapes In Sindhi Script

The connection between characters with its preceding or succeeding characters is called context sensitivity (Sattar *et al.*, 2009). A single character can change its shape according to its connecting character which may be the succeeding character, preceding character or both. Table 8 shows the connection of character "ب" as an initial with other characters**.** The development of Sindhi OCR is more challenging due to the large number of characters used for same base shapes. In Arabic script, base shape " ب " is used for only three characters, " ت " , " ب ", and " ٹ " whereas it is used for 9 characters in Sindhi script. Therefore, an Arabic OCR must be capable of recognizing only three characters for this base shape as well their changing character forms, while a Sindhi OCR must be capable of recognizing nine different characters for the same base shape.

**Table 8: Connecting shapes of "ب" as an initial**

| S.No: | Shape Set | Initial (ب) | | S.No: | Shape Set | Initial (ب) |
|---|---|---|---|---|---|---|
| 1 | ا | بَا | | 12 | ق | بق |
| 2 | ب ,ٻ ,پ ,ڀ,ت ,ٹ ,ٺ ,ٽ,ٿ | بت | | 13 | ڪ | بڪ |
| 3 | ج ,ڃ , ڄ ,ڇ ,چ ,ڃ ,ح ,خ | بج | | 14 | ک ,گ ,ڳ ,ګ | بک |
| 4 | جھ | بجھ | | 15 | گھ | بگھ |
| 5 | د,ڈ ,ڈ , ډ , ڊ ,ذ | بد | | 16 | ل | بل |
| 6 | ر , ڑ ,ز | بر | | 17 | م | بم |
| 7 | س ,ش | بس | | 18 | ن , ڻ | بن |
| 8 | ص ,ض | بص | | 19 | و | بو |
| 9 | ط ,ظ | بط | | 20 | ھ | بھ |
| 10 | ع , غ | بع | | 21 | ء | بء |
| 11 | ف ,ڦ | بف | | 22 | ي | بي |

**Context sensitive shapes of base shape " ا "**

The first letter of the Sindhi alphabet "ا" can be connected to one character whereas on some occasions it cannot. For example, it can be connected to a letter "ت " and "ک " as in the word " قَتَا کا " but the same letter "ا" cannot be connected with "ف " as in the word "لفافو ". In the letter, the character "ا" is not connected with the letter "ف ", which is a succeeding character but is connected to its preceding character "ف " which is already connected with the letter "ل ", which clearly shows that the letter " ا " can be connected with a preceding character but not with a succeeding character.

**Context sensitive shapes of base shape " ب "**

This group of letters comprises of nine characters in Sindhi script and the rule is applied to all of the characters in the same way. Any of these characters will remain in standalone (isolated) form if they cannot be connected to a preceding character and the character is the last character such as "ڪتاب", where the character "ب " is not connected with its preceding character " ا ". The character "ت" of the same family can be connected to both its preceding character "ک" and its succeeding character " ا ". Another example of connecting "ب" with the succeeding and preceding characters is "سبب"; where "ب" is connected to both characters, if it is in between. The character "ب" uses its final shape when it is connected as the last character.

In another case, the character "پ" in the word "سوپارو",is not connected with the preceding character " و ", and uses its first case and connects with the character " ا " as it is capable of connecting with a preceding character. Some examples of connecting these group shapes are given in Table 9.

**Table 9: Connecting shapes of base shape "ب"**

| S.No: | Character | Isolated | Final | Medial | First |
|---|---|---|---|---|---|
| 1 | ب | ڪتاب | سبب | سبب | بدڪ |
| 2 | ٻ | راڄ | ڇڇ | سٻاجھو | ٻڪ |
| 3 | پ | اپ | سٽپ | سٽيو | پٽ |
| 4 | ت | عورت | اُفت | رستا | تارو |
| 5 | ٽ | رٽ | منٽ | گٽو | ٽلھو |
| **S.No:** | **Character** | **Isolated** | **Final** | **Medial** | **First** |
| 6 | ٺ | اٺ | ٿونٺ | ننڍجڻ | ٿونٺ |
| 7 | ٽ | سربڻ | قٽ | قٽجڻ | ٽبي |
| 8 | ٿ | وارٿ | بحث | مٿبت | ٿواب |
| 9 | ڀ | آلاڀ | گڀ | ٿڀجڻ | ڀگ |

| Character | Isolated | Final | Medial | First |
|---|---|---|---|---|
| ا | ساوا | ترانا | بابا | انب |
| جھ | ہاجھ | ڳڇھ | جھجھو | جھرڪي |
| ر | مار | ماهر | مرڪڻ | رازو |
| ڙ | ڙ | ميهڙ | مڙهه | ----- |
| ز | راز | رمز | رازو | زمين |
| س | اس | بس | ماسڙ | ساز |
| ش | فاش | معاش | شوشو | شخص |
| ص | مخصوص | شخص | مخصوص | صاد ر |
| ض | عرض | مرض | مرضي | ضعيف |
| ط | خطاط | خط | طوطو | طرز |
| ظ | محظ | غليظ | مظلوم | ظلم |
| ع | شعاع | شمع | رعايا | عينڪ |
| غ | داغ | بالغ | شغل | غلام |
| ف | معاف | معاف | مفروزو | فروش |
| ق | اوطاق | متعلق | مقروظ | قلم |
| ڪ | اڪ | مهڪ | مڪتب | ڪعبو |
| ک | رک | مک | سکر | ڪٽ |
| گ | رنگ | ونگ | وانگر | گانو |
| گھ | اگھ | گھوگھو | اگھور | گھوڙو |
| ڳ | رڳ | ماڳ | مرڳو | ڳرو |
| ڱ | لڱ | سڱر | سڱ | ----- |
| ل | عيال | محل | مليل | لڪير |
| م | معلوم | رحم | رحمت | مفهوم |
| ن | ملعون | فن | منع | نالو |
| ڻ | وڻ | ٻڌڻ | وڻڻ | ----- |
| و | مهرو | ليڪو | آواز | وڏيڪ |
| ء | ----- | ----- | ----- | ----- |
| ھ | آگاھ | ملھ | وهندڙ | هڪڙو |
| ي | واري | آلي | ايڪو | يگانو |

**Context sensitive shapes of base shape "ح"**

There are a total of seven characters for this group shape and their shape changes when connecting to other characters. Examples are given in Table 10. Sindhi alphabet comprises of seven characters written with base shape "ح" and it has the highest number of characters based on this base shape among languages adopting the Arabic script. Characters with two dots in vertical direction as in the second entry of Table 10 are not available in any other languages and they pose another problem in recognition. The fifth entry in Table 10, the character "ڇ" with four dots inside poses additional challenge for Sindhi OCR to recognize.

**Table 10: Connecting shapes of base shape "ح"**

| S.No: | Character | Isolated | Final | Medial | First |
|---|---|---|---|---|---|
| 1 | ج | سماج | حج | حجم | جبل |
| 2 | ڇ | اڇ | ڇڇ | سڇي | ڇپ |
| 3 | چ | اچ | چچ | ---- | ---- |
| 4 | چ | اچ | مچ | خچر | چپ |
| 5 | ڇ | رڇ | هيڇ | مڇر | ڇڙي |
| 6 | ح | مداح | صبح | محل | حجم |
| 7 | خ | رخ | مريخ | سخي | خچر |

**Context sensitive shapes of base shape "د"**

The connecting examples of these group shapes are given in **(Table 11)**. This group contains six characters differentiated by the number, placement and orientation of dots and again this group has the largest number of characters among the languages adopting the Arabic script.

**Table 11: Connecting shapes of base shape "د"**

| S.No: | Character | Isolated | Final | Medial | First |
|---|---|---|---|---|---|
| 1 | د | داد | آباد | مدد | در |
| 2 | ڏ | اڏ | واڏ | مڏر | ڏوپي |
| 3 | ڊ | اڊ | گھنڊ | مندو | ڊيل |
| 4 | ڈ | اوڈ | ڈاڈو | گڈ ر | ڈيڈ ر |
| 5 | ڊ | رڊ | سنڊ | واڊ و | ڊور |
| 6 | ذ | محاذ | اخذ | مذاق | ذرو |

Each character in the alphabet changes its shape when connecting to another character according to the position of the succeeding or preceding characters. The character may have two to four shapes, so more training and testing are required to build an OCR for Sindhi.

The changing forms of aforementioned base shape group characters are given in **(Table 12)**. Two characters "جھ" and "گھ" pose a challenge in recognition as they can also represent two separate character instead of one. This creates ambiguity while classifying the characters. The characters with four dots are also a challenge for recognition in Sindhi script.

It can be concluded that more characters need to be recognized in Sindhi OCR, than any other OCRs. Along with the largest number of characters to be recognized, some additional character base shapes, characters with four dots and more characters in shape groups (very small difference between the characters) are the challenges to be tackled in the development of Sindhi OCR. **(Table 13)** shows the analysis of the character shapes to be considered for the Sindhi OCR.

**Table 12: Different no: of Character shapes for recognition**

| Character | No: of ch: shapes | Character | No: of ch: shapes | Character | No: of ch: shapes | Character | No: of ch: shapes | Character | No: of ch: shapes |
|---|---|---|---|---|---|---|---|---|---|
| ا | 2 | ج | 4 | ڊ | 2 | ض | 4 | گ | 4 |
| ب | 4 | ڄ | 2 | ڈ | 2 | ط | 2 | ڳ | 4 |
| ٻ | 4 | چ | 4 | ڍ | 2 | ظ | 2 | گھ | 4 |
| پ | 4 | ڇ | 4 | ذ | 2 | ع | 4 | گ | 4 |
| ڀ | 4 | چ | 4 | ر | 2 | غ | 4 | ل | 4 |
| ٽ | 4 | ڇ | 4 | ڑ | 2 | ف | 4 | م | 4 |
| ٿ | 4 | ح | 4 | ز | 2 | ڦ | 4 | ن | 4 |
| ٺ | 4 | خ | 4 | س | 4 | ق | 4 | ڻ | 4 |
| ٹ | 4 | د | 2 | ش | 4 | ڪ | 2 | و | 2 |
| ٹ | 4 | ڈ | 2 | ص | 4 | ک | 4 | ھ | 4 |
| ٹ | 4 | | | | | ء | 1 | ي | 4 |

## 3. SINDHI LIGATURES

To form a ligature or a word, a character is to be connected to its preceding or succeeding character or both. This connecting nature is called the cursive nature of the language. A ligature may be a word or it may not completely be a word by itself. A word may comprise isolated characters only; one, two or more ligatures or a combination of all. A word can be formed from a single ligature, multiple ligatures, one ligature and isolated characters, multiple ligatures and isolated characters, and only isolated characters. Table 14 shows some of the possible options to form a word in Sindhi. It is obvious that if there are more characters, then more combinations of ligatures as well as words would be formed. This aspect is more challenging when grammar, searching and sorting facilities are to be included in Sindhi OCR.

**Table 13: Different forms of ligature formation**

| | | |
|---|---|---|
| 1 | One ligature forms a word | ٿلهو پلو |
| 2 | More than one ligature form a word | موتمار صحتمند |
| 3 | Ligature and isolated characters to form a word | بهادر |
| 4 | More than one ligature and isolated character | هردلعزيز |
| 5 | Only isolated characters | رازو ادارو |

## 4. NOISE

In any OCR, noise is a main factor that may occur due to various sources like scanning quality, ink dots while printing and old document. A noise may occur on documents with low resolution scanning. So noise must be removed to recognize the characters. It is a challenging job to find the difference between the area of any object (character) and the noise (Introduction to Image noise, 2013). In Sindhi script it will be more challenging to find out the difference between the character dot and the noise because a single noise dot can change the shape of a character to another character as discussed earlier. Sindhi alphabet consists of 35 characters with dots and 17 without dots so it is even more challenging to differentiate between the character dot and noise. Some researchers used noise removal tools whereas Ajward et al. (2010) chose to experiment on high quality paper images which were noise free.

## 5. SKEW

Skew is the slant which exists in an image containing text that may occur due to scanning errors. So, it's necessary to put extra care at the time of scanning. Identifying this slant and eliminating it is called skew detection and correction. An arbitrary polygon method was used by An- Shatnawi and Omar (2009) on Arabic text images. Chaudhuri et al. (1997) proposed an algorithm for Indian script to detect and correct skew detection. Skew detection and correction was applied on Gujarati numbers by Desai (2010) in their work on handwritten Gujarati number recognition.

## 6. FONT

Font is an important factor in recognizing the characters especially when they are in printed form. The printed characters might be in different styles, families and sizes. The styles include italic, bold, underline, strikethrough and plain style. Fig. 6 presents the illustration of font styles in MB Latifee family of fonts in Sindhi of size 12.

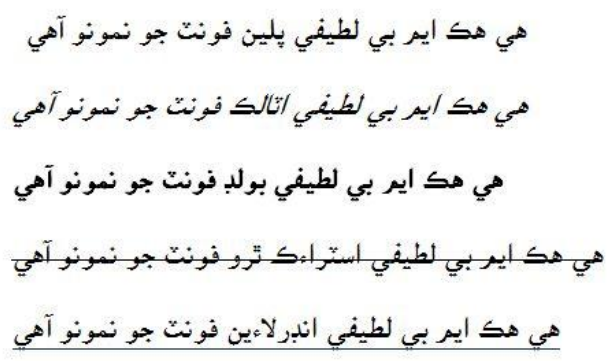



**Fig. 6: Font styles of MB Latifee in 12 font size**

Various font families have been created for Sindhi language writing, some of the fonts are good for recognition whereas some of them pose highly challenging jobs. Some of the font families available for Sindhi Language are MB Latifee, MB Khursheed, MB Sindhi web, MB Sania, MB Bhitai Sattar, MB Halai Shabir Kumbhar, MB Leeka Shabir Kumbhar, MB Shabir Kumbhar Web and MBSarang Sattar. **(Table 14)** illustrates some of the font families available for the Sindhi language writing. From Table 14, it is obvious that the writing style of MB Leeka Shabir Kumbhar MB Leeko Shabir Kumbhar and MB Leekyal Shabir Kumbhar should be avoided for the recognition purpose because it poses more challenges in separating the words, dots and other beautification marks.

**Table 14**: **Font size examples for Sindhi writing**

| S.No: | Font Name | Font Sample |
|---|---|---|
| 1 | MB Latifee | هي ايم بي لطيفي فونٽ جو نمونو آهي |
| 2 | MB Khursheed | هي ايم بي خورشيد فونٽ جو نمونو آهي |
| 3 | MB Sarnag Sattar | هي ايم بي سارنگ ستار فونٽ جو نمونو آهي |
| 4 | MB Preen Shabir Kumbhar | **هي ايم بي پرين شبير ڪمپار فونٽ جو نمونو آهي** |
| 5 | MB Lajpat Rai | هي ايم بي لجپت راء فونٽ جو نمونو آهي |
| 6 | MB Sania | هي ايم بي سانيا فونٽ جو نمونو آهي |
| 7 | AS Sagar | **هي اي ايس ساگر فونٽ جو نمونو آهي** |
| 8 | MB Leeka Shabir Kumbhar | هي ايم بي ليڪا شبير ڪمپار فونٽ جو نمونو آهي |
| 9 | MB Leeko Shabir Kumbhar | هي ايم بي ليڪو شبير ڪمپار فونٽ جو نمونو آهي |
| 10 | MB Leekyal Shabir Kumbhar | هي ايم بي ليڪيل شبير ڪمپار فونٽ جو نمونو آهي |

Font size also plays a vital role in the development of any OCRs and many of the OCRs are limited to single font as well as sizes. Al-Muhtaseb *et al.,* (2008) used eight various fonts for the recognition of Arabic script and reported varying degree of recognition accuracy for each font. Razzak *et al.,* (2009) presented an OCR system for multi-font numbers of Urdu and Arabic scripts. Slimane *et al.,* (2013) tested 100,000 Arabic text images in 10 various sizes of the font and used another 100,000 images for training purpose. The font size and font family pose challenges because these affect the accuracy rate of an OCR system. **(Fig. 7)** depicts the word "SINDH" in various font sizes written in MB Latifee font family.

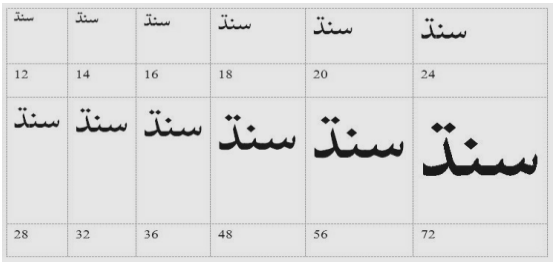



**Fig. 7: Various font sizes written in MB Latifee**

The complexity of OCRs is based on the difficulty level of the font family, size and styles, in which fixed font OCRs are relatively easier than the multifont OCRs whereas omnifont OCRs are based on the feature database of styles, sizes and family of the fonts which result in an acceptable accuracy for the recognition. Khorsheed, (2007) used a sliding window in decomposition technique for extraction of simple statistical features from offline Arabic text documents and Hidden Markov Models Toolkit for recognition of multifont computer generated sheets.

## 7. RESULTS AND DISCUSSION

Sindhi is a language with a rich history and literature and it was written in various scripts in different regions such as Shikarpuri, Khowjki, Khudaabadi. Three common scripts are used today and Arabic script is the most widely used script for writing Sindhi Language. The writing style follows the style of Arabic language; the characters are written starting from right to left whereas the numbers are written in the opposite direction. The challenges in Sindhi language OCR development are as characters for Arabic OCR but includes some additional challenges pertaining to Sindhi script because Sindhi possess more than Arabic and two more ligatures are formed in Sindhi script. Sindhi script comprises 35 dotted characters which is more than the Arabic and Persian full alphabets and the number of dots ranges from one to four. These dots pose more

challenges for Sindhi OCR as these dots differ in placement and orientation. Other challenges include additional orientations of two dots i.e. vertical and diagonal and three dots with two directions i.e. pointing upward and pointing downward. Additional characters in Sindhi are difficult to recognize as most of the characters possess the same base shape and differ in the number and placement and orientation of dots. A single base shape "ب" contains 9 different characters in Sindhi whereas there are only 3 characters in Arabic for the same base shape. Additional base shapes are used to extend the script for Sindhi and make it the script with the largest extension among languages adopting the Arabic script. Two additional characters "ڄ" and "ڳ" pose ambiguity challenge while recognizing the characters because these characters are in fact made up of two individual characters. The grammar, searching and sorting implementation of Sindhi OCR is even more difficult as the Sindhi script characters are scattered in various locations in Unicode. Characters in Sindhi script change their form according to the location in a word. In this way, one character may have two to four shapes. As Sindhi contains 52 distinct characters, there are more characters than any language for recognition. Skew and noise can make the recognition even more challenging as some of the fonts available for writing Sindhi is not suitable for recognition. As discussed a font named MB Shabir Kumbhar Leekyal, Leeko and Leeka should be avoided for the recognition of Sindhi script. Summary of issues and challenges are given in **(Table 15)**.

**Table 15: Summary of issues and challenges**

| Issues & Challenges | Explanation |
|---|---|
| **Many types of scripts** | Region based scripts such as Shikarpuri, Khowjki, Khudaabadi and others. Common script used today is Arabic, the most widely used for writing Sindhi Language. |
| **Styles of Writing** | Follows Arabic style; characters are written from right to left and numbers in the opposite direction. |
| **Arabic and additional challenges** | Arabic OCR challenges and additional challenges pertaining to Sindhi script. |
| **Additional characters** | More characters than Arabic script with more number, placement and orientation of dots. |
| **More Ligatures** | More characters so the more ligatures are formed because of more combinations are formed. |
| **Additional orientation of two and three dots** | Additional orientations of two dots i.e vertical and diagonal, three dots with two directions i.e pointing upward and pointing downward |
| **Additional base shapes** | More base shapes and more characters with single base shape. Arabic script contains three distinct characters with same base shape whereas in Sindhi script the same base shape contains 9 distinct characters. |
| **Characters formed by combining two individual characters** | Two characters "ڄ" and "ڳ" are formed from two individual characters and difficult to recognize the difference. |
| **Inconsistency in Unicode placement** | Sindhi characters possess various locations and difficult to search and sort. |
| **More characters than any other Arabic adopting language** | More characters than any other Arabic scripting adopting language and most of the characters change their shape according to placement in a word, so it has the most number of characters for recognition. |
| **Skew and noise pose additional challenges** | Skew and noise pose increased challenges because of the cursive style and increased number of dots. |
| **Some fonts not suitable for Sindhi OCR** | Font named MB Shabir Kumbhar Leekyal, Leeko and Leeka should be avoided because of the beautification marks. |

## 8. CONCLUSION AND FUTURE WORK

Sindhi script uses Arabic script for writing, and it is a script that extends Arabic alphabet the most compared to any other languages. In this paper issues and challenges of the OCR for Sindhi script are presented. More characters in Sindhi script form more ligatures and words which is a potential problem in recognizing Sindhi script. One single base shape contains multiple characters up to 9, and the characters differ by the placement and orientation of the dots and these individual characters possess multiple shapes when written in a word or sentence. Additional characters in Sindhi present ambiguity in recognition and results in challenges in implementation of grammar, searching and sorting. A good number of fonts are available for Sindhi script and some of the fonts are not suitable and should be avoided because of the beautification marks. A lot of attentions are needed to build a complete working OCR. For future work, the techniques used by researchers will be investigated and

a solution will be found for the development of Sindhi OCR. Font family, sizes and styles will be an additional consideration for capability enhancement of Sindhi OCR.

**REFERENCES:**

About Indus Civilization (2013).[Online] Available: http://en.wikipedia.org/wiki/ Indus_civilization (July 6, 2013)

Ajward, S., N. Jayasundara, S. Madushika, R. Ragel, (2010), "Converting printed Sinhala documents to formatted editable text", *in* 'Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on', 138--143.

Alana, G. A. (1993) Sindhi Sooratkhati, Sindhi Language Authority Hyderabad, Sindh. Fourth Edition.

Alana, G. A. (2004). Sindhi Boli jo Bunn Bunyad. First Edition, Sindhi Language Authority Hyderabad. Sindh. Pakistan.

Al-Muhtaseb, H. A.., S.A. Mahmoud, R. S. Qahwaji, (2008), 'Recognition of off-line printed Arabic text using Hidden Markov Models', *Signal Processing* 88(12), 2902 - 2912.

An-Shatnawi, M. K. Omar, (2009), Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity, *Journal of Computer Science* **5**(5), 363--368.

Chaudhuri, B.B.. U. Pal, (1997), Skew angle detection of digitized Indian script documents, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2), 182–186.

Desai, A. A. (2010), Gujarati handwritten numeral optical character reorganization through neural network, *Pattern Recognition* 43(7), 2582 - 2589.

Introduction to Image noise (2013). [Online] Available: http://en.wikipedia.org/wiki/Image_noise (July 7, 2013)
M., Jelodar, M. Fadaeieslam, N. Mozayani, M. Fazeli, (2005), A Persian OCR System using Morphological Operators, *in* 'The Second World Enformatika Conference, WEC', 137--140.

Khorsheed, M. (2007), 'Offline recognition of omnifont Arabic text using the HMM ToolKit (HTK)', *Pattern Recognition Letters* 28(12), 1563 - 1571.

Parvez, M. T., S. A. Mahmoud, (2013), Arabic handwriting recognition using structural and syntactic pattern attributes, Pattern Recognition 46(1), 141-154.

Pal, U., A. Sarkar, (2003), Recognition of printed Urdu script, in Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), COMPUTER SOCIETY, EDINBURGH, Scotland, 1183 -1187.

Razzak, M., S. Hussain, A. Belaid, M. Sher, (2009), 'Multi-font Numerals Recognition for Urdu Script based Languages', *International Journal of Recent Trends in Engineering (IJRTE)*. inria-00437121, version 1-29 Nov. 2009.

Rafael C. G. Richard E. Woods. Digital Image Processing 2/e, Thresholding, Image Segmentation.

Sattar, S. A., S. Haque, M. K. Pathan, Q. Gee, (2009), Implementation Challenges for Nastaliq Character Recognition'Wireless Networks, Information Processing and Systems', Springer Berlin Heidelberg, 279-285.

Soille, P. (2003). Principles and Applications: Morphological Image Analysis, 2nd ed., Springer-Verlag, NY.

Slimane, F., S. Kanoun, J. Hennebert, A. M. Alimi, R. Ingold, (2013), A study on font-family and font-size recognition applied to Arabic word images at ultra-low resolution, Pattern Recognition Letters 34(2), 209 - 218.

Tabassam N., S. Ammar, H. Shah, H- u- Rehman and A. Faiz,(2009), Optical Character Recogntion System for Urdu (Naskh Font) using pattern Matching Technique, International Journal of Image Processing, (IJIP), Vol. (3): Issue (3): 92Pp.