Australian Journal of Basic and Applied Sciences, 3(4): 4160-4169, 2009
ISSN 1991-8178

# Character Segmentation of Sindhi, an Arabic Style Scripting Language, using Height Profile Vector

[1]Noor Ahmed Shaikh, [1]Ghulam Ali Mallah, [2]Zubair A. Shaikh

[1]Assistant Professor and PhD student, Shah A. Latif University, Khairpur, Sindh, Pakistan
[2]Professor and Director, FAST-NU, Karachi, Sindh, Pakistan, zubair.shaikh@nu.edu.pk

**Abstract:** In this paper, a problem of sub-word segmentation of printed Sindhi, an Arabic style scripting language, into characters is addressed. Printed or handwritten Sindhi text is cursive in nature. In the cursive writing, mostly the subsequent characters in a word are joined with each other. In the proposed segmentation algorithm, first of all, Height Profile Vector (HPV) of thinned primary stroke of a sub-word is calculated and analyzed for the segmentation into its constituent characters. The number and locations of possible segmentation points (PSP) are determined. The number of PSPs gives a rough estimation of the number of characters in the sub-word. The data around the last PSP is further analyzed to determine the exact number of characters in the sub-word. As the characters' set of Sindhi is the superset set of Arabic characters' set hence the proposed segmentation algorithm may be used for the segmentation of text written in other Arabic scripting languages.

**Key words:** Sindhi OCR, Character Segmentation, Pattern Recognition.

## INTRODUCTION

Character recognition is one of the most important fields of pattern recognition has been around since the development of first version of OCR in 1950's (Mori, S., C.Y. Suen, 1992). Since then several character recognition systems have been proposed for English, Chinese, Japanese and other similar languages that use isolated characters (Badr, B.A. and S.A. Mahmoud, 1995). Character recognition systems for other languages like Arabic and Persian are not much robust and character recognition systems for Sindhi and Urdu are still mostly in research labs, primarily due to their property that such languages are cursive in nature (Kavianafar, M. and A. Amin, 1999). Recognizing unconstrained off-line cursive writing has proven to be a very difficult task, mainly due to the difficulty of character segmentation. Because of this difficulty, several attempts have been made to recognize the sub-words instead of characters (Liying Zheng, 2006; Mandana Kavianifar and Adnan Amin, 1999; Somaya Alma'adeed, 2006). This approach can narrow down the sub-word candidates, because in the large-vocabulary several sub-words may have the same global shape (Yannikoglu, B. and P.A. Sandon, 1998).

In this paper, we are addressing the segmentation of off-line printed Sindhi sub-word into characters. Sindhi is a language that uses Arabic scripting. So, the techniques used for the segmentation of Arabic script may be used for Sindhi script and vice versa. Since the segmentation of Arabic script into characters is more difficult. So, many segmentation systems do not segment into characters but some other units or parts which are easier to segment. Elgammal *et. al.* (2001) segmented the words into small connected segments called *'scripts'*.

The organization of this paper is as follows: Section 2 presents Introduction to Sindhi. In Section 3, the related work for Arabic character segmentation is presented. The proposed method for the segmentation of Sindhi characters is presented in Section 4. In Section 5, 6 and 7 conclusions, future work and the acknowledgements are presented respectively.

### *2. Introduction to Sindhi:*

Sindhi is an Indo-Aryan language having roots in Lower Indus River Valley. The name has been driven from Sindhu, the ancient name of the river Indus. Sindhi is the third major spoken language of Pakistan and over 30 million people in Pakistan and India speak Sindhi. Beyond the Indian sub-continent, it is also spoken

**Corresponding Author:** Noor Ahmed Shaikh, Assistant Professor and PhD student, Shah A. Latif University, Khairpur, Sindh, Pakistan
E-mail: noor.shaikh@salu.edu.pk

by large communities in the United Kingdom and the United States, and around the world. The language is versatile and has potential to expand and grow to fulfill the needs of the modern times. It is used in education, literature, office, court-business, media and in religious institutions of Sindh.

Evidence for Sindhi as a written language dates to a Sindhi translation of the Islamic Qur'an in 883 A.D., followed a century later by a Persian translation of the ancient Indian religious epic Mahabharata taken from a language thought to be Old Sindhi.

The starting point of current Sindhi *abjad* is the version of the Perso-Arabic script that was used to write Urdu, under British influence in 1852 the same was adopted to write Sindhi. In India Sindhi is also written with the Devanagri script (Pal, U. and B.B. Chaudhuri, ).

Both Arabic and Sindhi use same writing system i.e. *Semitic Abjad* that represents consonants plus some vowels, and use Naskh style of writing in which all text is written on an imaginary horizontal line called baseline. Both languages are cursive in nature in which letters are connected with each other in sub-words on the baseline. This is similar to Latin *'joined up'* handwriting, which is also cursive. Arabic as well as Sindhi characters that can be joined are always joined in both handwritten and printed text. Fig. 1 shows the samples of Sindhi and Arabic cursive writings. The shape of many letters varies depending on the positions (beginning, middle or end of a word, or on their own) where they appear in a word. The four forms of characters according to their position in the sub-word i.e. initial, middle, final and isolated are shown in Table: 1.

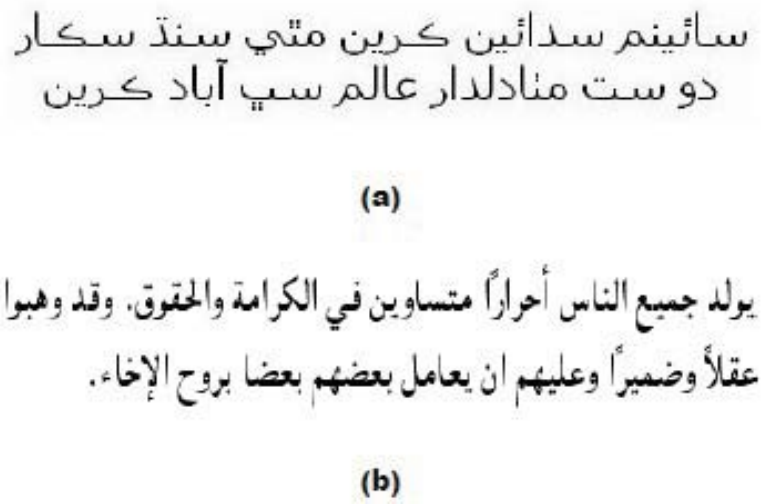


**Fig. 1:** Samples of Cursive writing (a) Sindhi (b) Arabic

**Table 1:** Sindhi Characters at different positions, S=Standalone, I=Initial, M=Medial, F=Final

| Character | S | I | M | F |
|---|---|---|---|---|
| ALEF | ا | ا | --- | ا |
| BEH | ب | بـ | ـبـ | ـب |
| JEEM | ج | جـ | ـجـ | ـج |
| SEEN | س | سـ | ـسـ | ـس |
| QAF | ق | قـ | ـقـ | ـق |
| KAF | ك | كـ | ـكـ | ـك |
| KEHEH | ک | کـ | ـکـ | ـک |

The only difference between the thematic structures of Arabic and Sindhi is the number of dots. The thematic structures also called orthographic structures define the geometrical and structural shape of the characters. In Arabic most letters have one to three dots that can be positioned above, below or inside them; however, the maximum number of dots used in Sindhi is four. As the dots are the diacritic marks, that are supposed to affect the sound or the phonemes. As there is increase in the number of sounds, so that number of characters is increased. Hence, the characters' set of Sindhi having 52 characters is the superset of the Arabic characters' set which comprises of 28 characters only as shown in Fig.2.

| | | | | | | |
|---|---|---|---|---|---|---|
| ڻ | ٽ | ت | ٻ | ب | ٻ | ا |
| ڇ | ڇھ | ح | ح | ٻ | ٿ | ٺ |
| ڈ | ڈ | د | خ | ح | ڇ | ڄ |
| س | ز | ڑ | ر | ذ | ڊ | ڊ |
| غ | ع | ظ | ط | ض | ص | ش |
| گ | گ | ک | ک | ق | ق | ف |
| و | ٹ | ن | م | ل | گ | گھ |
| | | ي | ء | ھ | | |

(a)

| | | | | |
|---|---|---|---|---|
| ج | ث | ت | ب | ا |
| ر | ذ | د | خ | ح |
| ض | ص | ش | س | ز |
| ف | ع | ع | ط | ط |
| ن | م | ل | ک | ق |
| | ي | و | ھ | |

(b)

**Fig. 2:** Characters' set (a) Sindhi (b) Arabic

### *3. Segmentation:*

Segmentation is one of the important processes of any OCR system and plays a vital role in the performance of an OCR system. Rejection or misrecognition rate will be very high if the characters are not segmented accurately.

### *3.1 Related Work:*

Sindhi is a language that uses Arabic style scripting. So, the techniques used for the segmentation of Arabic script may be utilized for Sindhi script and vice versa. Most of the segmentation methods separate diacritic marks before performing character segmentation.

Some of the early work on off-line Arabic word segmentation was carried by K.R. Pakker *et. al.* (1995) who segment Arabic handwritten words by calculating the thickness of the stroke on each text line and this value is considered as the threshold for the determination of the segmentation points between characters on the baseline.

B.A. Najoua *et. al.,* (1995) proposed the method of segmentation of printed and handwritten words into characters. In which the approximate limits of the characters in the PAW (Piece of Arabic Word) are estimated using vertical histogram modulated by the vertical width of writing. The distance equal to one and a half of the writing is considered suitable for the segmentation.

D. Motawa *et. al.* (1997) uses morphological operations to segment the handwritten words into singularities. Regularities are found by subtracting the singularities from the original image. Regularities hold the information that is necessary for linking a character to its subsequent character. Hence, these regularities are the entrants for segmentation. Regularities close to the baseline are segmented using the rules for long and short regularities accordingly.

Elgammal *et. al.* (2001) performed the segmentation of Arabic text by using the topological relationship b/w the baseline and the text line. The baseline is identified by using the histogram; however the text line is represented by line adjacency graph (LAG). The LAG formed for each sub-word is then transformed into the compressed LAG or simply c-LAG which is homomorphic to the LAG and has minimum number of nodes and each node is either labeled as path or junction. The adjacent nodes are considered as the junction nodes as the path nodes are never adjacent to each other.

Noor and Zubair (2008) proposed a segmentation method in which the vertical projection graph for each line is computed. The computed graph is then processed to generate a string indicating relative variations in pixels. The string is scanned further to produce the characters from the sub-words. Some rules have been developed to help in finding the segmentation points in the sub-words.

The type of methods that divide the text line into three regions or zones calculate the histogram and uses different functions on the histogram of a region as a tool to determine the connection points. In (Sarfraz, M., S.N. Nawaz, 2003) the value of the vertical profile of the middle zone is calculated. The area where it is less than 66% of baseline thickness is considered a connection area between two characters. The area which has a larger value is regarded as the beginning of a new character, as long as the profile is greater than one third of the baseline. A. Ymin *et al.* (1996) proposed a method for the segmentation of printed multi-font Uygur text that uses Arabic style scripting in writing. In this method text line is divided into three zones. From the edge of character strokes of upper zone the algorithm searches for possible break points along the vertical projection, which may cause the under segmentation of the characters having loops in them. The quasi topological segmentation bases on the first segmentation, to segment a character on a combination of feature-extraction and character-width measurements.

In (El-Khaly, F. and M.A. Sid-Ahmed, 1990) a thinned sub-word is segmented into characters by following the baseline of the sub-word. The sub-word is segmented when pixels start to move above or below the baseline.

The type of segmentation methods that uses contour of the binary image as input uses different functions on the contour of the sub-word as a tool to determine the connection points. In (Margner, V., 1992) these connection points are characterized to be the locations where curvature of upper contour of sub-word changes from positive to negative. In (Sari, T., L. Souici and M. Sellami, 2002) the probable segmentation points are characterized as Valid Segmentation Points (VSP) if local minimum in the lower outer contour qualifies the acceptance rules. The acceptance and rejection rules are the morphological rules which have been extracted from the Arabic text characters.

The segmentation of sub-words into characters is the most delicate stage. The methods, devised by the researchers to perform segmentation of sub-word into characters, are facing the problems of under segmentation or over segmentation.

### *4. Proposed Algorithm:*

In the proposed segmentation algorithm, a thinned primary stroke of Sindhi sub-word is obtained and segmented by using the following procedure:

- The acquired text image is thinned by using a thinning algorithm (Shaikh, Z.A. and N.A. Shaikh, 2006), which gives comparatively good results.
- Lines of text are segmented by using horizontal projection histogram.
- Base Line is detected by using the method described in section 4.1.
- Sub-words are extracted from the lines of text using connected component extraction method. A test is also conducted to characterize the extracted sub-word either a primary stroke or a secondary stroke. The process of extracting and testing of sub-words is described in section 4.2.
- Finally, the thinned binary image of the primary stroke of a sub-word is segmented into characters. The entire process of segmentation is described in section 4.3.

### *4.1 Base Line Detection:*

The baseline $\boldsymbol{b}$ for the extracted line of text $L(i, j)$ may be determined from the horizontal projection of the extracted line of text Ph, by initially letting $\boldsymbol{b}$=1, then

$$b = i if (P_h(i) > P_h(b)) 2 \le i \le G$$

where

$$P_h(i) = \sum_{j=1}^{H} L(i,j) \qquad for 1 \le i \le G$$

***4.2 Extraction of Sub-words:***

As Sindhi and other Arabic style scripting languages are written from right to left, so the sub-words are extracted from the same direction. The sub-word may consist of two parts called primary strokes and secondary strokes. Primary stroke have one or more connected root characters. However, the secondary stroke is either dots (Nuqtaas) or diacritic marks (Airabs) i.e. zair, zabr, pesh etc. associated with the root characters. Vertical Projection Histogram and Connected Component Analysis are the methods used to separate sub-words or words from the text lines. The former method is failed in a situation when boundaries of sub-words vertically overlap each other. So, we use connected component analysis method to extract out the connected parts of the sub-word i.e. primary and secondary strokes. This can be done by using Area Voronoi Diagram (Wang, Z., Y. Lu, C. L.Tan, 2003) or 8-connected component analysis (Gonzalez, R.C., R.E. Woods, 2005) method.

We use 8-connected component analysis method. This process starts by looking for the object pixels (foreground pixels) from the top-right end of each segmented line of text. When a foreground pixel is found, its *x-y* coordinates are inserted into the list, and the considered pixel will be set as background pixel. The process will look for the 8-connected pixels of the found pixel and place them into the list. This process will be continued until no 8-connected pixels are found.

In order to characterize the segmented component as a primary stroke or a secondary stroke; we determine the minimum and maximum value of *y* in the list. If ***b*** (baseline) lies between these values of *y,* the segmented component would be characterized as a primary stroke, otherwise a secondary stroke.

Primary stroke ➜ min(y) < b < max(y)

In case of a secondary stroke, its position needs to be identified. As dots above and below the characters play a major role in differentiating some characters that differ only by the number or location of dots. It would be above the primary stroke if ***b***>*min(y),* and it would be below the primary stroke if ***b***<*max(y).*

***4.3 Extraction of Characters:***

Various approaches have been used to extract out the characters from the sub-word. In the proposed method for the segmentation of sub-words into characters, first of all, we determine the Height Profile Vector (HPV) of the primary stroke of a sub-word. HPV is then analyzed to determine the locations of the Possible Segmentation Points (PSPs) as well as the number, *n,* of PSPs. The value of *n* gives a rough estimate about the number of characters in the sub-word. The data around the last PSP will further be analyzed to infer the exact number of characters in the sub-word.

***4.3.1 Determination of HPV:***

Height Profile Vector, HPV, of the primary stroke of a sub-word is determined by calculating the difference between the baseline ***b*** and the row number of topmost pixel in each column, or the maximum of the distances of each black pixel in the column from the baseline.

Let *W(i,j)* be the image of the primary stroke of a sub-word having size PxQ and ***b*** be the baseline of the text line from which it has been extracted, then the HPV for it is determined as follows:

$$HPV = \left[ b - i_j \,\middle|\, i_j \text{ is the position of Ist of } jth \text{ column} \right] 1 \le j \le Q$$

The graphical representation of HPV and corresponding primary strokes of sub-words are shown in Fig.3.

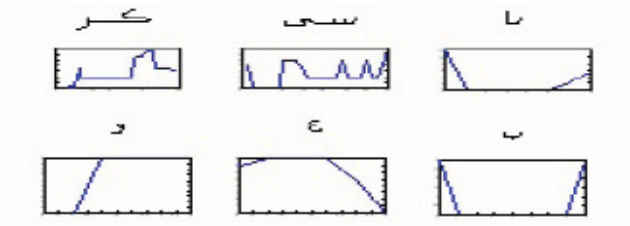

**Fig. 3:** Primary Strokes of Sub-words and graph of their corresponding HPV, top: connected characters, bottom: isolated characters.

***4.3.2 Determination of the Locations and Number (n) of PSPs:***

To determine the locations and the number, *n,* of PSPs we analyze HPV of the primary stroke of a sub-word from the right hand side, as Sindhi and other Arabic style scripting languages are written from right to left.

The consecutive pixels that remain on the baseline, i.e. zeros in the HPV, may be considered as the segmentation line, if the number of such consecutive pixels *(bs)* exceeds by some threshold (T). In Fig. 4 primary strokes and their corresponding HPV are shown. The pixels that remain on the baseline in Fig. 4(a) from column 12 to column 19, hence *bs* = 8.

In order to determine the location of PSP, we use the following formula:

PSP(*n*) = *i*-round(*bs*/2)

where *i* is the last position, from the right hand side, of the corresponding segmentation line. In this example for PSP(1), it is,

i = 20

∴ PSP (1) = *i*-round (*bs*/2) =16

Similarly, other segmentation points in the primary stroke can be determine, if exist, by analyzing its HPV. for PSP(2),

*bs* = 4

and *i* = 25

∴ PSP (2)=23

Similarly,

PSP(3)= 31

The number, ***n,*** of PSPs in the sub-word is determined by the length of vector PSP i.e. *n*=length (PSP). Here it is *n* =3.

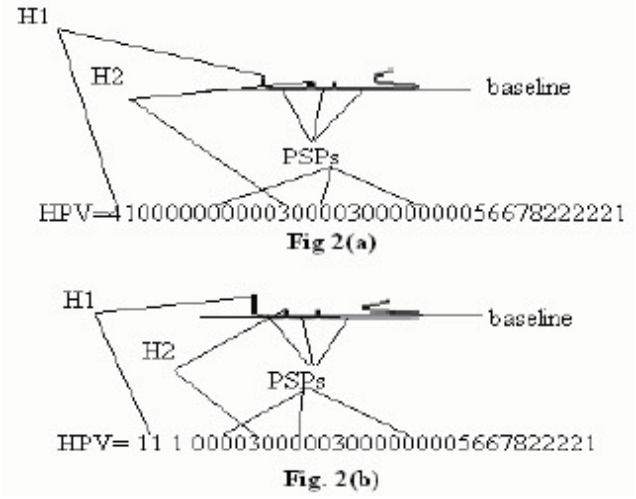


**Fig. 4:** Primary Strokes of Sindhi Sub-words and their corresponding HPV.

***4.3.3 Interpretation of n:***

- *n*=0 implies that the sub-word is an isolated character (IC) having no baseline (د، ا، ج، ر،).
- n=1 implies that sub-word is either an isolated character (ب، گ، ف) or a sub-word that consists of two characters (با، كا نو، مي). . It will be decided after analyzing data around the PSP.
- n>1 implies that the sub-word have n-1 definite characters up to the (n-1)th PSP (پټ، . پټي، پټو، مټ، منّا).The number of rest of the characters will be decided after analyzing data around the nth PSP, whether it is one or two.

***4.3.4 Analysis of Data Around nth PSP:***

The number of PSPs, n, gives the rough estimation about the number of characters in a sub-word; it is either n or n+1. To determine the exact number of characters in a sub-word, the following parameters are required that can be obtained by analyzing the data around the nth PSP.

- Difference (D) between last value (H1) and the peak before the nth segmentation point (H2) in HPV. H1 and H2 are depicted in Fig. 4.
- The position of last three heights in HPV.
- The number of PSPs i.e. n.

**Table 2:** Lookup table to identify the number of characters in a sub-word

| Case | D | Position of Last Pixel HPV(1) | Position of HPV(2)\| HPV(3) | No. of PSP (*n*) | No. of Characters (m) |
|---|---|---|---|---|---|
| 1 | >-3 | <=0 | -- | >1 | *n* |
| 2 | >-3 | <=0 | -- | =1 | *n+1* |
| 3.1 | >-3 | >0 | <0 | >1 | *n+1* |
| 3.2 | >-3 | >0 | >=0 | >1 | *n* |
| 4.1 | >-3 | >0 | <0 | =1 | *n+1* |
| 4.2 | >-3 | >0 | >=0 | =1 | *n* |
| 5 | <=-3 | <=0 | -- | -- | -- |
| 6 | <=-3 | >0 | -- | -- | *n+1* |

In other words, the exact number of characters may be determined by looking at Table 2. This table is prepared after performing experiments on different sub-words written in a variety of fonts and sizes.

***5. Experimental Results:***

After performing an experiment on variety of primary strokes of the sub-words, a total of six patterns of the connections of characters are identified that are illustrated in Table 2. The description of each is given below:

**Case 1:** The first case relates with those sub-words having *n* segmentation points apparently, but there are *n*+1 segmentation points may be viewed through *HPV,* as in the case of (i) and (ii) .The descenders of character Dal and Meem form a segmentation point. The other parameters for determining the number of characters in a sub-word suggest that the last segmentation point will not produce a character but the part of a last character.

**Case 2:** This case relates with those sub-words having *n* segmentation points and the last character having descenders below the base line as in the case of . The parameters for determining the number of characters in a sub-word suggest that such sub-words will have *n*+1 number of characters.

**Case 3:** This case relates with two different kinds of sub-words having *n* segmentation points and the last character having an ascender whose slope with the second last peak in the *HPV* is greater than –3, the second and/or third last heights in *HPV* determine the number of characters in sub-words. If any or both are below the baseline as in the case o f , the parameters for determining the number of characters in a sub-word suggest that such sub-words will have *n*+1 number of characters; otherwise number of characters is *n*.

**Case 4:** This case relates with such sub-words having one segmentation line and an ascender. Two different kinds of sub-words fall into this category as . and isolated characters like . Both kinds of sub-words have one segmentation line but different number of characters. Hence the positions of second and/or third last heights in *HPV* determine the number of characters. If any or both of such heights are below the base line, it suggests that the number of characters in such sub-word is two i.e. *n*+1; otherwise *m* is one. The later option signifies an isolated character.

**Case 5:** No such sub-words have been encountered during experimentation. This case has been considered in the table just to complete all possible entries.

**Case 6:** This case relates with those sub-words having *n* segmentation points and an ascender whose slope with the second last peak is less than –3. This kind of sub-words has Alif (~) in the end usually. The parameters for determining the number of characters in a sub-word suggest that such sub-words will have *n*+1 number of characters.

Table 3, depicts the primary strokes of sub-words of six discussed cases and their segmentation results. Note that there is an exception for the character (seen or sheen) whether in the sub-word or isolated, which is segmented into three parts and needs extra attention at the later stage of the recognition process.

***6. Conclusions:***

The proposed algorithm has been tested on printed Sindhi text of many types and sizes of fonts. In some cases, the segmentation algorithms perform under segmentation and in some cases perform over segmentation. The under segmentation is natural in some cases, for example SEEN, SHEEN, JHAY and GHAY, because these characters are formed by connected the shapes of some other characters. First two characters are formed by connected two BEH like characters and a NOON, when it is at the isolated or final position and two BEH like characters when it is at the initial or middle position of the sub-word. This under segmentation can not be avoided, but this error needs to be handled at the later stage of the character recognition by framing some rules. The next two characters are compound characters, which are formed by adding HEH in JEEM and GAF respectively. The under segmentation of these two characters will not be regarded as error, because these characters will not have even a single Unicode point but are represented by two code points in the memory.

**Table 3:** Experimental Results of Segmentation

| Primary Stroke of Sub-Word | PSP (*n*) | Case | *m* | C1 | C2 | C3 |
|---|---|---|---|---|---|---|
| ا | 0 | IC | 1 | ا | | |
| د | 0 | IC | 1 | د | | |
| ر | 0 | IC | 1 | ر | | |
| و | 0 | IC | 1 | و | | |
| لد | 2 | 1 | 2 | د | ل | |
| لم | 2 | 1 | 2 | م | ل | |
| كر | 1 | 2 | 2 | ر | ك | |
| بس | 2 | 3.1 | 3 | س | ـ | ب |
| مبى | 2 | 3.1 | 3 | ى | ـ | م |
| كب | 3 | 3.2 | 3 | ب | ـ | ك |
| س | 1 | 4.1 | 2 | س | ر | |
| ب | 1 | 4.2 | 1 | ب | | |
| ما | 2 | 6 | 3 | ا | ـ | م |
| عا | 1 | 6 | 2 | ا | ع | |

***Future Work:***

Some achievements have been made over the course of this research and further work in the areas presented here is promising i.e. fine tuning of the proposed algorithm to avoid the under and over segmentation of the characters and to determine a suitable feature set and a classifier to recognize the segmented characters and the diacritic marks associated with them.

## ACKNOWLEDGMENT

The authors would like to thank and acknowledge the Centre for Research in Ubiquitous Computing (CRUC) at FAST-NU, Karachi campus, where this entire work has been carried and Higher Education Commission (HEC), Pakistan who sponsored this work through its indigenous PhD program.

## REFERENCES

Badr, B.A. and S.A. Mahmoud, 1995.*"A survey and bibliography of Arabic optical text recognition",* Signal Processing, 41: 49-76.

Elgammal, A. and M.A. Ismail, 2001. A. Elgammal and M.A. Ismail, *"A graph-based segmentation and feature extraction framework for Arabic text recognition",* Proceedings of 6th International conference on Document Analysis and Recognition, pp: 622-627.

El-Khaly, F. and M.A. Sid-Ahmed, 1990. *"Machine recognition of optically captured machine printed Arabic text",* Pattern Recognition, 23(11): 1207-1214.

Gonzalez, R.C., R.E. Woods and S.L. Eddins, 2005. *"Digital image processing using MatLab",* 3rd Indian Reprint.

Kavianafar, M. and A. Amin, 1999. *"Pre- processing and structural feature extraction for multi fonts Arabic / Persian OCR",* Proceedings of 5th Intl. Conference on Document Analysis and Recognition.

Liying Zheng, 2006. "Machine Printed Arabic Character Recognition Using S-GCM", 18th International Conference on Pattern Recognition, pp: 893-896.

Mori, S., C.Y. Suen and K. Yamamato, 1992. *"Historical review of OCR research and development",* Proceedings of IEEE, pp: 1029-1058.

Mandana Kavianifar and Adnan Amin, 1999. "Pre-processing and Structural Feature Extraction for a Multi-Fonts Arabic/Persian OCR", Proceedings of 5th International Conference on Document Analysis and Recognition.

Margner, V., 1992. *"SARAT - A system for the recognition of Arabic printed text",* 1 1th International Conference on Pattern Recognition, pp: 561-564.

Motawa, D., A. Amin and R. Sabourin, 1997. *"Segmentation of Arabic cursive script",* Proceedings of the 4th International conference on Document Analysis and Recognition, pp: 625-628.

Noor A. Shaikh, A. Zubair and G. Ali, 2008. *"Segmentation of Arabic Text into Characters for Recognition",* IMTIC 2008, CCIS 20, Springer-Verlag Berlin Heidelberg, pp: 11-18.

Najoua, B.A., E. Noureddine, 1995. *"A robust approach for Arabic printed character segmentation",* International Conference on Document Analysis and Recognition, pp: 865-868.

Pal, U. and B.B. Chaudhuri, *"Automatic separation of machine printed and hand-written text lines",* Journal/ conference, Year.

Pakker, K.R., H. Miled, Y. Lecourtier, 1995. *"A new approach for Latin /Arabic character segmentation",* International Conference on Document Analysis and Recognition, pp: 874-877.

Somaya Alma'adeed, 2006. "Recognition of Off-Line Handwritten Arabic Words Using Neural Network" Proceedings of the Geometric Modelling and Imaging~ New Trends, pp: 141-144.

Sarfraz, M., S.N. Nawaz and A. Al-Khuraidly, 2003. *"Ofline Arabic text recognition system",* Proceedings of the Int. Conference on Geometric Modeling and Graphics, pp: 30-34.

Sari, T., L. Souici and M. Sellami, 2002. *"Of-line handwritten Arabic character segmentation algorithm: A CSA",* Proceedings of 8th International Workshop on Frontiers in Handwriting Recognition, pp: 452-457.

Shaikh, Z.A. and N.A. Shaikh, 2006. *"A universal thinning algorithm for cursive and non-cursive character patterns",* Mehran University Research Journal of Engg. & Tech., 25(2): 163-168.

Wang, Z., Y. Lu, C. L.Tan, 2003. *"Word extraction using area Voronoi diagram",* Proceedings of the Conference on Computer Vision and Pattern Recognition, 3: 31-36.

www.linguistics.uiuc.edu/jscole/...sindhi/3_Cole_Sindhi_ElsevEncycl.pdf

Ymin, A. and Y. Aoki, 1996. *"On the segmentation of multi-font printed Uygur scripts",* International Conference on Pattern Recognition.

Yannikoglu, B. and P.A. Sandon, 1998. *"Segmentation of of-Line cursive handwriting using linear programming",* Pattern Recognition, 31(12): 1825-1833.