Glyph Identification and Character Recognition for Sindhi OCR

NISAR AHMED MEMON*, FATIMA ABBASI**, AND SHEHNILA ZARDARI***

RECEIVED ON 18.04.2016 ACCEPTED ON 21.02.17

ABSTRACT

A computer can read and write multiple languages and today's computers are capable of understanding various human languages. A computer can be given instructions through various input methods but OCR (Optical Character Recognition) and handwritten character recognition are the input methods in which a scanned page containing text is converted into written or editable text. The change in language text available on scanned page demands different algorithm to recognize text because every language and script pose varying number of challenges to recognize text. The Latin language recognition pose less difficulties compared to Arabic script and languages that use Arabic script for writing and OCR systems for these Latin languages are near to perfection. Very little work has been done on regional languages of Pakistan. In this paper the Sindhi glyphs are identified and the number of characters and connected components are identified for this regional language of Pakistan. A graphical user interface has been created to perform identification task for glyphs and characters of Sindhi language. The glyphs of characters are successfully identified from scanned page and this information can be used to recognize characters. The language glyph identification can be used to apply suitable algorithm to identify language as well as to achieve a higher recognition rate.

Key Words: Optical Character Recognition, Glyphs, Sindhi, Character Identification

1. INTRODUCTION

Various input devices get input of text, voice and image data. ORC is an input method which takes less time [1] and using less or no time in getting more text as input. The text available on scanned paper or document is converted into editable text which is much faster alternative to type such amount of text with keyboard. The OCR applies set of algorithms to identify, segment, extract and recognize text (printed by machines). The text written on an image might be in different type of language or script and these scripts demand suitable algorithms to segment and recognize glyphs and characters [2]. The non-cursive script such as Latin and Cyrillic are posing less challenges compared to Arabic and Indian scripts and easy to recognize whereas the later pose more challenges and still need a lot of attention of the researchers [3].

Corresponding Author (E-Mail: nmemon@kfu.edu.sa)

* Department of Computer Engineering, King Faisal University, Al-Ahsa, Kingdom of Saudi Arabia.

** Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah.

*** Department of Computer Science & Software Engineering, NED University of Engineering Techology, Karachi.

1.1 Regional Languages and Sindhi

Urdu is the NL (National Language) of Pakistan and it must be announced within fifteen years, this was stated in 1973 constitution of Pakistan in article 251 in which Urdu must be made NL within fifteen years but law remain unimplemented. However, in 2015 Government of Pakistan has announced the plans to make Urdu as only NL [4]. Several other languages such as Kashmiri, Hindko, Sindhi, Balti, Shina, Marwari, Pashto, Wakhi and Punjabi are commonly spoken in Pakistan having millions of speakers and some of them are called provincial languages namely Sindhi, Punjabi, Pashto and Balochi. Punjabi language is a regional language spoken by 44% of the total population of Pakistan [4] with estimated 76 million of speakers having majority in Punjab province of Pakistan. Pashto and Sindhi are also the official languages of KPK (Khyber Pakhtunkhwa) and Sindh provinces respectively. Sindhi language is the official language of Sindh province of Pakistan spoken by 60 million people in Sindh and other areas of the world [2]. Sindhi adopts Arabic script for writing and inherits all of the posed problems of Arabic scripts with addition of new problems and challenges. Arabic possesses less number of character variations using same base shape whereas Sindhi extends the same base shape with more dots, creating more number of characters. The dot placement and orientations are altered so that more characters can be formed in Sindhi. A complete list of issues and challenges are identified in [5-6].

2. RELATED WORK

Latin OCRs are at their peak level of accuracy whereas research on Arabic adopting languages are still need attention. The OCR research on national and regional languages of Pakistan is at the very initiating stage. Few of the researchers are engaged in regional languages such as [7-11] for Urdu; [2,5,6,12,13] for Sindhi; [14-16] for Pashto language. Very little work has been done on Gurumukhi Punjabi which is written and spoken in India [17] whereas no any work is found on Shahmukhi Punjabi which is written and spoken in Pakistan and Balochi language OCR (to the best of knowledge of authors).

In [18] authors have presented various languages which use Arabic script for writing as Arabic script is a cursive script so the difficulties and challenges are posed at every stage of the character recognition process. The two writing styles of Urdu writing namely Naskh and Nastaliq have been explained with examples. The paper also presents a glance of researches done on various scripts like Sindhi, Yughur, Kashmiri, and Persian.

For the complete list of implementation challenges during a lifecycle of an OCR is presented in detail [6,16] for Sindhi language. The challenges faced by researchers for every language OCR such as Noise, Font, scanning resolution and others are presented in detail. The paper discusses every detail of issues and challenges for Sindhi OCR is in [5,16] such as writing style of Sindhi script, Dots and their importance in language and character definition and their use in pronunciations. The style of dots, nature of dots, direction of dots position and orientations of dots are given in detail. The character Shape groups, the number of characters and their analysis with respect to OCR and the Unicode representation is also discussed in details.

In [15] authors have reported a technique in which they have created a Pushto image database for Pushto character recognition. The differences of Latin script and Pashto script are also presented. The synthetically created text image database is a combination of ligatures selected from a novel. Only one font karor has been used for the creation of the Pashto image database with the help of software “Aisha soft” which is a legal property of Pashto academy situated at University of Peshawar, Pakistan. The created text images are in the form of .bmp format and four font sizes (12,14,1,18) have been used. The images containing 1000 ligatures along with 4 font sizes totaling 4000, are in 100x100 resolutions. This Pashto text image database has been named as FAST-NU.

3. PROPOSED SYSTEM

The complete description of proposed system is defined as under. First description for "how the glyphs are identified" is given and then different steps for character identification and recognition are defined.

3.1 Glyph Identification

Glyphs can be a presentations of a character shapes to be presented or displayed when any character is rendered. Glyph can be a presentation of one or more characters. For natural languages characters are written whereas digital fonts are represented by glyphs [18]. For any multilingual OCR, it is necessary to understand character shape and their glyphs so that the language can be identified and an efficient OCR can be implemented with ease.

3.2 Character Identification and Recognition

Character identification is a preliminary step for multi-script recognition when multiple language characters are optically recognized [2]. For the glyph identification and recognition of Sindhi script, the proposed system is illustrated in Fig. 1. The text image of Sindhi language is loaded and some preprocessing steps such as conversion into gray level and binary format is performed. The converted text image is then segmented for the line word and characters but as proposed system is limited to isolated characters so there is no word segmentation in the proposed system only the isolated characters are to be segmented. The lines are extracted from the text images and words (in our case, the characters) are extracted from the segmented lines. In classification, the glyphs are identified, the number of characters are found and then the output stage is containing the OCR process to recognize the character. The identified and recognized character is then exported to any Unicode supported editor.

3.3 Design and Development

For the implementation of the proposed system the MATLAB 2015 has been used and an interactive graphical User Interface has been created is shown in Fig. 2. The steps of proposed system are aligned in sequence so that the process can be understood easily. The stages of proposed system are described in following subsections.

3.3.1 Load Image

By pressing the load image button the various image formats can be loaded into the image box where it is fit according to the size. The image box contains the resized photo of the text whereas the system holds the original image also. The supportable image formats in MATLAB can be used for loading in the system. The loading of an image is shown in Fig. 3(a).

3.3.2 Preprocessing

The first stage will result in loading of an image and this loaded image is the input of preprocessing stage where

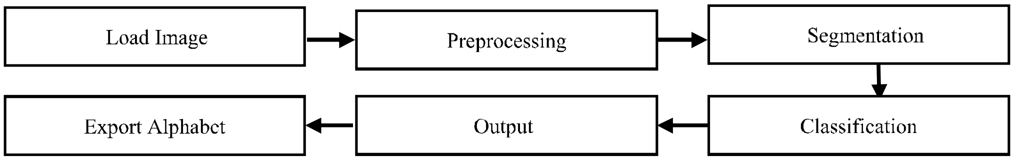


FIG. 1. PROPOSED SYSTEM FOR GLYPH IDENTIFICATION AND RECOGNITION

image noise is removed and the color channels are checked. If the image is in color, then the text image is converted into grey level format. The image is checked with the threshold value of 127 and further converted into binary format because for OCR two tone binary images are preferred as foreground and background are separated in binary images [7] as shown in Fig. 3(b).

3.3.3 Segmentation

The segmentation stage is considered the most important and crucial stage of any OCR. Here we used horizontal and vertical segmentation for separation of text lines from an image and characters from text lines. Free space is an indicator for segmenting text lines from a text image and characters from text lines. After

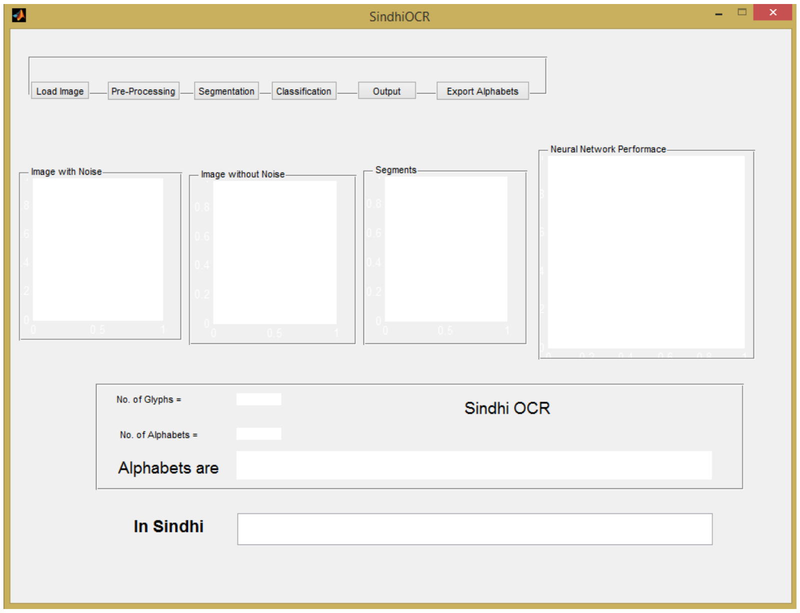


FIG. 2. INTERACTIVE GUI CREATED FOR GLYPH IDENTIFICATION

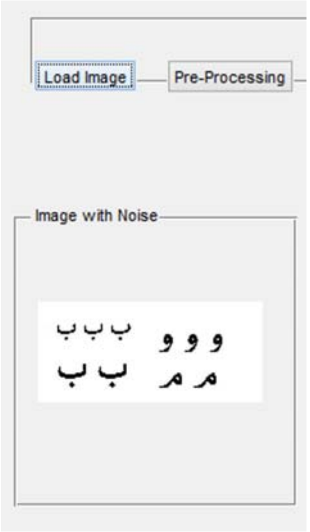


FIG. 3(a). LOADING OF AN IMAGE

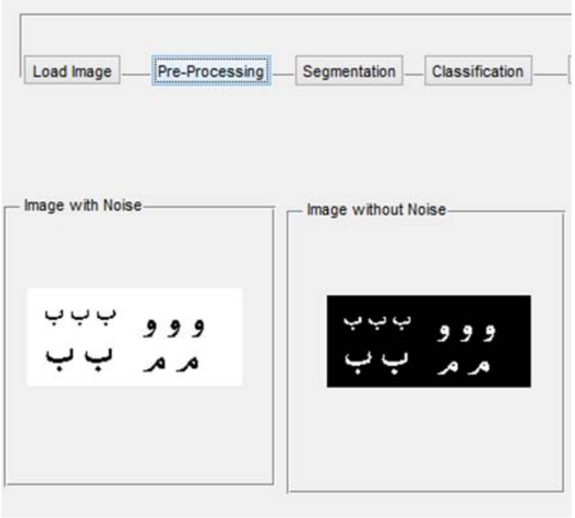


FIG. 3(b). PREPROCESSING STAGES

completely running segmentation algorithm the total number of glyphs are calculated and a number of glyphs presented in input image is written in proposed application. This process is shown in Fig. 4. The segmentation algorithm uses the formula to count the moments as shown in Equation (1).

$$HrI = \Sigma fi\ j(I) \qquad (1)$$

The Equation (1) will add the values of the individual locations in an image called I in horizontal locations or row wise, where value of individual location in I(image) might be zero or one. The white pixel represents background and black as text whereas the vice versa rule can be applied for segmenting characters from the image. The number of glyphs are displayed here in our case the number of glyphs are fifteen in number. It can be observed from Fig. 4. Where five characters are without a dot (مر and و) and another character of Sindhi language (ب) are five in number and each character is displayed along with its constituting dot so a total number of glyphs are fifteen in number.

3.3.4 Classification

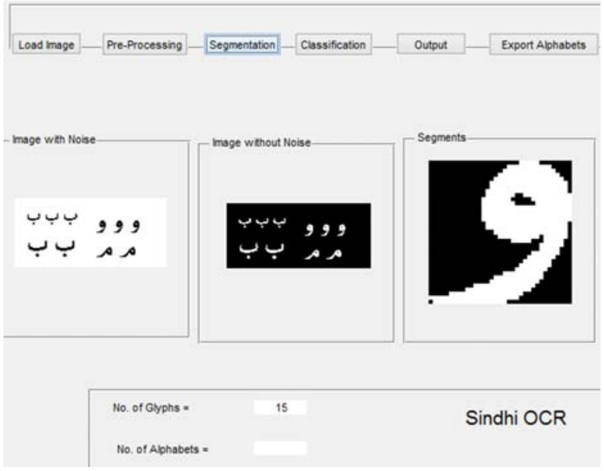
In this stage the segmented characters are used as input and then by using feature extraction algorithm [19], features are extracted to use along with neural network feedforward algorithm for the recognition of characters as shown in Fig. 5.

3.3.5 Output

The features extracted have been used to recognize the characters. With the language mapping the characters have been identified that which character is available in a specific language. We mapped Sindhi language script according to their number of characters and number of glyphs. The number of glyphs is the indicator for a particular language. After the glyph identification an OCR for isolated character has been applied so that the characters can be recognized. The recognized characters are displayed on User interface and with the help of export alphabet button these recognized characters can be exported to any text editor.

3.3.6 Export Alphabet

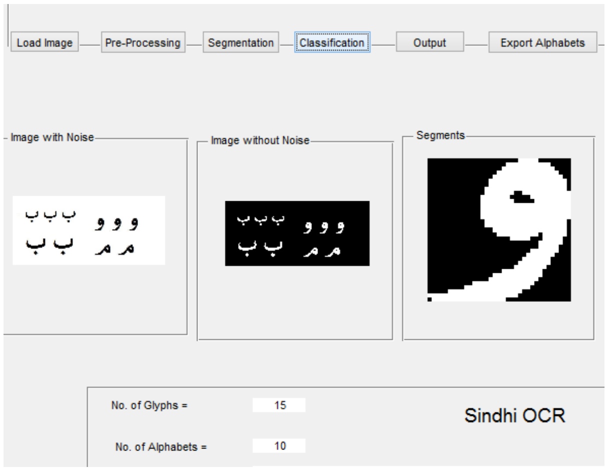
The last stage of the proposed system is to export alphabet in the same sequence to text editor supporting Unicode. In our case we are exporting a text file containing recognized characters. This text file can easily be opened in any text editor supporting Unicode.



No. of Glyphs = 15

No. of Alphabets =

FIG. 4. SEGMENTATION OF CHARACTERS AND IDENTIFICATION OF GLYPHS



No. of Glyphs = 15

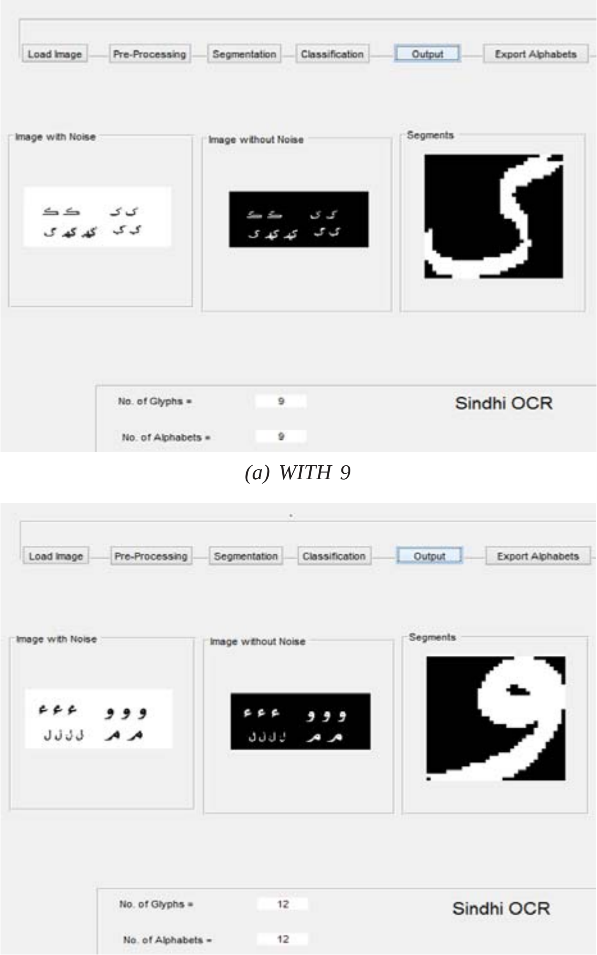
No. of Alphabets = 10

FIG. 5. IDENTIFICATION OF GLYPHS AND NUMBER OF CHARACTERS

4. RESULTS

Our customized application can identify various language characters and recognize with an acceptable accuracy. Some of the characters pose potential problems in recognition if their shapes are identical or with slight difference. The problems are increased if character shape groups [5-6] are the same. The system successfully identifies the glyphs and number of characters which information is sent to further algorithms. The proposed system successfully identifies the glyphs of Sindhi script of many other fonts also. The experimental results show that the system also identifies glyphs of other fonts. Fig. 5 shows that the 15 number of glyphs are available whereas they combine and form 10 characters of Sindhi script. The identification is done on the basis of shapes of characters.

Some of the characters in Sindhi script are without any dots and these characters are well identified by the proposed system and accurately identify the glyphs and the characters. Fig. 6 illustrates the identification of glyphs and characters accurately recognized. In Fig. 6(a) 9 glyphs and characters have been successfully identified and Fig. 6(b) also shows the 12 glyphs and characters. The system produces some of the logical errors when ambiguous dots have been found on the text image as shown in Fig. 7 where the two dots of character (ٻ) have been identified as one dot resulting 13 dots instead of 17 dots. This problem may be because of the connecting pixels in low resolution image whereas this problem can be easily handled when image is in high resolution and the font where dots have space between the dots.



No. of Glyphs = 9

No. of Alphabets = 9

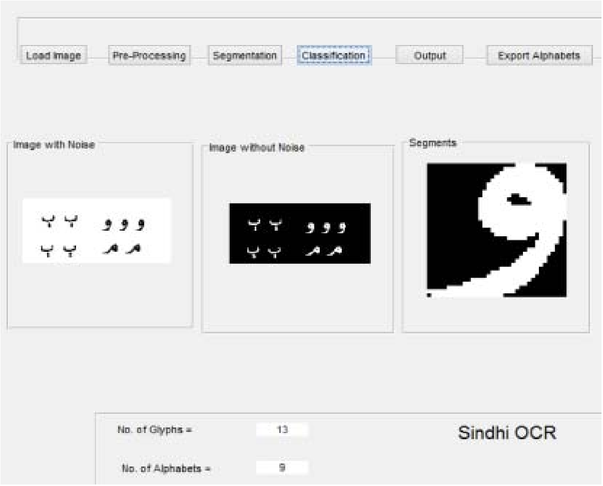
(a) WITH 9

No. of Glyphs = 12

No. of Alphabets = 12

(b) WITH 12

FIG. 6. IDENTIFICATION OF GLYPHS (CHARACTERS WITHOUT DOTS)



No. of Glyphs = 13

No. of Alphabets = 9

FIG. 7. IDENTIFICATION OF GLYPHS (CHARACTERS WITH DOTS)

5. CONCLUSION

The work on Sindhi OCR and handwritten recognition is at infant level and the glyph identification is an addition to the Sindhi OCR research. The glyph identification has many other applications in OCR and handwritten character recognition. The glyphs identification process can ease the process of OCR as many of the characters can be identified and differentiated prior recognition algorithms or classification. The proposed system identifies glyphs of various fonts.

ACKNOWLEDGEMENT

The authors are very grateful to the Department of Computer Engineering, King Faisal University, Al-Ahsa, Saudi Arabia, Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, NED University of Engineering & Technology, Karachi, Pakistan, for providing their resources to carry out this research. Valuable suggestions given by Dr. Akhter Hussain Jalbani, QUEST Nawabshah, are also highly acknowledged.

REFERENCES

[1] Pal, U., and Sarkar, A., "Recognition of Printed Urdu Script", Proceedings of 7th International Conference on Document Analysis and Recognition, Computer Society, Volume 2, pp. 1183-1187, Edinburgh, Scotland, Edinburgh, UK, 2003.

[2] Hakro, D.N, "Enhanced Segmentation and Feature Extraction for Sindhi Optical Character Recognition", Ph.D. Thesis, University Sains Malaysia, Malaysia, 2015.

[3] Sattar, S.A., Haque, S., Pathan, M.K., and Gee, Q., "Implementation Challenges for Nastaliq Character Recognition", Wireless Networks, Information Processing and Systems, pp. 279-285, Springer Berlin Heidelberg, 2009.

[4] (Languages, 2016) https://en.wikipedia.org/wiki/Languages_of_Pakistan (Accessed on 24th April, 2016).

[5] Hakro, D.N., Ismaili, I.A., Talib, A.Z., and Mojal, G.N., "Issues and Challenges in Sindhi OCR", Sindh University Research Journal (Science Series), Volume 2, No. 46, pp. 143-152, Jamshoro, Pakistan, 2014.

[6] Hakro, D.N., Ismaili, I.A., Talib, A.Z., Mojal, G.N., "A Study of Sindhi Related and Arabic Script Adapted Languages Recognition', Sindh University Research Journal (Science Series), Volume 3, No. 46, pp. 323-334, Jamshoro, Pakistan, 2014.

[7] Lehal, G.S., and Rana, A., "Recognition of Nastalique Urdu Ligatures", 'Proceedings of 4th International ACM Workshop on Multilingual OCR, pp. 7:1-7:5, New York, NY, USA, 2013.

[8] Akram, M., and Hussain, S., "Word Segmentation for Urdu OCR System", Proceedings of 8th Workshop on Asian Language Resources, pp. 88-94, Beijing, China, 2010.

[9] Pal, U., and Sarkar, A., "Recognition of Printed Urdu Script", Proceedings of 7th International Conference on Document Analysis and Recognition, Computer Society, pp. 1183-1187, Edinburgh, Scotland, UK, 2003.

[10] Sattar, A.S., "A Technique for the Design and Implementation of an OCR for Printed Nastalique Text", Ph.D. Thesis, NED University of Engineering & Technology, Karachi, Pakistan, 2009.

[11] Shamsher, I., Ahmad, Z., Orakzai, J., and Adnan, A., "'OCR for Printed Urdu Script using Feed Forward Neural Network", Proceedings of World Academy of Science, Engineering & Technology, 2007, URL: http://dx.doi.org/10.1007/978-3-540-89853-5 30

[12] Nizamani, A., and Janjua, N., "Sindhi OCR Using Back Propagation Neural Network", International Journal of Computer Science and Security, Volume 1, No. 3, pp. 1-8, 2011.

[13] Shaikh, N., Mallah, G., and Shaikh, Z., "Character Segmentation of Sindhi, an Arabic Style Scripting Language, using Height Profile Vector", Australian Journal of Basic and Applied Sciences, Volume 3, No. 4, pp. 4160-4169, 2009.

[14] Ahmad, R., Amin, S., and Khan, M., "Scale and Rotation Invariant Recognition of Cursive Pashto Script Using SIFT Features", 6th International Conference on Emerging Technologies, pp. 299-303, 2010.

[15] Wahab, M., Amin, H., and Ahmed, F., "Shape Analysis of Pashto Script and Creation of Image Database for OCR", International Conference on Emerging Technologies, pp. 287-290, Islamabad, Pakistan, 2009.

[16] Decerbo, M., MacRostie, E., and Natarajan, P., "The BBN Byblos Pashto OCR System", Proceedings of 1st ACM Workshop on Hardcopy Document Processing, pp. 29-32, 2004.

[17] Rani, R., Dhir, R., and Lehal, G., "Identification of Printed Punjabi Words and English Numerals Using Gabor Features", World Academy of Science, Engineering & Technology, 2011.

[18] Sattar, S.A., and Shah, S., "Character Recognition of Arabic Script Languages", 1st International Conference on Computing and Information Technology, pp. 502-506, Taibah University, Al-Madinah Al-Munawwarah, Saudi Arabia, 2012.

[19] Dileep, D., "A Feature Extraction Technique Based on Character Geometry for Character Recognition", arXiv Preprint arXiv:1202.3884, 2012.