Sindh Univ. Res. Jour. (Sci. Ser.) Vol.48(4) 839-844 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Isolated Optical Character Recognition

D. N. HAKRO, M. MEMON*, S. A. AWAN**, Z. A. BHUTTO, M. HAMEED

Institute of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan

Received 12rd January 2016 and Revised 5th July 2016

Abstract: Optical Character Recognition (OCR) is considered the fastest method of input which facilitates the data entry as easy and versatile. Glyph images yield a character code in OCR process. Sindhi language is one of the oldest languages of the world and much of the valuable work has been done in Sindhi language and literature. The OCR for Sindhi language is a need of time as almost all of languages are enriched with OCRs. Sindhi OCR is one of the most difficult systems because of large number of characters, four dots and connecting nature of characters. The paper discusses the process of development of Sindhi OCR for isolated characters. The system is designed for recognizing isolated Sindhi characters. Different Sindhi alphabet images were used for experiments. The system recognizes isolated characters of Sindhi alphabet images into editable text.

Keywords: Optical Character Recognition, Feature Extraction, Neural Networks, script

1. INTRODUCTION

Character Recognition is used in machine learning as a fastest method of input. It is important aspect for machine learning because significant attention has been given to this technology during last previous years (Pal and Sarkar, 2003). Considerable techniques of cursive as well standalone are proposed by researchers for the recognition (Jelodar *et al*., 2005). Arabic script is still focus point for the researchers and research efforts are continued on (Gillies *et al*.2003; Hamid and Haraty, 2001) as Arabic script recognition is not as mature as Latin and Japanese script OCRs. Many efforts have been taken and also the work is continued on Arabic (Parvez and Mahmoud, 2013; Slimane *et al*., 2013), Urdu (Pal and Sarkar,2003), Punjabi (Rani *et al*.,2011), Devanagari (Bansal and Sinha, 2002; Huanfeng and Doermann, 2003), Gujrati (Dholakia *et al*, 2007), Pashto (Decerbo *et al*., 2004), Persian (Pourasad *et al*., 2011; Jelodar *et al*.2005; Alaei *et al*., 2010), English (Saha and Som, 2010; Gupta and Long, 2007; Ganapathy and Liew, 2008), Chinese (Ding, 1997; L and She, 2004; Kim and Bang, 1992; Matic *et al*., 1990), Uighur (Li *et al*., 2012; Li *et al*.,2010), Korean (Kim and Kim,1996; Kim and Bang,1992), Japanese (Nagata,1998), kanji (Wakahara and Kimura,1999; Kobayashi *et al*.,1983), Sinhala (Ajward *et al*., 2010; Premaratne and Bigun, 2004), Thai (Tangwongsan and Sumetphong, 2008), Bangla (Basu *et al*., 2009; Angshul, 2007; Zhou *et al*., 2006), Telugu (Jawahar *et al*., 2003; Lakshmi and Patvardhan, 2004; Lakshmi, and Patvardhan, 2003), Hindi (Jawahar *et al*.,2003; Huanfeng and Doermann, 2003), Oriya (Chaudhuri *et al*., 2002), Kannada (Sagar, *et al*., 2008), Tamil (Hewavitharana and Fernando, 2002) and some other Indian Scripts. Indian languages possessing same script as Sindhi are also lucky enough that a huge work is done and this research is continued on. Sindhi Language efforts are at initial point on behalf of our knowledge only we are going to recognize the machine printed script of Sindhi language. Numbers of script are in use for writing Sindhi Language whereas Arabic script is practiced widely. Arabic script differs from European languages in that it is a purely language with cursive handwriting (Sagar *et al*.2008), characters are connected not separated individually in a word. Sindhi language is extended form of Arabic script so it is written in Arabic style of script which demands the character-level segmentation. Firstly, we start by introducing the differentiating features of Sindhi script, then steps to recognize isolate characters, which comprise of the scanning of an image, converting into two-tone binary images and image is segmented to lines, lines to characters (as characters are available in alphabet).

1.1 Introduction to Sindhi Language (Sindhi Text Features)

In a book titled "Sindhi Boli jo Bunn Bunyad ", According to Moulana Ubedullah Sindhi, "All the Holy books were sent in seven languages from heaven and remaining all world languages are derived from these seven languages. Hebrew and Sindhi are two languages among these heavenly languages (Alana, 2004). In fact,

++Corresponding emails: dill.nawaz@gmail.com,
E-mail: Dill.Nawaz@gmail.com,mmemon@, Zabhutto@usindh.edu.pk, shafique.neduet@gmail.com
*Institute of Business Administration, University of Sindh, Jamshoro,Pakistan
**Benazir Bhutto Shaheed University, Cheel Chowk, Lyari Karachi, Sindh, Pakistan.

it is the language of rich civilization of 5000 years old. According to Dr. Baloch (Alana, 2004) different opinions are available for the originity of Sindhi Language, one of them is the Sindhi Language might has remained the branch of Sansakrit via Varchada Apabhransha. Following are the characteristics of Sindhi language script.

1. Sindhi language set contains 52 characters and some of the characters possess two to four shapes because Sindhi language is cursive like Arabic. The detail is presented in (Hakro *et al*., 2014)

2. Sindhi is a cursive as it is extended form of Arabic script like, Persian, Urdu, and Pashto. Characters are connected and make a component.

3. Characters form a ligature by connecting to one or two characters results a cursive nature.

4. Fifteen groups can be formed according to character geometry, shape and characteristics.

5. Sindhi script consists of standalone (isolated) and ligatures whereas words may be the combination of isolated and ligatures.

6. The characters have no dots, one, two, three or four dots inside, above or below the character.

7. Writing style starts from right and follows to left whereas numbers follow inverse style.

8. A character used in a ligature changes its form according to its connecting possibility with succeeding or proceeding character. This characteristic is called context sensitivity.

9. Sindhi script has total 52 characters and as the cursive nature a character has more shapes shown in **(Fig-1a)**, which means more characters for recognition.

10. The letters have same base but dots make differences which make recognition process more difficult as shown in **(Fig-1b)**.

Isolated	Start	Middle	End
ڀ	ڀ	ڀ	ڀ
ڇ	ڇ	ڇ	ڇ
ڦ	ڦ	ڦ	ڦ

Fig-1(a): Different forms of letters according to their positions

Fig-1(b): Same base with different number and position of dots

1.2 Related Work

Nizamani and Janjua (2011) worked on isolated Sindhi character recognition with back propagation neural network. Their system takes input by drawing with mouse on the drawing control developed in Visual Basic. The systems start work by normalization then feature extraction with the lattice which collects feature vector by omitting free spaces. The next stage is the classification, which classifies the feature vector into training sets. The final step is recognition with the help of back propagation neural networks.

Shaikh and Mallah (2009) proposed a height profile vector based segmentation algorithm for Sindhi language. The cursiveness problem is addressed and thinned sub-words are segmented in several steps. Using extraction methods, the sub-words are extracted after Detecting baseline. Finally thinned primary strokes of characters are extracted. The six different connection patterns were identified in their research.

Nawaz *et al*. (2009) proposed an xml based algorithm for the recognition of isolate Urdu characters. The system has two sections namely training which contains height and width calculation, diacritic removal, chain code creation and saving to an xml file. The recognition phase contains the line, word and character segmentation, then class and character identification. The recognition accuracy is reported as 89% at the rate of 15 characters per second.

Almohri and Gray (2008) worked on hardware focused a DSP-based system for isolated Arabic characters using neural networks. The lines segmentation is done by horizontal projection and stored in a separate array. A total of 14 features are extracted and 4 of them are used in feature extraction phase in whole image.700 Arabic samples used to create the database. Fuzzy ART neural network with 14 inputs and 1 output is used for the recognition. A sum total of more than 20,000 trainings on the various samples in database resulted in 95% accuracy. The system can result 98% accuracy if the font and size is same as in database.

Parvez and Mahmoud (2013) proposed structural and syntactic based techniques for Arabic handwritten system. Preprocessing steps contain word, sub word extraction as well as slant correction. Slant correction further produces possible segments which are recombine in the last during recognition. Primary and secondary components are identified and labeled. Near vertical strokes are used for the writing slant correction and claimed that slant correction may improve the recognition rate. The body and dots of characters are recognized separately. Nearest neighbor is used for recognition with a rate of 79.58%.

Pourasad *et al*. (2011) proposed a novel method for Persian language based on spatial matching with contours points. Every pixel is a Cartesian coordinate in gray level image. On the matching of pixel brightness value in spatial domain the character images are said to be same. On increasing the number of points in the same size image, the efficiency is also increased. All Persian characters with variant sizes and fonts are stored, points extracted and descriptor vector stored. Euclidean distance and gradient value for every point is extracted in MATLAB. Character accuracy reported ranging from 91% to 100%.

In Sinhala OCR (Premaratne and Bigun, 2004) Premaratne and Bigun have proposed a novel approach for Brahmi Sinhala script by using Linear Symmetry. They proposed a segmentation free technique in which they used the orientation features so effectively that the basic alphabet may be recognized without segmenting to basic shapes. Linear symmetry has been used as detecting agent for skew detection as well as edge detection. The authors claim their proposed method has a slightly better performance as compared to traditional projection and frequency domain methods. Experiments were performed with identical font and different fonts. The images containing from 600 to 1200 characters with varying noise, some of from newspapers have been used in their experiments. The accuracy rate has been claimed from 84% to 93% for the same shape character set while from 75% to 88% for the different set.

2. PROPOSED SYSTEM

The proposed algorithm for Sindhi Optical Character Recognition has the following steps. First is Gray to binary conversion (binarization), word and character segmentation, feature extraction, and recognition shown in **(Fig-2)**.

2.1 Digitization and Noise Removal

The paper was first scanned through the flatbed scanner and the resolution was given 300 dpi, the page can also be scanned through the handheld scanner but handheld scanner is not suitable for the high resolution. Some of the testing papers were also scanned on low resolution. The page scanned is in colored form or grayscale which was converted into binary form containing only two values 0 or 1 in its digital values, this process is also called binarization. For the noise removal, there are some noise filters like, median filtering, maximum and minimum filtering. Median filter (using Matlab) was used to save the high frequency details of image and remove the salt and pepper noise.

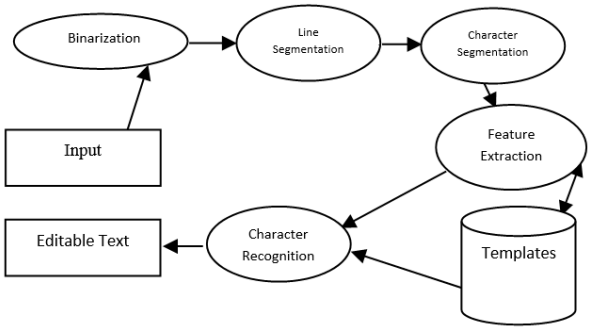

Fig-2: Block diagram of proposed algorithm

2.2 Line Segmentation

Line segmentation has been performed with the help of statistical moments, free space or a very low score between the lines is an indicator that the lines are separated and can be segmented very easily. The low score of moments is the indicator that there may be a little bit overlapping between the lines as shown in **(Fig-4)** below between line 5 and 6. Horizontal projection (Histogram) was the choice to separate lines. **(Fig-3a)** depicts the side by side view of the horizontal projection (Histogram). The peak shows the highest pixels in a line. Both ways can be used for line segmentation, white background and black foreground and vice versa, here the black background has been selected in which white pixels are shown with the peaks.

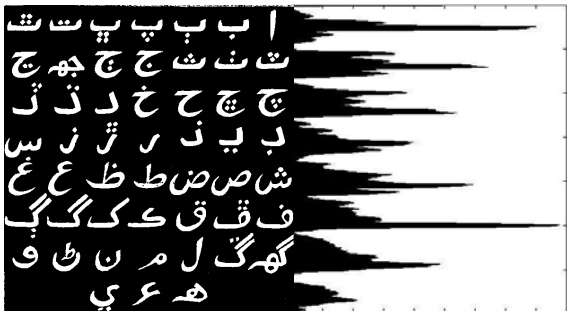

Fig-3(a): Line detection Using Histogram

Fig-3(b): Character detection Using Histogram

2.3 Character Segmentation

The word and character segmentation performed by the similar way, as the line segmentation performed but only the difference was column, base to segment the characters instead of the rows in previous step. Vertical projection was the technique to perform character segmentation and the space between the characters is the indicator. **Fig-(3b)** shows character detection and segmentation by vertical histogram of first line available in image of **Fig-(3a)**.

2.4 Feature Extraction

Feature analysis deal with the determination of feature set and descriptors to describe letters of alphabet whereas the feature extractor deals with the extraction of features relating to a character. These character features can then be given as input to the classifier (Haider and Gray, 2008). The two typical stages in character recognition are the feature extraction and the classification. The character is represented in such a way that it may differ to other characters while in the other technique it is given a suitable class according to its characteristics by a suitable classification method (Trier *et al*., 1996). For this purpose, an enhanced algorithm based on (Dileep, 2012) has been employed to for the feature extraction where various number of features have been extracted. These features are local zone based and some of the features are global and these extracted features are combined in a complete feature vector for Sindhi characters.

2.5 Recognition

The character recognition is the most crucial step for the OCR in which the extracted character is matched with the database of character vector. On the basis of features extracted in previous step namely feature extraction, classifier decides for the character class and on successfully matching process, the Unicode of corresponding image is produced as the output. The process is depicted in the following **(Fig-4)**.

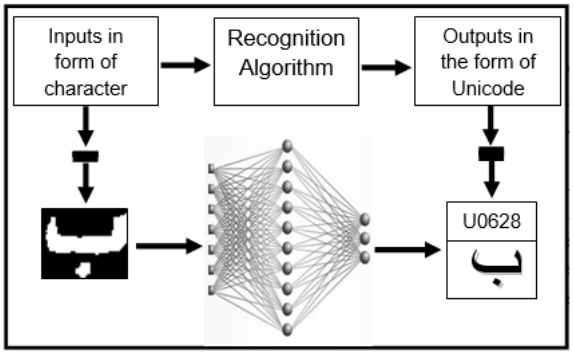

Fig-4: Recognition algorithm from image vector to Unicode

Neural networks are understood as efficient due to their generality, simplicity and learning ability (Zhou, 1999) and considered as good choice to use in handwritten character recognition (Khatatneh, 2006). The Artificial Neural Networks (ANN) has the natural tendency to store experimental knowledge (Haykin, 1994) and feed forward back propagation may produce reasonable results if configured properly in the situations when the input has never seen before (Bozinovic and rihari ,1989). Back propagation algorithm is suitable for the situations and produces better results especially when the input is unknown (Bozinovic and rihari ,1989). Feed Forward Back propagation for the recognition of Sindhi characters has been employed due to the fact of unknown inputs. For the experimental purpose different image vectors were normalized for the input defined in feature extraction but the algorithm can be applied not only these vector maps but it can be applied to other vectors with a very slight modification. The Network applied consists of input layer, a middle layer (hidden) and resulting output layer contains the vector with only 1 value in a particular location leaving others as 0. This is repeated up to 1 available for the last character "ye" (ي) in Sindhi alphabet, which is the one of the output character code out of 52 characters available in Sindhi language.
Hidden layer contains the different number of neurons adjusted during the experiments according to varying inputs. In output layer the winning neuron is giving one variable showing the number which is to be corresponded as the recognized character. The complete process is shown in **(Fig-5)**.

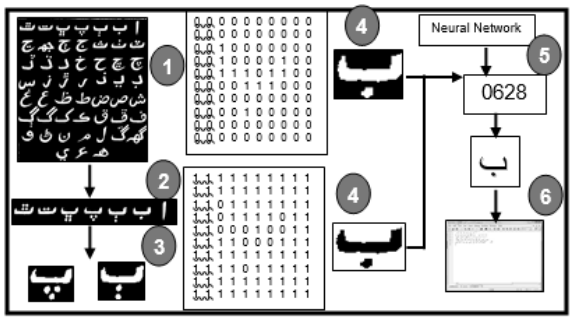

Fig-7: Complete process for Sindhi OCR

3. RESULTS

Binarization is considered as essential process because of the robustness and any image can be easily processed. The experiments were done on various sizes such as 5x7, 10x10, 10x14,24x42, 100x100,and 128x128. The alphabet images were scanned through scanner used for training and testing. The image is loaded in Matlab and processed in different routines, every routine called in sequence such as binarizing routine, which converts into binary image and noise

removal routine removes noise. The segmentation routines segment the lines and characters and those characters are resized into different sizes. The feature vectors are given to neural network which produces a target and it is converted into corresponding character. Various number of alphabet of machine printed Sindhi characters were scanned. The characters were used for testing and training. The results were calculated based on correct and incorrect recognized characters. The average accuracy achieved by proposed system was 93%. The characters with same base shape and only difference in dots are the reason for the incorrect recognition. The system was also tested on template matching an old feature extraction method and produced 100% accuracy but the size must be same for the input characters.

4. CONCLUSION

The new milestone has been set for others to set their research base on this background provided to them which is the utmost idea of the research. The system has been tested on various scanned alphabet images. The dots and noise sometimes create ambiguity in recognition that becomes a reason for decreasing the accuracy as well as the efficiency of the algorithm. The system recognized successfully the isolated characters from alphabet images. The system achieved 93% accuracy for the isolated Sindhi machine printed characters. The degraded performance was a result of the characters with same base shape and slight difference in dots. The research carried on has been a preliminary step towards the Sindhi optical character recognition and builds the new foundations for the connected characters or ligatures of Sindhi Language. This is the window for the extending this work to Sindhi Optical Character Recognition of Ligatures and grammar and so on.

REFERENCE:

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

Alaei, A., P. Nagabhushan, U. Pal, (2010), A Baseline Dependent Approach for Persian Handwritten Character Segmentation, *in* Proceedings of the 2010 20th International Conference on Pattern Recognition, 1977--1980.

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

Angshul, M., (2007), Bangla basic character recognition using digital curvelet transform, Journal of Pattern Recognition Research 2(1), 17--26.

Almohri, H., J. S. Gray, (2008), A Real-Time DSP-Based Optical Character Recognition System for Isolated Arabic characters using the TI TMS320C6416T, Proceedings of The 2008 IAJC-IJME International Conference, ISBN 978-1-60643-379-9.

Bansal, V., (2002), Segmentation of touching and fused Devanagari characters. Pattern Recognition 35, 875-893.R.

Chaudhuri, B. B., U. Pal, M. Mitra, (2002), Automatic recognition of printed Oriya script, Sadhana, Volume.27, Part 1, 23-34. (c) Printed in India.

Decerbo, M. E. MacRostie, P. K. Natarajan, (2004), The BBN Byblos Pashto OCR System. HDP' 04, November 12, 2004,Washington DC, USA. Copyright 2004 ACM 1-58113-976-4/04/0011.

Dileep, D., (2012). A feature extraction technique based on character geometry for character recognition, arXiv preprint arXiv: 1202. 3884.

Dholakia, J., A. Yajnik, A. Negi, (2007), Wavelet Feature Based Confusion Character Sets for Gujarati Script, Conference on Computational Intelligence Multimedia Applications, 2007. International Conference on Volume 2, Issue, 13-15 Page(s):366 – 370 Digital Object Identifier 10.1109/ ICCIMA.. 230Pp.

Gillies, A., E. Erlandson, J. Trenkle, S. Schlosser, (2003), Arabic Text Recognition System, Pattern Recognition letters, Vol. 24 , 284-299.

Ganapathy, V., K. Liew, (2008). Handwritten character recognition using multiscale neural network training technique, World Academy of Science, Engineering and Technology 39: 32–37.

Gupta, A., L. Long, (2007), Character recognition using spiking neural networks, Neural Networks, IJCNN 2007. International Joint Conference on, IEEE, 53–58.

Hamid, A., R. Haraty, (2001), A Neuro-Heuristice Approach for Segmenting Hand written Arabic Tex, ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, 25-06-2001 – 29-06-2001, 110-113.

Hakro, D. N., I. A. Ismaili, A. Z. Talib, Z. Bhatti, G. N. Mojai, (2014), 'Issues and Challenges in Sindhi OCR', Sindh University Research Journal (Science Series) 46(2), 143-152.

Hewavitharana, S., H. Fernando, (2002), A two stage classification approach to Tamil handwriting recognition, Tamil Internet, 118--124.

Huanfeng, M. A., D. Doermann**,** (2003). Adaptice Hindi OCR Using Generalized Hausdorff Image Comparison. ACM Transactions on Asian Language Information Processing, Vol.2 No.3, Pages 193-218.

Jelodar, M. S., M. J. Fadaeieslam , N. Mozayani, M. Fazeli, (2005), A Persion OCR System using Morphological Operators, Transactions on Engineering, Computing and Technology v4, ISSN 1305-5313.

Jawahar, C. V., P. Kumar, S. S. K. Ravi, (2003), A bilingual OCR for Hindi-Telugu documents and its applications, Proceedings of Seventh International Conference on Document Analysis and Recognition, 2003 Volume, Issue, 3-6 Page(s): 408 - 412 vol.1

Khatatneh, K., (2006), Probabilistic Artificial Neural Network for Recognizing the Arabic. Hand Written Characters, Journal of Computer Science 3 (12), 881-886.

Kim, H. J., P. K. Kim, (1996), Recognition of off-line handwritten Korean characters", Pattern Recognition, Volume: 29(2), 245 – 254.

Kobayashi, K., F. Yoda, K. Yamamoto, H. Nambu, (1983), Recognition of hand printed Kanji characters by the stroke matching method, Pattern Recognition Letters 1, 481 - 488.

Lakshmi, C. V., C. Patvardhan, (2003), A high accuracy OCR System for Printed Telugu Text, TENCON, Conference on Convergent Technologies for Asia-Pacific Region Volume 2, Issue, 15-17 . Page(s): Vol.2, 725 – 729, Digital Object Identifier 10.1109/ TENCON. 2003.1273274.

Lakshmi, C. V., C. Patvardhan, (2004), An optical character recognition system for printed Telugu text , Pattern Analysis & Applications, 1433-7541 (Print) 1433-755X (Online), 190-204

Li, G-h., S. Peng-fei, (2004), An approach to offline handwritten Chinese character recognition based on segment evaluation of adaptive duration. J. Zhejiang University SCIENCE, ISSN 1009-3095,J Zhejiang Univ SCI, 5(11):1392-1397.

Li, J., Z. Lu, A. Yimiti, F. Tan, (2012). Handwritten Uighur character segmentation and performance evaluation, Fourth International Conference on Machine Vision (ICMV 11), International Society for Optics and Photonics, 83491E–83491E.

Li, J., Z. Lu, A. Yimiti, F. Tan, W. Wang, (2010). Multiple feature cooperation based handwritten Uighur character segmentation on mobile phone, Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on, Vol. 9, IEEE, 384–388.

Matic, N. P. J. C. Platt, T. Wang, (1990), Quick stroke: An Incremental On-line Chinese Hand writing Recognition system. In D. Tourestsky, editor, NIPS, Vol. 2, 415-422. Morgan Kaufmann.

Nagata, M., (1998), Japanese OCR Error Correction using Character Shape Similarity and Statistical Language Model. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume: 2, COLING-ACL: 922-928, 1998.

Nizamani, A. M. N. U. H. Janjua, (2011), Sindhi OCR using Backpropagation Neural Network, International Journal of Computer Science and Security (IJCSS) , Vol. (1) : Issue (3). 65Pp

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

Premaratne, H., J. Bigun, (2004), A segmentation-free approach to recognize printed Sinhala script using linear symmetry, Pattern recognition 37(10), 2081--2089.

Sagar, B. M., Dr. G. Shobha, Dr. P. K. Ramakanth, (2008), OCR for printed Kannada text to Machine editable format using Database approach, WSEAS TRANSACTIONS on COMPUTERS, Issue 6, Vol. 7, ISSN: 1109-2750