Handwritten Sindhi Character Recognition Using Neural Networks

SHAFIQUE AHMED AWAN*, ZAHID HUSSAINABRO*, AKHTAR HUSSAIN JALBANI*,
DIL NAWAZ HAKRO**, AND MARYAM HAMEED**

RECEIVED ON 17.02.2017 ACCEPTED ON 21.08.2017

ABSTRACT

OCR (OpticalCharacter Recognition) is a technology in which text image is used to understand and write text by machines. The work on languages containing isolated characters such as German, English, French and others is at its peak. The OCR and ICR (Intelligent Character Recognition) research in Sindhi script is currently at in starting stages and not sufficient work have been cited in this area even though Sindhi language is rich in culture and history. This paper presents one of the initial steps in recognizing Sindhi handwritten characters. The isolated characters of Sindhi script written by thesubjects have been recognized. The various subjects were asked to write Sindhi characters in unconstrained form and then the written samples were collected and scanned through a flatbed scanner. The scanned documents were preprocessedwith the help of binary conversion, removing noise by pepper noise and the lines were segmented with the help of horizontal profile technique. The segmented lines were used to extract characters from scanned pages.This character segmentation was done by vertical projection. The extracted characters have been used to extract features so that the characters can be classified easily. Zoning was used for the feature extraction technique. For the classification, neural network has been used. The recognized characters converted into editable text with an average accuracy of 85%.

Key Words: Character Recognition, Handwritten, Sindhi, Segmentation, Feature Recognition.

1. INTRODUCTION

Optical Mark Recognizer is commonly used in examining candidates in entry test for the universities and other objective type of tests in various job examinations. An OCR is also a type of recognition in which text of image is recognized which is also considered as the faster method to input the text. ICR is an effective recognizer for the text which is written by human hand and this form of recognition is considered as challenging job as every writer has different style of writing and it is a very challenging job when recognizing skewed, overlapping, disconnected text. The handwritten recognition introduces more challenging problems while recognizing Arabic script adopting languages as Latin script languages have already peaked to the perfection [1]. Many of the systems such as Sindhi Dictionary [2-3], Sindhi Unicode based word processor [4], Sindhi OCR [5-7], and Sindhi text image databases have been proposed but to the best of our knowledge no work has been conducted yet on Sindhi handwritten character recognition.The recognition of handwritten Sindhi

Corresponding Author (E-Mail: shafique.neduet@gmail.com)
* Department of Information Technology, Quaid-e-Awam University of Engineering, Sciences & Technology, Nawabshah.
** Institute of Information & Communication Technology, University of Sindh, Jamshoro.

characters is an effort to open a new window of research for the researchers working on pattern recognition and with the minor modification the generalized algorithms can be used with the other languages adopting the Arabic script.

2. RELATED MATERIAL

The OCR is comparatively easier as compared to handwritten recognition [8] as OCR recognizes the machine printed characters whereas handwriting recognition is considered typical job as every person writes with his/her own style and writing is also constrained or unconstrained with limits or without limits. Languages using Latin scripts are relatively easy as these languages use non-cursive scripts where characters are not connected witheach other and every single character is in its isolated form. The OCRs for these languages are understood as comparatively easier than the languages using Arabic script where characters are connected witheach other and characters change their shape according to the placement inside a sentence.

Asabi and Bigun[8]recognized Ethiopian official language Amharic words by using Hidden Markov Models. The Amharic words are written unconstrained. The structural features have been used for the recognition. The methods have been used for the recognition and both are using structural features. Feature design, character feature lists and feature level concatenation are three components for the recognition model. The structural features comprise primary strokes of Amharic characters and spatial relationships of these strokes. The character segmentation is not performed whereas the whole single line has been detected using direction field tensor. The features have been extracted using structural features. The recognition rate is reported for different datasets. The recognition rate decreases from 74-41% when samples are increased.

Pal and Sarkar [11] made use of water reservoir for isolated Urdu characters. The text lines are segmented using projection approach to extract Urdu isolated characters from the lines. The indication of free space is used for segmenting lines from each other. The testing of system wascarried out on clean pages as well as cheap alphabet children books.The line segmentation of the system produces 98.3% and the character segmentation accuracy have been reported as 96.9%. Testing was performed on 3050 characters and 97.8% is the recognition rate reported for basic characters and numerals. Two classes for the system errors were defined namelysegmentation error with a rate of 0.7% and tree classification errors with a rate of 1.1% of overall errors.

Hakro and Talib [12] created a text image database for Sindhi OCR containing 15 billion of Sindhi character images forming 4 billion words. The words and character text images were created in more than 150 fonts, 4 styles, 4 font angles and font weights. A custom built software has been created to convert text to text images in various fonts, colors, sizes, anglesand various other features. To make the database versatile and extendable the text was collected from various sources such as books, web portals, theses and other sources of Sindhi language.

3. METHODOLOGY

3.1 Peculiarities in Sindhi Character Recognition

Sindhi language has 52 characters nearly double as many characters available in Arabic.Sindhi is the largest extension of all languages that adopted Arabic script. The characters change shapes according to their position in a word. The characters have one to four shapes in writing. The single character base shape represents multiple characters with only difference in the number of dots, position and orientation of dots.

There are additional base shapes in Sindhi script. Four dotted characters are available in Sindhi alphabet. A detailed discussion is present in [5] including the number of dots, importance, orientation of dots and other peculiarities of Sindhi OCR.

Preprocessing: The handwritten papers written in Sindhi characters were scanned and passed from preprocessing techniques. The Sindhi handwritten paperswere scanned at 300dpi and some of the images were scanned at 200dpi. The high dpi was selected due to the some of the errors and faded writings and disconnected characters during writing process. One of the writing samples is shown in Fig. 1. The pen for writing was used black as well as blue. So, the paper which was scanned in colored image was converted into greyscale and then followed by grey scale to black and white conversion. The process of convertingtext image into black and white is also called binarizationin which only two colors are available namely black and white. Blackis represented by 0 and whereas white represented by 1. After scanning and conversion to black and white image, noise removal has been applied so that small dots during scanning called noise can be removed and pepper noise removal has been used for this purpose. Fig. 2 shows the process of conversion. Fig. 2(a) shows the conversion from color image into grey scale and Fig. 2(b) shows the conversion of grey scale to black and white image. The reason for conversion to black and white or two tone coloring is the need for the identification of text and subtraction of the background. The further OCR processing can be done on both type of images such as white foreground with black background or black foreground with white background as the ultimate target is to identify the text and recognize the characters.

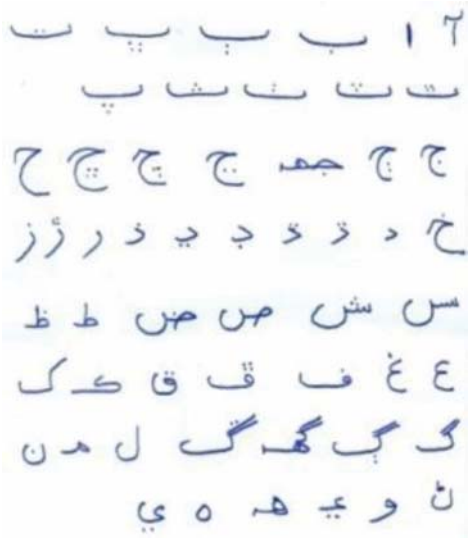


FIG. 1. SCANNED WRITTEN SINDHI TEXT IMAGE

(a) GREY LEVEL *(b) BINARY IMAGE (TWO TONE)*

FIG. 2. IMAGE CONVERSION

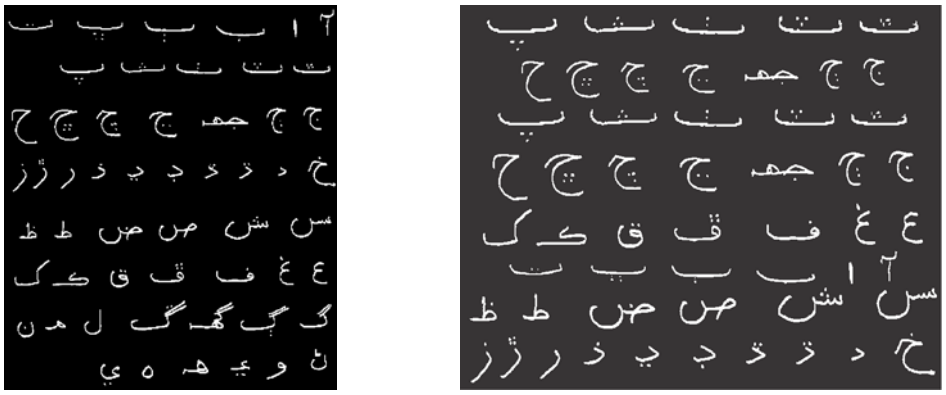
3.2 Segmentation

The Sindhi handwritten text imageswere scanned, and converted and they are now ready to be segmented for the characters. The handwritten imagesweresegmented into text line containing Sindhi characters. The text lines are segmented by counting the moments or white spaces between two text lines. These white lines are counted and the availability of white dots and unavailability of text indicates that there is a free space and lines can be easily segmented. The same process can be applied for the word segmentation but this study is recognizing of isolated characters, So these standalone characters can be considered as words.

The segmentation approach which was applied for line segmentation was basically applied in horizontal whereas the characters available in single segmented line are segmented by using the same approach except it is a vertical order. The process of segmenting lines is shown in Fig. 3 and the segmentation of characters is shown in Fig. 4.

3.3 Feature Extraction

Feature extraction is the important section of the OCR process. A good feature extraction approach can increase the rate of recognition accuracy [8]. A large amount of calculation can be saved and other advantages are available with an efficient feature extraction method. A modified feature extraction method based on zoning [13] was used to extract features of Sindhi characters written by various subjects. This feature extraction method has provided an acceptable accuracy rate [9] and this was the reason to select this approach for extracting Sindhi character features. The zoning approach is based on the character's geometry. The image of the extracted character is divided into 3x3 zones and a total of 9 zones have been formed to represent an isolated Sindhi character. The line and stroke based features are extracted from a zone or a segment from segment 0-8. The local features are extracted and represented in a feature vector. The local features are then combined into global vector as shown in Fig. 5.



a) SINDHI TEXT IMAGE *(b) SEGMENTED TEXT LINE*

FIG. 3. SEGMENTATION OF LINES

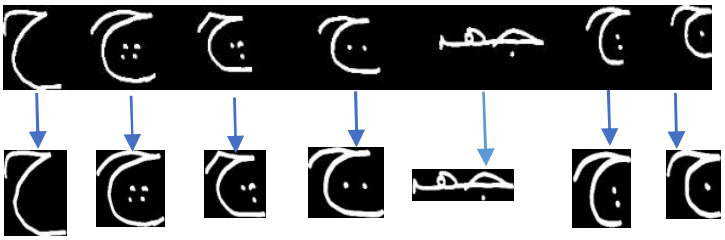


FIG. 4. SEGMENTATION OF CHARACTERS: (A) SINDHI TEXT LINE: (B) SEGMENTED CHARACTERS

The zoning technique for the feature extraction has been selected due to its suitability of the Sindhi script as it can be observed in Fig. 5 that a Sindhi character is divided (zoned) into nine zones. These zones have differentiating characteristics such as white to black pixel ratio or black to white pixel ratio. These differences are because the Sindhi characters vary in size and shapes and their shapes sometimes differ by only one or two dots [5]. The changing shape of Sindhi script challenge is suitably handled by zoning algorithm.

3.4 Recognition

The extracted features were matched with trained samples of handwritten characters. For classification and recognition,the ANN (Artificial Neural Network) is used.Back propagation FFN (FeedforwardNetworks) tend to produce good results specially when inputs are unknown. The intense training can increase the accuracy and performance of recognizing Sindhi characters written by various subjects.

The various subjects were asked to write Sindhi isolated characters on a paper and then these written papers were scanned through scanner. These images were passed through preprocessing. The images were segmented for lines and then the characters were extracted from text images. These character images were used to extract features and then with the help of ANN the features were matched with existing training images. The recognition accuracy achieved is 85%. The accuracy was calculated manually with the help of correctly recognized numbers and their average. Table 1 illustrates the percentages of correctly recognized characters. FromTable 1, it can be observed that the character produced highest accuracy because of absence of dots and the shape is different from other characters. The characters without dots have produced high accuracy and the examples are the characters such as (ا) (ح) (د) (ر) (ڪ) (ک) (ل) (م) and (و). The two characters without dots (س) (ص) were difficult to recognize due to the shape containing angular form.

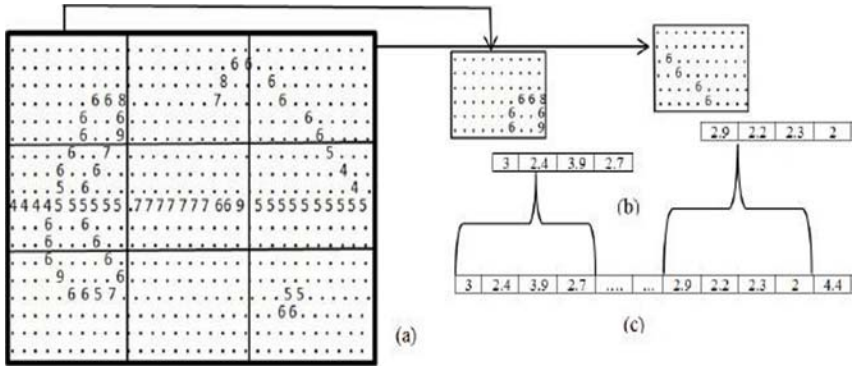


FIG. 5. FEATURE VECTOR FORMATION: (A) IMAGE DURING PROCESSING (B) ZONES AND LOCAL FEATURES (D) FINAL FEATURE VECTOR

TABLE 1. PERCENTAGE OF CORRECTLY RECOGNIZED HANDWRITTEN CHARACTERS (ISOLATED)

Character	%	Character	%	Character	%	Character	%	Character	%
ا	100	ج	80	ڊ	78	ض	87	گ	93
ب	92	جھ	86	ڈ	86	ط	94	گھ	77
پ	86	چ	79	ډ	77	ظ	77	ڳ	80
ڀ	70	ڇ	90	ذ	86	ع	95	گ	82
ت	85	ڃ	82	ر	90	غ	93	ل	90
ٹ	81	ڄ	70	ڑ	77		88	م	95
ٽ	72	ح	90	ز	89	ق	89	ن	88
ٿ	85	خ	88	س	71		90	ڻ	84
ث	89	د	97	ش	60	ک	95	و	96
ٺ	78	ڈ	80	ص	74	ک	94	ء	85
						ھ	92	ي	88

4. CONCLUSION

A lot of work has been done on various languages of the world regarding the improvement of OCR technology as every script pose different challenges for the researchers. A lot of work has also been done on other Sindhi computing but very little work has been done on Sindhi Script and its recognition. The work on handwritten character recognition is still in infancy. This is the first step towards the recognition of Sindhi handwritten words and sentences and ultimately the Sindhi handwritten text images. Text imageswere written by various subjects containingSindhi charactersand scanned through flatbed scanner.The experiments have been performed in Matlab 2015a. The scanned images were preprocessed and isolated characters were segmented preceded by line segmentation. The isolated Sindhi characters written by various subjects were recognized and the average accuracy rate obtained is 84%. A window is open for this system to be implemented on full text images of Sindhi script written by various subjects.

ACKNOWLEDGEMENT

The authors are very grateful to the Institute of Information Technology, University of Sindh, Jamshoro, and Department of Information Technology, Quaid-eAwam University of Engineering, Sciences & Technology, Nawabshah, Pakistan, for providing their resources to carry out this research.

REFERENCES

[1] Hamid, A., and Haraty, R., "A Neuro-Heuristice Approach for Segmenting Hand written Arabic Tex", ACS/IEEE International Conference on Computer Systems and Applications, Beirut, Lebanon, pp. 110-113, 25-26, June, 2001.

[2] Bhatti, Z., Ismaili, I.A., Hakro, D.N., and Waqas, A., "Unicode Based Bilingual Sindhi-English Pictorial Dictionary for Children", American Journal of Software Engineering, Volume 2, No. 1, pp. 1-7, 2014.

[3] Bhatti, Z., Waqas, A., Ismaili, I.A., Hakro, D.N., and Soomro, W.J., "Phonetic Based SoundEx&ShapeEx Algorithm for Sindhi Spell Checker System", arXiv Preprint arXiv, 1405.3033, 2014.

[4] Bhatti, Z., Ismaili, I.A., Soomro, W.J., and Hakro, D.N., "Word Segmentation Model for Sindhi Text", American Journal of Computing Research Repository, Volume 2, No. 1, pp. 1-7, 2014.

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

[6] Hakro, D.N., Talib, Z., and Mojai, G.N., "Multilingual Text Image Database for OCR", Sindh University Research Journal (Science Series), Volume 47, No. 1, pp. 181-186, Jamshoro, Pakistan, 2015.

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

[8] Assabie, Y., and Bigun, J., "Offline Handwritten Amharic Word Recognition", Pattern Recognition Letters, Volume 32, No. 8, pp. 1089-1099, 2011.

[9] Hakro, D.N., "Enhanced Segmentation and Feature extraction approaches for Sindhi Optical Character Recognition", Ph.D. Thesis, Universiti Science Malaysia, Malaysia.

[10] Zaafouri, A., Sayadi, M., and Fnaiech, F., "Printed Arabic Character Recognition using Local Energy and Structural Features", 2nd International Conference on Communications, Computing and Control Applications, pp. 1-5, 2012.

[11] 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, Scotland, Edinburgh, UK, 2003.

[12] Hakro, D.N., and Talib, A.Z., "Printed Text Image Database for Sindhi OCR", ACM Transactions on Asian Low-Resource Language Information Process, Volume 15, No. 4, pp. 1-18, 2016.

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