Optical Character Recognition (OCR) System For Saraiki Language Using Neural Networks

[1]M. T. Jan, [2]Y. Saleem
[1,2]Department of Computer Science & Engineering, University of Engineering and Technology, Lahore
[2]ysaleem@gmail.com

***Abstract*-**Saraiki language is one of the local languages of Pakistan. It is spoken and understood over a large geographical part of Pakistan. Little work has been done to develop Optical Character Recognition systems for local languages due to the complex writing system. The OCR system for Saraiki language can help to digitize the language literature. This work presents an OCR system that uses the Neural Network to recognize the printed text images of Saraiki (Urdu/Arabic/Punjabi) language generated in MS Word. Neural Network is trained with the segmented and isolated character set. At first, characters are extracted from the text image using segmentation approach. These segmented characters are then fed to the Neural Network in order to be recognized. MATLAB is used for the implementation of the OCR system that at present shows about 85% accuracy.

***Keywords*-**Saraiki OCR (SOCR), Feed Forward Neural Networks (FFNN), Machine Learning, Pattern Recognition.

I. Introduction

Saraiki language belongs to the Indo-Aryan (Indic) languages. Saraiki is majorly spoken in Pakistan but it is also spoken in some areas of India and Britain.

A. Saraiki Language

The areas of "Saraiki belt" includes Southern Punjab, Indus Valley, Northern Sindh, Bannu, Tank, Jampur, Dera Ghazi Khan and Dera Ismail Khan [i] . It is transliterated as "Sirāikī".Saraiki language was recognized at the national level of Pakistan for the first time in the 1981 Census. This is the fourth most spoken local language in Pakistan after Punjabi, Pashto, and Sindhi. It is in the 61st number as a world's largest language [ii]. Saraiki is usedas a medium of expression by more than 18 million people of southern Pakistani Punjab, adjacent border regions of Khyber Pakhtunkhwa (KPK) and northern Sindh, and nearly by some 70,000 emigrants and their descendent in India [iii].

B. Saraiki Writing System

Saraiki language is written from right to left using Arabic and Urdu language scripts as the majority of alphabets of Saraiki language are present in both the languages[iv]. Saraiki language has some special characters that are additional to the Urdu language. These special characters are ټ (ṭë), ڄ (ĵë), ڈ (ḍäl), گ (ğäf) and ڻ(ṇŭn). Figure 2 shows a detailed character set for the Saraiki language with symbolic representation and hexadecimal Unicode for each character.

II. Literature Review

Saraiki script has a cursive nature. Alphabets change their shape when combined with other alphabets to form differentwords.Figure 1 shows the shapes of a letterع('ain) at different positions in a word when used in combination with some other letters [v].

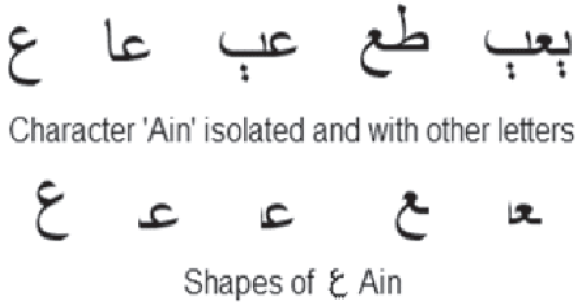

Fig. 1. Shapes of ع "Ain" in formation of Saraiki words.

Reference [vi] suggested there should be some centralized repository of words, usually known as a lexical database for cross-lingual processing. Natural language processing or natural language engineering has many tasks such as word sense disambiguation, machine translation, part of speech tagging [vii]. All these tasks also need large-scale lexical databases. NLP is a theory-motivated range of computational techniques for the automatic analysis and representation of human language. NLP research has evolved from the era of punch cards and batch processing (in which the analysis of a sentence could take up to 7 minutes) to the era of Google and the likes of it in which millions of webpages can be processed in less than a second [viii]. Saraiki is also written using different font styles like Urdu. Most popular fonts used are Nastalique style and Naskh style [ix] .

The efficiency of an OCR system is highly dependent on the writing style used for the particular language. An OCR system designed for one writing style might not work for the other writing styles due to thedifferent cursive structure.Despite a large character set Saraiki/Urdu has a small set of character classes which are easily distinguishable from one another.The characters belonging to a class are similar in shape but differ from each other due to the use of dots or symbols above or below these shapes [x].

III. METHODOLOGY

In this work, the methodology used for the development of the OCR system is word segmentation based. Characters from the words or ligatures of the language are extracted and then fed to the neural network in order to get recognized. Saraiki language has forty-four basic characters and a special character'Do-Čashmī Hë'. Many characters take multiple shapes based on their position in the ligatures. Each character is assigned a unique bit sequence for representation. The SOCR system also works well for some other languages like Urdu, Arabic, Persian and Punjabi [xi].This system has been designed to work with the Arial writing style used in the Microsoft Word, but it can be used for some other writing styles like Shamsheer, Aasaar, Batool, Unwan and Bombay Black used in Inpage Urdu as shown in Table I.

Table. I: Character For Saraiki Language for Symbolic Representation and Unicode [iii]

تھ	ٹ	پھ	پ	ٻ	بھ	ب	آ	ا
تھے	ٹے	پھے	پے	ٻے	بھے	بے		الف
thë	të	phë	pë	b̲ë	bhë	bë		Alif
th	t	ph	p	b̲	bh	b	ã	Ã
	062A		067E	067B		0628	0622	0627
ح	چھ	چ	ڃ	جھ	ج	ث	ٹھ	ٹ
وڈی حے	چھے	چے	ڃے	جھے	جیم	ثے	ٹھے	ٹے
vaḍḍī ḥë	čhë	Čë	ĵë	jhë	jīm	t̲ë	ṭhë	ṭë
ḥ	čh	Č	ĵ	jh	j	t̲	ṭh	ṭ
062D		0686	0684		062C	062B		0679
رھ	ر	ڎ	ڈ	ڈھ	ڈ	دھ	د	خ
رھے	رے	ڈال	ڈال	ڈھے	ڈال	دھے	دال	خے
rhë	rë	ḍäl	ḍäl	ḍhë	ḍäl	dhë	däl	ẖë
rh	r	ḍ	ḍ	ḍh	ḍ	dh	d	ẖ
	0631	0630	0759		0688		062F	062E

ط	ض	ص	ش	س	ژ	ز	ڑھ	ڑ
طوے	ضاد	صاد	شین	سین	ژے	زے	ڑھے	ڑے
ṭ̈ë	żuäd	ṣuäd	shīn	sīn	ẑë	zë	ṛhë	ṛë
ṭ	ż	ṣ	sh	s	ẑ	z	ṛ	ṛ
0637	0636	0635	0634	0633	0698	0632		0691
گھ	گ	کھ	ک	ق	ف	غ	ع	ظ
گھے	گاف	کے	کاف	قاف	فے	غین	عین	ظوے
ghë	gäf	Khë	käf	qäf	fë	ġain	'ain	ẓ̈ë
gh	g	Kh	k	q	f	ġ	'	ẓ
	06AF		0643	0642	0641	063A	0639	0638
ٹ	ں	نھ	ن	مھ	م	لھ	ل	پ
ون	نونغنہ		نون		میم		لام	ےاف
n̆ün	nün ġunna	Nh	nün	mh	mīm	lh	läm	Ğäf
n̆		Nh	n	mh		lh	l	G
0768	06BA		0646		0645		0644	06B3

ے	ی	ء	ھ	ہ	و
بڑی ے	چھوٹی ے	ہمزا	دوچشمیے	چھوٹی ھے	واؤ
vaḍḍī yë	čhöṭī yë	hamza	dö-čhashmī hë	čhöṭī hë	Väö
e	y	´	h	h	v
06D2	0649	0621	06BE	0647	0648

The character segmentation approach used in this system works for any font size, but in this work, the system has been trained to work with "24" font size of Arial writing style used in Microsoft Word documents. A detailed architecture of the Saraiki OCR system is given in Fig. 3. The system is divided mainly into three modules.

A. Preprocessing
B. Segmentation
C. Recognition

A. Pre-processing

Pre-processing phase involves the procedures like skewing the text image, noise removal and extracting the individual lines of text. Microsoft Word is used for the generation of input Saraiki text to be used in the system as input image files. For the Saraiki OCR system, individual text lines are provided as input (bitmap image files) to the system which is assumed to

be noise-free and skewed. The extra white area above and below the text image is removed, and the image is then converted to the binary bitmap image of pure black and white pixels. This pre-processed image text is used in the later stage for character segmentation.

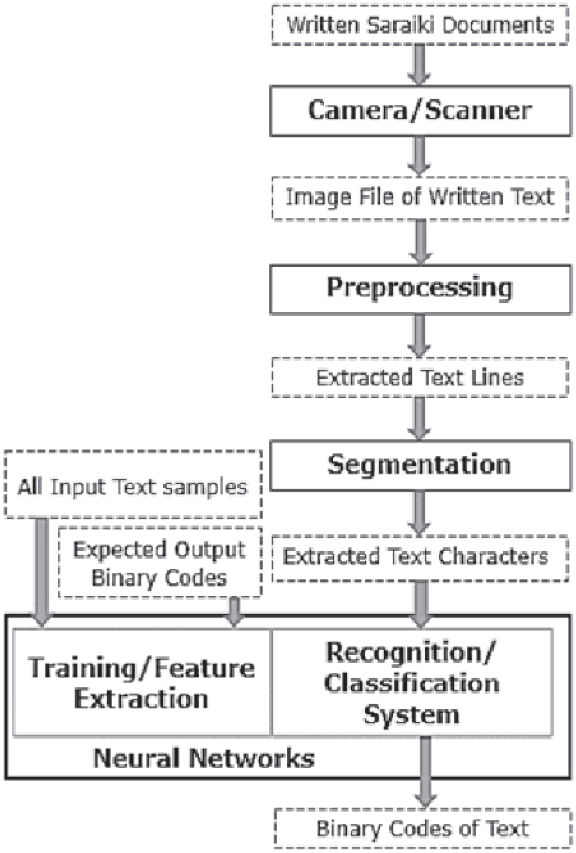

Fig. 2. Saraiki OCR (SOCR) System architecture.

B. *Segmentation*

In this phase, the pre-processed text image is used to extract each possible character constituting this text. In Arabic like languages, words are formed by the combination of one or more ligatures [xii].Each ligature is an isolated letter or a combination of multiple letters. For the used font, characters join each other on the baseline to form a ligature. Segmentation algorithm exploits the nature of the font.

It works on the principle of measuring the vertical and horizontal pixel strength variations above and below the baseline of the text image while traversing the image horizontally. Figure 4 shows the text structure of Saraiki language using different ligatures and characters.

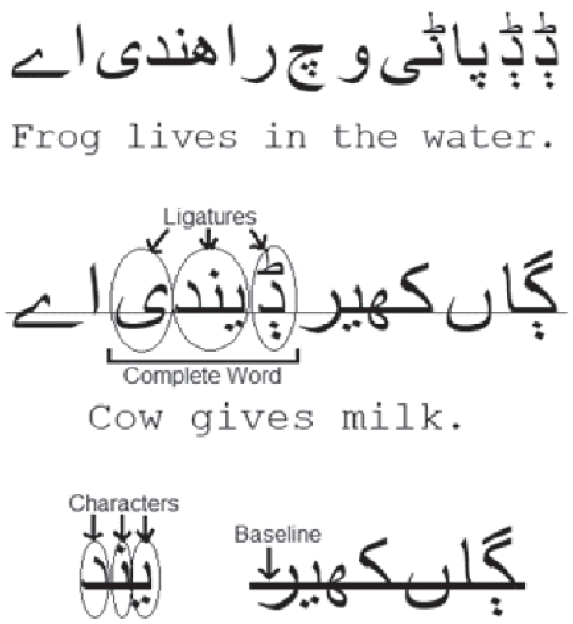

Fig. 3. Saraiki text formation using Ligature and Characters

In segmentation, firstly, the ligature words are isolated by exploiting the vertical gap present between the ligatures which results in the zero-pixel strength of black pixels. Secondly, segmentation algorithm finds the minimum vertical pixel strength which is mostly consistent throughout the ligature. This minimum vertical pixel strength is actually the height of the baseline on which joining of characters takes place for the used writing style. This minimum vertical pixel strength is found between every two joined characters in order to isolate the possible characters forming the ligature.

The result ofsegmentation module is the text image with isolated characters constituting the text which are fed to the Feed Forward Neural Network (FFNN) in order to get recognized. Figure 4 shows the segmented text image window of MATLAB after the segmentation process.

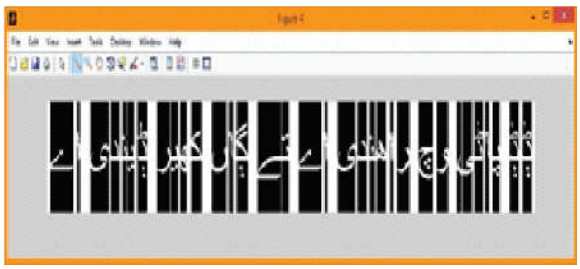

Fig. 4. MATLAB window showing segmented Saraiki text

C. *Recognition*

A three-layer (Input, Hidden and Output) Feed Forward Neural Network is made usingMATLABtools to train and simulate character recognition. A total number of sixty-one charactersand shapes are identified and generated for the Saraiki OCR system. A hundred samples of all the character shapes are extracted from training the network. Each input (training and simulation) sample is resized to 30x21 dimension to match the input layer of the Neural network which consist of 630 (30x21) neurons. The output layer of the network consists of six neurons which constitute the 6-bit target number for every possible input character shape. This target sequence is then converted to the 16-bit Unicode used in Microsoft Office for every character.

The Hidden layer of the neural network consists of 4000 neurons which are chosen based on the different trial results. Transfer functions used for supervised learning are 'Tansig' for the second layer (Hidden) and 'Logsig' for the third layer (Output). 'Trainscg' is the training function used to generate bias values and weights for the network.FFNN is generated with above settings that are trained with sixty-one hundred training samples for the SOCR system.

All the segmented characters generated in the segmentation phase are resized to 30x21 dimensions to match input layer of FFNN. These resized characters are fed to the Neural Network using the simulation 'sim' function of MATLAB which results in the 6-bit code for each character. This code is then converted to the hexadecimal code (Unicode) of each character used in MS Office. If the character is not recognized, then "Character Not Matched" message is displayed. Otherwise, each character and its recognized Unicode is displayed as output.

IV. Conclusion

The research aims to provided that SOCR System has been tested with different text samples, and it gives about 85% accurate recognition. Segmentation procedure shows above 90% accurate results. In this improved segmentation approach, garbage characters are minimized by considering the character joins over the baseline only. Characters like س, ش, ص and ض produce a character ں as garbage when occurring at the end of a ligature or in isolated form. Otherwise, the garbage produced in the other scenarios is left unrecognized and can be controlled in the recognition phase. In future, this system will be enhanced for the recognition of the other regional as well as international languages.

References

[i] N. B. Jumani, R. Rahman, F. Rahman, and M. J. Iqbal, "Effects of Native Language Saraiki on English Language Pronunciation," International Journal of Business and Social Science, vol. 2, 2011.

[ii] W. P. Thomas and V. P. Collier, "A national study of school effectiveness for language minority students' long-term academic achievement," 2002.

[iii] S. Saleem, "Flood and socio-economic vulnerability. New challenges in women's lives in northern Pakistan," 2013.

[iv] R. Kausar and M. Sarwar, "The History of the Urdu Language Together with Its Origin and Geographic Distribution," 2015.

[v] M. Y. Potrus, U. K. Ngah, and B. S. Ahmed, "An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online Arabic text recognition," Ain Shams Engineering Journal, vol. 5, pp. 1129-1139, 2014.

[vi] E. Hasan, M. M. Iqbal, et al. , "AN ONLINE PUNJABI SHAHMUKHI LEXICAL RESOURCE.," Sci.Int.(Lahore), vol. 27, pp. 2529-2535, 2015.

[vii] I. El Maarouf, J. Bradbury, V. Baisa, and P. Hanks, "Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing," in LREC, 2014, pp. 1001-1006

[viii] E. Cambria and B. White, "Jumping NLP curves: a review of natural language processing research [review article]," IEEE Computational Intelligence Magazine, vol. 9, pp. 48-57, 2014.

[ix] A. Rana and G. S. Lehal, "Offline Urdu OCR using Ligature based Segmentation for Nastaliq Script," Indian Journal of Science and Technology, vol. 8, 2015.

[x] C. Wen, A. Rebelo, J. Zhang, and J. Cardoso, "A new optical music recognition system based on combined neural network," Pattern Recognition Letters, vol. 58, pp. 1-7, 2015.

[xi] R. Srivastava and R. A. Bhat, "Transliteration Systems Across Indian Languages Using Parallel Corpora," Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development, p. 390, 2013.

[xii] R. Fabri, M. Gasser, N. Habash, G. Kiraz, and S. Wintner, "Linguistic introduction: The orthography, morphology and syntax of Semitic languages," in Natural Language Processing of Semitic Languages, ed: Springer, 2014, pp. 3-41