Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 48 (1) 213-216 (2016)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

A Novel Approach for Online Sindhi Handwritten Word Recognition using Neural Network

A. A. CHANDIO, M. LEGHARI, D. HAKRO*, S. A. AWAN**, A. H. JALBANI

Department of Information Technology, Quaid-e-Awam University of Engineering, Science and Technology Nawabshah Pakistan

Received 22nd April 2014 and Revised 16th July2015

Abstract:Online and Offline Handwritten Recognition has become an important part of the field of Pattern Recognition since it is considered as a technological revolution between man-machine interfaces. The handwriting has sustained to be a continuous mode of communication for recording information in day-to-day life. The exciting nature of Handwriting Recognition and Segmentation has attracted the concentration of researchers from academic as well as industry domains. Research for Online and Offline Handwritten Sindhi Word Recognition is at very beginning stage as compared to Latin, Chinese and Arabic languages. Sindhi language is among the ancient languages of the world. This language contains fifty two alphabet characters with different shape, cursive style and position of characters which increases the complexities and difficulties for recognition as compared to other Unicode based languages. This research paper has addressed a novel approach for recognizing Sindhi words using Artificial Neural Network (ANN). Self-Organizing Map (SOM), an ANN algorithm has been used for Sindhi Word Recognition entered on the surface of touch screen device such as Tablet PC or Smart Phone on real time. Unsupervised learning method has been used to train the proposed system that randomly alters the weight of the matrix closer to the input. A dataset consisting of 1200 words has been collected from 60 native writes of Sindhi language. An accuracy rate of 83% has been achieved with recognition time of 20-30 milliseconds.

Keywords: Online Sindhi Handwritten Word Recognition; Handwritten Recognition; Sindhi Language, ANN; SOM

1. INTRODUCTION

Online handwriting recognition systems will become important for the use of cursive languages such as Sindhi language, specifically for Tablet PC and smart mobile devices, where the usage of keyboard has not become common due to the large number of alphabet characters available in these languages. However such systems have been already in use with consistent performance for English and other languages (Plamondon, *et al.,* 2000). Handwriting is one of the universal method of communication among people and a lot of researchers are doing their research for Handwritten Recognition systems. The literature reviewed shows that considerable work has been done for recognition of handwritten characters and words of English, Chinese and Arabic as well as Indic languages (Pradeep, *et al.,* 2014) (Sundaram, *et al.,* 2015). However, there has been a very little research on Sindhi handwritten character recognition. Handwritten Sindhi words recognition is a complex task due to numerous writing styles, personal style of each writer since each writer may write the same word in many writing styles and even a same writer may write the words in different writing styles. Another problem for Sindhi words recognition is the similarity of shapes of different characters, position of dots and number of dots such as **"ٹ", "ٺ", "چ","ح", "ک","ڪ"** and others. Traditionally, handwritten recognition field is categorized into two fields: Online and Offline. In online handwritten recognition the movements of stylus are recorded in a series of time while the character is under creation and in Offline recognition the text is available in the form of an image (Zanchettin, *et al.,* 2012). The handwritten recognition can further be distinguished as isolated character or word recognition and sentence recognition. Handwriting style can also be classified as constraint or unconstraint. This research study is based on unconstraint handwriting style where the writer can write in any direction as well as connect characters during the writing process. Regardless of more than 30 years of research on handwriting recognition (Thompson, *et al.,* 2009) (Bunke, 2003), the development of general purpose and reliable system for unconstrained word and sentence level recognition is an open and challenging task. In this research article an ANN algorithm called SOM has been implemented for unconstrained Sindhi Word Recognition at real time which provides more efficient results. This algorithm is based on unsupervised learning and does not need the interaction of human beings during the learning process, reduces the amount of training data, speedup the learning process and compress the transmitted information more effectively (Kohonen, *et al.,* 1996).

Section 2 describes the basic characteristics of the Sindhi language, Section.3 highlights the Neural Network use for Sindhi words recognition, Section.4 shows proposed system's block diagram, Section.5 outlines the experimental results and their discussions and Section.6 outlines the conclusions and future work.

++Corresponding Author'sasghar.ali@quest.edu.pk,legharimehwish@quest.edu.pk
*Department of Information Technology, University of Sindh, Jamshoro
**Department of Computer Science & I.T, Benazir Bhutto Shaheed University, Lyari Karachi Pakistan

2. BASIC CHARACTERISTICS OF SINDHI LANGUAGE

Sindhi language is the second largest spoken language in Pakistan with more than 40 million speakers approximately in Sindh province and some areas of India (Nizamani, *et al.,* 2013). Sindhi language is an Indo-Aryan language written as a Perso-Arabic in Pakistan and both as Devanagari and Perso-Arabic scripts in India (Leghari, *et al.,* 2010). The writing style of Sindhi characters is from right to left. Sindhi language is based on fifty two alphabet characters and seven diacritic signs. Among fifty two characters of alphabet, three characters are derived from Persian alphabet, twenty nine characters are derived from Arabic characters and twenty characters are modified which make the difference between Sindhi alphabet characters and Arabic alphabet characters. **(Table1)** shows the list of Sindhi alphabet characters.

Table 1: List of Sindhi Alphabet Characters

Characters derived from Arabic alphabet	**29 characters**
Characters derived from Persian alphabet	**03 characters**
Modified characters	**20 characters**
Total	**52 characters**

The modified characters of Sindhi language are developed by varying the structure of few existing characters such as "ڳ"character is developed by combining "گ" and "ھ" respectively. Also dots and other diacritics signs are used to create new characters for example "ڇ ", "ڦ" and "ٻ" and few others do not belong in the Arabic alphabets. As Sindhi language has a nature of cursive style so characters are connected with each other in order to create a word. Therefore, there are four forms of styles of characters such as isolated form, initial form, medial form and final form. Table 2 shows all the forms of the Sindhi alphabet characters.

Table 2: Forms of Sindhi alphabet Characters

Character	Isolated Form	Initial Form	Medial Form	Final Form
ب"Beh"	ب	بـ	ـبـ	ـب
ع "Ain"	ع	عـ	ـعـ	ـع
ق "Qaf"	ق	قـ	ـقـ	ـق
ص "Sad"	ص	صـ	ـصـ	ـص
ي "Yeh"	ي	يـ	ـيـ	ـي

A complexity of the Sindhi character is the change of the shape of a character depending upon its position in the word. The context in which the characters are used also play an important role for determining the specific shape of the characters at specific positions. The shape of the characters in isolation, initial, media and final forms when used in words may also vary significantly. **(Fig. 1)** shows the various shapes of Sindhi characters when used in words. The word in the screen shot is a handwritten and the same word in red border is type written. The word is a combination of four individual characters in various forms as described at the right side of the snap shot. Dots and diacritics are used to give different meanings and identities to the characters. The diacritics also help for creation of more Sindhi language characters as well give oral and written appearance to sounds. It is also a challenging task to associate the dots and diacritics to their base characters in Sindhi language because the dots and diacritics have very small size.

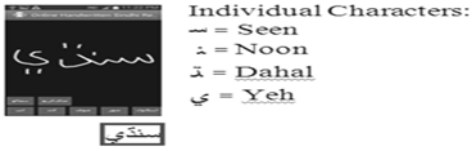

Fig.1: Different shapes of Sindhi characters when used in words Seen_i,Noon_m, Dahal_f and Yeh_iso show the initial form, medial form, final form and isolated forms of the characters in a word

3. NEURAL NETWORK APPROACH FOR SINDHI WORD RECOGNITION

Many research findings have shown that the use of Neural Networks create efficient machines having capability of making decisions. The machines are trained using any model or algorithm. Commonly supervised and unsupervised training methods are used. Unsupervised learning has more accuracy for handwritten characters and words recognition. In this method if the input value is not matched with any value in the dataset then the Neural Network will try to find the best comparative result Ganchev, 2003). In order to recognize the Sindhi language words, a self-organizing map algorithm is used for training the system. The SOM algorithm is being used in various pattern recognition applications (Khosrowabadi, *et al.,* 2010) Ghorpade, *et al.,* 2010) (Hanafizadeh, *et al.,* 2011). The training procedure of SOM is started with random weights which minimize the rate of errors. SOM algorithm learns from the training patterns. In training step an input pattern is given to the algorithm which is disseminated forward (Kohonen, 1998). The input training can be given in any writing style and font size. The algorithm has three layers input, hidden and output. The input layer normalizes the input pattern and hidden layer adjusts the weights and the output of the network is given on the output layer by the neurons.

4. BLOCK DIAGRAM OF PROPOSED SYSTEM

(Fig.2) shows the recognition process of the proposed system.

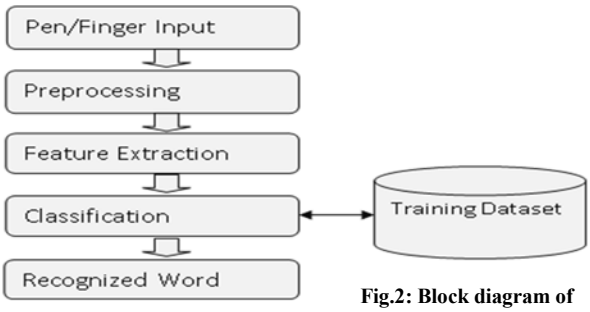

Fig.2: Block diagram of proposed system

Each step of the proposed system is described shortly as follows:

Step 1: Pen / Finger Input:

This is the initial step of the proposed system where the pattern of input word given on the surface of the touch sensitive device like smart mobile phone or tablet pc is captured. The input was taken on the surface of the touch screen device. The input word is represented in a form of matrix of pixels along with other attributes such as number of dots and direction of strokes. The input pattern is then passed to the next step for further processing.

Step 2: Preprocessing:

During the phase of preprocessing, the word that has been input is normalized in order to define it in its baseline form for increasing the accuracy of recognition. The samples of words taken for input were in varying styles and sizes, so normalization technique was used to remove the noise and other irrelevant patterns from the input. Smoothing techniques have also been applied for making the appearance of the word flat and the captured word has been normalized for an optimum size.

Step 3: Feature Extraction:

During this step the required features for recognition process are obtained from the input word. The white spaces of the touch screen are skipped. Those features include dots, base stroke and secondary strokes etc. In this step each word is represented as a feature vector which describes the identity of the word. It can be useful for increasing the accuracy of the classification.

Step 4: Classification:

In this stage of system the word along with its extracted features is assigned to a most relevant class/group from already defined classes with the help of an artificial neural networks (ANNs) algorithm. The input pattern of word received by the classification step is forwarded to the input neurons. The input neurons further compile the input pattern of word and match it with the words existing in the dataset.

Step 5: Recognized Word:

In this step the input word vector in normalized form maps its weighting values with the values given in the matrix and discover the matching pattern. The SOM algorithm consists of two layers of the nodes such as input layer and mapping layer. The input layer and each node of the mapping layer is described as a vector that includes nodes. The mapping layer also contains certain random numbers. So the weight of each input word pattern is compared with the weight of mapping layer and the node that has smallest "Euclidian" distance is considered as the winning node. The results of the winning node are finally presented for output as a recognized word.

5. EXPERIMENTAL RESULTS

A total of 1200 samples of words were obtained by 60 native users of the Sindhi language on the surface of the touch sensitive device with the help of stylus pen or their finger. Each user was allowed to write 20 samples of different Sindhi words in any style and size. The proposed system has been implemented in Android Platform. A total of 30 Sindhi language words were used for sampling and each sample of the word was tested 5 to 10 times by 25 different users and an average accuracy of 83% has been achieved. The rejection rate of recognition was also measured which was 10% to 40% due to the complexity in writing style. **(Fig. 3)** shows the results of few input words and their recognized words.

Fig.3: Recognition results of few handwritten Sindhi words

There are various words in Sindhi language which have similar style. The recognition rate of words matching in style was up to 70%. From 30 different input word patterns, few words were recognized as 100% by the system. This ratio may possibly be decreased if there are thousands of training samples in the dataset. It was also observed that the system recognized word patterns with 100% accuracy for some input patterns and up to 90% for other input patterns when the training and testing were performed by the same user as shown in figure 4. However the recognition accuracy of the system was up to 70% when training samples were taken from one user and testing were performed by other user.

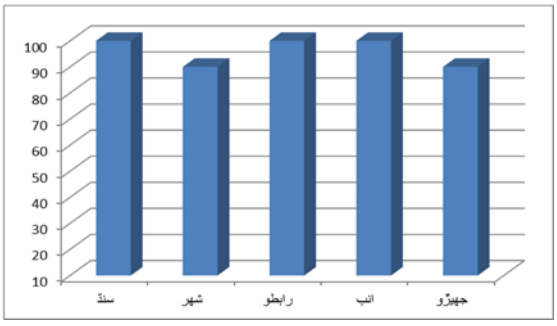

Fig.4: Recognition results when training and testing were performed by the same user

(Table 3) shows the recognition and rejection ratio of 10 different input word patterns out of training samples of 30 words.

Table 3: Recognition and Rejection ratio of 10 input word patterns

Input Word	Recognition Ratio	Rejection Ratio
شهر	80%	20%
جهيڙو	70%	30%
انب	90%	10%
قسمت	80%	20%
سنڌي	80%	20%
سنڌ	90%	90%
شڪست	80%	20%
اسڪول	70%	30%
نوجوان	70%	30%
رابطو	80%	80%

6. CONCLUSIONS AND FUTURE WORK

Handwritten Sindhi recognition field is in its very basic level and a little work has been done for offline handwritten recognition. In this research work an android platform based application has been developed for the recognition of handwritten Sindhi words entered at the screen of touch sensitive device. Self-Organizing Map, a library of Neural Network was used to train and test the proposed system. A total of 1200 samples of input word patterns for training purpose were obtained from 60 different native users of Sindhi language. The writers were allowed to write the input patterns in any writing style and font size. The system was tested from 25 users and an accuracy of 83% was achieved. Sindhi language has various characters with different number of dots at different positions. Therefore the recognition of Sindhi language characters and words is complex task. Currently this system is capable to recognize only isolated or single words of Sindhi language and can be extended in future for recognizing multiple words as well as sentences. The proposed system may also be implemented using Hidden Markov Model, Support Vector Machine or other pattern recognition techniques.

REFERENCES:

Bunke, H., (2003), Recognition of cursive roman handwriting: past, present and future. Proceedings of IEEESeventh International Conference on Document Analysis and Recognition, 448-459.

John T., M. Blumenstein, V. Nguyen, and T. Hine (2009), Offline cursive character recognition: A state-of the-art comparison. Proeedings of 14thConference of the International Graphonomics Society, 1-4.

Ganchev, K., (2003), Language segmentation for Optical Character Recognition using Self Organizing Maps. In Class of Senior Conference on Natural Language Processing, Computer Science Department, Swarthmore College, USA, 109-115.

Ghorpade, S., J. Ghorpade, S. Mantri, D. Ghorpade, (2010), Neural Networks for face recognition Using SOM. International Journal of Computer Science and Telecommunications.

Gorjizadeh, S., S. Pasban, S. Alipour (2015), noisy image segmentation using a self-organizing map network.Advances in Science and Technology Research Journal Volume 9, No. 26, 118–123.

Hanafizadeh, P., M. Mirzazadeh, (2011), Visualizing market segmentation using self-organizing maps and Fuzzy Delphi method–ADSL market of a telecommunication company. Journal of Expert Systems with Applications, 38(1), 198-205.

Kohonen, T., E. Oja, O. Simula, A. Visa, J. Kangas, (1996), Engineering applications of the self-organizing map. Proceedings of the IEEE 84, no. 10 (1996), 1358-1384.

Kaski, S., T. Honkela, K. Lagus, T. Kohonen, (1998), WEBSOM–self-organizing maps of document collections. Neurocomputing, 21(1), 101-117.

Khosrowabadi, R., H. C. Quek, A. Wahab, K. K. Ang, (2010), EEG-based emotion recognition using self-organizing map for boundary detection.Proceedings of IEEE 20th International Conference on Pattern Recognition(ICPR), 4242-4245.

Leghari, M., M. U. Rahman, (2010), "Towards Transliteration between Sindhi Scripts by using Roman Script". In the Conference on Language and Technology, National Language Authority Islamabad, Pakistan

Nizamani, A. M., N. U. H. Janjua, (2013), Sindhi OCR using Back propagation Neural Network. International Journal of Advanced Computer Science, 3(3)

Plamondon, R., S. N. Srihari, (2000), Online and off-line handwriting recognition: a comprehensive survey. Proceedings of IEEETransactions onPattern Analysis and Machine Intelligence, 22(1), 63-84.

Pradeep, J., E. Srinivasan, S. Himavathi, (2014), An investigation on the performance of hybrid features for feed forward neural network based English handwritten character recognition system. Proceedings of WSEAS Transactions on Signal Processing, 10(1), 21-29.

Sundaram, S., A. G. Ramakrishnan,(2015), Bigram language models and reevaluation strategy for improved recognition of online handwritten Tamil words. ACM Transactions on Asian and Low-Resource Language Information Processing, 14(2), 8Pp

Zanchettin, C., B. L. D. Bezerra, W. W. Azevedo, (2012), A KNN-SVM hybrid model for cursive handwriting recognition. Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN), 1-8.