Multi-Digit Handwritten Sindhi Numerals Recognition using SOM Neural Network

ASGHAR ALI CHANDIO*, AKHTAR HUSSAIN JALBANI*, MEHWISH LEGHARI*,
AND SHAFIQUE AHMED AWAN**

ABSTRACT

In this research paper a multi-digit Sindhi handwritten numerals recognition system using SOM Neural Network is presented. Handwritten digits recognition is one of the challenging tasks and a lot of research is being carried out since many years. A remarkable work has been done for recognition of isolated handwritten characters as well as digits in many languages like English, Arabic, Devanagari, Chinese, Urdu and Pashto. However, the literature reviewed does not show any remarkable work done for Sindhi numerals recognition. The recognition of Sindhi digits is a difficult task due to the various writing styles and different font sizes. Therefore, SOM (Self-Organizing Map), a NN (Neural Network) method is used which can recognize digits with various writing styles and different font sizes. Only one sample is required to train the network for each pair of multi-digit numerals. A database consisting of 4000 samples of multi-digits consisting only two digits from 10-50 and other matching numerals have been collected by 50 users and the experimental results of proposed method show that an accuracy of 86.89% is achieved.

Key Words: Sindhi Handwritten Numerals Recognition, Multi-Digits Recognition, Multi-Font Digits Recognition, Self-Organizing Map

1.   INTRODUCTION

Handwritten digits recognition is a technique to segment, classify, detect and recognize a digit from the image. It is called a subfield of pattern recognition and AI (Artificial Intelligence). The process for recognition of digits involves the phenomenon of detection of digits from an image and then converts them into machine readable format such as ASCII (American Standard Code for Information Interchange) for recognition purpose [1]. The digits recognition techniques can be classified as either offline or online. In offline technique, the document with handwritten or typewritten digits is generated first, converted into digital form, stored in the disk and then processed. However, in online recognition technique the digit is processed for recognition during its creation [2]. A lot of research is being carried out for digits recognition systems since last few decades due to their use in various common applications such as cheque numbers in bank, vehicle number plates, barcode numbers, postal codes and others.

Corresponding Author (E-Mail: asghar.ali@quest.edu.pk)
*   Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah.
**   Department of Computer Science & Information Technology, Benazir Bhutto Shaheed University, Lyari, Karachi.

Handwritten digits recognition systems can also be important in order to make historical books and documents in machine readable and editable format so that they may easily be accessed. So these systems can be helpful for the automation processes and they can be used to enhance the interface between human beings and machine in various applications. Handwritten numerals recognition is one of the challenging task which is under research since many years. The numerals written by hand are not uniform, they may be written in many different writing styles by different writers and even the same writer may write in different ways at different times [3].

Sindhi Language is one the ancient Indus valley language having hundreds of years history and is widely spoken by approximately 40 million people in Sindh province of Pakistan, many states of India as well as various other areas of the world [4,5]. It is taught as a basic language in almost all the primary schools of Sindh province and is second largest spoken language in Pakistan.

It is difficult to extract accurate features for online Sindhi handwritten numeral recognition because Sindhi numerals are written like Farsi and Arabic numerals. However, using digital pen or stylus it is easy to write digits on the surface of the touch screen device. Sindhi numerals are similar to Arabic and Farsi scripts, but most of the digits are written like the old Arabic digits style. Some of Sindhi digits such as "۴", "۵" and "٧" are different in style as compared to Arabic digits. These digits are also similar with Urdu and Hindi digits but have minor differences [6]. The writing style and direction of Sindhi digits is also different from Arabic digits as well as Urdu digits. Fig. 1 shows the list of isolated numerals used for Sindhi scripts and Fig. 2 shows few of the multi-digit numerals.
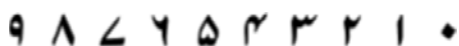
In isolated digits each numeral can be represented among one of the ten classes from ' to ۹. For multi-digit recognition, a string of digits is separated into sub images, each consisting of a single digit. This process is done by the recognition component called segmenter. Each separate digit is then recognized by the recognizer. Fig. 3 shows the multi-digit recognition technique.

The organization of the paper is shown as: Section 2 outlines the existing work done for numerals recognition for different languages. Section 3 shows the block diagram used of multi-digit handwritten Sindhi numeral recognition. Section 4 shows the experimental results. Section 5 outlines the discussions and section 6 outlines the conclusions and future work.

2.   LITERATURE REVIEW

Sindhi language has not received proper attention of the researchers even this language has millions of speakers and writers. However, a lot of research work has been done for many other languages of the world. Singh, et al. [3] presented a technique called fusion of global and local features for handwritten Devanagari digits recognition and achieved a result of 95% or better and they also reported that by combining global and local features the confusion value has been decreased [3]. A remarkable work has been done for isolated digits recognition but not a more work is done for multi-font numerals recognition in near past [7]. The techniques which are so far used for multi-font numerals recognition suffer from many problems such as increased computation time and a huge collection of training set for each sample per font. A huge collection of training set provides the facility to recognize multi-font digits but the accuracy of the system

۱۵،۱٦،۱۷،۱۸،۱۹،۲۰،۲۱،۲۲،۲۳،۲۴،۲۵،۲٦،۲۷،۲۸،۲۹،۳۰،۳۱،۴۰،۴۱،۵۱،٦۱،۷۱

FIG. 2. LIST OF FEW SINDHI MULTI-DIGIT NUMERALS

| ۵۵ | → Segmentation → | ۵۵ | Recognizer | → "55" |

FIG. 3. A MULTI-DIGIT RECOGNITION TECHNIQUE

۹ ۸ ۷ ٦ ۵ ۴ ۳ ۲ ۱ ۰

FIG. 1. LIST OF ISOLATED SINDHI NUMERALS

Mehran University Research Journal of Engineering & Technology, Volume 36, No. 4, October, 2017 [p-ISSN: 0254-7821, e-ISSN: 2413-7219]

902
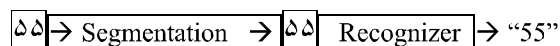
is decreased rapidly [8]. Montazer et. al. [9] have proposed a multi-font Persian digits recognition system using neuro-fuzzy interference engine and obtained an accuracy rate of 97%. In this proposed system structural features are used along with Mamdani fuzzy interference engine for defining fuzzy rules with 33 different types of fonts and styles. Arjun et al. [10] have proposed a technique using Euler number that recognizes 17 different types of fonts with various sizes ranging from 8-72 points and achieved an accuracy rate of approximately 99.76% on a dataset of 2890 images of digits.

Goyal and Koolagudi [11] have proposed a method for Hindi number recognition that uses Gaussian mixture models. The proposed system has been implemented for the recognition of Hindi numerals using mobile recorded speech and microphone. The data was collected from male, female and children with the help of microphone or mobile devices and claimed an accuracy rate of 98.9% for microphone recorded speeches and 96.4% for mobile phone device recorded speeches. Numerals recognition system is mixture of both ANN (Artificial Neural Network) and Pattern Recognition. Different ANN algorithms are used for digits and characters recognition, however back propagation algorithm is easy and has better speed. This algorithm has more accurate results as compared to other algorithms. In back propagation algorithm the dataset is divided into two different sets called training set and test set. The size of training set is larger as compared to the test set. It is also important to consider that the network should not be over trained, if the network is over trained for many epochs then it memorizes the patterns of training and may not recognize other patterns except the trained ones [12].

Bazrafkan and Broumandnia [13] have proposed a new string matching algorithm for handwritten numerals recognition based on linked list data structure. With the help of chain code technique the handwritten numerals are first transformed into string patterns then are recognized using refer algorithm which is implemented with the help of linked list data structure. The refer algorithm is used to calculate the distance among the chain code string patterns. The proposed algorithm reduces the time complexity as well as memory required for computation and enhances recognition accuracy. The recognition accuracy of 94.8% has been achieved over 3000 samples of handwritten digits. Rani et. al. [14] have used zone based hybrid feature extraction techniques for handwritten Gurmukhi numerals recognition and have achieved an accuracy of 99.37%. In the proposed system a total of 1500 samples were collected, 150 samples for each of the Gurmukhi digit. SVM (Support Vector Machine) has been used for digits classification and recognition along with zone based feature extraction techniques. Digits of variant size of images were created and it is also proposed that the accuracy of recognition depends upon the number of features and the size of the numeral image.

The available literature also shows a remarkable work done for handwritten Arabic digits/numerals recognition [15-18] and not much work is done for Farsi and Urdu handwritten numerals recognition [19-22]. This research work presents the handwritten Sindhi language multi-digit recognition using neural network.

3. BLOCK DIAGRAM OF PROPOSED SYSTEM

A Multi-Digit Sindhi Numeral Recognition System is presented that can recognize multi-digits of Sindhi language written with human figure or digital pen on the surface of touch screen device like smart mobile device or laptop. Fig. 4 shows the block diagram for multi-digit handwritten numeral recognition system.

The proposed system is implemented using NN algorithm called SOM or Kohonen Network and is based on five most common phases described below:

Mehran University Research Journal of Engineering & Technology, Volume 36, No. 4, October, 2017 [p-ISSN: 0254-7821, e-ISSN: 2413-7219]

903

Phase-1: Mulit-Digit Data Acquisition: In this phase a collection of multi-digit handwritten Sindhi numerals samples were obtained from the users on touch screen device such as smart mobile device and tablet pc. For the creation of dataset of numerals each user was allowed to write the sample of digits in any style with different image sizes. Each user was asked to write two samples of twenty multi-digit pairs in Sindhi script. A total of 4000 samples were collected from 50 users.

Phase-2: Preprocessing: In input process, some noise is added with the digits due to the writing style and variant size of input image given by different users, so a filter may be used that could remove the noise. Few preprocessing techniques are performed in order to improve the quality of the input digit such as noise removal which removes any unwanted or irreverent bit patterns from the input digit that simply cause a noise. Normalization is also performed in this phase that scale size of the input digit to the standard size that helps in recognition accuracy.

Phase-3: Feature Extraction: Feature Extraction is one of the important phase that enhances the final recognition accuracy rate. In this process each digit is represented like a feature vector that describes its identity. So the extracted features should represent each class of digits and may also be able to record unique features for every class of digits. Therefore, each pair of multi-digit should have a similar
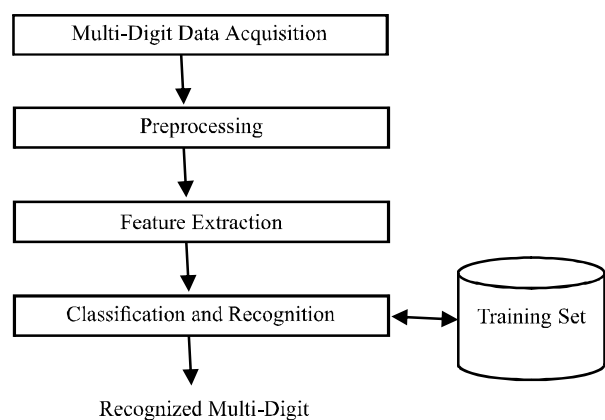
feature vector despite their writing styles and various font sizes. This will also help in the accuracy of classification of digits as explained in next phase.

Phase-4: Classification: The classification phase matches the input pair of digit with various samples already available in the dataset. The extracted features such as speed of writing, direction and duration of writing and pattern of writing are used by the classifier to match the digit. A neural network bases classification technique is used to perform this task.

Phase-5: Recognition: This is the last phase of recognition process that recognizes the final pair of multi-digit and displays it on the screen of output device in string format.

5.   EXPERIMENTAL RESULTS

The proposed system has been applied multi-digit database which was developed by collecting samples from 50 different writers. Developed system can further be amended to recognize multi-digits consisting of multiple pairs. The accuracy of the system varies from 60-100% due to variations in writing styles as well as the size of input multi-digit pair.

Fig. 5 shows the results of few multi-digit numerals developed using the proposed system.

A new experiment was performed while conducting this research study showing that this system was easily capable of adapting handwriting of any user. The samples provided by all the users were separated and then for the purpose of experiment the system was presented with a small number of samples provided by a single user. Then the system was tested by that same user (called User1) and by another different user (called User2). Then the results for both users were compared and found that the results for the same user were 100% while system could recognize only 60-70% handwritten digits from the different user. Fig. 6 shows the comparisons of the results for same user (User1) and different user (User2).

FIG. 4. BLOCK DIAGRAM OF PROPOSED SYSTEM

Individual multi-digit recognition accuracy is given in Table 1. It is also shown that few multi-digits like "٣٠" and "٤٠" have 100% accuracy rate. The accuracy rate of matching multi-digits like "١٥" and "٥١" or others of same shape but different order of digits was up to 60%. The results in Table 1 also shows that from the 4000 samples of training available in the dataset, the proposed system has weekly recognized the multi-digits "١٣" and "١٩".

Fig. 7 shows the recognition rate of few handwritten multi-digits of Sindhi language which are similar in shape and style.

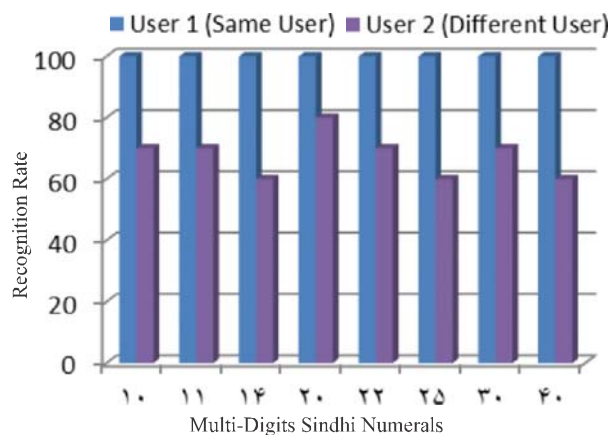FIG. 5. RESULTS OF FEW MULTI-DIGIT HANDWRITTEN SINDHI NUMERALS

FIG. 6. COMPARISON OF RESULTS FOR SAME USER (USER 1) AND DIFFERENT USER (USER 2)

TABLE 1. ACCURACY RATE OF 29 MULTI-DIGIT SINDHI NUMERALS

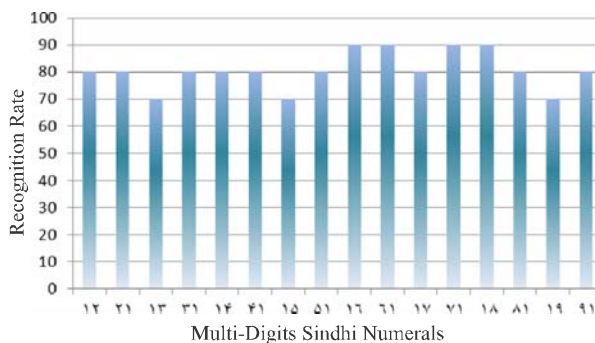| Digit	Recognition (%)	Rejection (%)
١٠	80	20
١١	90	10
١٢	80	20
١٣	60	40
١٤	80	20
١٥	70	30
١٦	90	10
١٧	80	20
١٨	90	10
١٩	60	40
٢٠	80	20
٢١	70	30
٢٢	90	10
٢٣	70	30
٢٤	80	20
٢٥	80	20
٢٦	80	20
٢٧	70	30
٢٨	80	20
٢٩	70	30
٣٠	100	00
٣١	70	30
٤٠	100	00
٤١	70	30
٥١	80	20
٦١	70	30
٧١	70	30
٨١	70	30
٩١	70	30

FIG. 7. RECOGNITION RATE OF SIMILAR MULTI-DIGIT NUMERALS

5. DISCUSSIONS

In order to measure the accuracy of the proposed system, the performance of the system was tested on 4000 samples of 29 different multi-digit Sindhi language numerals. Initially the system was trained and tested on two pair multi-digits, however more than two digit pairs can also be trained and tested using the proposed system. The samples of digits were collected using touch screen smart mobile device. The users were asked to write the multi-digit pair using their fingers. The proposed system is implemented on Android Platform. Each pair of multi-digit was tested 4 times by 20 different users and the average accuracy rate of 86.89% was achieved as shown in Table 2. The rejection rate of the system varied from 10-40% depending upon the writing style and direction of the typed multi-digit pair.

6. CONCLUSION

In this research study a system was developed for recognition of handwritten multi-digit Sindhi numerals using Kohonen neural network. The SOM also called Kohonen Network is simple to use and provides high accuracy recognition rate for multi-digit Sindhi numerals as compared to other techniques. Sindhi digits are based on Arabic and Urdu digits. However the writing style and direction of Sindhi digits is different from Arabic and Urdu digits. A total of 4000 multi-digit samples consisting only two digits pairs for training were collected from 50 different users. Each user was allowed to write a sample in any style and size. The developed system was tested from 20 users and an accuracy rate of 86.89% was achieved. However, the recognition rate for style and shape matching multi-digits was low. There are many issues in Sindhi script yet as this language is complex. The proposed system works only for multi-digit consisting of two digits pair numerals, however it is not able to separate the digits from the words which is still a big problem. The proposed system will further be implemented to recognize multi-digits consisting of multiple pairs. This system can further be extended to recognize the multi-digits of other languages. The proposed algorithm can be used to recognize the handwritten characters and words without considering their font, size and style limitations.

ACKNOWLEDGEMENT

REFERENCES

[1]   Charles, P.K., Harish, V., Swathi, M., and Deepthi, C.H.. "A Review on the Various Techniques Used for Optical Character Recognition", International Journal of Engineering Research and Applications, Volume 2, No. 1, pp. 659-662, 2012.

[2]   Kapoor, V., and Gupta, P., "Digit Recognition System by using Back Propagation Algorithm", International Journal of Computer Applications, Volume 83, No. 8, pp. 33-36, 2013.

TABLE 2. AVERAGE ACCURACY RATE OF PROPOSED SYSTEM

	True Positive	True Negative
Predicted Positive	1723	202
Predicted Negative	103	299
Overall Accuracy	86.89%	

[3] Singh, P., Verma, A., and Chaudhari, N.S., "Handwritten Devnagari Digit Recognition using Fusion of Global and Local Features", International Journal of Computer Applications, Volume 89, No. 1, pp. 6-12, 2014.

[4] Sindhi Language, "In Sindhi Language Authority", Retrieved May 12, 2015, from http://sl.sindhila.org/en/about-sindhi-language

[5] Leghari, M., and Rahman, M.U., "Towards Transliteration between Sindhi Scripts by using Roman Script", Conference on Language and Technology, National Language Authority, Pakistan, 2010.

[6] Sadri, J., Suen, C.Y., and Bui, T.D., "Application of Support Vector Machines for Recognition of Handwritten Arabic/Persian Digits", Proceedings of 2nd Iranian Conference on Machine Vision and Image Processing, Volume 1, pp. 300-307, 2003.

[7] Dhandra, B.V., Malemath, V.S., Mallikarjun, H., and Hegadi, R., "Multi-Font Numeral Recognition without Thinning Based on Directional Density of Pixels", IEEE 1st International Conference on Digital Information Management, pp. 157-160, 2006.

[8] Hassanpour, H., and Samadiani, N., "Recognition of Multi-Font English Numerals using SOM Neural Network", International Journal of Computer Applications, Volume 98, No. 4, pp. 37-41, 2014.

[9] Montazer, G.A., Saremi, H.Q., and Khatibi, V., "A Neuro-Fuzzy Inference Engine for Farsi Numeral Characters Recognition", Expert Systems with Applications, Volume 37, No. 9, pp. 6327-6337, 2010.

[10] Santosh, A.N., Navaneetha, G., Preethi, G.V., and Babu, T.K., "An Approach to Multi-Font Numeral Recognition" IEEE Conference on TENCON Region, pp. 1-4, 2007.

[11] Goyal, H.R., and Koolagudi, S., "Hindi Number Recognition using GMM", International Journal of Computer Applications, Volume 63, No. 21, 2013.

[12] Kapoor, V., and Gupta, P., "Digit Recognition System by using Back Propagation Algorithm", International Journal of Computer Applications, Volume 83, No. 8, pp. 33-36, 2013.

[13] Bazrafkan, M., and Broumandnia, A., "A New String Matching Algorithm and its Application in Hand-Written Digits Recognition", International Journal of Computer Applications, Volume 70, No. 17, pp. 13-16, 2013.

[14] Rani, G.S.R., and Dhir, R., "Handwritten Gurmukhi Numeral Recognition using Zone-Based Hybrid Feature Extraction Techniques", International Journal of Computer Applications, Volume 47, No. 21, pp. 24-29, 2012.

[15] Zaghloul, R.I., Enas, D.M.B., and AlRawashdeh, F., "Recognition of Hindi (Arabic) Handwritten Numerals", American Journal of Engineering and Applied Sciences, Volume 5, No. 2, pp. 132, 2012.

[16] Charfi, M., "A New Approach for Arabic Handwritten Postal Addresses Recognition", arXiv Preprint arXiv, Volume 1204, No. 1678, 2012.

[17] Salimi, H., and Giveki, D., "Farsi/Arabic Handwritten Digit Recognition Based on Ensemble of SVD Classifiers and Reliable Multi-Phase PSO Combination Rule", International Journal on Document Analysis and Recognition, Volume 16, No. 4, pp. 371-386, 2013.

[18] Diem, M., Fiel, S., Garz, A., Keglevic, M., Kleber, F., and Sablatnig, R., "Competition on Handwritten Digit Recognition", IEEE 12th International Conference on Document Analysis and Recognition pp. 1422-1427, 2013.

[19] Azad, R., Davami, F., and Shayegh, H.R., "Recognition of Handwritten Persian/Arabic Numerals Based on Robust Feature Set and K-NN Classifier", arXiv Preprint arXiv, Volume 1407, No. 6492, 2014

[20] Javidi, M.M., and Sharifizadeh, F., "A Modified Decision Templates Method for Persian Handwritten Digit Recognition", Journal of American Science, Volume 8, No. 1, 2012.

[21] Sagheer, M.W., He, C.L., Nobile, N., and Suen, C.Y., "A New Large Urdu Database for Off-Line Handwriting Recognition", Image Analysis and Processing – ICIAP, pp. 538-546, Springer Berlin/ Heidelberg, 2009.

[22] Javed, L., Shafi, M., Khattak, M., and Ullah, N., "Hand-Written Urdu Numerals Recognition using Kohonen Self Organizing Maps", Sindh University Research Journal (Science Series), Volume 47, No. 3, Jamshoro, Pakistan, 2015.