PASHTO OPTICAL CHARACTER RECOGNITION USING NEURAL NETWORK

Sahibzada Abdur Rehman Abid[1*], Muhammad Naeem[2], Asma Gul[3], Nasir Ahmad[1]

ABSTRACT

In this paper, an Optical Character Recognition system for printed/scanned Pashto continuous text is presented. The proposed Pashto optical character recognition system uses a Feed Forward Neural Network (FFNN), consisting of an input layer, a hidden layer and an output layer. The input layer is composed of 315 neurons, which receive the pixels data i.e. binary data from a 21x15 symbol pixel matrix. The hidden layer contains 2000 neurons which has been chosen after testing based on optimal result, while the output layer is composed of 6 neurons. As the joinable Pashto characters on different locations in text change its size and shapes, as a result 60 Pashto characters with 110 samples for each Pashto character has been used to train the network.

KEYWORDS: *Optical Character Recognition, Pashto OCR, Neural Network, Multilayer Feed Forward Neural Network*

INTRODUCTION

The Optical character recognition, abbreviated as OCR, is the process of converting printed/scanned images of handwritten, typewritten, or printed text into machine-readable text. Several approaches for OCR have been proposed in literature. However, these studies are conducted mostly on the widely spoken languages such as Chinese, English, Arabic and French (Yang *et al*., 2015; Poznanski and Wolf, 2016). More recently, OCR research on other regional and local languages has also been reported such as in the work on handwritten Devnagari character recognition (Arora *et al*., 2008) has been reported using Neural Network based classifier. In the said work, four features extraction approaches i.e. shadow features, intersection, chain code histogram, and straight line fitting features, have been used. The classification decisions of the four multi-layer perception based classifiers are combined using a weighted majority voting approach. In (Al-Jawfi, 2009) a handwritten Arabic character recognition system has been proposed, utilizing LeNet neural network. The designed LeNet neural network consists of two main steps. In the first step, the shape of the Arabic character is learnt through a pixel matrix of size 16×16, while in the second step the number and position of dots, as well as its zigzag or dot nature is recognized by utilizing back propagation algorithms. It is reported that the performance of the back propagation algorithm greatly depends on the accuracy of segmentation and noise removal stage.

In (Zafar et al., 2006), a system for online handwriting recognition has been explained. In their system, the useful information about character recognition has been extracted bypassing the extensive pre-processing. It is claimed that the system achieves a recognition rates of 51% to 83% for different sets of character samples, using Back Propagation Neural network (BPN).

A printed Urdu OCR system is presented in, where isolated Urdu characters are recognized by using a combination of character features such as topological, contour along with features based on water reservoir concept (Pal and Sarkar, 20032).

An Urdu Character Recognition system based on Principal Component Analysis has been proposed (Khan et al., 2012). The authors have developed their own databases, namely for training and testing of the system, named as 'TrainDatabase' and the 'TestDatabase'. In the training step, all the images in the training database are projecting onto the Eigen space using PCA approach. In recognition step, the test images are projected onto the same space for recognition.

A medium size database for Pashto optical character recognition research has been developed (Ahmad *et al*., 2013). In the same work, they have also reported the development of an optical character recognition system for recognition of isolated Pashto characters. The classification is performed at two levels, i.e. High level classification and Low level classification, and the K

1 Department of Computer Systems Engineering, University of Engineering & Technology Peshawar, Pakistan*
2 School of Computer Science, University of Guelph, Canada
3 Department of Mathematical Sciences, University of Essex, UK

nearest neighbor (K-NN) classifier has been utilized for low level feature classification. Beside a few reported researches on Pashto OCR, the research on Pashto OCR is still in the initial stage and a lot of research work is needed to develop a Pashto OCR system deployable for practical applications.

The remainder of the paper is structured as follows. The next section introduces the Pashto language and Pashto script. The following section then explains the working of proposed Pashto OCR system while the last section presents the results of Pashto OCR and a discussion on the obtained results.

PASHTO LANGUAGE AND PASHTO SCRIPT

The estimated number of speakers of Pashto language around the globe is between 50 to 60 Million (Paul, 2009). Pashto is one of the most widely spoken languages of Pakistan and one of the national languages of Afghanistan. The three dialects of Pashto language are; Northern dialect, Central dialect and Southern dialect (David, 2014).

Like others languages, Pashto language has its specific alphabets set. Pashto script is a customized version of the Arabic script, with some additional characters which are not present in the character set of Arabic language. As each language has its own specific phonemes that are not present in other language, similarly the Pashto Alphabets (ګ, پ, ې, ړ, څ, ږ, ډ, ښ, ې, ړ, ۍ, ټ, ډ) are not present in the Arabic language. Some of these alphabets have a larger similarly with other languages written in the Arabic script such as Urdu and Persian. But both of these languages have their own specific phonemes too which may or may not be present other language and vice versa. The character set /alphabets of Pashto language, known as Haroof-e-Tahajee, consist of 44 characters given in Figure 1.

ا ب پ ت ټ ث ج چ ځ خ ح څ د ډ ذ ر ړ ز ږ ژ

س ش ښ ص ض ط ظ ع غ ف ق ك ګ ل م ن

ه و ۀ ي ې ۍ ئ

Figure 1: Pashto Characters Set.

In Pashto language, characters have different appearance when they appear at different location in a word. Majority of the Pashto characters have different orientation in the start, middle and end of a word, but some of the characters remain the same at any place i.e. don't change in the start, middle and end of the word. All Pashto characters with associated changes in orientations at different position i.e. in the start, middle and end of the word has been shown in Table1.

PASHTO OPTICAL CHARACTER RECOGNITON

The Pashto OCR system proposed in this work has two main modules; segmentation module, and recognition module. The block diagram of Figure 2 shows the working of the proposed Pashto OCR system. Segmentation of the Pashto Text in the images as well as feature extraction is performed by the segmentation module while recognition module uses neural network algorithm to classify the segments as characters.

Segmentation / Features Extraction

Pashto Text image in Bitmap (.bmp) format is given as an input to the OCR system and the preprocessing step converts the Pashto text line in the image into black and white colors. The parts of image, below and

Table 1 Pashto Characters Appearance at Different Positions.

S. No.	Contextual Form		
	Beginning	Middle	End
1.	-	-	ا
2.	ب	ب	ب
3.	پ	پ	پ
4.	ت	ت	ت
5.	ټ	ټ	ټ
6.	ټ	ټ	ټ
7.	ج	ج	ج
8.	چ	چ	چ
9.	ح	ح	ح
10.	خ	خ	خ
11.	څ	څ	څ
12.	څ	څ	څ
13.	-	-	د
14.	-	-	ډ
15.	-	-	ذ
16.	-	-	ر
17.	-	-	ړ
18.	-	-	ز
19.	-	-	ژ
20.	-	-	ږ
21.	س	س	س
22.	ش	ش	ش
23.	ښ	ښ	ښ
24.	ص	ص	ص
25.	ض	ض	ض
26.	ط	ط	ط
27.	ظ	ظ	ظ
28.	ع	ع	ع
29.	غ	غ	غ
30.	ف	ف	ف
31.	ق	ق	ق
32.	ک	ک	ک
33.	ګ	ګ	ګ
34.	ل	ل	ل
35.	م	م	م
36.	ن	ن	ن
37.	ڼ	ڼ	ڼ
38.	-	-	و
39.	ه	ه	ه
40.	ې	ې	ي
41.	ۍ	ۍ	ی
42.	-	-	ی
43.	-	-	ی
44.	ئ	ئ	ئ

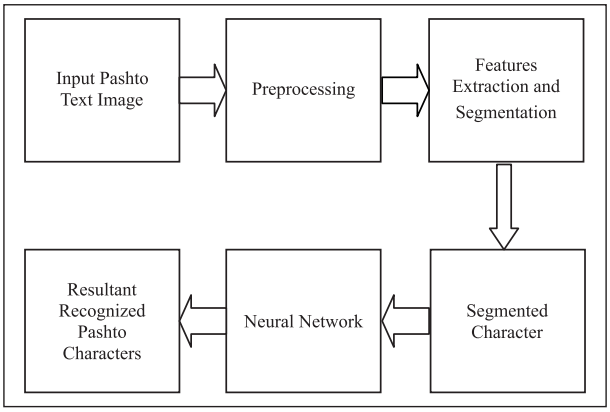

Figure 2: Pashto OCR Block Diagram.

above the text area are removed and only the text part is passed onwards for features extraction.

The segmentation results of Pashto text in the images is greatly dependent on the level of complexity, presented by the characters during scanning. The characters have been divided into three levels of complexity groups, simple, semi complex and complex on basis of the topological characteristics, the number of holes, the height, width of holes as well as the direction of these holes as shown in the following Table 2.

Table 2: Pashto text on the basis of complexity.

S. No.	Category	Pashto Characters
1	Simple	ا, ب, پ, ت, ټ, س, ش, ښ, ک, ل, ن
2	Semi Complex	ذ, ر , ړ, ز , ږ, ح, څ, چ, ځ, خ, ې, ی, ي
3	Complex	د, ړ, ف, ق, ګ, ص, ض, ط , ظ, ع , ه , و, ټ, م, ۍ, ڼ, غ

One of the main challenges in both the training and testing phase is the segmentation of image. The character/symbols in the image are detected by examining the color value of individual pixels. The dataset used in this research the colors are either white or black. The images containing the Pashto text are in bitmap format and have the Pashto text characters in only single line with no border lines. All the pixels containing text characters are detected in the images and information about the text pixels are saved in a matrix. The extreme points of all the characters i.e. left, right, top and bottom are detected and noted.

Neural Network Architecture & Training.

In this work, Multilayer Feed Forward Neural Network (FFNN) is used to classify the characters detected in segmentation module. Implemented for the proposed research is composed of three (03) layers, input, hidden and output layer. The input layer is composed of 315 neurons, which receive pixels data from identified characters i.e. binary values from a 21x15 symbol pixel matrix. The hidden layer consists of 2000 neurons, which is chosen on the bases of optimal result on the test data, while the output layer is composed of 6 neurons. To make all the input of the neural network of the same size, all the characters are resized to the same size of 21x15 arrays.

All the samples of the 44 Pashto characters/Haroof-e-Tahajee were resized, normalized and formed vectors. The training samples are then provided to the neural network for training. The Multilayer Feed Forward Neural Network (FFNN) with the mentioned parameters took 8-10 hours, on a 2.2 GHZ, Core 2 due System with 2 GB RAM. The goal of the goal of 0.0005 was achieved with 2000 epochs.

As all joinable Pashto characters on different locations in text change its size and shapes, as a result 60 different Pashto characters with 110 samples for each Pashto character has been used to train the network. Some of the training samples are shown in Figure 3.

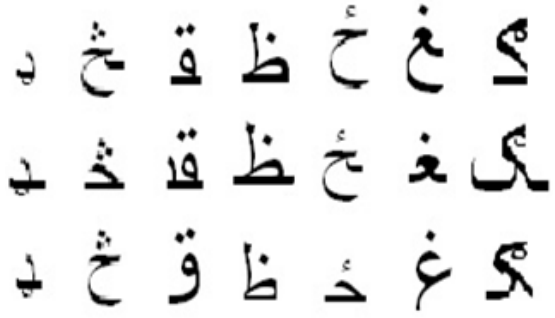

Figure 3: Sample of some of the Pashto Characters for Training.

The images of the Pashto test text are segmented and resized to the same size arrays as used in the training. These are then provided to the trained network which gives a 6 digit binary number, if the 6 digit binary number matches to any of the 60 character set that have been

used as target during the training, the relevant character is displayed, if not "Not Matched" message is displayed.

RESULTS AND DISCUSSION

The experimental results show that the segmentation algorithm gives an accuracy of 92% on test data, while the correct recognition rate of the Neural Network averages to 84%. The basic reason for this drastic drop in the recognition stage is due to the reason that in the joining form the character shapes changes at different position in the Pashto text, this makes its recognition much difficult. Moreover, most of the Pashto characters are in the form of identical pairs and this makes them un-differentiable.

For some Pashto character, such as ا, ب, پ, ت, ټ, س, ش, ښ, ک, ل, ن, ف , ق, د , ر , م , ح, ه, و, خ, ج, ي, ی , ع , غ recognition rate is 88%, while some of the Pashto character such as ذ, چ, څ, خ , ز, ژ , ړ, ډ, ږ, گ, ض gives a recognition is 85%. Character such as ژ, ږ, ډ, ړ , ذ, ز confuse with د , ر and چ, څ , خ , confuse with ج, څ and گ, ض is confuse with ص, ک where the rest of characters such as ط , ظ , ڼ, ئ , ټ , recognition rate is 73% and they are mostly confuse with ی, ص, ا, ي, ی, ض, ت, ل, ن.

REFERENCES

1. *Ahmad, N., Khan, A. A., Abid, S. A. R., Yasir, M., and Ullah, N. (2013), "Pashto Isolated Character Recognition Using K-NN Classifier", Sindh Univ. Res. Jour. (Sci. Ser.), Vol. 45 Issue. 4, pp. 679-682.*

2. *Al-Jawfi R., (2009), "Off Handwriting Arabic Character Recognition LeNet Using Neural Network", The International Arab Journal of Information Technology, Vol. 6, No. 3, pp. 304-309.*

3. *Arora, S., Bhattacharjee, D., Nasipuri, M., Basu, D. K., and Kundu, M. (2008), "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition", IEEE Region 10 and the Third International Conference on Industrial and Information Systems (ICIIS 2008), pp. 1-6.*

4. *David, A. (2014). "Descriptive grammar of Pashto and its dialects", Vol. 1, Walter de Gruyter, Berlin.*

5. *Khan, K., Ullah, R., Ahmad, N., and Naveed, K. (2012), "Urdu Character Recognition using Principal Component Analysis", International Journal of Computer Applications, Vol. 60, Issue. 11.*

6. *Pal, U., and Sarkar, A. (2003), "Recognition of Printed Urdu Script", IEEE Seventh International Conference on Document Analysis and Recognition (ICDAR), pp. 1183-1187.*

7. *Paul, L. M. (ed.), (2009), "Ethnologue: Languages of the World", (16th eddition), Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/(retrieved on 12 March 2018).*

8. *Poznanski, A., and Wolf, L. (2016), "Cnn-n-gram for handwriting word recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2305-2314.*

9. *Yang, W., Jin, L., and Liu, M. (2015), "Chinese character-level writer identification using path signature feature, DropStroke and deep CNN", IEEE 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia pp. 546-550.*

10. *Zafar M., Dzulkifli M., and Razid M. (2006), "Write Independent Online Handwritten Character Recognition using A Simple Approach", The International Arab Journal of Information Technology, Vol. 5, No. 3, pp. 476-484.*