Urdu Character Recognition using Neural Network

Nighat Naaz Ansari¹, Ajay Kumar Singh²

¹M.Tech Student, CSE, AKTU, U.P
²Professor, CSE, AKTU, U.P

Meerut Institute of Engineering and Technology
Meerut, Uttar Pradesh, India 250005

Abstract — In this paper, recognition of the isolated Urdu characters is proposed. Method used for the recognition of Urdu characters is based on artificial neural network. In this method, first the image of Urdu alphabets is taken as input, then preprocessing steps such as conversion to Gray-scale image and later to binary image, noise removal is done. Then each character in the input image is segmented. Segmentation is done on the basis of connected components. Then after segmentation, recognition of characters is done by the help of neural network. In the last step, a text file shows the Unicode of the recognised characters and then the characters are matched with these codes. Simulation of the project is done in MATLAB.

Keywords — Urdu characters, neural network, Gray-scale image, binary image, segmentation, Unicode.

I. INTRODUCTION

Optical character recognition (OCR) is technique in which all the handwritten or printed is converted into editable form. So in Urdu OCR, the characters of the Urdu alphabets are recognised and written in a text file which can be edited. Urdu OCR is an important field of research as after English, Urdu is the most widely used language in the world. So the recognition of Urdu characters is very important.

Urdu is written from right to left unlike English which is written from left to right. Urdu script is a mixture of Persian (or Farsi) and Arabic script. Arabic and Urdu script have many characters in common. Urdu is written in Nastaliq script whereas Arabic written in Naskh style. Fig.1 shows the character set in Urdu.

Fig. 1 Character set of Urdu alphabets

The research work in the field of Urdu OCR has been done previously and the recognition of the characters has been done using different methods. The characters are segmented using structural features and are recognised using fuzzy logics as in [1]. Ligature based segmentation is done in [2] using support vector machine (SVM) and K-nearest neighbour (kNN). Separation of primary and secondary strokes by con- centrated centroid distance method has already been applied in [3]. Ref [4] describes a five method algorithm for segmentation of characters from a word. [5] paper uses Freeman Chain Code for the separation of primary and secondary strokes and vertical scanning is done to extract primary ligature. For extracting features from the characters Direct Cosine Transform (DCT) is applied in [6] and then classification of ligatures is done by Hidden Markov Model (HMM). The method used in [7][8] is based on the template matching approach where the structural features of the characters are used for segmentation of primary and secondary ligatures. In [9] the recognition of Urdu ligatures is done in two phases. In first phase, diacritics are recognised and the second phase employs the recognition of the primary ligature.

There are several methods that exist for the recognition of Urdu characters as mentioned above. In this paper, we propose a different method for the recognition of isolated Urdu characters which uses artificial neural network for the classification of characters. For character segmentation, we have considered the connecting components which uses region shrinking algorithm for separating each character from the input image. Noise removal is done by passing the input image to a median filter for obtaining better result and thus improving the performance of the proposed system.

Rest of the papers is organizes as follows. Section II throws light on challenges in Urdu OCR. In section III, we gave an overview of proposed method for character recognition. Section IV discusses the methodology used for the recognition of Urdu characters. Section V discusses classification done by artificial neural network in detail. Section VI shows the results of the simulation in MATLAB. Finally, we conclude our work in section VI with direction for future research in section VII.
II. CHALLENGES IN URDU CHARACTER RECOGNITION

The recognition of Urdu characters is difficult because of the following reasons:

i. Similar structure: Many characters in Urdu alphabets have similar structure with only the difference of dots between them. So their recognition is difficult as these characters lie in the same class. For example characters ب, ت and ث have similar structure as in fig. 2.

ii. Secondary strokes or diacritics: Urdu characters consist of two part – primary stroke and secondary stroke. Primary stroke is the main body of the character whereas secondary stroke or commonly called as diacritics consists of dots and vowels. For example ح is the main body of the character “kha” whereas خ is the complete character with dot above the main body. The recognition of this dot or diacritics is difficult as after the segmentation it becomes individual character shown in fig.2.

iii. Occurrence of characters at different positions: The shape of character changes when the position of their occurrence changes within a word. The shape of isolated character is different from the shape when it occur either at the start, in the middle and at the end of a word. Not only the shape of character changes at different position, its size also changes as shown in fig.3.

iv. Presence of loop in the character: Some of the Urdu characters such as، گ etc. contain loops within them. Characters contain openings which are called as loop as shown in fig.4. Sometimes this loop is so small that the internal opening is disappeared thus making the recognition process complex.

v. Overlapping of characters: Urdu is written in cursive calligraphic style in which the characters overlap each other making the process of recognition more difficult by humans as well as by machines also. Overlapping of characters is shown in fig.5.

vi. Connectivity of characters: characters are connected to each other to form words but sometimes characters remains in their original form in the word i.e. they are not connected yet they still form a word. For example in fig.6 (a) characters are disjoint whereas in fig.6 (b) characters are connected.

Thus these are some problems which are related to the recognition of Urdu characters.

III. PROPOSED METHOD FOR CHARACTER RECOGNITION

As we have already mentioned in the literature review that there are different methods that exists for the recognition of Urdu characters. In this paper, the proposed method for the recognition of characters is achieved by using neural network. In this method, the characters from the input image are segmented on the basis of connected components i.e. pixels belonging to same component or character are separated from the other characters. Since the characters are isolated, so segmentation is made easy by using labelling property of MATLAB. This type of segmentation is called region shrinking segmentation. In this approach, the regions of interest from the input image i.e. each character is segmented by taking connectivity property into account and using bwlabel function in MATLAB.
All the steps in the proposed method are shown with the help of a flow chart given in Fig. 7.

In the proposed method, the first step is the acquisition of image either by scanner or a high resolution camera. Second step is the pre-processing done to the input image. It includes noise reduction so as to improve the rate of recognition. After pre-processing, segmentation is done based on connectivity of the components or characters (region shrinking segmentation). Then the next step after segmentation is the classification and recognition of characters using a neural network. Then last step is the matching of characters with its Unicode of 4 digits. The character with the maximum value of correlation function matched with the template is considered. Thus characters are recognised with the template matching algorithm. Therefore our goal is achieved that is the Unicode of characters obtained as output which is written to a text file. Thus the task of recognition of Urdu characters is accomplished.

Fig. 8 Sample image with 7 Urdu Characters

Fig. 8 shows a sample of input image and the output is the corresponding Unicode of these characters.

IV. PROPOSED METHODOLOGY

Our proposed methodology includes several steps that we have followed for Urdu character recognition. These steps are as follows:

i. Image acquisition: In the first step, the acquisition of image takes place. The input image is then converted to a gray scale image as shown in fig. 9(a).

ii. Pre-processing: The gray scale image is then passed to the median filter for the removal of salt and pepper noise. Contrast of the image is also enhanced by adjusting threshold value of the gray scale image. The gray scale image is then converted to binary image as in fig. 9(b).

iii. Segmentation: In the next step, image segmentation is done by exploiting the connectivity of the components. In it the connected components of a character are counted and labelled using bwlabel function of MATLAB. Isolated characters are extracted from this step as in fig. 9c).

iv. Classification: After segmentation and extraction of each character, classification is done using a neural network. The segmented characters are then fed to a neural network which matches them with the target characters.

v. Output: In the final step, the Unicode of matched or the recognised characters are written in a text file which is editable. Thus the task of recognising the characters of Urdu is accomplished.

V. RECOGNITION USING NEURAL NETWORK

The step of classification and recognition is very crucial as it is the step which will finally achieve the target. In this paper, the recognition is done with the help of a neural network. The segmented characters are fed to feed-forward artificial neural network and they are matched with the training set. The flow chart in fig. 10 explains the steps that were followed in the algorithm:

The performance of the neural network is plotted as a graph which shows the matching of character with ‘*’. Fig.11 shows the performance of the neural network for the character laam.

Fig. 11 Performance plot of neural network for character laam

Thus the character is located with the asterisk (*) sign.
VLSIMULATION RESULT

The simulation of this paper is done in MATLAB version R2012a (7.14.0.739). The character from the image is recognised as the Unicode corresponding to it and also the name of the character is also written. The algorithm is tested on the small dataset of 50 images. All the images of the Urdu characters are obtained from the internet.

Fig.12 Final output of recognition of character laam

Fig.12 shows the output of the recognition of the letter laam.

<table>
<thead>
<tr>
<th>No of images tested</th>
<th>No. of images recognised</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>43</td>
<td>86%</td>
</tr>
</tbody>
</table>

Table.1 Simulation Result

VII. CONCLUSION

Urdu OCR is a very challenging field of research owing to the difficulties in recognition of Urdu characters as discussed above in section II. Since most of the work has been done on Arabic naksh script but not much has been done on nasta'leeq style. In this paper we have proposed a method which recognizes Urdu characters. The accuracy of recognition using the proposed method is 86%.

VIII. FUTURE WORK

Since recognition of Urdu characters is very complex yet we have achieved an accuracy rate of 84% for isolated characters. The proposed method is unable to recognize the characters with similar structure and same number of dots on them but at different position. For example character fe and bay are recognized as same character by the above method. So, future work can be carried out in this direction.

REFERENCES