Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (4) 679-682 (2013)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Pashto Isolated Character Recognition Using K-NN Classifier

N. AHMAD, A. A. KHAN, S. A. R. ABID, M. YASIR, NASIM-ULLAH

University of Engineering & Technology Peshawar, Pakistan

Abstract- This paper presents the development of Optical Character Recognition (OCR) system for printed Pashto text. The problem of the unavailability of the standard database for Pashto language has also been addressed by developing a medium size database with 25 different variations with a total number of 1125 entries in the final database. In the proposed approach, individual Pashto characters are recognized utilizing both high and low level features. High level features are based on the structural information from the characters and the resulting binary trees uniquely classify each of the characters. The approach though quite robust is affected slightly by the variation in size, orientation and writing style. An alternative low level feature approach based on K-Nearest Neighbors has been used giving an overall word recognition of 74.8%.

Keywords: Optical Character Recognition, Feature extraction, K-Nearest Neighbors, Pashto OCR

1. INTRODUCTION

Pashto is a major language of Pakistan and the official language of Afghanistan with around 55 million speakers but little work has been done so far on the automatic recognition of Pashto text. Pan Localization project (Pal and Sarkar, 2013) is the regional initiative for the development of local language computing capacity in Asia. The objective of the project is, not only to develop the capacity for R&D in Asia in the local languages but also to advance the policy for local language contents creation and access across the Asia. This creation and access of contents in the local languages is termed as “localization”. Localization of ICT’s requires definition and implementation of standards. OCR is also a part of those standards. Around 14 Asian languages have been worked on in the Pan Localization project including Urdu.

Work has also been reported on other languages of the world such as in (Arora et al., 2008), Neural Classifier has been used to classify handwritten Devnagari Character. They use four feature extraction techniques that are intersection, shadow feature, chain code histogram and straight line fitting features. Weighted majority voting technique is used for combing the classification decision obtained from the four multi layer perception based classifiers. Experimental results show that this approach achieves better result and accuracy rate. In (Pal and Sarkar, 2013), OCR system is presented for printed Urdu characters, which a popular Indian script and national language of Pakistan. In their proposed system, individual character is recognized by using a combination of topological, contour and water reservoir concept based features. A prototype of the system has been tested on printed Urdu characters. The character is recognized in two stages. First the characters are divided into subsets by feature based tree classifier and then more sophisticated features are used to classify characters belonging to the leaf nodes.

In (Nagy, 1998), a precise system for the classification and recognition of the Chinese and Japanese handwritten characters is presented. Before extracting directional element feature (DEF) from each image of character they use transformation based on partial inclination detection (TPID) to reduce undesired effects of degraded images. In the recognition process city block distance with deviation (CBDD) are proposed for rough classification and fine classification. The system achieves very high accuracy rate. A technique for the recognition of printed Arabic characters is presented in (Amin, 1998). First of all a word is segmented, each character is transformed into a feature vector containing significant information about the characters that help in onward recognition of the characters. The features of printed characters include strokes and byes in varies directions, end point, intersection points, loops, dots and zigzags. The words skeletons are decomposed into a number of links in orthographic order, and then it is transformed into a sequence of symbols using vectors quantization. Single hidden Markov model is used for the purpose of classification. OCR has also been developed for other languages like Bangla (Chaudhuri and Pal, 1998), Oriya (Chaudhuri, Pal and Mitra, 2002), Greece etc. OCRs are applicable and useful in check sorting and verification, office automation, scanning and preserving the historical information, converting the scanned document into editable form etc. This paper presents the development of an OCR system for Pashto language. The rest of paper is organized as follows. Section 2 describes Pashto script and its various properties.

Section 3 and 4 discusses the development of Pashto isolated character database and Pashto OCR, respectively. The results and discussion on the OCR system are presented in section 5 while directions for future work are outlined in section 6.

2. MATERIAL AND METHODS

Pashto Script

Pashto alternatively spelled as Pakhtu, Pashtu, Pukhtu, Pushto also known as Afghani is an Indo-European language and belongs to the Eastern Iranian branch of Indo-Iranian language family. It has two major dialects, hard (northern) dialect and soft (southern) dialect. The difference between these two is phonological. For example, people of hard dialect pronounce Pashto as 'Pukhto' or 'Pakhto' while in soft Pashto; it is called 'Pashto'. In the paper, the word Pashto refers to both hard and soft dialects. The Kandahari form of Pashto which is another dialect, is reflected in the spelling system, and is considered to be the "standard". In this paper, Peshawari form of Pashto is discussed.

Pashto is written in a variant of the Persian language script which in turn is a variant of Arabic script. Pashto script consists of 45 alphabets shown in **Figure 1**. In Pashto, two or more characters are combined to make a word or combine character. The shape of the characters varies according to its position in the word. For example the shape of isolated character is different from the same character when combine with other characters, e.g. it is different in start, middle and end of a word as shown in **Table 1.** This makes the overall recognition process a bit difficult. In this paper, only the isolated characters classification problem has been addressed.

Most of the Pashto alphabets are derived from Persian alphabets. To account for Pashto's own sounds, certain alphabets has been added and modified to it. Pashto language has been written in both Naskh and Nasta'liq scripts but Naskh has been adapted as a standard and is discussed here.

Pashto language has adapted all 32 alphabets of Persian language which consist of 28 Arabic alphabets plus 4 additional Persian specific alphabets. Urdu language which is also a variant of Persian language has 38 alphabets, 32 Persian alphabets plus 6 special Urdu alphabets. Pashto owns the 6 Urdu special alphabets with only representational difference in some characters as shown in **Table 2** and adds 7 Pashto special alphabets shown in **Table 3,** making a total of 45.

3. DATABASE DEVELOPMENT

Pashto language has no standard database for research for research on OCR. To develop the OCR for Pashto script, there must be some standard database to evaluate the performance of the OCR system. In this paper, a medium size database has been developed. The database consists of 45 isolated Pashto character with 25 variations of each character making a total of 1125 entries in the database. The text used in this work is provided by the Pashto Department of the University of Peshawar, and after preprocessing and resizing have been added to the final database. The size of each alphabet in the database is kept to 80 x 60.

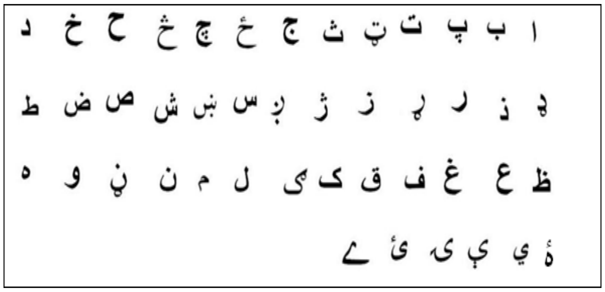
ا ب پ ت ټ ث ج ځ چ څ ح خ د
ډ ذ ر ړ ز ژ ږ س ش ښ ص ض ط
ظ ع غ ف ق ک ګ ل م ن ڼ و ه
ۀ ي ې ی ۍ ئ ے

Fig 1. Pashto Alphabets

Table 1. Variations of characters according to positions

End	Middle	Start	Isolated
ـب	ـبـ	بـ	ب
ـح	ـحـ	حـ	ح

Table 2. characters representation in Urdu and Pashto

Pashto representation	Urdu representation	Alphabet
ټ	ٹ	ٹ
ډ	ڈ	ڈ
ړ	ڑ	ڑ
ګ	گ	گ
ی	ے	ے

Table 3. Pashto special Characters

ئ	ې	ي	ښ	ږ	څ	ځ

4. OCR SYSTEM FOR PASHTO

The developed Optical Character Recognition system takes the images of alphabets and classifies them automatically. The input to OCR is an alphabet and the output is the label provided to that character. Generally, any pattern recognition process is composed of four major steps. Among these, the first step is data acquisition. In this case, the data may be any characters set from the database. The next step is pre-processing. In this step, operations like noise removal, thresholding, scaling etc are performed. Thresholding is performed on the alphabets to convert them into binary and separate the background and foreground information. The next step is feature extraction. In this paper, two types of features have been extracted and used for classification namely High level features and low level features. Finally, the extracted features are provided to the classifier in order to classify the different alphabets.

Features Extraction

Feature extraction is the most important step of any recognition system. The purpose of feature extraction is to take the important characteristics of the image and classify the overall image using this small set of information. The selection of features directly effects the classification operation. Good features results in a higher success rate in the process of recognition and vice versa. In this paper, two types of features have been extracted.

a) Structural features/High level features
b) Statistical features/Low level features

Classification Using High Level Features

At higher level, characters are represented by structural features. Since these features belong to the structure of the character, therefore these are largely invariant to style variation and distortion. Structural features are base on shape and geometry of the alphabets such as branch point, strokes and their direction, inflection between two points, horizontal curve at top or bottom and cross point.

The geometric properties used in this study are discussed below:

1) Number of labels: First of all, the alphabets have been divided into 3 different categories based on the number of labels. A label is a connected component in a character. For example, the 1, 2 and 3 labels alphabets are shown in **Figure 2**.

2) Area: After dividing the alphabets in three main categories, the next step is to further differentiate the alphabets in the same category. To do so, the area covered by each character has been used. For example, both the alphabets in Figure 3 have two labels but differ in area.

3) Euler Number: To further classify the alphabets, the Euler number property has been used. This gives the difference of objects to the number of holes in it. For example, the character in Figure 4 gives -1 because it is a single object having two holes.

4)Extent and orientation: Two more geometric properties i.e. extent and orientation have been used to further separate the charectors still lying on the same node. The extent property gives the area of the object divided by the area of the bounding-box. Bounding-box is the smallest rectangle containing the whole object. The orientation property gives the angle between the x axis and major axis of the ellipse.

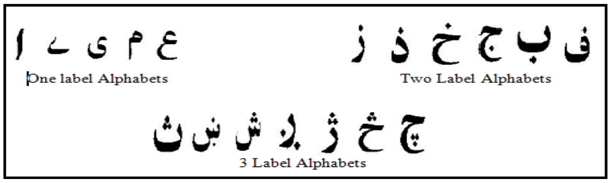

Fig 2. No of label in Different characters

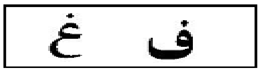

Fig 3. Alphabets with same labels and different area

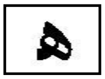

Fig 4. character with euler number equal -1

Using the above mentioned properties, all the alphabets of Pashto script have been classified. This overall classification makes a tree form of structure. The partial classification tree is shown in Figure 5.

Classification Using Low Level Features

Classification on high level feature is very robust and easy but in some cases when the properties of the characters are same, this approach fails to classify the characters. Consider two label characters with same area, extent and orientation. Low level feature are those which are obtain after some mathematical operations are performed on the image. In this case, we subdivide the image into small windows of size 5x5 and take the average of those values thus computing 192 features for images having dimensions 80 x 60. These features are then provided to the classifiers both for the training and testing purposes. The K-Nearest Neighbors classifier is used for the classification of these features.

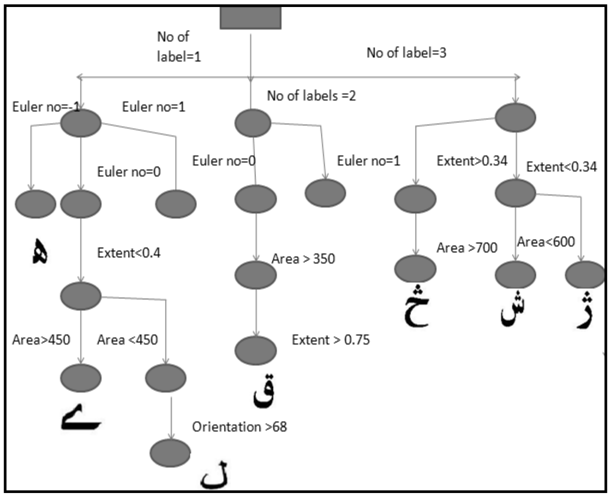

Fig 5. Classification Using high level features

Description of the classifier-

In this study the K-NN (Cunningham et al., 2007) is used for classification. KNN is one of the simplest classifiers which classifies an unknown observation by looking at the nearest neighbors of the unknown point among the known data, and assigns a class label to the unknown observation based on the majority vote among the nearest neighbors.

The determination of the neighbors and thus the output label depends mainly on the distance metric. Let the observations in the training data be represented by x_i, i=1,2,3…N, where N is the number of observations in training data. Each observation x_i is described by a set of features F. If the unknown observation is represented by q, then the distance between test observation q and training point x_i is given by:

$$d(q,xi) = \sum_{f \in F} wf\delta(qf,Xif)$$

A number of possibilities for the distance metric have been used, however a typical version is given as:

$$\delta(qf,Xif) = \begin{cases} 0 & f\ discrete\ and\ qf = xif \\ 1 & f\ discrete\ and\ qf \neq xif \\ |qf - xif| & f\ continuous \end{cases}$$

The k-neigbors are selected on this distance metric. To find the class of the query, the simplest approach is to assign the majority class to the query.

5. RESULTS AND DISCUSSION

This section presents the performance of the developed OCR system on the classification of Pashto alphabets. In case of high level feature, the input to the classifier is the image of the alphabet while the result is the recognized text output. The classification scheme developed uniquely identifies all the alphabets. However, this approach is affected by the variation in style and size. In KNN based classification, the input is training data, test data and the label vector containing the classes of the training data. The classifier classifies each instance of the test data to one of the class in the label vector. The classifier has been run for different sets of training and testing data. The average accuracy of the classifier was found to be 74.8%.

6. CONCLUSION AND FUTURE WORK

In this paper, we developed a medium size database for Pashto language and a basic OCR system which recognizes isolated Pashto characters. We performed two types of classification based on the nature of the features namely High level classification and Low level classification. We used K-NN for low level feature classification. In future, we will work on the recognition of connected words and will use different classifiers for result comparison.

REFERENCES:

Amin, A. (1998) “Off-line Arabic character recognition: The state of the art”, Pattern Recognition, (31), 517-530.

Arora, S., D. Bhattacharjee, M. Nasipuri, D. K. Basu, and M. Kundu (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), 1-6.

Chaudhuri, B. B., and U.Pal, (1998) “A complete printed Bangla OCR system”, Pattern Recognition, (31), 531-549.

Chaudhuri, B. B., U. Pal, M. Mitra, (2002) “Automatic Recognition of Printed Oriya Script”, Sadhana, 27(1), 23-34.

Cunningham, P., S. J. Delany, (2007) “k-Nearest neighbour classifiers”, Technical Report UCD-CSI-2007-4, University College, Dublin.

Nagy, G. (1988) “Chinese character recognition-A twenty five years retrospective”, 9th IEEE International Conference on Pattern Recognition, 163-167.

Pal, U., and A Sarkar,. (2003), “Recognition of Printed Urdu Script”, Seventh International Conference on Document Analysis and Recognition (ICDAR), 1183-1187.