DOI: http://dx.doi.org/10.26483/ijarcs.v8i9.4759

ISSN No. 0976-5697

Volume 8, No. 9,November-December 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A REVIEW ON RECOGNITION OF HANDWRITTEN URDU CHARACTERS USING NEURAL NETWORKS

Mohd Jameel
School of Comp. & System Sciences
Jaipur National University, Jaipur
Rajasthan, India.

Sanjay Kumar
School of Engineering & Technology
Jaipur National University, Jaipur
Rajasthan, India.

Abdul Karim
Deptt.of Computer Science
Govt. PG College, Rajouri
Jammu &Kashmir, India.

Abstract: Character recognition being one of the most interesting and attractive areas of pattern recognition and artificial intelligence has got additional consideration during last decade due to its wide range of applications. It contributes immensely to the computerization process and enhancing the man-machine interaction in many applications. It is an art of detecting and recognizing the characters from input image and converting them into ASCII or other corresponding machine editable form. There are four main phases of Character Recognition – Data acquisition and Preprocessing, Segmentation, Feature extraction and Classification. Several research studies have been carried out for recognition of scripts like Chinese, Japanese, English, Devanagari, etc. but the research regarding Urdu Script is still immature due to cursive, variable and overlapping nature of Urdu characters and different writing styles. Research studies on printed Urdu characters have shown good recognition rate but the Handwritten Urdu Script Recognition is still an open and challenging area for researchers. This paper presents a review of Urdu handwritten character recognition methods with special reference to neural networks and includes information regarding the various operations that may be performed on the image for the recognition of Urdu characters. In literature, it has been found that B-Spline curves are not yet applied in combination with Neural Networks for Urdu script recognition. The current research work intends to use B-Splines curves for feature extraction with Neural Network as classifier and focuses on isolated characters in offline domain.

Keywords: Handwritten, Urdu, Character recognition, neural network, B-Spline curve,Pattern recognition, features.

1. INTRODUCTION

Character recognition being one of the most interesting and attractive areas of pattern recognition and artificial intelligence has got additional consideration during last decade due to its wide range of applications. It contributes immensely to the computerization process and enhancing the man-machine interaction in many applications. It is an art of detecting and recognizing the characters from input image and converting them into ASCII or other corresponding machine editable form. There are four main phases of Character Recognition – Data acquisition and Preprocessing, Segmentation, Feature extraction and Classification. The Script Recognition technique is very useful in making paperless environment in many major organizations as far as the preservation of their previous record is concerned. Urdu is a cursive language having connected characters making words. Urdu language is famous and spoken by a majority of population in Indian Subcontinent and also in parts of Europe and America. A lot of work has been done in Urdu poetry and literature from many centuries. Creation of recognition system for Urdu language shall play an important role in converting all those texts from physical Libraries to electronic libraries. Most of the stuff already placed on internet is in the form of images having text which consumes a lot of memory space to transfer or to read online. As far as the Recognition of Handwritten Urdu text is concerned, the problem is more challenging and needs extensive research as there is no significant performance achieved as compared to printed Urdu character recognition and other scripts. The major difficulty in recognition of Handwritten Urdu language is its cursive nature and different font styles. The intensity of the problem further increases when it comes to writing styles and moods of different writers. Research in this area has got more focus during last decade as its applications are increasing. It helps in improving Human-Computer Interaction, Paperless environment, online newspapers, online availability of old literature, paper checking, automating official tasks, reading bank receipts, postal addresses and data entry forms. The importance of Handwritten Urdu Recognition System further increases where Urdu is an official language like Jammu Kashmir and it can go a long way in digitization of land records and maps which are written in Urdu and are on the verge of degradation.

2. URDU LANGUAGE CHARACTERICS AND CHALLENGES

The Urdu language after being evolved from many languages like Arabic, Farsi and Sanskrit, has the characteristics of all these languages. All the characters in this language have been taken from these languages. Urdu language is a more cursive and complex than other languages as it contains the connected characters to make words. This cursive and variable nature makes it very difficult to be recognized through usual Character Recognition Methods. It is a context sensitive language and is written in the form of ligatures that may comprise a single or many different characters to form a word. Urdu character set consists of 39 characters. The characters may be single looped, double looped and incompletely looped. Dots and diacritics are also a part of character set. Dots include single, double and triple dots. Many characters are similar in shape and only the number and position of dots differentiates them. The shape of a character changes depending upon its position in the word. Every character is of minimum 2 shapes and maximum 4 shapes and its positions in the word are isolated, initial, middle and last. Another characteristic of Urdu characters is that they are written from right to left whereas numerals are written from left to right. There is no baseline in Urdu writing rather the text is Centre justified and some characters overlap with one another. This makes it very difficult to be recognized and in case of handwritten Urdu characters, the problem becomes more challenging and open to the researchers..

ادڈذرڑزژوے

ب پ ت ٹ ث ج چ ح خ س ش ص ض ط ظ ع غ

ف ق ک گ ل م ن ہ ی ھ

Fig.1 Non joiner and joiner Typewritten Urdu characters

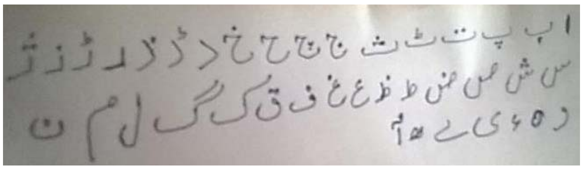

Fig.2 Handwritten Urdu characters

ب ،بح ،بت ،بد ،بص ،بط ،بن ،بو ،بق ،بف ،بم ،بز ،بل ،با ،بی

Fig.3 Shapes of Urdu characters based on position in word

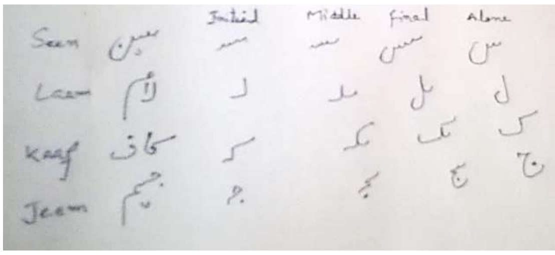

Fig.4 Variable shapes of Handwritten Urdu Characters

3. REVIEW OF EXISTING WORK

***3.1Image Acquisition*: ** Process of feeding Image of characters into the computer is known as Image acquisition. Earlier, the image acquisition was performed with the help of flat-bed scanners which took clear images of standard quality from which characters were to be recognized. Advantages of scanner are low noise levels, less blurring and less text Skewness. With the advancement in technology, high resolution cameras have been brought in smaller devices such as mobile phones which help in capturing high quality images thus relieving us from slow operation and physical connections of scanners. The images acquired through these cameras too have problem of shadows, blurring and Skewness to some extent but can be overcome by applying suitable image enhancement techniques.

3.2 Pre-Processing:ThePre-processing of the image is the process which involves changes and alterations to the image to make it suitable for recognition. The following techniques may be used for image enhancement:

3.2.1 Conversion of RGB to Gray-Scale:The image acquired before preprocessing is usually a coloured image which means that the pixel value of the image contains a three colour components; Red, Green and Blue. First of all this coloured image is converted into a standard gray-scale image and is represented through a single matrix as the detection of characters on a coloured image is more difficult than on a gray-scale image.

3.2.2 Skew Correction:Images obtained through Camera may be skewed and distorted due to improper image capturing techniques. This type of effect can be reduced by rotation of the image by a certain degree. The rotation of the image has been calculated by A.F Mollah et al in [7].

3.2.3 Binarization:Binarization is the process of selecting a threshold value is for conversion of pixel values into 0's and 1's. The 0's represent the black pixels whereas 1's represent the white pixels. There are several methods of selecting threshold value. One of the simplest methods of selecting threshold value is to find out the median value of the maximum and minimum intensity values in the image which can be written as:

Threshold value, $T = (I_{max} + I_{min})/2$

3.2.4 Noise Reduction:The presence of unnecessary pixels in an image is called noise. Noise may be in the in the form of Salt and Pepper noise or Gaussian noise. Low pass filtering is used to remove the Gaussian noise from the image [1] and there is no need to filter Salt and Pepper noise as it is very low as compared to the Gaussian noise.

3.2.5 Thinning: Thinning is the process by which the reduction of width of foreground pixels in the image takes place. While thinning, it is mandatory to preserve the form of the characters on the image. Thinning removes the extra pixels around axis along which the shape of the character is preserved. If a skeleton of a character is 5 pixels wide, the extra 4 pixels are removed to make the character 1 pixel wide.

3.3 Segmentation:When the character image is completely preprocessed, it is passed to the segmentation stage where each character is separated from one another. The Image at this stage is divided into two regions namely background region and foreground region. This segmentation stage separates the foreground region from the background region. The foreground region is a collection of text characters on one or more lines. The segmentation involves two main steps:

1. Line Segmentation. 2. Character Segmentation.

Line segmentation is the separation of the different lines of characters present in the image. Each line is defined by a minimum vertical gap between the characters present on a line and on the line above and below it. This gap can be used for the detection and separation of different lines of characters. Character Segmentation is the separation of characters present in the same line. Once the lines are separated, each character is extracted from the line. There is a constant horizontal gap between characters which is used for the separation of characters [1].

3.4 Feature Extraction: Feature extraction is the process which retrieves the most relevant set of parameters that uniquely and precisely define the properties of characters to be recognized. This set of parameters is known as feature vector of a character. The main objective of the feature selection and extraction is to maximize the recognition rate with least number of parameters and generate a similar set of features for the same class of characters. The features are to be chosen as suggested by Lippman as: "Features should contain information required to distinguish between classes, be insensitive to irrelevant variability in the input, and also be limited in number, to permit, efficient computation of discriminant functions and to limit the amount of training data required." Feature extraction is done after the completion of preprocessing stage of character recognition system. It plays a crucial part in recognition of characters with high accuracy and poorly extracted features will definitely reduce the recognition rate. In literature, many feature extraction methods have been found to be applied such as Template matching, Deformable templates, Unitary Image transforms, Graph description, Projection Histograms, Contour profiles, Zoning, Geometric moment invariants, Zernike Moments, Fourier descriptors, Gradient feature and Gabor features and many of them have been used for recognition of Urdu characters but B Spline Curve approximation has not been used for Urdu character recognition so far (to our knowledge) inspite of the fact that they are more robust and continue in nature. The proposed work intends to make use of B Spline curves extract the features of segmented Handwritten Urdu Characters to form the feature vector.

3.5 Classification:The classification is the process of identifying each character and assigning a suitable character class to it. A good classification depends mainly on how the characters are segmented and their features extracted. While classifying the characters, two different approaches are followed. First is decision based approach in which description of character is numerically expressed in the feature vector and second approach is applied when characteristics are derived from the physical structure of the character which is not easily quantified. In this case a relationship between the characteristics may be of importance when deciding on class membership. SaeedaNaz et al [1] have made an efficient use of Multi-Dimensional Long Short Term Memory (MDLSTM) Recurrent Neural Networks for recognition of printed Urdu text-lines written in the Nasta'liq writing style and obtained a recognition accuracy of 98% for the unconstrained Urdu Nasta'liq printed text, which gives highest performance among the state-of-the-art techniques but it has not been applied for Handwritten Urdu characters.In fact, the character recognition systems should also be able to read handwritten text. M. Farhad et al. [5] proposed a methodology for Character Recognition of English alphabets using Artificial Neural Network for classification with curvature features of characters as input to the network. They applied different seeking angles using predetermined features for the recognition of characters and achieved 90% accuracy but in that case the feature extraction was very time consuming. Amit Choudhary et al. have shown an extensive work on offline handwritten English character recognition using multilayer feed forward neural network and an accuracy of 85.62% has been reported. Sarmad Hussain has worked on Urdu text to speech system. The work focused on the use of Urdu phonological processes and divided it into three stages. In the first stage, he has converted text into its respective phonemes. In second stage, these phonemes were converted into numerical parameters and at third stage; speech was synthesized through these parameters but this work is also based on typed Urdu characters. KashifShabeeb and D.S Singh have developed a GUI for Urdu Text to Speech Converter and was based on the recognition of isolated printed Urdu characters using Artificial Neural Networks, however, its accuracy has not been mentioned. Shamsher et al. [8] have applied Feed Forward Neural Network to recognize the Urdu characters and got successful result for isolated characters consuming minimum processing time. Feed Forward Neural Network was also applied by Ahmad et al. [9] to recognize the joined Urdu characters but this system could not merge small segments of the character. Haider et al. [10] have proposed an online method for the recognition of Urdu handwritten characters and it could handle only single stroke handwritten characters. U. Pal and A.Sarkar [11] have presented an Optical Character Recognition system for printed Urdu script. They used Hough transforms for the detection and correction of the skew in the text. Hussain et al.[16] have presented a method for recognition of Urdu script using Kohonen Self-Organizing Map as classifier. Another efficient effort has been made by Tariq et al. for isolated Urdu characters in which soft converter is used to recognize isolated Urdu characters. Shahzad et al. [13] presented an online system for recognition of handwritten isolated Urdu characters which are drawn on a Tablet PC. Sagheer et al. [14] proposed a handwritten Urdu words recognition method which recognizes and classifies the candidate words Support Vector Machine. I.K. Pathan et al. [4] presented a recognition system for offline handwritten isolated characters based upon the Invariant moments which separated the characters into primary and secondary components and Support Vector Machine was used for classification.

4.PROPOSED SYSTEM FOR HANDWRITTEN URDU CHARACTER RECOGNITION USING FEED FORWARD NEURAL NETWORK.

After a complete review of the literature, the use of Neural Networks in recognition of printed characters has given a promising result but their application in recognition of Handwritten Urdu Characters has not been investigated (to our knowledge) in combination with B-Spline curve approximation as feature extraction method inspite of the fact that the B-Splines are the continuous curve representations and affine invariant. It is fact that the characters are formed by certain curves and hence each letter or character may be represented by a curve. A neural network is a computational model and finds applications in highly complex problems and statistically variable data. Several associative neural network models are available and multi-layered feed-forward back propagation neural network is best suited for our problem. In neural network the set of inputs is mapped to a set of outputs and this mapping becomes a multidimensional mapping surface. The goal of learning is to direct the mapping surface according to a desired output. The overview of whole process of recognition of handwritten Urdu language characters is shown in Fig.5

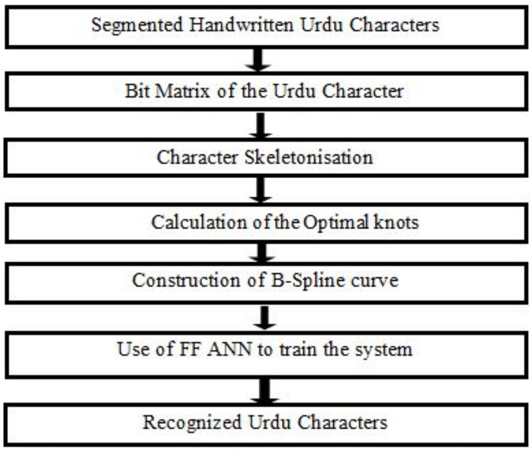
Segmented Handwritten Urdu Characters
Bit Matrix of the Urdu Character
Character Skeletonisation
Calculation of the Optimal knots
Construction of B-Spline curve
Use of FF ANN to train the system
Recognized Urdu Characters

Fig.5 Urdu HCR using Feed Forward Neural Network

5. CONCLUSION

This paper provides a detailed review of published research work in all necessary stages of Urdu handwritten character recognition system. It has been observed that significant progress is made in case of printed Urdu character recognition but handwritten Urdu script recognition system is quite immature due to discussed complexities. Several feature extraction, preprocessing, segmentation and recognition methods are employed by the researchers and reported different accuracy levels but the application of B-Spline Curve approximation in combination with Feed forward Neural Network as classifier has not been made for handwritten Urdu characters recognition. In current research work, the use of this technique has been proposed to enhance the accuracy and efficiency of Urdu Handwritten Character Recognition considering curve like structure of Urdu characters and continuous and affine invariant nature of B-splines and highly adaptive capability of Neural Networks

6. REFRENCES

[1] SaeedaNazetal “Urdu Nasta’liqtext recognition using implicit segmentation based on MDLSTM neural networks” Springer 2016.

[2] Sabahat Mir, S.ZamanM.W.Anwer”Printed Urdu Script Recognition Using Analytical Approach’ 13th Intel Conference on Frontiers of IT 2015.

[3] R.K Jambekar “A Review of Optical Character Recognition System for Recognition of Printed Text “IOSR Journal of Comp. Engineering 2015.

[4] I,K. Pathan, A.A. Ali and R.R. J, Recognition of offline handwritten isolated Urdu characters, Advances in Computational Research, 2012.

[5] M. M. Farhad, S M N Hossain, Ahmed S Khan,“An Efficient Optical Character Recognition Algorithm using Artificial Neural Network by Curvature Properties of Characters”, 3^{rd} International Conference on info, elect & vision 2014.

[6] S. Shastry, Gunasheela G, ThejusDutt, Vinay D S and S.RaoRupanagudi, “ A novel algorithm for Optical Character Recognition ”. IEEE, 2013.

[7] A. F. Mollah, S. Basu, N. Das, R. Sarkar, M. Nasipuri, M. Kundu, “Text/Graphics Separation and Skew Correction of Text Regions of Business Card Images for Mobile Devices”, Journal of Computing, Vol. 2, Issue 2, February 2010

[8] I. Shamsher, Z. Ahmad, J.K. Orakzai, and A. Adnan “OCR For Printed Urdu Script Using Feed Forward Neural Network, Proceedings of World Academy of Sc., Eng and Tech. volume 23, 2007.

[9] Z. Ahmad J.K. Orakzai, and I. Shamsher “Urdu Compound Character Recognition Using Feed Forward Neural Networks” IEEE 2009.

[10] I. Haider, and K.U. Khan, Online Recognition of Single Stroke Handwritten Urdu Characters, IEEE,2009.

[11] U. Pal and A. Sarkar “Recognition for Printed Urdu Script”Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)

[12] J. Tariq, U. Nauman, M.U. Naru, Softconverter: A Novel Approach to Construct OCR for Printed Urdu Isolated Characters, IEEE, 2010.

[13] N. Shahzad, B. Paulson, T. Hammond, Urdu Qaeda: Recognition System for Isolated Urdu Characters, UI Workshop on Sketch Recognition, Sanibel Island, Florida, 2009.

[14] M.W. Sagheer, N. Nobile, C.L. He, C.Y. Suen, A Novel Handwritten Urdu Word Spotting Based on Connected Components Analysis, 2010.

[15] Afzal, M. and Hussain, S., “Urdu Computing Standards: Urdu ZabtaTakhti (UZT) 1.01”, in the Proceedings of International IEEE Multi topic Conference (INMIC), Lahore University.

[16]Mohd Jameel, Sanjay Kumar “Offline Recognition of Handwritten Urdu Characters using B Spline Curves: A Survey”International Journal of Computer Applications V.157–No 1, 2017