ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

ISOLATED CHARACTER RECOGNITION USING HIERARCHICAL APPROACH WITH SVM CLASSIFIER

Shaina*, Harpreet Kaur Bajaj
[*] M. Tech (CSE) (Sem6) DAV Institute of Engg and Technology
Asstt. Professor, Comp. Dept. DAV Institute of Engg and Technology

DOI: 10.5281/zenodo.154564

ABSTRACT

This paper proposes a method for Urdu language text. Character recognition is obtained by OCR. This paper represents the effectiveness of characters with SVM Classifier using Hierarchical approach. SVM is a useful technique for data classification. The objective of SVM is to generate a model which predicts the target value. The work is done on Sindhi Character Set. The experiment shows that character recognition with SVM Classifier achieves a recognition rate of 93.0481%.

KEYWORDS: Optical Character Recognition, SINDHI Character Set, SVM Classifier.

INTRODUCTION

Optical Character Recognition is valuable and significant in office mechanization as well as spontaneous data access in banks. There are some OCR systems for Indian scripts like Gurumukhi, Devnagari and Bangla, Malayalam. Naz et al [1] described the optical character recognition (OCR) literature with reference to the Urdu-like cursive writings. For this, various attempts are grouped into three parts, namely: printed, handwritten and online character recognition. Husain et al [2] described the facility of text input through keys that are an inconvenient and slow way of input. The design of an online Urdu handwriting recognition system was recognized for about 850 single character, 2 character and 3 character ligatures, enabling input of about 18000 common words from the Urdu Dictionary. Khan et al [3] proposed a method for Urdu language text founded in Urdu Text and the recognition degree obtained as 96.2 % for isolated characters. Akram et al [4] observed the outline of Urdu document images having font size between 14 to 44 has 86.15% ligature recognition correctness tested on 224 document images. Shamsher et al [5] proposed an Optical Character Recognition scheme for published Urdu, a general Pakistani/Indian writing. Khan, K., Siddique [6] described an effective system for Urdu text and results demonstration was 100 % accuracy for 4, 5-character ligatures, 87 % for 3-character ligature and 78 % for 2-character. Ahmad et al [7] discussed the Urdu script characteristics; the characters recognition method obtainable here was inherited the complexity of Urdu script to crack the problem. A word was scanned and examined for the level of its complexity and it achieved 93.4% correctness on the average. Rani et al [8] presented the efficiency of Gabor Filter banks with KNN, SVM and PNN classifiers to classify the writings at line level from such trilingual booklets. The experiments presented that Gabor sorts with SVM classifier achieve a recognition degree of 99.85% for trilingual forms. Singh et al [9] presented a relative performance analysis of feature(s)-classifier mixture for Devanagari optical character recognition system and was originated to be 96.69%. Khan, H.I [10] discussed around a hint to identify Kannada vowels by chain code features and the level of correctness touched to 100%. Ahmed Lawgali [11] in his paper described a review of Arabic character recognition schemes which are categorized into the character recognition groups: printed and handwritten. Sahlol and Suen [12] proposed new approaches for handwritten Arabic character recognition which is founded on original preprocessing processes including dissimilar kinds of noise removal also dissimilar kind of structures like structural. Sharma and Jain [13] offered the growth of Gurumukhi character recognition system of isolated handwritten characters by using Neocognitron at the first time and accuracy for both learned and unlearned Gurumukhi characters were 92.78 %. Dara and Panduga [14] described offline HCR by removing features using 2D FFT and using the provision vector machines for Telugu documents. The best percentage recognition correctness for Telugu handwritten characters was 71%. Ali and Shaout [15] presented a survey of the research published in this area. The paper analyzed and compared the various algorithms with task-specific respect to different stages of the offline AHCR. Khan et al[16] discussed in detail method to identify Kannada vowels using chain code features. 8–neighborhood method has

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

been effected which permits generation of eight different codes for each character.Aram et al [17] discussed the structure of Urdu Nastalique OCR. The existing scheme held the recognition of Urdu Nastalique document images taking font size 14 to 44. The system was 86.15% ligature recognition correctness verified on 224 document images. Khan et al [18] described a well-organized technique for Urdu text search in computer generated and handwritten scanned images. Investigational results showed 100 % accuracy for 4, 5-character ligatures, 87 % for 3-character ligature and 78 % for 2-character ligatures.

In this paper, basic focus is on accuracy of OCR for character recognition using Sindhi script. The experiments show that character recognition with SVM Classifier achieves a recognition rate of 93.0481%.

INTRODUCTION TO SINDHI CHARACTER SET

Figure1: Sindhi character set

Sindhi Language script has a total of 52 letters in its alphabet and its neighboring languages such as Persian (32), Urdu (39) and Pashto (44). Sindhi constitutes the largest extension of the original Arabic script. Sindhi starts from right to left for the characters, numbers, the system is from left to right Profitable OCR systems can mainly be gathered into two sorts: Task-specific readers and general-purpose page readers. A reader grips only exact document kinds [19]. Some of the most mutual task-specific readers read bank cheque, letter mail, or credit-card slips. These readers typically apply custom-made image-lift hardware that internments only a few predefined text regions. For example, a bank-cheque reader may test just the quantity field and a postal OCR system may test just the report block on a mail piece. Such systems underline high throughput charges and low error charges. General-purpose page readers are intended to grip a extreme variety of papers such as commercial letters, technical writings, and newspapers. These systems capture an image of a text page and distinct the page into script regions and non-text regions. Non-text regions such as graphics and line pictures are repeatedly saved distinctly from the text and connected recognition grades. Text regions are segmented into lines, words, and characters and the characters are conceded to the recognizer. [20]

OCR TECHNIQUES

Following are the steps which are used in OCR.

Preprocessing: It is an attempt to improve the performance of OCR. All those processes which improve the image quality and prepare it for next stages are come under this step.

Segmentation: Under this, text is subsequently segmented into paragraphs, lines, words, characters and sub-characters. For each connected component in a word, accents or separate dots are merged to form a character, with the supposition that a character won't be too extensive or too thin. Some of the approaches obtainable for the segmentation are labeled below [3].

Thresholding: The humblest way of image segmentation is called the thresholding way. This technique is founded on a clip-level to turn a gray-scale image into a binary image. There is also a composed histogram thresholding. The key of this technique is to choice the threshold value. Numerous popular methods are used in manufacturing.

Compression based Method: Compression based approaches guesses that the optimal segmentation is the one that reduces, over all imaginable segmentations, the coding distance of the data. . The method defines each

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

segment by its surface and border shape. Each of these mechanisms is showed by a probability circulation function.

Histogram based Method: Histogram-based approaches are very well-organized likened to other image segmentation approaches since the characteristically need only one pass done the pixels. In this technique, a histogram is added since all of the pixels in the image. Histogram is used to discover the bunches in the image.

Edge Detection: Edge detection is a strong arena on its individual within image processing. Region boundaries and edges are faithfully associated, since there is regularly a piercing adjustment in strength at the region boundaries. Edge detection methods have consequently been used as the base of another segmentation technique.

Partial Differential Equation-Based Methods: Using a partial differential equation (PDE)-based technique and solving the PDE equality by a numerical scheme, one can segment the image. The essential idea is to grow an original curvature near the lowest potential of a cost function, where its definition reproduces the task to be spoken.

Graph Partitioning Methods: Graph partitioning approaches are real tools for image segmentation since they model the influence of pixel areas on assumed cluster of pixels or pixel, under the supposition of homogeneity in images. In these approaches, the image is modeled as a weighted, undirected graph.

Feature Extraction: Feature study controls the descriptors, or feature set, used to define all characters. Agreed a character image, the feature extractor arises the features that the character holds.

Classification: Classification is made by relating an input character image with a set of patterns from separately character class. Each comparison consequences in a resemblance amount between the input character and the template.

Post processing: This stage includes to rise the recognition degree by falling the number of errors specially the condensed the rate of confused characters [21].

Proposed System: Character recognition is obtained by OCR. The work is done on Sindhi character set with SVM Classifier using Hierarchical approach.

Following Steps are followed during proposed system:

Step1: First Step is preprocessing. During this stage noise is removed from images of train databases and test databases
Step2: A database of images named ‘Train Database’ has been created. Different writing styles are chosen so that there is no problem in classification stage.
Step3: A Second database of images named ‘Test Database’ also has been created.
Step4: A matrix L (M*M) has been calculated. Eigen vectors and Eigen values are found.
Step5: Feature vector is created for each image. This value is used for classification.
Step6: A threshold value is chosen and that value is used for classification purposes.
Step7: Feature vector of a character to be recognized.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

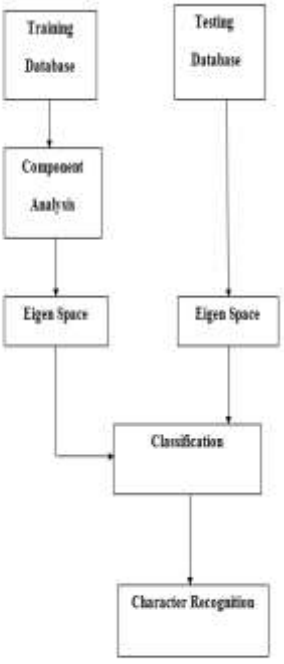

Figure 2: Flow Chart of proposed system

RESULTS AND ANALYSIS

Old and newly scripts were used to assess outcome on good and bad class paper which formed outcome on the usual as 93.4%, which can further be enhanced by a lexicon and focusing further onto recurrently used characters in the script.

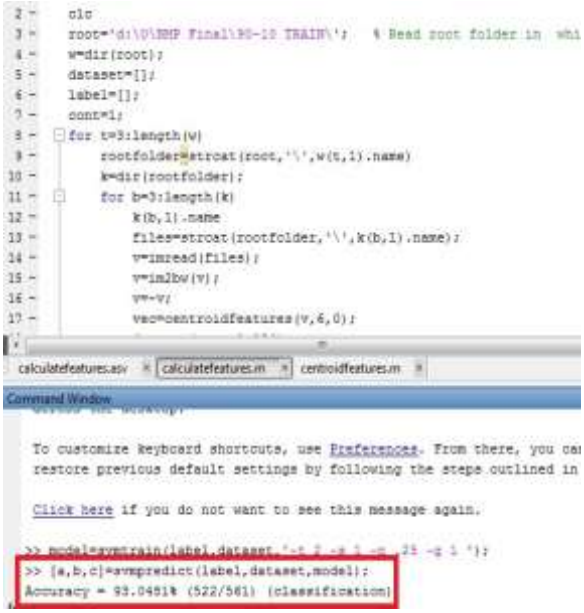

Figure 3: Accuracy using classification

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

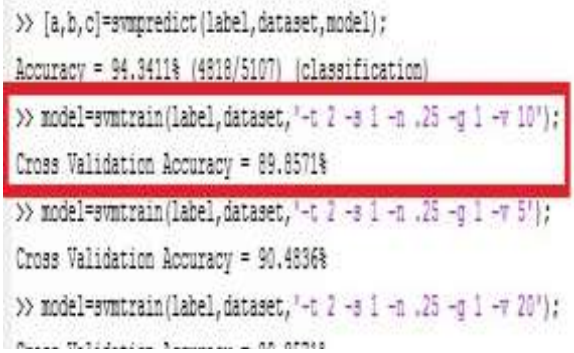

Figure 4: Cross Validation Accuracy 1

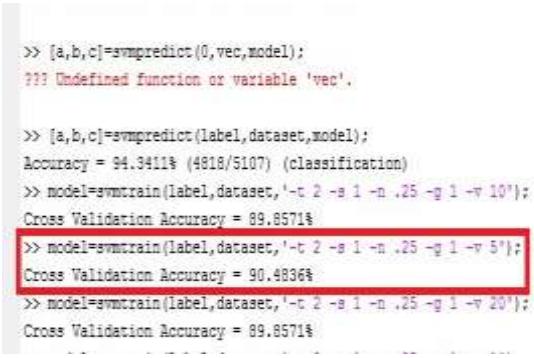

Figure5: Cross Validation Accuracy 2

Figure 3 represents the accuracy using classification is 93.0481%.
Figure 4 represents the cross validation accuracy using SVM Classifier is 89.8571%.
Figure 5 represents the cross validation accuracy using SVM Classifier is 90.4836%.

CONFUSION MATRIX

Characters	غ	ق	ژ	ه	ث	ع	ا	ي	ع	عو	عي	عيَ	عو	عوُ
ق		1											1	
گ		1												
ث					2									
ژ			1											
ه				1										
ن										1				
و													2	
ئ										1				
ي													5	
ع						2								1
عوُ														1
عي						2	6		1	2				
عيَ							2	1			1			
عو									1			1		

Figure 6: Confusion Matrix

The confusion matrix has been generated which shows the accuracy of characters. This confusion matrix works on recognition of isolated characters.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

CONCLUSION AND FUTURE SCOPE

In proposed system, character recognition with SVM Classifier achieves a recognition rate of 93.0481%. This method works on recognition of isolated characters only. The proposed method can be combined with artificial neural network.

REFERENCES

[1] Naz, S., Hayat, K., Razzak, M. I., Anwar, M. W., Madani, S. A., & Khan, S. U. (2014). "The optical character recognition of Urdu-like cursive scripts. Pattern Recognition", vol.47 (3), pp.1229-1248.Chicago
[2] Husain, S. A., Sajjad, A., & Anwar, F. "Online Urdu Character Recognition System. In MVA" pp. 98-101,2007 May.Chicago.
[3] Khan, K., Ullah, R., Khan, N. A., & Naveed, K. "Urdu character recognition using principal component analysis". International Journal of Computer Applications, vol.60 (11). 2012. Chicago
[4] S.Hussain, S., & Ali, S. Nastalique"segmentation-based approach for Urdu OCR". International Journal on Document Analysis and Recognition (IJDAR), vol. 18(4), pp. 357-374.2015. Chicago
[5] Shamsher, I., Ahmad, Z., Orakzai, J. K., & Adnan, A. "OCR for printed Urdu script using feed forward neural network". the Proceedings of World Academy of Science, Engineering and Technology, vol23.2007. Chicago
[6] Khan, K., Siddique, M., Aamir, M., & Khan, R. "An Efficient Method for Urdu Language Text Search in Image Based Urdu Text". International Journal of Computer Science Issues, vol. 9(2), pp. 523-527, 2012.
[7] Ahmad, Z., Orakzai, J. K., Shamsher, I., & Adnan, A. "Urdu Nastaleeq optical character recognition". In Proceedings of world academy of science, engineering and technology Vol. 26.2007, December.
[8] Rani, R., Dhir, R., &Lehal, G. S. "Gabor features based script identification of lines within a bilingual/trilingual document". International Journal of Advanced Science and Technology, vol. 66, pp. 1-12,2014.
[9] Singh, J., &Lehal, G. S. "Comparative Performance Analysis of Feature (S)-Classifier Combination for Devanagari Optical Character Recognition System". Editorial Preface, vol 5(6),2014.
[10] Khan, H. I. (2013). "Isolated Kannada Character Recognition using Chain Code Features".
[11] Lawgali, A. "A Survey on Arabic Character Recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition", vol 8(2), pp. 401-426,2015.
[12] Sahlol, A., &Suen, C. "A Novel Method for the Recognition of Isolated Handwritten Arabic Characters".arXiv preprint arXiv:1402.6650,2014.
[13] Sharma, D., & Jain, U. "Recognition of isolated handwritten characters of Gurumukhi script using Neocognitron". International Journal of Computer Applications, vol 10(8), pp. 10-16,2010.
[14] Dara, R., &Panduga, U. "Telugu Handwritten Isolated Characters Recognition using Two Dimensional Fast Fourier Transform and Support Vector Machine". International Journal of Computer Applications, vol 116(5),2015.
[15] Ali &Shaout"Isolated Arabic Handwritten Character Recognition: A Survey".2014.
[16] Khan, K., Ullah, R., Khan, N. A., & Naveed, K. "Urdu character recognition using principal component analysis". International Journal of Computer Applications, vol60(11).2012.
[17] "Framework of Urdu Nastalique Optical Character Recognition System"
[18] Khan, K., Siddique, M., Aamir, M., & Khan, R." An Efficient Method for Urdu Language Text Search in Image Based Urdu Text. International Journal of Computer Science Issues",vol 9(2),pp. 523-527,2012.
[19] Ahmad, Z., Orakzai, J. K., Shamsher, I., & Adnan, A. "Urdu Nastaleeq optical character recognition". In Proceedings of world academy of science, engineering and technology, Vol. 26.2007, December.
[20] Shamsher, I., Ahmad, Z., Orakzai, J. K., & Adnan, A. "OCR for printed Urdu script using feed forward neural network". the Proceedings of World Academy of Science, Engineering and Technology, vol23,2007.
[21] Husain, S. A., Sajjad, A., & Anwar, F. "Online Urdu Character Recognition System". In MVA ,pp. 98-101,2007, May.