Advances in Computational Sciences and Technology
ISSN 0973-6107 Volume 10, Number 5 (2017) pp. 903-914

http://www.ripublication.com

Writer Identification From Offline Isolated Handwritten Gurumukhi Characters

Kanu Kalra

M. Tech. Research Scholar (Department of Computer Science & Engineering), Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India

Simple Rani

Associate Professor, Department of Computer Science & Engineering, Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India

Abstract

Optical character recognition is commonly used as an arrangement of data entry from published paper data proceedings, statements, bank declarations, electronic receipts, business postcards, mail printouts, or any suitable credentials. This paper deals with feature extraction based classification approach for writer identification from offline isolated handwritten Gurmukhi characters. Features extracted are Zoning features, Open and end point intersection feature. For classification, KNN and Multilayer Perceptron model classification approach is used to detect the writer in Gurumukhi script. This paper also presented comparisons with state of the art methodologies between the classifiers. The database of 30 writers is taken in which whole data is divided into 70 percent of training data and 30 percent of testing data and total database is 10,500. The whole simulation is taken place in WEKA environment.

Keywords: Optical Character recognition, Feature extraction, classification.

I. INTRODUCTION

Artificial intelligence is an area of computer science that emphasizes the creation of intelligent machines that work and reacts like humans and has remained the most challenging research area after the introduction of digital computers. Giving machines the power to see to interpret and ability to read is one of the major tasks of AI. We

humans have the ability for optical character recognition. In other words, we can differentiate between different typescripts and identify them as an A, B etc... Can we insert such ability in software and if we can, how can we? A great deal of activities has been performed on such function but still the results are not 100% and the researches are going on to improve results. Suppose if we wanted to digitize a magazine article or a printed contract. You could spend hours retyping and then modifying misprints. Or we could change all the required resources into digital arrangement in several minutes using a scanner (or a digital camera) and Optical Character Recognition software.

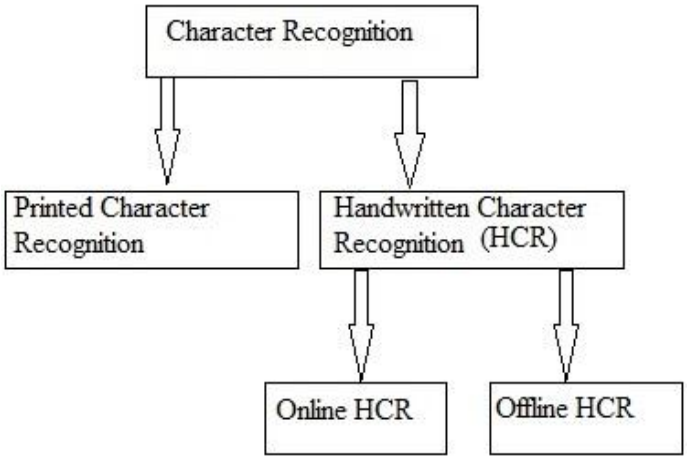

Figure 1: Hierarchy of character Recognition.

There are various types of OCR which are given for various applications.

1. Optical character recognition (OCR) – targets typewritten text, one glyph or character at a time.
2. Optical word recognition – targets typewritten text, one word at a time (for languages that use a space as a word divider). (Usually just called "OCR".)
3. Intelligent character recognition involves machine based learning
4. Intelligent word recognition (IWR) – also targets handwritten script or text, one word at a period. This is especially useful for languages where glyphs are not separated in cursive script.

OCR is generally an "offline" process, which analyzes a static manuscript. Handwriting movement investigation can be recycled as input to handwriting recognition. Instead of merely using the shapes of glyphs and words, this technique is able to capture motions. This additional information can make the end-to-end process more accurate.

II. LITERATURE SURVEY

OCR software often "pre-processes" images to improve the chances of successful recognition. It is stage used to improve the quality of the document image. Dhaval *et al.*[1] Proposed a method that finds the text segmentation with the maximum average likeliness for the resulting typescripts. A graph model is recycled that designates the possible positions for segmenting neighboring typescripts, and then an average longest path algorithm is applied to identify the globally optimal segmentation. Adak *et al.* [2] Proposed an efficient method for the identification of writer for bangla characters. They have used various key points for the structural analysis if the image and scale invariant key point descriptors. Then they have calculated feature sets with various neighbor key points and also used support vector machines for the proper classification. Kumar *et al.*[3] proposed a method that works on different intensity values for extraction of text-lines. The nature of script makes the development of text line subdivision very interesting. Kumar *et al.*[4] proposed a graph based technique to detect the common touching and proximity errors in handwritten text lines. In this method, In refinement step, Expectation Maximization (EM) is used to iteratively split the error segments to correct text-lines. Kumar *et al.*[5] Proposed a method that works directly on gray-scale document images. This algorithm constructs distance transform directly on gray-scale image, which is used to calculate two types of seams: medial seams and separating seams. Garg et al. [6] Proposed line segmentation techniques review with less error probability and high accuracy. The aim of their research work is to deliver some statistics of problems that typically comes through line separation and works on approaches of line division on Indian writings.

In OCR, recognition is a very crucial step as the result of feature extraction step directly affects the result of recognition step. It is difficult to recognize handwritten character due to the variability in writing styles of individual and similarity in the shape of characters is also a major problem. Due to modifiers and diacritics in the Punjabi language, there occurs touching and overlapping in the text. The problems occur mostly in handwritten text as compared to printed text. Some more problems like varying font size, skewness also makes recognition difficult in handwritten text. The aim of the research is to develop a robust system which identifies the write on offline Gurmukhi character. Many algorithms have been proposed but the results are not satisfactory. Some techniques fail to determine the connected components and some fails to segment them at proper place . We have chosen a research topic writer identification from offline isolated handwritten scanned Gurumukhi character from the existing or self-collected data base using feature extraction approach and classification algorithm in WEKA tool.

III. PROPOSED METHODOLOGY

The research work deals with the implementation of the feature extraction approach to extract Zoning features and Open and end point intersection feature and the classification approach using K means Nearest Neighbor and Multilayer perceptron model. Each writer has its own writing style and to recognize this different writing

style is a tedious job. Writing style of Gurumukhi script is from left to right of a page. Some individual writes all the character first and then make header line on those characters which connects character together and that creates a single word whereas some writer writes characters closely with their header line that represents a word but actually characters are not connected. It is very difficult to recognize the writer of the whole document or a single line or a single word at once. So for the proper recognition of Punjabi text, we need to segment document in as small as possible parts and this must be taken care that the parts should be meaningful. The aim of this work is to identify the writer of the Gurumukhi text document into lines so that the recognition can be done accurately. In optical character recognition, recognition is a significant phase and accuracy of character recognition highly depends on feature extraction and classification. Here firstly the text document scanning is done and an image is prepared. Text line detection and separation in digital image documents is a challenging job for handwritten document analysis and character recognition. The problem becomes more difficult if the text lines in the text image are connected or overlapped. Skewness and varying font sizes also makes the character recognition difficult in handwritten text. The goal of OCR is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze and then the classification is done based on feature extraction of the zoning and open and end point intersection using K Nearest Neighbor and Multilayer perceptron model. The performance is evaluated in terms of accuracy to identify the writer for the Gurumukhi script. The methodology flow steps are given below:

STEP 1. First we have collected the database of Gurumukhi characters

STEP 2. Then we have input the Gurumukhi characters for the training process

STEP 3. Then we have applied the feature extraction approach and extracted the Zonal features and Open and end point intersection

STEP 4. After training process we have evaluated the testing approach and we have taken 30 percent of data for testing purpose

STEP 5. Then we have classified the writers based on feature extraction using KNN
STEP 6. Then we have classified the writers based on feature extraction using Multilayer Perceptron model .

STEP 7. Then performance parameters are evaluate

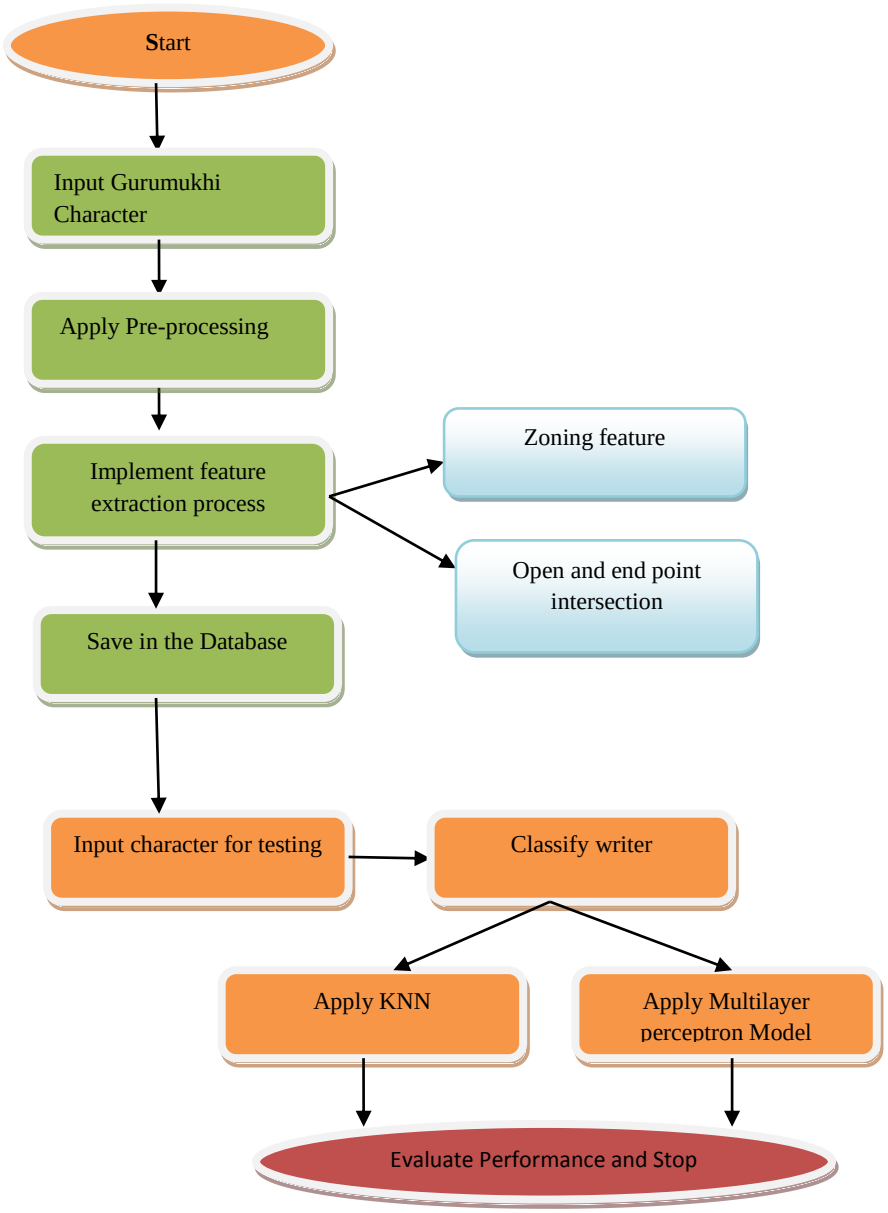

Figure 2: Methodology Flow Diagram.

IV. RESULTS AND DISCUSSION

The accuracy of writer detection based on our proposed approach for Gurumukhi character image highly depends upon the normalization and recognition. The classification accuracies are evaluated using two classifiers and their performance is given below mentioned Table 1. The accuracies are evaluated for each character by 30 writers written each character 10 times which means 30*10*35=10,500 characters.

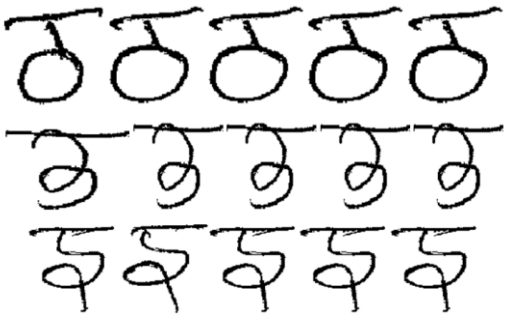

Figure 3: Input samples from different writers.

Where 10 are the number of iterations, 35 are the Gurumukhi characters and 30 are the total number of writers

Table 1: Classification accuracies of different classifiers

Character	Accuracy in (%age) using K-NN	Accuracy in (%age) using MLP
Urha	47.77	53.33
Aara	56.66	67.77
Eeri	45.55	45.55
Sassa	42.22	44.44
Hahha	52.22	53.33
Kakka	52.22	56.66
Khkha	53.33	54.44
Gagga	43.33	40.00
Ghaga	51.11	65.55
Naiya	46.66	47.77
Chcha	50.00	57.77
Chachcha	40.00	38.88
Jajja	55.55	53.33
Jhaja	65.55	53.33
Nana	33.33	41.11
Tainka	42.22	43.33
Thaththa	41.11	33.33
Dadda	55.55	54.44
Dhada	60.00	57.77
Naana	48.88	50.00
Tata	44.44	43.33
Thatha	45.55	54.44
Dada	55.55	55.55
Dhadha	53.33	50.00

Nanna	42.22	47.77
Pappa	45.55	55.55
Phpha	54.44	55.55
Babba	45.55	57.77
Bhaba	36.66	52.22
Mamma	50.00	61.11
Jaiyaa	60.00	55.55
Rarra	45.55	52.22
Lalla	65.55	63.33
Vava	47.77	62.22
Rarha	46.66	55.55
Average	49.20	52.40

Table 2: Classification accuracies of different writers using KNN

Writers	**Accuracy in(%age)using Zoning _ KNN**	**Accuracy in(%age)using Intersection _ KNN**
Writer_1	56.57	61.71
Writer_2	44.76	45.71
Writer_3	35.23	42.86
Writer_4	25.71	31.43
Writer_5	48.57	67.14
Writer_6	40.57	38.29
Writer_7	39.29	44.29
Writer_8	80.00	81.43
Writer_9	46.67	47.62
Writer_10	61.43	61.43
Writer_11	57.14	60.00
Writer_12	40.00	37.14
Writer_13	41.43	42.86
Writer_14	37.14	40.00
Writer_15	45.71	40.00
Writer_16	60.01	60.01
Writer_17	42.86	44.57
Writer_18	85.71	82.86
Writer_19	55.71	65.71
Writer_20	64.77	62.87
Writer_21	49.29	47.14
Writer_22	72.86	77.14
Writer_23	44.76	41.9
Writer_24	34.29	35.71
Writer_25	48.57	43.43
Writer_26	38.1	36.19

Writer_27	71.43	74.29
Writer_28	0	0
Writer_29	46.86	44.57
Writer_30	26.67	28.58

Table 2 shows the classification accuracies of thirty writers based on features extraction in terms of KNN. First column indicates the total number of writers which we have taken for testing purpose. The first column shows the accuracies using K Nearest Neighbor based on Zoning feature extraction for different writers. The third column indicates the Intersection features extracted based KNN classification accuracies. Figure 4 shows the classification accuracies for KNN based feature extraction for the Open and end point intersection. The above accuracies are given for the 30 writers using K Nearest Neighbor.

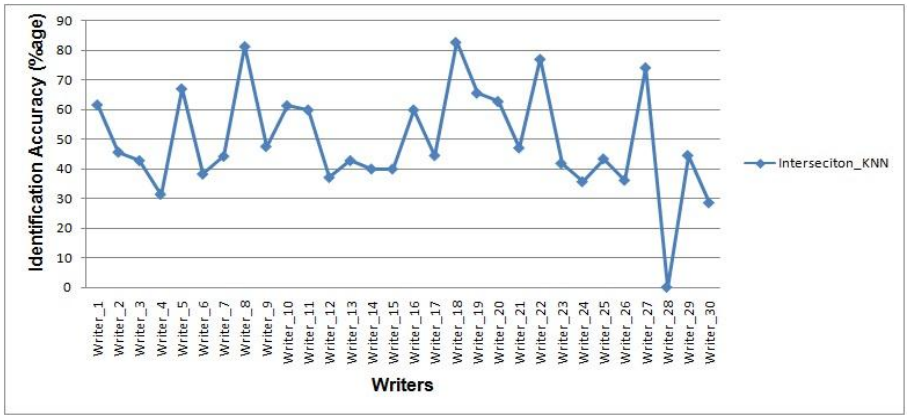

Figure 4: Accuracies using KNN for intersection feature

Figure 5 shows the classification accuracies for KNN based feature extraction for the Zoning features extraction. The above accuracies are given for the 30 writers using K Nearest Neighbor and shows the variations for different writers which are detected during testing phase.

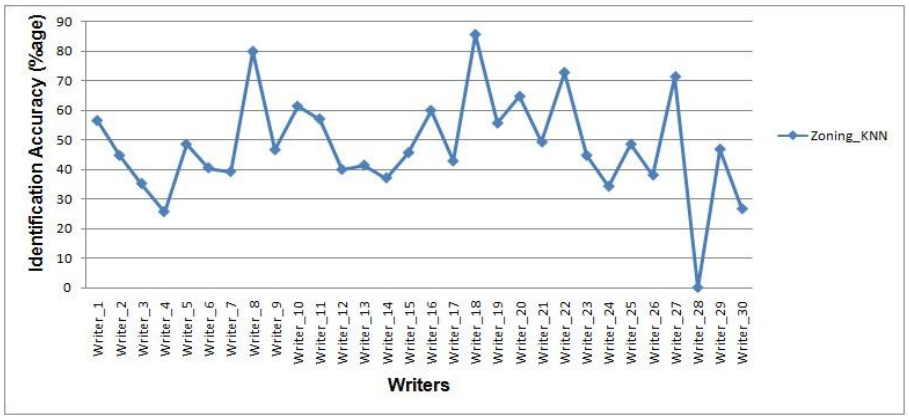

Figure 5: Accuracies using KNN for zoning features

Table 3: Classification Accuracies of different writers using MLP

Writer	Accuracy in(%age)using Zoning_MLP	Accuracy in(%age)using Intersection_MLP
Writer_1	59.43	65.14
Writer_2	55.24	47.62
Writer_3	46.67	49.53
Writer_4	42.86	25.71
Writer_5	72.86	68.57
Writer_6	46.29	42.86
Writer_7	36.43	47.14
Writer_8	84.29	91.43
Writer_9	56.2	50.48
Writer_10	65.71	60.00
Writer_11	60.00	55.00
Writer_12	42.86	51.43
Writer_13	50.00	48.57
Writer_14	45.71	52.86
Writer_15	52.86	47.14
Writer_16	67.63	63.81
Writer_17	49.71	44.57
Writer_18	80.00	75.71
Writer_19	68.57	70.00
Writer_20	69.53	68.57
Writer_21	60.00	52.86
Writer_22	81.43	74.29
Writer_23	49.53	43.81
Writer_24	46.43	37.86
Writer_25	52.57	43.43
Writer_26	46.66	41.91
Writer_27	78.57	78.57
Writer_28	0	0
Writer_29	56.57	53.71
Writer_30	35.24	31.43

Table 3 shows the classification accuracies of thirty writers based on features extraction in terms of MLP. First column indicates the total number of writers which we have taken for testing purpose. The first column shows the accuracies using MLP based on Zoning feature extraction for different writers. The third column indicates the intersection features extracted based MLP classification accuracies. Figure 6 shows the classification accuracies for MLP based feature extraction for the Open and end point intersection features. The above accuracies are given for the 30 writers using Multi-layer perceptron learning model classification.

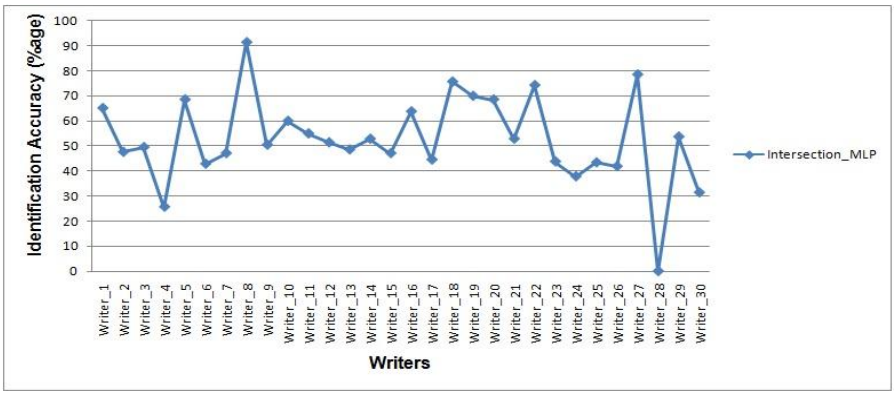

Figure 6: Accuracies using MLP for open and end point intersection

Figure 7 shows the classification accuracies for MLP based feature extraction for the zoning. The above accuracies are given for the 30 writers using Multi-layer perceptron model classification.

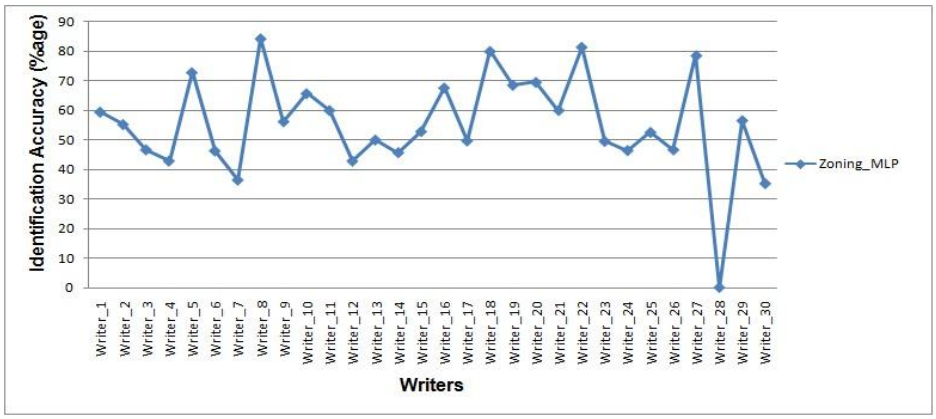

Figure 7: Accuracies using MLP for zoning features.

Figure 8 shows the classification approach based on feature extraction and shows that the Multi layer perceptron model is performing better feature extraction process than KNN

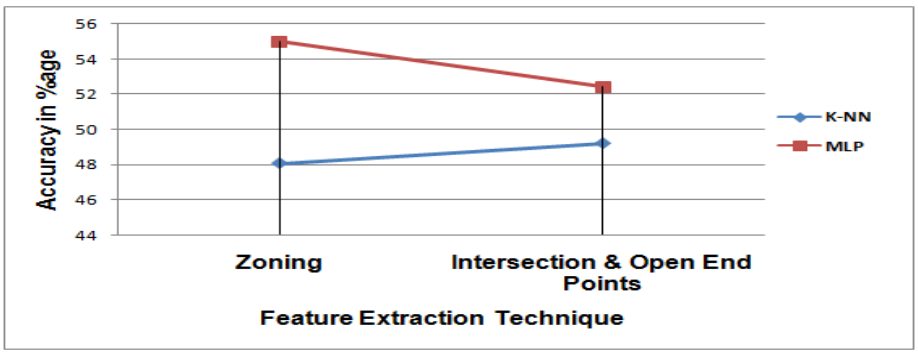

Figure 8: Accuracy based on feature extraction.

Figure 9 shows the classification between two classifiers based on zoning and intersection and open end points features and shows that MLP is performing better with accuracy of 55 % recognition rate then KNN of 53% recognition rate.

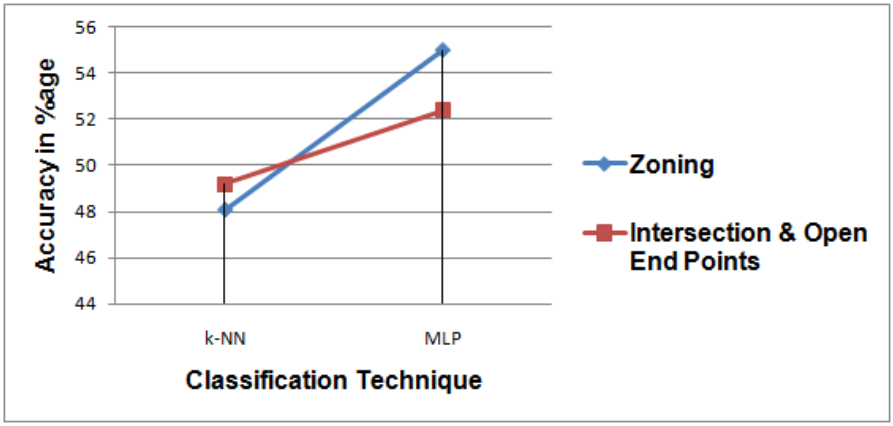

Figure 9: Accuracy based on classifiers

V. CONCLUSION

Offline handwritten character recognition involves the development of such a system which can generate description of handwritten characters from the scanned document image. This is a very challenging task due to many problems like recognizing of handwriting of different individuals. Every person has its own style of writing, so such a system must be able to recognize characters in each and every style of a writer. So for successful solution of this, one must concentrate on feature extraction step before recognition as the accuracy of recognition highly depends on feature extraction.

In this paper, we have proposed a robust feature extraction and classification approach for Gurumukhi characters has some complexities like connected components, overlapped and skewed text which makes the recognition very difficult in handwritten Gurmukhi Script. From the results it is clear that the proposed work is very useful for the identification of writer. Since the approach proposed in this research only considers the word spotting within the different writing style from different type of writers. To make it more practical, in future one has to plan to introduce learning mechanism into the approach in order to cope with more different writing styles and increase more different dataset. More dataset from more writers has also been introduced in future

REFERENCES

[1] S. Dhaval, J. Zhou, J. Waggoner and S. Wang. "Handwritten text segmentation using average longest path algorithm."In Applications of

Computer Vision (WACV), 2013 IEEE Workshop on, pp. 505-512. IEEE, (2013).

[2] A. Chandranath and B. B. Chaudhuri. "Writer Identification from offline isolated Bangla characters and numerals." In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pp. 486-490. IEEE, (2015).

[3] M. Ravi, Nayana N. Shetty, and B. P. Pragathi. "Text line segmentation of handwritten documents using clustering method based on thresholding approach." In International Journal of Computer Applications (0975–8878) on National Conference on Advanced Computing and Communications-NCACC, pp. 9-12. (2012).

[4] Y. Peng, J. Kumar, Le Kang, and D. Doermann. "Real-time no-reference image quality assessment based on filter learning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 987-994. (2013).

[5] N.K. Garg, L. Kaur, and M.K.Jindal, "A New Method for Line Segmentation of Handwritten Hindi Text," Seventh International Conference on Information Technology, (2010).

[6] G. Rahul, and N.K. Garg. "Problems and Review of Line Segmentation of Handwritten Text Document."International Journal 4, no. 4, (2014).

[7] A. Nicolaou, and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines," 10th International Conference on Document Analysis and Recognition, IEEE Computer society, pp. 626-630, (2009).

[8] G. Utpal , and B.B. Chaudhuri. "Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multi factorial analysis." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), no. 4, pp.449-459, (2002).

[9] A. Zahour, B. Taconet, L.L.Sulem, and W. Boussellaa, "Overlapping and multi-touching text line segmentation by Block Covering analysis," Pattern Analysis and Applications, Vol. 12, pp. 335-351, (2008).

[10] V. Papavassiliou, T. Stafylakis, V. Katsouros, and G. Carayannis, "Handwritten document image segmentation into text lines and words," Pattern Recognition, Vol. 43, pp. 369 – 377, (2010).