Recognition of Offline Handwritten Hindi Text Using SVM

Naresh Kumar Garg *naresh2834@rediffmail.com*
GZSPTU Campus, CSE Department
Bathinda-151001, India

Dr. Lakhwinder Kaur *mahal2k8@yahoo.com*
UCOE, Punjabi University,
Patiala, India

Dr. Manish Jindal *manishphd@rediffmail.com*
Punjab University Regional Centre,
Muktsar, India

Abstract

Handwritten Hindi text recognition is emerging areas of research in the field of optical character recognition. In this paper, a segmentation based approach is used to recognize the text. The offline handwritten text is segmented into lines, lines into words and words into character for recognition. Shape features are extracted from the characters and fed into SVM classifier for recognition. The results obtained with the proposed feature set using SVM classifier is very challenging.

Keywords: Handwritten Hindi Text, Segmentation, Shape Based Features, Recognition Rate, SVM Classifier.

1. INTRODUCTION

Devanagari is the script for writing Hindi language. Hindi is the official language of India. Offline handwritten Hindi text recognition is need of the hour due to large number of application of Hindi OCR. Development of handwritten OCR is very difficult due to different writing styles of the individuals. The techniques developed for recognition of printed characters can not be directly applied on Handwritten text. Due to large number of characters and presence of half characters makes the recognition process even more complex.

There are mainly two approaches for recognition of text- Holistic approach and segmentation based approach. Due to different writing styles of writers and various shapes of characters it is very difficult to use the holistic approach. We have used the segmentation based approach to develop the recognition system for handwritten Hindi text.

Further the paper is divided into following sections- section 2 discussed the related work, section 3 explains the database taken for experimental work, section 4 is about proposed technique used for the recognition of handwritten Hindi text, section 5 discusses the results and last section is about future scope. References are given at the end of the paper.

2. PRELIMINARIES

A lot of work has been done in the past on recognition of printed Hindi text and Hindi numeral recognition. A few research reports are available in the field of handwritten text recognition. Most of the work done in handwritten Hindi text recognition is on recognition of isolated characters. To the best of author's knowledge, no commercial OCR for handwritten Hindi text is available, yet.

A good survey about OCR is given in [1]. The performance of any classifier depends upon the quality of features fed into it. A very good survey about recognition of Devanagari script is given in [2]. It is mentioned in this paper that a lot of research has been done in the past in the recognition of printed text and isolated characters of handwritten Devanagari text, but only few research reports are available on recognition of handwritten text. Work on recognition of printed devanagar text is explained by veena bansal in [3].

A good survey about feature extraction is given in [4]. Trier et al. [5] present an interesting survey of feature extraction method for off-line recognition of segmented characters. The authors describe important aspects that must be considered before selecting a specific feature extraction method.

To the best of author's knowledge, no commercial OCR for handwritten Hindi text is available, yet. The structural and statistical features are very useful for character recognition [6].

In [7], Hanmandlu et al. had used Fuzzy model based techniques for recognition of Handwritten Hindi Characters and the recognition rate of 90.65% was reported at character level.

In [8], Kumar and Singh had used Zernike moments for recognition of Devnagari handwritten characters and reported recognition rate of 80%. Shaw et al.[9] worked on recognition of handwritten devnagari words using segmentation approach.

The work on line segmentation, consonant segmentation, upper modifier segmentation and lower modifier segmentation in Handwritten Hindi text were explained by us in [10, 11]. The algorithm for segmentation of Half characters in handwritten Hindi text is explained in [12]. We have explained a method based on structural features for segmentation of half characters in handwritten Hindi text. Recognition of non compound handwritten Devanagari characters using MLP and minimum edit distance is explained in [13].

3. DATABASE

All experiments were conducted on database constructed by taking handwritten data from fifteen writers. Documents are scanned at 300 dpi. The handwritten documents were reduced in size in paint to 35% to increase the speed of execution. The percentage of stretching of the document in horizontal and vertical direction was same. The sample database is shown in figure 1.

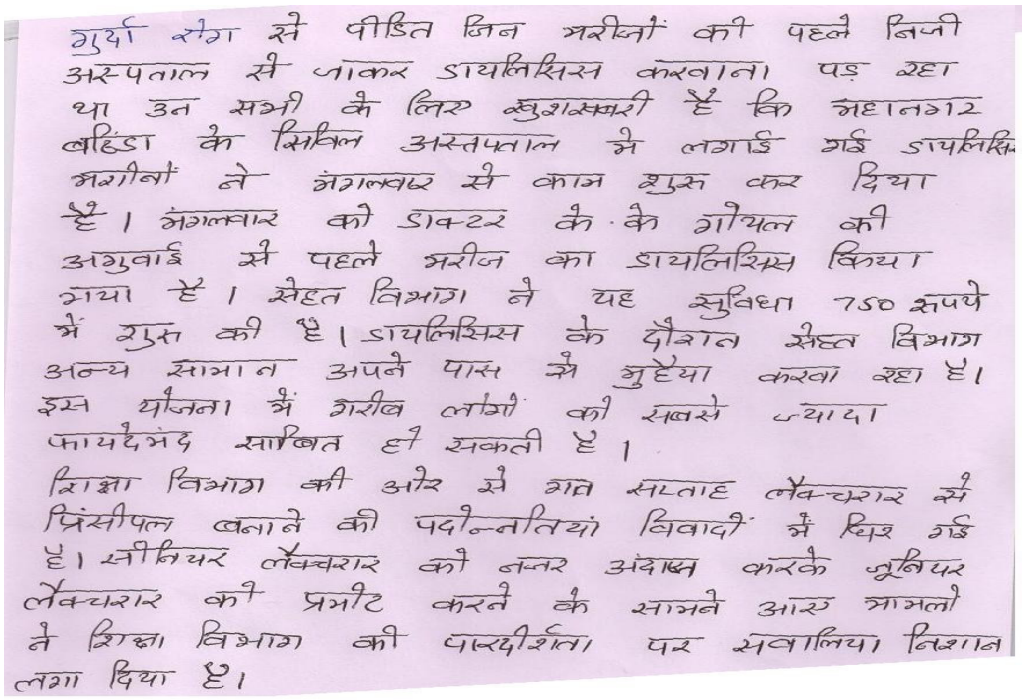
गुर्दा रोग से पीड़ित जिन मरीजों को पहले निजी
अस्पताल से जाकर डायलिसिस करवाना पड़ रहा
था उन सभी के लिए खुशखबरी है कि महानगर
बठिंडा के सिविल अस्पताल में लगाई गई डायलिसि
मशीनों ने मंगलवार से काम शुरू कर दिया
है। मंगलवार को डाक्टर के.के गोयल की
अगुवाई से पहले मरीज का डायलिसिस किया
गया है। सेहत विभाग ने यह सुविधा 750 रुपये
में शुरू की है। डायलिसिस के दौरान सेहत विभाग
अन्य सामान अपने पास से मुहैया करवा रहा है।
इस योजना में गरीब लोगों को सबसे ज्यादा
फायदेमंद साबित हो सकती है।
शिक्षा विभाग की ओर से गत सप्ताह लैक्चरार से
प्रिंसीपल बनाने की पदोन्नतियां विवादों में घिर गई
है। सीनियर लैक्चरार को नजर अंदाज करके जूनियर
लैक्चरार को प्रमोट करने के सामने आए मामलों
ने शिक्षा विभाग की पारदर्शिता पर सवालिया निशान
लगा दिया है।

FIGURE 1: Sample Database.

4. PROPOSED TECHNIQUE

Handwritten Hindi text written by different persons was scanned and binarized in Matlab. Segmentation of the text was performed in the following sequence:-

1. Text was segmented into lines.
2. Lines were segmented into words.
3. Upper modifiers were segmented from words.
4. Lower modifiers were segmented from words.
5. Consonants, half characters, matras and joint characters were segmented from words.

The techniques used for segmentation was explained in [6][7]. The strip wise vertical projection method was used for line segmentation. Word segmentation was done using vertical projection method. For character segmentation after upper and lower modifier removal from the word, a header line was detected again for each word and then vertical projection along with other constraints for joining characters were used for segmentation. Segmentation of text is very tedious task. The segmentation error propagates to recognition and reduces the recognition rate. Holistic approach was not used due to heavy character set and large number of compound characters available in handwritten Hindi text.

After segmentation, feature extraction was another tedious task performed on each character. The recognition rate of characters mainly depends upon the correctness of the features used for recognition. The efforts were made on the correctness of the features. The shape based features were extracted by applying many heuristics depending upon the shape of the character for each feature. The programming was done to extract each feature by applying many heuristics to make the feature unique for each character.

Total 59 features are selected to make a unique feature set for recognition of handwritten Hindi text. After carefully analyzing the characters set of Hindi language, different features are selected. Feature set include bars (End bar, Middle bar), end points, loops, crossings, presence of

particular horizontal and vertical lines, groves, curves and projection profiles in front, back, bottom and top of the character.

Many heuristics are applied in the extraction of each feature. The heuristics are applied to differentiate similar shaped characters. Some of the similar shaped characters are given in table 2.

Total 41 characters are considered for character recognition. These are most commonly used characters of handwritten Hindi text shown in figure 2.

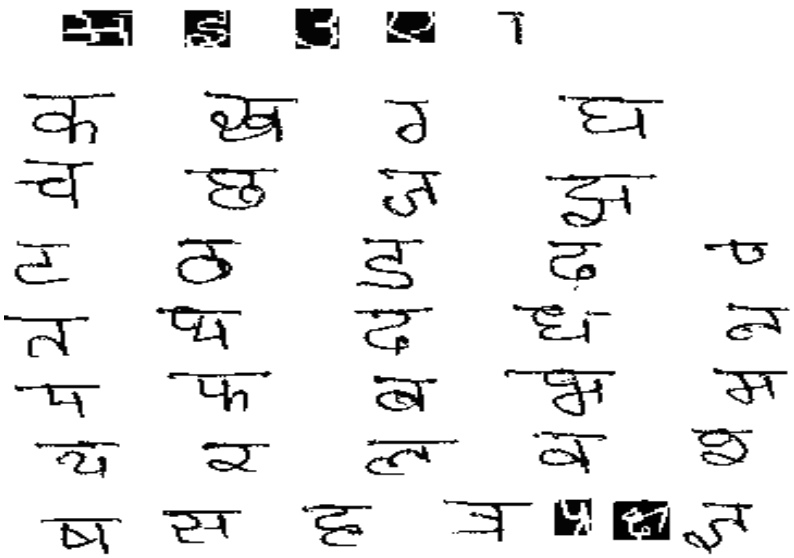

FIGURE 2: Most Commonly Used Characters.

5. RESULTS

The results obtained with shape based features and SVM classifiers are given in table 1.

Ten samples of each character are used for training the classifier. For 41 characters 410 samples are used for training purpose. All the other characters are used for testing purpose.

TABLE 1: Recognition Accuracy of Characters.

No of characters	% of characters correctly recognized (including segmentation errors)	% Accuracy of Characters recognized from correctly segmented characters
2016	76.4	89.6

The errors in segmentation propagate to recognition. The overall recognition rate is less due to segmentation errors. The recognition rate obtained from correctly segmented characters is 89.6%, which is very promising. Some of the similar shaped characters which create confusion during recognition are given in table 2.

Till now most of the work is done on recognition of isolated characters. The feature set for isolated characters can not be directly applied on the handwritten text.

TABLE 2: Similar Shaped Characters.

S No.	Character	Confused with
1.	p	Jai
2	k	f
3	l	t
4	r	sh
5	r	tt
6	a	m
7	adh	e
8	th	dh
9	s	m
10	j	n
11	ch	b
12	d	b
13	s	kh

Similar characters like r ‘ ’ , g ‘ ’ and sh ‘ ’ are very much confusing and difficult to recognize. They can be recognized with the help of complete word only. Also characters ch ‘ ’ , jai ‘ ’ and p ‘ ’ are very much confusing due to different writing style used by the

different writer's. Characters m '  ' and bhh '  ' are very similar in shape. If the upper left loop of character 'bhh' is very small and merges with the character than it looks like character 'm'. Shapes of these characters are very similar and minor differences in shapes are difficult to detect even with human eye. These types of problems can be solved during post processing stage.

The obtained results can not be compared with the literature work because most of the work available in literature is on recognition of isolated characters. The results of recognition of handwritten text can not be compared with the results of recognition of isolated characters due to non availability of standard database for handwritten Hindi text. The results obtained in our work are still comparable with results of recognition of isolated handwritten Hindi characters.

6. DISCUSSION AND FUTURE SCOPE

From the results it is clear that shape based features and SVM classifier are very useful to develop an OCR for handwritten Hindi text. The segmentation errors affect the recognition rate. The similar shaped characters creates problem in recognition. The post processing can reduce the errors in recognition that occur due to similar shaped characters and improve the recognition rate. The efforts can be made in the future in the following direction:

1) Segmentation techniques can be improved to reduce the segmentation errors and recognition rate.

2) More features can be added in the feature set to differentiate similar shaped characters.

3) Other classifiers can be tried with shape based features.

7. REFERENCES

[1] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR Research and development", Proceedings of the IEEE, Vol. 80, No. 7, pp. 1029-1058, 1992.

[2] R. Jayadevan, S.R. Kohle, P.M. Patil and U. Pal, "Offline recognition of Devanagari script: A survey." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions,* Vol.41, No. 6, pp:782-796, 2011.

[3] V. Bansal, "Integrating knowledge sources in Devanagari text recognition", Ph.D. thesis, IIT Kanpur, INDIA, 1999.

[4] N. Arica and F. T. Y. Vural, "An overview of character recognition focused on offline handwriting", *IEEE Trans. On Systems, Man, and Cybernetics – Part C: Applications and Reviews,* Vol. 31, No. 2, pp. 216-233, 2001.

[5] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition: A survey", *Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.

[6] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier, and C. Olivier, "A structural/statistical feature based vector for handwritten character recognition", *Pattern Recognition Letters*, Vol. 19, No.7, pp:629-641, 1998.

[7] M. Hanmandlu, O.V. Ramana Murthy and Vamsi Krishna Madasu, "Fuzzy Model based recognition of handwritten Hindi characters", Digital Image Computing Techniques and Applications, pp:454-461, 2007.

[8] S. Kumar and C. Singh,"A Study of Zernike Moments and its use in Devnagari Handwritten Character Recognition", International Conference on Cognition and Recognition, pp:514-520, 2005.

[9] B. Shaw, S. K. Parui, and M. Shridhar, "A segmentation based approach to offline handwritten Devanagari word recognition," Proceedings of IEEE International Conference on Information Technology, pp: 256–257, 2008.

[10] N. K. Garg, L. Kaur and M. K. Jindal, "Segmentation of Handwritten Hindi Text", *International Journal of Computer Applications (IJCA)*, Vol. 1, No. 4, pp.22-26, 2010.

[11] N. K. Garg, L. Kaur and M. K. Jindal, "A new method for line segmentation of Handwritten Hindi Text", Proceedings of the 7th International IEEE Conference on Information Technology: New Generations (ITNG), pp.392-397, 2010.

[12] N. K. Garg, L. Kaur and M. K. Jindal, "The Segmentation of Half Characters in Handwritten Hindi Text", Proceedings of the ICISIL 2011, Springer, pp.48-53, 2011.

[13] Sandhya Arora et al. "Recognition of Non-Compound Handwritten Devanagari Characters using a Combination of MLP and Minimum Edit Distance", International Journal of Computer Science and Security (IJCSS), Vol. 4, Issue 1, pp. 107-120, 2010.