International Journal of Engineering Research and General Science Volume 3, Issue 6, November-December, 2015
ISSN 2091-2730

AN OVERVIEW OF HANDWRITTEN GURMUKHI CHARACTER RECOGNITION

Ambuj, Nishant Anand
M.tech. Student, CBS Group of Institutions, Jhajjar, Haryana;ambuj965@gmail.com;9999429018

Abstract- In this thesis work we have suggested offline recognition of isolated handwritten characters of Gurmukhi script. We have also prolonged the work by applying the same methodology to recognize handwritten Gurmukhi numerals. In our work we have considered 35 fundamental characters of Gurmukhi script all assumed to be isolated and bearing header lines on top to recognize. In numerals, handwritten elements of ten digits from different writers are considered.

Gurmukhi numerals using three characteristic sets and three classifiers. Among three characteristic sets, first characteristic set is comprised of distance profiles having 128 characteristic . Second characteristic set is comprised of different types of projection histograms having 190 characteristic. Third characteristic set is comprised of zonal density and Background Directional Distribution forming 144 features. The three classifiers used are:- SVM , PNN & K-NN The SVM classifier is used with Radial Basis Function kernel. We have observed the 5-fold cross verification accuracy in the case of each characteristic set and classifier. We have obtained the optimized result with each combination of characteristic set and classifier by adjusting the different limits. The results are compared and trends of result in each combination of characteristic set and classifier with varying limits is also discussed. With PNN and K-NN the highest results are obtained using third characteristic set as 98.33% and 98.51% respectively while with SVM the highest result is obtained using second characteristic set as 99.2%. The results with SVM for all characteristic sets are higher than the output with PNN & K-NN.

Keywords— Handwritten Gurumukhi Character Recognition, Diagonal characteristic, SVM classifier with RBF kernel.

I. INTRODUCTION

The storage of scanned data have to be cumbersome in size and many processing applications as searching for a content, checking, maintenance are either hard or impossible. Such data require human beings to process them manually, for example, postman‟s manual action toward acceptance and sorting of postal addresses and zip code. Optical character recognition translates such scanned images of printed, typewritten or handwritten documents into machine predeterminded text. This translated machine predeterminded text can be easily edited, searched and can be processed in many other ways according to needs. It also requires small size for storage in comparison to scanned documents. Optical character recognition helps humans ease and reduce their jobs of manually handling and processing of documents. Computerized processing to identify individual character is required to change scanned data into machine encoded form.

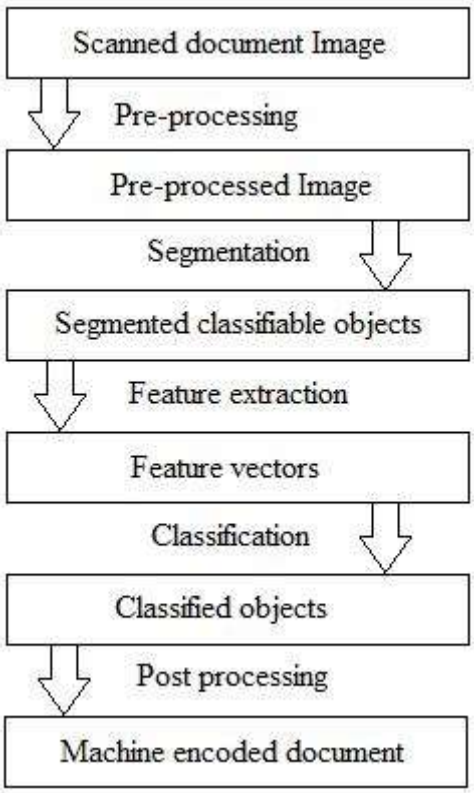
In comparison to languages like English and Japanese, the recognition research on Indian languages and scripts is relatively lagging behind. Among available research work on Indian languages, most of the work is on Devnagari and Bangla script. The work on other Indian languages is in fewer amounts

1.1 CLASSIFICATION OF CHARACTER RECOGNITION SYSTEM

Earlier OCR was widely used to identify printed or typewritten data. But recently, there is an increasing trend to identify handwritten data. The identification of handwritten data is more difficult in comparison to identification of printed data. It is because handwritten data contains unconstrained variations of written styles by different writers even distinct writing styles of same writer on different times and moods. Sometimes, even a writer can‟t identify his/her own handwriting, so it is very difficult to gain acceptable identification accuracy involving all possible changes of handwritten samples.

1.2 **RECOGNITION SCHEME**

OCR involves many steps to completely identify and produce machine encoded text. These stages are termed as: Pre-refinement, Segmentation, characteristics extraction, Classification and Post processing. The composition of these stages is shown in figure 1.2 and these stages are listed below with brief description. These stages, except post processing are elaborated in next section of overview of OCR stages .



Pre-processing

The pre-processing is a series of operations performed on the scanned input image. It essentially improves the image rendering it suitable for segment formation. The following steps are used to create data-set. Entire work has been done in Mat lab.

Handwritten sample is scanned in RGB format.

RGB image is converted into gray-scale image.

Gray-scale image is converted into binary image by using a suitable threshold value by Otsu's method.

Other preprocessing method like median filtration, dilation, some morphological operations are applied to join separate pixels, to remove isolated pixels, to set neighbor pixel values in majority and to remove the spur pixels.

Segmentation

In the segmentation stage, an image of sequence of characters is decomposed into sub-images of individual feature [14]. In the proposed system, the pre-processed input image is segmented into isolated features by assigning a number to each feature using a labeling process. This labeling provides data about number of features in the image. Each individual feature is uniformly re sized into 10^2 x 10^2 pixels for extracting characteristics .

Feature Extraction

Diagonal characteristics are very important characteristics in order to achieve higher recognition accuracy and reducing miscalculation. These characteristics are extracted from the pixels of each zone by moving along its slant as shown in Fig 2. Following procedure describes the computation of Diagonal characteristics for each character image of size 10^2*10^2 pixels having 10*10 zones and thus each zone having 10*10 pixel size. Each of these zones are having 19 slants. The number of foreground pixels along each diagonal are summed up to get 19 characteristics from each zone, then these characteristics for each zone are averaged to extract a single characteristics from each zone.

ISSUES OF HANDWRITTEN CHARACTER RECOGNITION

To identify handwritten documents, either online or offline, the character identification is much affected by style changes of handwriting by different writers and even different styles of same writer on distinct times. alteration and noise incorporated while digitization is also a major issue in character identification that affects the recognition accuracy negatively. The many character identification issues regarding handwritten feature identification are listed below:

Handwriting Style Variations

Constrained and Unconstrained Handwriting

Writer Dependent or Independent Recognition

Personal and Situational Aspects

PROPOSED WORK

In our proposed work we have identified isolated handwritten characters of " Gurmukhi " script. In our work we have used some statistical characteristics like projection histograms, distance profiles and zonal density; and one direction characteristics Background Directional Distribution in different combinations to construct different characteristics vectors. We have presented detailed comparative analysis of the recognition with these characteristics vectors using three types of classifiers- Support Vector Machines , KNN and PNN . The best result obtained is 95.07% with zonal density and BDD characteristics in combination and SVM classifier for character recognition. We have also recognized "Gurmukhi" numerals with different features and best recognition rate obtained is 99.2%. The detailed description of the proposed work for character and numeral recognition is given in chapter 3 and chapter 4 respectively.

REFERENCES:

1. G.G. Rajput, S.M. Mali, “Fourier Descriptor Based Isolated Marathi Handwritten Numeral Recognition”, *International Journal of Computer Applications*, Vol. 3, No. 4, pp. 9-13, June 2010

2. Chih-Chung Chang and Christian Lindig, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

3. Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, “A Practical Guide to Support Vector Classification”, [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

4. Sarbajit Pal, Jhimli Mitra, Soumya Ghose, Paromita Banerjee, "A Projection Based Statistical Approach for Handwritten Character Recognition," *in Proceedings of International Conference on Computational Intelligence and Multimedia Applications*, Vol. 2, pp.404-408, 2007

5. N. Araki, M. Okuzaki, Y. Konishi, H. Ishigaki, "A Statistical Approach for Handwritten Character Recognition Using Bayesian Filter" *3rd International*

Conference on Innovative Computing Information and Control (ICICIC), pp.194-198, June 2008

6. Wang Jin, Tang Bin-bin, Christian Lindig, Piao Chang-hao, Lei Gai-hui, "Statistical method-based evolvable character recognition system",*IEEE International Symposium on Industrial Electronics (ISIE)*, pp. 804-808, July 2009
7. Apurva A. Desai, "Gujarati Handwritten Numeral Optical Character Reorganization through Neural Network", *Pattern Recognition*, Vol. 43, Issue 7, pp. 2582-89, July 2010
8. D. Singh, S.K. Singh, M. Dutta, "Handwritten Characterr Recognition Using Twelve Directional Feature Input and Neural Network", *International Journal of Computer Applications*, Vol. 1 No.2, pp. 86-87,March 2010
9. Gurpreet S Lehal, C. Singh, "A Gurmukhi Script Recognition System", *Proceedings of 15th International Conference on Pattern Recognition*, Vol. 2, pp. 557-560, 2000
10. Gurpreet S Lehal, C. Singh, "A Complete Machine printed Gurmukhi OCR", *Vivek*, 2006
11. Gurpreet S Lehal, C. Singh, "Feature Extraction and Classification for OCR of Gurmukhi Script", *Vivek* Vol. 13, pp. 2-12, 1999
12. Gurpreet S Lehal, C. Singh, "A Post Processor for Gurmukhi OCR", *Sadhana*, Vol. 25, Part 1, pp. 99-111, 2002
13. D. Sharma, Gurpreet S Lehal, "An Iterative Algorithm for segmentation of Isolated Handwritten Words in Gurmukhi Script", *The 18th International Conference on Pattern Recognition (ICPR)*, Vol. 2, pp. 1022-1025, 2006
14. Vijay. Goyal, Gurpreet S Lehal, "Comparative Study of Hindi and Punjabi Language Scripts", *Nepalese Linguistics*, Vol. 23, pp. 67-82, 2008
15. G.S. Lehal, Nivedan Bhatt, "A Recognition System for Devnagari and English Handwritten Numerals", *Proc. ICMI, Springer*, pp. 442-449, 2000
16. D. Sharma, Preety Kathuria, "Digit Extraction and Recognition from Machine Printed Gurmukhi Documents", *Proceedings of the International Workshop on Multilingual OCR MORC Spain*, 2009
17. Anuj Sharma, Rajesh Kumar, R. K. Sharma, "Online Handwritten Gurmukhi Character Recognition Using Elastic Matching",*Conference onImage and Signal Processing (CISP)*, Vol.2, pp.391-396, May 2008
18. Anuj Sharma, R.K. Sharma, Rajesh Kumar, "Online Handwritten Gurmukhi Character Recognition", Ph.D. Thesis, Thapar University, 2009 [Online]. Available: http://dspace.thapar.edu:8080/dspace/bitstream/10266/1057/3/Thesis_AnujSharma_SMCA_9 041451.pdf