

AN OVERVIEW OF HANDWRITTEN GURMUKHI CHARACTER RECOGNITION

Ambuj, Nishant Anand

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

Abstract :-This paper represent Handwritten Gurmukhi feature Recognition system using some statistical characteristics like zone density, projection histograms , 8 directional zone density characteristics in combination with some geometric characteristics like area, perimeter, eccentricity, etc. The image text is first pre-processed by using many methods like binarization, morphological operations (erosion and dilation) applied to delete noise and then segmented into isolated characters. The highest accuracy obtained by using these characteristics and back propagation classifier is 98%.

Keywords: Handwritten character recognition, Feature extraction, Diagonal features, Intersection and open end points features, SVM.

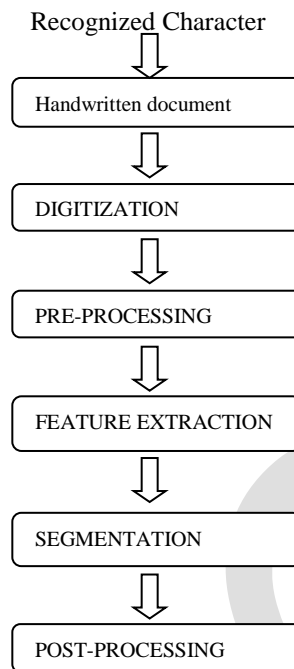
Introduction: Nowadays, computers have a great effect on us and we process almost all the important job of our lives electronically. Keeping in mind the usage of computers these days, we must to develop efficient, easy and fast ways for information transfer within human beings and computers. DAR systems play a major role in data transfer between human beings and computers.OCR system is an essential part of a DAR system. OCR systems have been developed to identified printed texts as well as handwritten texts. Handwritten text recognition systems essentially given an link for improving communication between users and computers. These empower computers to read and process handwritten texts. These systems shall further help significantly in bridging the gap between man and machine. Although, many researchers have worked to identify the features of Indian scripts, the problem of data exchanging within people and machines is still a problem in these scripts. The work carried out in this thesis focused the problem of handwritten characteristics identification for Gurmukhi script. Gurmukhi script is used to write Punjabi language. This language is one from the official languages of India. Gurmukhi script is the tenth most widely used script in the world.

The proposed recognition system

. A typical offline handwritten feature identification system include activities, namely, digitization, pre-processing, segmentation, character extraction, classification and post processing. The series of these activities is shown in Figure 1.2.

1.2.1 **Digitization** Converting a paper based handwritten text into an electronic form is referred as digitization. The electronic conversion is carried out by forming the bitmap image after checking the text. Digitization yields the digital picture which is then fed to the pre-action phase.

Figure 1.2: Block diagram of offline HCR system



1.2.2 PRE-PROCESSING

PRE-PROCESSING is the initial stage of character recognition. It encompasses skew detection and correction, skeletonize, and noise reduction/removal. Skewness means the tilt of the bit mapped picture of the scanned text. It usually surfaces when the text is wrong fed to the scanner. Skeletonization is applied in way to decrease the line width of the text from many pixels to a single pixel. Noise elimination is carried out to delete those unnecessary bits that do not play a substantial role in the document. After pre-processing, we have the digital text that is inputted to the Segmentation phase.

1.2.3 Segmentation In feature recognition, the process of segmentation plays a very important role. division is used to break the text into lines, words and akhars. For the work of division, an algorithm is used to find the division points in a handwritten data. Gurmukhi script data can be segmented into paragraphs, lines, words and characters. The challenge of a division method lies in the detection of the best division point for lines, words and characters in isolation. wrong segmentation can lead to the wrong recognition. Segmentation of a handwritten text is a difficult task owing to a variety of writing styles.

1.2.4 Feature extraction

Feature extraction is an important task of the recognition process which is used to calculate the applicable shape contained in the character. In the feature extraction stage, one can extract the characteristics of the character. The performance of the identification system depends on characteristics which are being extracted. In OCR applications, it is important to extract those characteristics that will make easy the system which can differentiate within all the character classes that exist. The extracted characteristics may be structural or statistical based. Structural characteristics depict a pattern in ways of its topology and geometry by provide it local and global Offline Handwritten Gurmukhi Script identification 8 properties. Characteristics of the distribution of pixel values on the bitmap image are taking as statistical characteristics.

1.2.5 Classification

Classification stage is the stage of an OCR system wherein one makes the decisions. It uses the characteristics extracted in the characteristics extraction stage, for making class membership in the identification system. The preliminary aim of the classification stage of an OCR system is to develop a constraint that can help to reduce the classifications issue relevant to characteristics extraction. Effectiveness of any character identification system is highly dependent on the capability of identify the rare characteristics of a character and the capability of the classifier to relate characteristics of a character to its class. different classification methods, namely, k-NN, HMM, SVM and Bayesian etc. exist in literature.

1.2.6 Post-processing

OCR results, in general, contain mistake since classification phase does not ever give one hundred percent correct results. To further refine the results of classification, post action applied. There are two most commonly used post-processing methods for mistake correction. These are (i) dictionary lookup and (ii) statistical way .

Gurmukhi script is the script used for writing *Punjabi* language and is derived from the old *Punjabi* term “*Guramukhi*”, which means “from the mouth of the Guru”. *Gurmukhi* script is the 10th most widely used script in the world [Source: Growth of Scheduled Languages: 1971, 1981, 1991, 2001 and 2011, Census of India, Ministry of Home Affairs, Government of India]. The writing style of the *Gurmukhi* script is from top to bottom and left to right. *Gurmukhi* script has three vowel bearers, thirty two consonants, six additional consonants, nine vowel modifiers, three auxiliary signs, and three half characters. In *Gurmukhi* script, there is no case sensitivity.

Objectives of this work

The objectives of the proposed reading are outlined as :

1. To reading and implement existing algorithms and ways like Bengal Engineering and Science University Shibpur algorithm, Democritus University of Thrace- Adaptive Run Length Smearing Algorithm , Institute of Language and Speech Processing-Line and Word division algorithm, and University of Athens-Hough conversion for line and word division of a handwritten document. A new algorithm will be proposed for division of lines for the offline handwritten Gurmukhi script data.
2. To explore existing characteristics (structural and statistical) and to suggest innovative character for offline handwritten Gurmukhi script identification.
3. To explore HMM, ANN, k-NN and SVM classifiers and to suggest efficient combinations of these in the form of many classifiers. In order to achieve these aims, a detailed survey of literature on different steps of a HCR system has been done. Statistical characteristic have been used for constructing a feature vector for identification purpose. many classifiers such as k-NN, HMM, SVM, Bayesian and MLP have been employed for identification purpose. Combinations of these classifiers have also been used for offline handwritten Gurmukhi character indentification in the thing carried out for this thesis.

Assumptions

We have considered this constraints while performing experiments in this thesis.

1. The handwritten documents have been checked at 300 dpi resolution.
2. The information considered in this work is free from noise.
3. The data considered in this work does not contain any non-text items such as images, characteristics etc.

4) Recognition of equations, line drawings, and gestural symbols.

5) Noisy tablet data.

REFERENCES:

- [1] Lorigo, L. M., and Govindaraju, V.: Offline Arabic handwriting recognition: a survey. IEEE Transactions on PAMI, 28, 5 (2006) 712-724
- [2] Plamondon, R. and Srihari, S. N.: On-line and off-line handwritten character recognition: A comprehensive survey, IEEE Transactions on PAMI, 22, 1 (2000), 63-84
- [3] Lehal, G. S. and Singh, C.: A Gurmukhi script recognition system, In Proceedings of 15th ICPR, 2 (2000), 557-560
- [4] Wen, Y., Lu, Y. and Shi, P.: Handwritten Bangla numeral recognition system and its application to postal automation, Pattern Recognition, 40 (2007), 99-107
- [5] Swethalakshmi, H., Jayaraman, A., Chakravarthy, V. S. and Sekhar, C. C.: Online handwritten character recognition of Devanagari and Telugu characters using support vector machine, In Proceedings of 10th IWFHR, (2006), 367-372
- [6] Pal, U., Wakabayashi, T. and Kimura, F.: A system for off-line Oriya handwritten character recognition using curvature feature, In Proceedings of 10th ICIT, (2007), 227-229
- [7] Hanmandlu, M., Grover, J., Madasu, V. K. and Vasikarla, S.: Input fuzzy for the recognition of handwritten Hindi numeral, In Proceedings of ITNG, (2007), 208-213
- [8] Rajashekaradhya, S. V. and Ranjan, S. V.: Zone based Feature Extraction algorithm for Handwritten Numeral Recognition of Kannada Script, In Proceedings of IACC, (2009), 525-528
- [9] Tripathy, J.: Reconstruction of Oriya alphabets using Zernike Moments, International Journal of Computer Applications, 8, 8 (2010), 26-32
- [10] Jindal, M. K.: Degraded Text Recognition of Gurmukhi Script", PhD Thesis, Thapar University, Patiala, India, 2008