Volume 3, No. 6, June 2012

Journal of Global Research in Computer Science

RESEARCH PAPER

Available Online at www.jgrcs.info

ISSN-2229-371X

RECOGNITION OF SIGNBOARD IMAGES OF GURMUKHI

Er.Puneet kaur[*1] and Er.Balwinder Singh[2]

[*1]Assistant professor Computer science Akal degree college , Mastuana Sangrur,Punjab,India
kaurbanu@yahoo.com

[2]Assistant Professor Computer science Yadwindra Institute of engg. And tech. Talwandi Sabo,Punjab,India
puyce@yahoo.com

Abstract: Recently there is growing trend among worldwide researchers to recognize handwritten characters of many languages and scripts. Much of research work is done in English, Chinese and Japanese like languages. However, on Indian scripts, the research work is comparatively lagging. The work on other Indian scripts is in beginning stage.

In this thesis work I have proposed recognition of isolated handwritten numerals of Gurumukhi script. In numerals, handwritten samples of 10 digits from different writers are considered.

I have taken all these samples on white papers written in an isolated manner. The dataset used to recognize numerals is collected from 15 different writers each contributing 10 samples of each digit.

After scanning, in preprocessing stage, the samples are converted to gray scale images. Then gray scale image is converted into binary image. I also applied median filtration, dilation, removal of noise having less than 30 pixels, some morphological operations to bridge unconnected pixels, to remove isolated pixels, to smooth pixel boundary, and to remove spur pixels. We segmented these samples in isolated and clipped images of each character based on white spaced pixels used for separation. In this paper, I have proposed recognition of isolated handwritten numerals of Gurumukhi script. In numerals, handwritten samples of 10 digits from different writers are considered.

I have taken all these samples on white papers written in an isolated manner. The dataset used to recognize numerals is collected from 15 different writers each contributing 10 samples of each digit.

INTRODUCTION

Optical character recognition is the prominent area of research in the world. OCR is the translation of scanned images of handwritten, typewritten or printed document into machine encoded form. This machine encoded form is editable text and compact in size. Character recognition can be applied on printed, type-written or handwritten text. Character recognition for handwritten characters is more complex due to varying writing styles of people.

Further the optical character recognition can be classified as offline recognition and online recognition. The offline recognition is associated with static applications in which entire document first scanned and then processed to recognize, while the online recognition is associated with dynamic application as web application where we need recognized result simultaneously or within a fraction of time. Optical Character Recognition (OCR) is the process of converting scanned images of machine printed or handwritten text into a computer processable format. The practical importance of OCR applications, as well as the interesting nature of the OCR problem, has led to great research interest and measurable advances in this field. But these advances are limited to English, Chinese and Arabic languages [l-3] and there has been very limited reported research on OCR of the scripts of Indian languages [2].

CLASSIFICATION OF CR

Earlier OCR was widely used to recognize printed or typewritten documents. But recently, there is an increasing trend to recognize handwritten documents. The recognition of handwritten documents is more complicated in comparison to recognition of printed documents. It is because handwritten documents contains unconstrained variations of written styles by different writers even different writing styles of same writer on different times and moods. Sometimes, even a writer can‟ t recognize his/her own handwriting, so it is very difficult to gain acceptable recognition accuracy involving all possible variations of handwritten samples. Traditionally OCR is considered to recognize scanned documents in offline mode. Recently, due to increased use of handheld devices online handwritten recognition attracted attention of worldwide researchers. This online handwritten recognition aims to provide natural interface to users to type on screen by handwriting on a pad instead of by typing using keyboard.

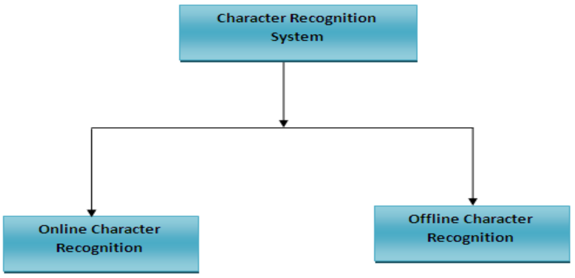

Character Recognition System
Online Character Recognition
Offline Character Recognition

Figure 2.1 shows the classification of Character Recognition

Off-line handwriting recognition refers to the process of recognizing characters in a document that have been scanned from a surface (such as a sheet of paper) and are stored digitally in grey scale format. After being stored, it is conventional to perform further processing to allow superior recognition. In case of online handwritten character recognition, the handwriting is captured and stored in digital form via different means.

BACKGROUND STUDY

For Indian languages most of research work is performed on firstly on Devnagari script and secondly on Bangla script. U. Pal and B.B. Chaudhury presented a survey on Indian Script Character Recognition [4]. This paper introduces the properties of Indian scripts and work and methodologies approached to different Indian script. They have presented the study of the work for character recognition on many Indian language scripts including Devnagari, Bangla, Tamil, Oriya, Gurumukhi, Gujarati and Kannada.

In particular to Gurumukhi script, Earliest and major contributors founded are Chandan Singh and G. S. Lehal. G. S. Lehal and Chandan Singh proposed Gurumukhi script recognition system [2]. Later they developed a complete machine printed Gurumukhi OCR system [36]. Some of their other research works related to Gurmukhi Script recognition are [37], [38] in which they proposed feature extraction, classification and post processing approaches for Gurumukhi scripts.

D. Sharma and Lehal proposed an iterative algorithm to segment isolated handwritten words in Gurumukhi script [39]. G.S. Lehal used four classifiers in serial and parallel mode and combined the results of classifiers operated in parallel mode. Combining multiple classifiers their individual weaknesses can be compensated and their strength are preserved.A comparative study of Gurumukhi and Devnagari Script is presented in [7]. In this paper closeness of Devnagari and Gurumukhi script,consonants, conjunct consonants, vowels, numerals and punctuation is discussed.

For recognition of handwritten Gurmukhi characters following approaches are found in literature survey. Naveen Garg et al. [8] have recognized offline handwritten Gurmukhi characters using neural network and obtained 83.32% average recognition accuracy.Puneet Jhajj et al. [9]used a 48×48 pixels normalized image and created 64 (8×8) zones and used zoning densities of these zones as features.

DATA SETS AND DATA COLLECTION

Data collection is the first phase in online handwritten recognition that collects the sequence of coordinate points of the moving pen. We have different phases for data collection:

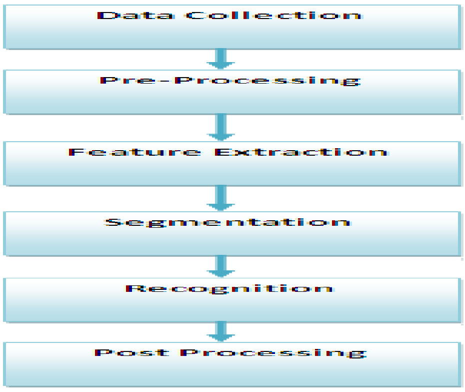


Figure 4.1 shows the phases of Handwritten Recognition.

Data sets:

The dataset of Gurumukhi numerals for our implementation is collected from 15 different persons. Each writer contributed to write 10 samples of each of numeral of 10 different Gurumukhi digits. These samples are taken on white papers written in an isolated manner. The table 1 shows some of the samples of our collected dataset. These samples are transformed in gray image. Among these samples, some distortions and irregularities are also incorporated by writers.

Table 1

GURMUKHI DIGIT ZERO	੦
GURMUKHI DIGIT ONE	੧
GURMUKHI DIGIT TWO	੨
GURMUKHI DIGIT THREE	੩
GURMUKHI DIGIT FOUR	੪
GURMUKHI DIGIT FIVE	੫
GURMUKHI DIGIT SIX	੬
GURMUKHI DIGIT SEVEN	੭
GURMUKHI DIGIT EIGHT	੮
GURMUKHI DIGIT NINE	੯

PROPOSED SYSTEM

In this paper, I am describing various methods of recognizing the Gurumukhi numerals .I have adopted here three sets of features, and SVM classifiers in the recognition process.

Optical character recognition involves many steps to completely recognize and produce machine encoded text. These phases are termed as: Pre-processing, Segmentation, Feature extraction, Classification and Post processing. The architecture of these phases is shown in figure and these phases are listed below with brief description. These phases, except post processing are elaborated in next section of overview of OCR phases.

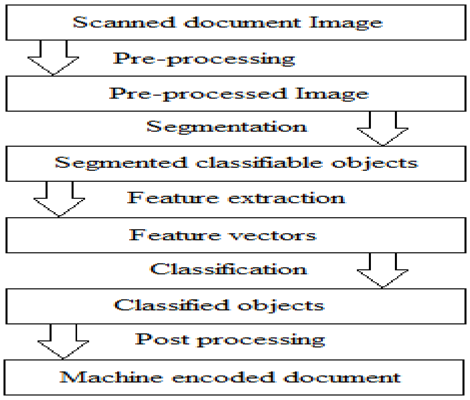


Figure 5.1 shows the various steps to follow for the character recognition system.

Pre-processing: The pre-processing phase normally includes many techniques applied for binarization, noise removal, skew detection, slant correction, normalization, contour making and skeletonization like processes to make character image easy to extract relevant features and efficient recognition.

Segmentation: Segmentation phase, which sometimes considered within pre-processing phase itself, involves the splitting the image document into classifiable module object generally isolated characters or modifiers. Generally practiced segmentations are line segmentation, word segmentation, character segmentation and horizontal segmentation to separate upper and lower modifiers particularly in context to most Indian scripts.

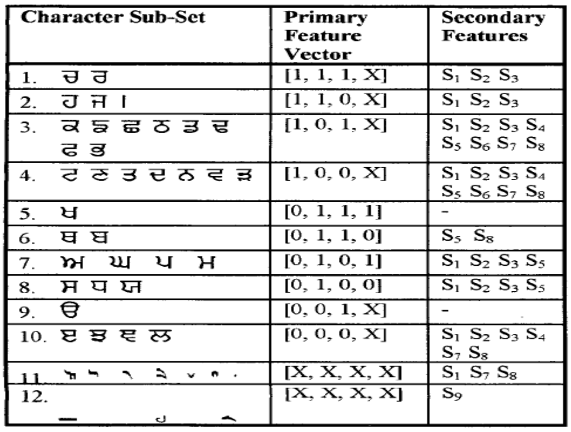
Feature Extraction: Feature extraction is used to extract relevant features for recognition of characters based on these features. First features are computed and extracted and then most relevant features are selected to construct feature vector which is used eventually for recognition. The computation of features is based on structural, statistical, directional, moment, transformation like approaches.

Classification: Each pattern having feature vector is classified in predefined classes using classifiers. Classifiers are first trained by a training set of pattern samples to prepare a model which is later used to recognize the test samples. The training data should consist of wide varieties of samples to recognize all possible samples during testing. Some examples of generally practiced classifiers are- Support Vector Machine (SVM), K- Nearest Neighbour (K-NN), Artificial Neural Network (ANN) and Probabilistic Neural Network (PNN).

Post Processing: In post processing step we bind up our work to create complete machine encoded document through the process of recognition, assigning Unicode values to characters and placing them in appropriate context to make characters, words, sentences, paragraphs and finally whole document. We also correct misclassified character based on some linguistic knowledge. We can use dictionary and other language grammatical tools to match the classification results and correct the syntax or semantic of word or sentence. By using dictionary we can restrict the possible combination of characters to form a word, or the combination of words to form a sentence. The recognition and segmentation error can be corrected using lexical information using a lexicon.

FEATURE SETS FOR CLASSIFICATION

Table 5.2 shows the Feature sets for classification.



Character Sub-Set	Primary Feature Vector	Secondary Features
1. ਚ ਰ	[1, 1, 1, X]	S_1 S_2 S_3
2. ਹ ਜ ।	[1, 1, 0, X]	S_1 S_2 S_3
3. ਕ ਙ ਛ ਠ ਡ ਢ ਫ ਭ	[1, 0, 1, X]	S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8
4. ਟ ਣ ਤ ਦ ਨ ਵ ੜ	[1, 0, 0, X]	S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8
5. ਖ	[0, 1, 1, 1]	-
6. ਥ ਬ	[0, 1, 1, 0]	S_5 S_8
7. ਅ ਘ ਪ ਮ	[0, 1, 0, 1]	S_1 S_2 S_3 S_5
8. ਸ ਧ ਯ	[0, 1, 0, 0]	S_1 S_2 S_3 S_5
9. ੳ	[0, 0, 1, X]	-
10. ੲ ਝ ਞ ਲ	[0, 0, 0, X]	S_1 S_2 S_3 S_4 S_7 S_8
11. ਿ ੀ ੇ ੈ ੋ ੌ ਂ	[X, X, X, X]	S_1 S_7 S_8
12. ੁ ੂ ੍	[X, X, X, X]	S_9

The classification stage is the main decision making stage of an OCR system. The classification stage uses the features extracted in the previous stage to identify the text segment according to preset rules. Binary classifier trees and nearest neighbor classifiers is the two most commonly used classifiers.

The binary tree classifier has the advantage of speed but is sensitive to noise. The nearest neighbour is less sensitive to noise and can easily be trained for more fonts and sizes but is computationally space and time intensive. The ***primary features*** are noise, font and size invarient and so the binary classification tree was used for them. The ***secondary*** features were found to be sensitive to font in some cases and so the nearest neighbour classifier with a variant sized vector was used.

IMPLEMENTATION TOOL

The work is done in MATLAB. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numerical computation. Using MATLAB, you can solve technical computing problems faster than with traditional programming languages, such as C, C++, and FORTRAN.

KEY FEATURES

a. High-level language for technical computing
b. Development environment for managing code, files, and data
c. Interactive tools for iterative exploration, design, and problem solving
d. Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration
e. 2-D and 3-D graphics functions for visualizing data
f. Tools for building custom graphical user interfaces

CONCLUSION

An attempt has been for the development of a system for recognition of Gurmukhi script.

OCR has been used to enter data automatically into a computer for dissemination and processing. The earliest of systems was dedicated to high volume variable data entry. The first major use of OCR was in processing petroleum credit card sales drafts. This application provides recognition of the purchaser from the imprinted credit card account number and the introduction of a transaction. The early devices were coupled with punch units which made small holes to be read by the computer. As computers and OCR devices became more sophisticated, the scanners provided direct access into the CPU (computer processing unit). This quickly lead to the payment processing of credit card purchases, known as "remittance processing". These two applications are still the two major applications for OCR.

In our paper, we have recognized Gurmukhi characters and numerals both separately. We have used isolated Gurmukhi samples and numerals written on plain paper. Zonal density, distance profiles, projection histograms, and background directional distribution (BDD) features are used in different combinations to construct feature vectors.

FUTURE WORK

The work can be extended to increase the results by using or adding some more relevant features. We can use some features specific to the mostly confusing characters, to increase the recognition rate. We can divide the entire character set to apply specific and relevant features differently. More advanced classifiers as MQDF or MIL can be used and multiple classifiers can be combined to get better results.

REFERENCES

[1]. G S Lehal, Ritu Leha and Chandan Singh, "A Shape Based Post Processor for Gurmukhi OCR" in the Department of Computer Science and Engineering, Thapar Institute of Engineering & Technology, 2001.

[2]. G **S** Lehal and Chandan Singh, "A Gurmukhi script recognition system", in Proceedings **IS'"** International Conference **on** Pattern Recognition, Vol 2, pp. 557-560 (2000).

[3]. Kartar Siddharth, Renu Dhir, Rajneesh Rani, “Handwritten Gurumukhi Charater Recognition Using Zoning Density and Background Directional Features”, International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 2, Issue 3, pp. 1036-1041, May-June 2011

[4]. U. Pal, B.B. Chaudhury, “Indian Script Character Recognition: A Survey”, Pattern Recognition, Elsevier, pp. 1887-1899,2004.

[5]. D. Sharma, G. 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.

[6]. G.S. Lehal, C. Singh, “Feature Extraction and Classification for OCR of Gurmukhi Script”, Vivek Vol. 12, pp. 2-12, 1999.

[7]. V. Goyal, G.S. Lehal, “Comparative Study of Hindi and Punjabi Language Scripts”, Nepalese Linguistics, Vol. 23, pp. 67-82, 2008 .

[8]. Naveen Garg, Karun Verma, “Handwritten Gurmukhi Charcter Recognition Using Neural Network”, M.Tech. Theis, Thapar University, 2009 [online]. Available: http://dspace.thapar.edu:8080/dspace/bitstream/10266/788/1/thesis+report+final.pdf.

[9]. Puneet Jhajj, D. Sharma, “Recognition of Isolated Handwritten Characters in Gurmukhi Script”, International Journal of Computer Applications, Vol. 4, No. 8, pp. 9-17, August 2010.