Comparative Study of Different Classifier for Gujarati Off- Line Text Recognition

Mr. Ashwin R. Dobariya[1] Prof.(Dr.) V. R. Rathod[2]
[1]Faculty of Computer Applications [2]Department of Computer Science
[1]Marwadi Education Foundation's Group of Institutions, Rajkot, Gujarat (India)
[2]Bhavnagar University, Bhavnagar, Gujarat (India)

Abstract---During the last fifty years, handwritten character recognition (HCR) has most often been targeted under the research of character recognition (OCR). This paper presents an experimental review of the efficiency of various classifiers for text recognition in terms of accuracy. The paper gives an overall idea about recognition results of different classifiers and provides new information & benchmark for future research for improvement in field of recognition. Here we have given comparative study of different Classifier for Gujarati offline text recognition.

Key words: Classifiers, Recognition, Handwritten character, Gujarati OCR,

I. INTRODUCTION

Gujarati language is deriving to Devnagari family of languages, which originated and used in Gujarat state of India by over 50 million people. Handwriting recognition is a process of a computer to translate human writing data into computer editable text. This can be done by two different ways, the first one is offline recognition technique, known as optical character recognition (OCR) & it is the most successful. The second technique of handwriting recognition often referred to as on-line recognition. Recognition Methods and rates are depended on types of handwriting, the number of writers & their writing style, the size of the character set etc. It is vary from person to person because a same paragraph is written by a one person again then also there is a variation in writing style. From past many years, many academic laboratories, researchers and companies are involved in research on handwriting recognition for trying to improvement in accuracy. In Character recognition process, we required to follow two things i.e. preprocessing on data set and decision making algorithms or methods.

In the first step of pre processing, it perform processes on scan image for segmentation. In second method, there is various kinds of decision methods have been used such as: various statistical methods, neural networks, structural matching and stochastic processing. [1]

The style of Gujarati script writing is differ from those of many other Indian languages not having any shirolekha (headlines)[2]. HCR is the recognition of handwritten text using computers system. The following are the basic steps for recognition process.

- Each and every step is having a specific purpose and directly or indirectly affect to accuracy in recognition process.
- *Binarization:* Basically, it is a process which separate the background and foreground of a scanned image of handwritten document, is called binarization.

- *Noise removal :* It reduce the unnecessary area from the scanned document so it can keep minimum memory resource and easy to identify the actual text.
- *Skew detection :* It is an error of scanning process. It means, during scanning the document it may be placed with some slope in either of the direction or it may happen during scanning.
- *Character Segmentation:* It is very important process to divided document data in line by line and word by word and character by character for final identification.
- *Feature Extraction:* Every character has some special and distinct characteristics / attributes to represent and define them. It is also possible that, a single character can be used for multipurpose or combination with other character. So one needs to have such set of characteristics in which each character is discriminated to its maximum level. To find a set for attributes that defines the character is called feature extraction while subset of characteristics which can define a character to its maximum level is called as feature selection.

II. GUJARATI SCRIPT & CHARACTERS

The Gujarati script was adapted from the Devanagari script to write the Gujarati language.

The earliest known document in the Gujarati script is a manuscript dating from 1592, and the script first appeared in print in a 1797 advertisement.

Gujarati has 11 vowels and 34+2 consonants. Apart from these basic symbols, other symbols called vowel modifiers are used to denote the attachment of vowels with the core consonants.

Consonant-Vowel combinations occur very often in most of the Indic languages including Gujarati. This is denoted by attaching a symbol, unique for each vowel, to the consonant, called a Dependent Vowel Modifier or Matra. [3]

Consonants
ક ખ ગ ઘ ઙ ચ છ જ ઝ ઞ ટ ઠ ડ ઢ ણ ત થ દ ધ ન પ ફ ભ
બ મ ય ર લ વ શ ષ સ હ ળ ક્ષ જ્ઞ
Vowels
અ આ ઇ ઈ ઉ ઊ એ ઐ ઔ ઓ ઓં ઔ અં
Some Conjuncts
ક્ર જ્ય ક્લ ચ્છ દ્ર ત્ર સ્પ સ્ત વ્ ક્ક ત્ય ચ્ય ત્ત ૬
Consonant – vowel
જા ગી જી બુ ખૂ હ્ ૐ કૈ પૈ પો ડો કૌ કં
Conjunct -vowel
જ્જા કૈ ક્લે ચ્છે દ્વી તૃ સ્ન્
Vowels modifiers
ા િ ી ુ ૂ ે ૈ ો ૌ ં

Table. 1: Describes the classification of Gujarati characters

III. DETAILS ABOUT CLASSIFIERS

In this research work, we have done a comparative study on different classifiers for Gujarati Text recognition process which is directly affect on accuracy of recognize Gujarati Text.

A. *Template matching and Fringe distance classifier*

Shah & Sharma have used template matching and Fringe distance classifier as distance measure.

Initially the sample images were filtered using low pass filter. Then the binarization was done by considering the optimal threshold method. Skew detection and correction is done within 0.05. They applied on printed characters and found recognition rate 78.34%. For upper modifier recognition rate was 50% where as for lower modifier it was 77.55% and for punctuation marks it was 29.6%. Cumulative for overall it was 72.3%.[4]

B. *K-NN classifier and minimum hamming distance classifier*

In 1999, Antani and Agnihotri have applied this classifier on printed Gujarati Text. They have created the data sets from scanned images, at 100 dpi. Scan Image pixel values are used as features creating 30x20= 600 dimensional binary feature space and they have used these two classifiers for recognition process. The best recognition rate was for 1-NN for 600 dimensional binary features space i.e. 67% K-NN in regular moment space gave 48% while minimum distance classifier had the recognition rate of 39%. [5]

C. *K-Nearest Neighbor (KNN) and Principal Component Analysis (PCA)*

Baheti M.J. has compared to methods to classify Gujarati handwritten numerals in which they have observed to have higher recognition rate using KNN compare to PCA and have achieved recognition rate of 90.04% with KNN and 84.1% for PCA.[6]

Same authors have extended this research with SVM Classifier [7]. They compared Gaussian distribution function and Support vector machine along with KNN and PCA and found that SVM have highest recognition rate of 92.28% in comparison to Gaussian distribution which is 87.2%.

D. *Wavelet features, GRNN classifier and KNN classifier*

J. Dholakia, A. Yajnik, A. Negi have applied these classifier on the printed Gujarati text of font sizes 11 to 15 with styles regular, bold and italic by finding the confusing sets of the characters. They collected many samples with initial size 32x32 and 16x16 wavelet coefficients have been extracted creating the feature vector. They have applied two classifiers GRNN and KNN with Euclidean distance as similarity measure and got 97.59 and 96.71 respective recognition rates. [8]

E. *Feed forward back propagation neural network*

Apurva A. Desai has used Feed forward back propagation neural network to classify Gujarati numerals and have demonstrated multi layer neural network with three layers (94, 50 and 10) neurons respectively and he used this concept with 278 sets of various digits. Out of these 278 sets, 11 sets were created by a standard font. From the 265 sets the author recorded the success rate for standard fonts as 71.82%, for handwritten training sets as 91.0% while for testing sets as a score of 81.5% was recorded. [9]

F. Hidden Markov Model (HMM)

It is suitable for handwriting recognition for a number of reasons. HMM's have also been successfully applied to image pattern recognition problems such as shape classification. HMM's qualify as suitable tool for cursive script recognition for a number a reasons. These algorithms are fast and can be implemented with reasonable effort for improvement in recognition rate. [10] [11]

Comparative analysis Of Classifier for Gujarati Text Recognition			
Sr.No	Researcher Name	Classifier	Recognition Rate (%)
1	S K Shah, A Sharma	Template matching and Fringe distance classifier	78.34 %
2	S. Antani, L. Agnihotri	K-NN classifier & Minimum hamming distance classifier	48.00% 39.00%
3	M. J. Baheti, K. V. Kale, M. E. Jadhav	K-Nearest Neighbor (KNN) and Principal Component Analysis (PCA)	84.10%
4	M. J. Baheti, K. V. Kale, M. E. Jadhav	SVM Classifier	92.28%
5	J. Dholakia, A. Yajnik, A. Negi	Wavelet features, GRNN classifier and KNN classifier	97.59 96.71
6	Apurva A. Desai	Feed forward back propagation neural network	91.00%
7	J. R. Prasad, U.V. Kulkarni,	Template Matching & Feed forward back propagation Neural Network	71.66%

Table. 2: Comparative analysis of Classifier for Gujarati Text Recognition

IV. CONCLUSION

A study is made on different feature extraction and recognition techniques & Classifiers used for Gujarati scripts. In this paper, the different characteristics of the scripts, their components and the methods used to segment and identify the characters or scripts are being elaborated. Moreover the detailed different study of Gujarati script is given with recognition rat. This paper studies various methodologies proposed by researchers to recognize off-line Gujarati text. The text recognition is a wide area and still there is more research to be required to achieve a good recognition result with the enhance techniques and methods.

REFERENCES

[1] Comparative Study Of Different Classifiers For Devanagari Handwritten Character Recognition By: Anilkumar N. Holambe & Dr.Ravinder.C.Thool, International Journal of Engineering Science and Technology Vol. 2 (7), 2010, 2681-2689

[2] Gujarati handwritten numeral Optical character recognization through neural network. By Apurva A.Desai Veer Narmad South Gujarat University, Surat, Gujarat, India.

[3] Technical Project Report Referred: "Development of Robust Document Analysis and Recognition System for Printed Indian Scripts ", Project sponsored by: Ministry of Communication and Information Technology, Jul – 2008

[4] "Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching" By: S K Shah and A Sharma ,IE(I) Journal-ET Vol.86 2006 pgs. 44-49.

[5] "Gujarati Character Recognition". By: S. Antani, L. Agnihotri., Fifth International Conference on Document Analysis and Recognition, (ICDAR'99), 1999 p. 418.

[6] COMPARISON OF CLASSIFIERS FOR GUJARATI NUMERAL RECOGNITION By: M. J. Baheti, K. V. Kale, and M. E. Jadhav, International Journal of Machine Intelligence, vol. 3. 2011

[7] Gujarati Numeral Recognition: Affine Invariant MomentsApproach, By: M. J. Baheti, Kale K.V., 2012 Soft Computing, pp. 140-146.

[8] "Wavelet Feature Based Confusion Character Sets for Gujarati Script", By: J. Dholakia, A. Yajnik, A. Negi ,ICCIMA, 2007 p 366-371.

[9] Gujarati handwritten numeral optical character reorganization through neural network, Pattern Recognition, By: Apurva A. Desai, 2010 ,vol. 43, pp. 2582-2589.

[10] Offline Cursive Handwriting Recognition using Hidden Markov Models. Pattern Recognition, By: H. Bunke, M. Roth, and E. G. Schukat-Talamazzini 28(9):1399–1413, 1995.

[11] 2-D shape classification Using Hidden Markov Model, By: Y.He, A.Kundu: IEEE Trans. On PAMI, vol.13,1991,pp.1172-1184.

[12] Offline Handwritten Character Recognition of Gujrati script using Pattern Matching By: Prasad, J. R.; Kulkarni, U. V. ; Prasad, R. S. ASID 2009. 3rd International Conference

[13] Gujarati Script Recognition: A Review –By Mamta Maloo, Dr. K.V. Kale IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

[14] "A Bilingual OCR for Hindi-Telugu Documents and its Applications." By : Jawahar C. V., M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran International Conference on Document

[15] Handwritten Character Recognition System Using Artificial Neural Networks By : Pelin GORGEL Oguzhan O, ZTAS, Journal Of Electrical & Electronics Engineering

[16] "A comparative study of neural network algorithms applied to optical character recognition" By : Smagt, P. ACM/ IEEE,2000.

[17] Offline Handwritten Character Recognition of Gujrati Script using Pattern Matching, By : J. R. Prasad, U.V. Kulkarni, 2003. Computer Engineering.

[18] Web URL : http://en.wikipedia.org/wiki/Artificial_ neural_ network