© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A Survey on Gujarati Handwritten OCR using Morphological Analysis

Vaidehi Patel, Prof. Abhinay Pandya

Department of Computer Engineering, LDRP Institute of Technologya and Research, Gandhinagar,Gujarat, India

ABSTRACT

Handwriting recognition has been one of the active and challenging research areas in the field of imageprocessing and pattern recognition. Handwriting recognition can be differentiated into two categories i.e.online handwriting recognition and Offline handwriting recognition. I mainly focus on offline handwritingrecognition. It is also observed that majority of work for this language is for printed form rather thanhandwritten form. One can obtain very less OCR related research work for Gujarati script, especially forhandwritten form. I use Freeman chain code for feature extraction and then used Hidden Markov method.There is no benchmark dataset is available so it is generated by collecting sample from more than 100writers of different age group and gender.

Keywords: Handwritten Recognition, Freeman Chain Code, Pre-processing, Classification

I. INTRODUCTION

The form of the text is varied from the scanneddocument to the typed text in various fonts. For humans,recognition of text is a trivial task, but to make amachine that recognizes characters is extremelydifficult. The current study is being focused onexploration of possible methods to develop an OCRsystem for Gujarati language when noise is present inthe signal. In order to understand the core challenges, athorough analysis of Gujarati Writing System has beendone. To know and understand the research in this field,existing OCR systems are studied. The Prominence wason finding workable segmentation technique anddiacritic handling for Gujarati words, and building arecognition module for these ligatures.

To develop an OCR system Gujarati text, a completeprocedure is proposed as well as a testing application isalso made. Test results from this are reported andcompared with the previous work done in this area.

II. METHODS AND MATERIAL

A. Features of Gujarati Language

Gujarati is regional language of state Gujarat in India. Gujarati is name of script which is written and spokenby people in Gujarat. Gujarati script has 34+2 constantsand 11 vowels. The structures of some Gujaratiharacters are very similar to Devnagri script. Devngari script has shirolekha on top of the characters, but Gujarati script has not. Gujarati script does not have the distinction of Lower and Upper Cases like English script. Gujarati script has combination of constants and vowels. Every vowel has a unique symbol, called vowels modifiers. Table 1 shows Gujarati vowels, consonants, matras, other symbols and some of the conjuncts [5].

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 ઍ ઐ ઓ ઑ ઔ અં
Some Conjuncts
ક્ક જ્ય ક્લ ચ્છ દ્ધ ત્ર સ્પ સ્ત ર્વ ક્ક ત્ય ત્ર્ય ત્ત દ્ર
Consonant – vowel
જા ગી જી બુ ખૂ હે કૃ કૅ પૈ પો ડૉ કૌ કં
Conjunct -vowel
જ્જા કે ક્લે ચ્છિ દ્ધી ત્રુ સ્નૂ
Vowels modifiers
ા િ ી ુ ૂ ૃ ે ૅ ૈ ૉ

Figure 1 : Gujarati Script[5]

B. Dataset Generation

There is no benchmark dataset is available so it isgenerated by collecting sample from more than 100 people of different age group and gender.

Figure 2 : Handwritten Characters

C. Feature Extraction

The objective of feature extraction is to capture theessential characteristics of the symbols, and it isgenerally accepted that this is one of the most difficultproblems of pattern recognition. There are belowtechniques used.

1. Zoning
2. Projection Histogram Feature
3. Moments
4. Fourier Transform
5. Freeman chain code

Table1 - Comparison of different feature extractiontechniques

Techniques	Advantages	Disadvantages
Zoning	Simple, Easy to implement	Density of object pixel in each zone is calculated. Efficiency is low
Moments invariants	Easy recognize pattern field	Higher order moments are sensitive to noise and variation of writing style
Freeman Chain code	Process time small, Storage is lossless ,easy to represent easy to implement	Length of FCC depend on starting point, may be revisit same node
Fourier Transform	Give valuable info about character structure. Recognize position shifted character	Not give accurate result. Tough to implement
Projection Histogram	Original histogram can be recovered, easy to implement	It is indiscriminate. It may increase the contrast of background noise while decreasing the usable signal

III. RESULTS AND DISCUSSION

CLASSIFICATION

The Extracted features are given as the input to theClassification process. A bag-of-key point extracted from thefeature extraction approaches are used for classification.

1. Template matching
2. Neural Networks
3. Support Vector Machine (SVM)
4. K-nn Classification
5. Hidden Markov Model

See all of these techniques in details.

Template matching is simple technique ofcharacter recognition; depend on matching thestored templates with the character or word to berecognized. The matching operation finds out thesimilarity between two vectors .An input image ismatched with set of already stored templates. Therecognition rate of template matching isproportional to noise and image deformation. [9]

Neural Networks is composed of interconnectednodes that are connected via links. Learning is providedby example via training, or exposure to a set of inputoutputdata (patterns), where the training algorithmadjusts the link weights.

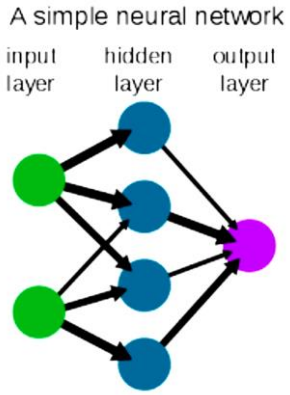

Figure 3 : Neural network

K-nn Classifiers is a nonparametric method used forclassification. It is a Statistical method. So, basically thek-nearest neighbor algorithm (k-NN) is a method forclassifying objects based on closest training examples inthe feature space[4].

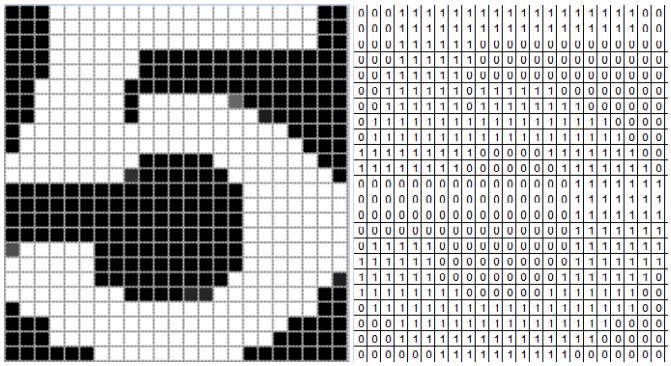

Figure4: The kNN classifier that is used here uses Euclidean distance method

Support vector machines (SVM), when applied totext classification provide high accuracy, but poor recall. Onemeans of customizing SVMs to improve recall, is to adjust thethreshold associated with an SVM. SVMs have achievedexcellent recognition results in various pattern recognitionapplications.

Hidden Markov Model is a finite set of states,each of which is associated with a probabilitydistribution. Transitions among the states aregoverned by a set of probabilities called transitionprobabilities. In a particular state an outcome orobservation can be generated, according to theassociated probability distribution. The probabilitiesfor each candidate character are calculated. Then,the probabilities are counted to obtain a final bestcharacter-list for character recognition[1].

Table2: Comparison Table

Method	Advantages	Disadvantages
Template Matching	High Speed. Simple to implement No need to extract Features	Not effectivewhen there arefont discrepancy, font slant,font defilement.The method isnot invariant to changes inillumination
Support Vector Machine	Good for small category dataset Easy to compute. Less time	High Complexityof training andexecution. If dataset is big hen results large number of SVMs which required more storage and time
Neural Networks	Higher recognizing ratio	Take more time for training
	with more training able to adapt to changes in the input data Accuracy is Greater	
K-NN algorithm	Simple Non parametric	Each step of distance is calculated so takes more time
Hidden Markov Model	Freedom to manipulate training and verification. The probability of observing sequence model and computed for each word.	Not completely automatic

IV. CONCLUSION

In this paper many techniques explained for the scanneddocument .Different OCR methods relevant to thesurvey of many papers. For developing faster OCR than Hidden markov model method is very suitable becauseit is probabilistic model and not requires more training.

V. REFERENCES

[1] Jinhong K. Guo and Matthew Y. Ma "Separating Handwritten Material from Machine Printed Text Using Hidden Markov Models" Published by IEEE 2001.

[2] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu*, Mahantapas Kundu,'Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition' published by IEEE 2008.

[3] Debi Prasad Bhattacharya, Susmita Koner "English alphabet recognition using chain code and LCS" published by Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 3 No. 2 Apr-May 2012.

[4] Chhaya Patel and Apurva Desai , "Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And KNearest Neighbour" published by International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181 Vol. 2 Issue 6, June - 2013

[5] Dholakia, J.; Yajnik, A.; Negi, A., "Wavelet Feature Based Confusion Character Sets for Gujarati Script," Conference on Computational Intelligence and Multimedia Applications. International Conference on, vol.2, no., pp.366,370, 13-15 Dec. 2007

[6] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu*, Mahantapas Kundu,'Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition' 978-1-4244-2806-9/08/$25.00© 2008 IEEE

[7] Anitha Mary M.O. Chacko and P.M. Dhanya 'A Comparative Study of Different Feature Extraction Techniques for Offline Malayalam Character Recognition '© Springer India 2015

[8] Chaudhari Shailesh A., and Ravi M. Gulati. "An OCR for separation and identification of mixed English—Gujarati digits using kNN classifier."Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on. IEEE, 2013.

[9] Prasad, J.R.; Kulkarni, U.V.; Prasad, R.S., "Template Matching Algorithm for Gujrati Character Recognition," Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on , vol., no., pp.263,268, 16-18 Dec. 2009