A Review on offline Handwritten Gujarati Character Feature Extraction and Recognition

Vishwas Goyal[1], Vivek Naik[2], Mitali Desai[3], Hiral Desai[4], Shreyash Salian[5]
[1, 2, 3, 4, 5] Babu Madhav Institute of Information Technology, Uka Tarsadia University

Abstract- *Optical Character Recognition (OCR) is a technical method of recognition of character from an image. Even large amount of information is preserved on paper and in the era of digital world it requires all preserved paper document to store their information in electronic format. So as operations like searching adding deleting and updating text from an image can be carried out. For English language great work is done but in case of Indian language the research is very finite. Another problem with handwriting character recognition is different users have different writing style while this is not a problem in printed text. This paper contains analysis of different features extraction and classification methods for Gujarati script. We have explained the fundamentals of Handwritten Character Recognition (HCR).*

Keywords- Feature extraction, Gujarati character recognition, Handwritten Character Recognition, Optical character recognition.

I. INTRODUCTION

Character recognition is an art of identifying characters from a text image. Character recognition algorithm varies as per different language and its characteristics like: direction of writing (i.e. left to right – English, Hindi, Gujarati), set of alphabets (i.e. English: A- Z, a-z), nature of writing that defines how sentence are written (cursive script: English, Gujarati script: single and individual Character). Handwritten Character Recognition (HCR) has lots of variations as different people have different handwriting styles. Several researches have been trying to evolve new techniques and methods which would reduce the image processing time while providing higher recognition accuracy. [2]

II. RELATED WORK

A. HCR

Whenever a handwritten document is thought for recognition, there are enumerable factors involved with it. Firstly the document is scanned so that the text on paper becomes the image on computer. Then this image is preprocessed and converted into either machine-editable format. To handle the image, preprocessing involves various steps [3] for enhancing the rate of recognition.

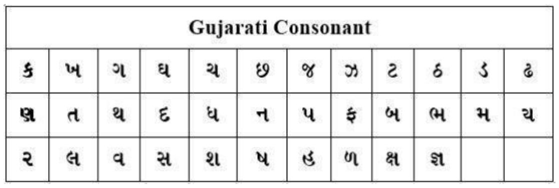
These general processing steps are summarized as under

- Binarization of scanned image
- Removal of Noise from scanned image
- Thinning of binarized image
- Skew detection and correction of scanned image,
- Segmentation of image
- Feature Extraction Techniques
- Recognition on the basis of Classifiers

The images are binarize to get the image in only two colors which make recognition of character more easy. Then after the process of binarization noise is removed from the image to extract only character from the image. Image thinning is the process where the entire image is thin to one pixel and extra pixels are removed for better processing. If capture image is tilt then skew correction is required. After all that preprocessing steps segmentation is applied. Then the process of feature extraction comes where the features are extracted from image and character form image is recognizing according to that feature. Feature extraction plays the important role for character recognition because if right and important feature are not selected then it leads to wrong recognition of character.

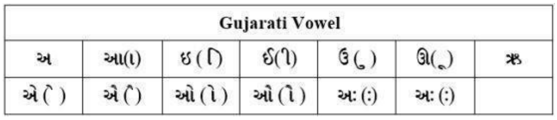
B. Gujarati Script

Gujarati is official language of Gujarat, India and very less OCR work done on it but from last few years many people are researching on it. Gujarati script have very large character set almost total 47 which include 34 consonant (Vyanjans) (Fig.1) and 13 vowel (Swar) (Fig.2).



Gujarati Consonant											
ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ	ટ	ઠ	ડ	ઢ
ણ	ત	થ	દ	ધ	ન	પ	ફ	બ	ભ	મ	ય
ર	લ	વ	સ	શ	ષ	હ	ળ	ક્ષ	જ્ઞ		

Figure 1: Gujarati consonant



Gujarati Vowel						
અ	આ(ા)	ઇ (િ)	ઈ(ી)	ઉ (ુ)	ઊ(ૂ)	ઋ
એ (ે)	ઐ (ૈ)	ઓ (ો)	ઔ (ૌ)	અ: (:)	અ: (:)	

Figure 2: Gujarati vowel

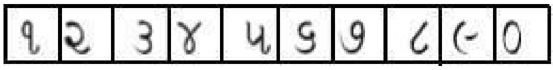


Figure 3: Gujarati Digits

C. Difficulties in Gujarati character recognition:-

Gujarati OCR have a various complication from it some of the problem in Gujarati OCR is as listed:

1) Gujarati characters have variety of feature.
2) Each person writing style is different and hard to get features from it.
3) Many characters of Gujarati are same as each other like ર (ra in consonant) and ૨ (numerical two in digits).
4) The Gujarati character like ગ (gh), ણ (na), શ (sha), લ (L) reduce the accuracy of the segmentation as they are closely connected so they are difficult to recognize.
5) Gujarati scripts have many modifiers like etc. so they also create a problem in recognition.
6) Gujarati scripts have one more problem with a half characters like “ ” where “ત” is used as half character.
7) One another problem in Gujarati language is with diacritic marks like in
8) Gujarati word “ થ ” diacritic mark used with character “પ”.

II. LITERATURE REVIEW

Feature extraction act as a major role in HCR. Adopting distinct pattern or information from an image which is helpful for character recognition is known feature extraction in character recognizing. [3]

Hetal R. Thaker proposed structural feature extraction and classification methodology based on decision tree they applied this methodology on 5 different characters where 150 samples of each character is used, result obtained by this methodology is shown in below table. Success ratio for ‘ણ’ is 98%, ‘શ’ is 80.6%, ‘ગ’ is 94.66%, ‘લ’ is 87.33% and ‘જ’ is 83.33%. [4]

Kamal Moro et al. used neural network technique in his paper. Neural network performance is 85.33% while in the neural network 80.5% performance is trapped with skeleton method and without skeleton 85.33% performance is trapped. Effective and efficient digitization process and segmentation process is required for classification of character. Extraction of characteristic plays an important role in the performance of each method of classification e.g. the unequal space between the character and words require more processing to divide them.[6]

Prachi et al. proposed a Gujarati OCR system for the identifying of basic characters in printed Gujarati text. Principal Component Analysis (PCA) was used to extract the properties of printed Gujarati characters. For the grouping of characters based on features hopfield neural classifier had been used by them. The system acquires the 93.25% accuracy. [7]

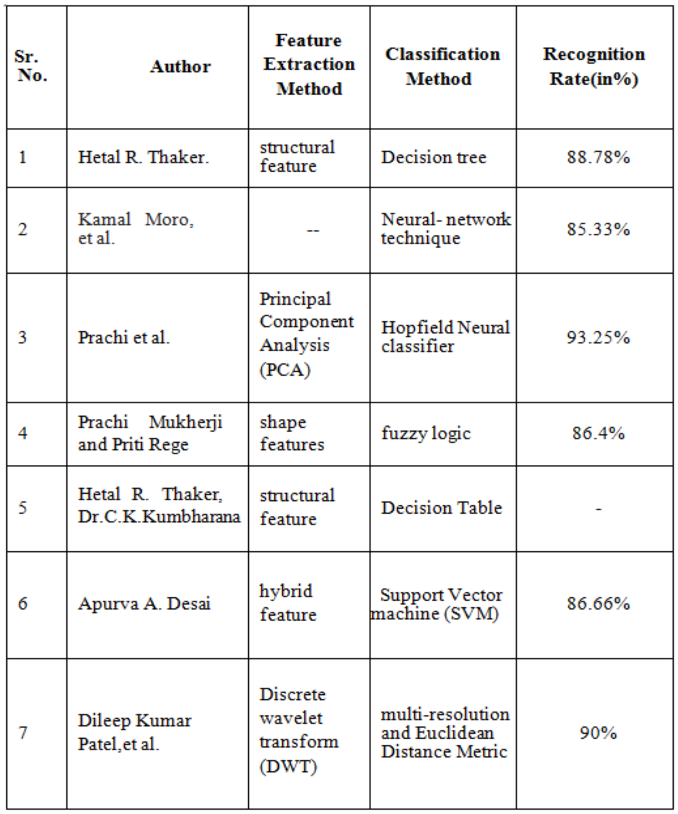
Hetal R. Thaker, Dr. C. K. Kumbharana. Use structural feature like horizontal line, curve, slope etc. They identify different structural feature and categorized it into 22 different groups. They classified it in form of decision table where row represent feature and column represent Gujarati consonant. By tracing the column they identify set of feature for each consonant and to identify particular feature is in how many consonant row have to be trace. [8]

Prachi Mukherji and Priti Rege, used shape features and fuzzy logic to identify offline Devnagari character. They segmented the thinning of characters into strokes using structural features. They rank the segmented shapes or strokes as left curve, right curve, horizontal stroke, vertical stroke, slanted lines etc. They used tree and fuzzy logic and obtained average of 86.4% accuracy. [10]

Dr. Mamta Baheti's, works on Gujarati numeral feature extraction and recognition.For that they apply an algorithm for Gujarati numeral recognition in that first take image from database, resize image, after then complement image, binarize and dilate it, also applied thinning for extra pixels renovation, image slicing applied (Matrix 4*4, 5*5, 8*8) then apply invariant moments approach and use gaussian function and at last compute the recognition rate on the basis of classified and miss-classified numerals. They consider noisy numerals and not implemented skew correction techniques. The recognition result is less because data set has poor quality of numerals with no constraints for pen, ink or numeral size.They receive accuracy for ૦=92.5%, ૧=71.25%, ૨=22.5%, ૩ =32.5%, ૪=75.5%, ૫=55%, ૬=53.75%, ૭=55%, ૮=86.25%, ૯=61.25% average accuracy is 60%.૦ and ૧ has given better result than other.[9] Apurva A. Desai, work on handwritten Gujarati character recognition using support vector machine with hybrid feature space. Gujarati character 'ja' is give result

97.97% and 'tha' gives poor result.'tha','pa','gha','dha','ya' and 'va' are confusing character because of their similar shapes. Character 'gha','pa' and 'ya' are confusing with 'tha'.There are three major works done KNN (Numbers - 96.99%), KNN, tree classifier (Alphabets - 63%), artificial neural network (Numbers - 81.66 %).This recognition shows that for handwritten character recognition hybrid feature is better than simple structural feature set. SVM with hybrid feature give good result of character recognition compared to KNN and SVM with gaussian kernel. [1] Though of this work they achieved 86.66% accuracy. Database collect from different age groups and gender of one hundred ninety nine writers. [1]

Dileep Kumar Patel et al. use the multi-resolution and euclidean distance metric for handwritten character recognition. In this paper characters are divided into 26 classes based on their properties. Using discrete wavelet transform (DWT) features of handwritten character images are extracted. In multi- resolution recognition accuracy is very high up to 90%. [13]



Sr. No.	Author	Feature Extraction Method	Classification Method	Recognition Rate(in%)
1	Hetal R. Thaker.	structural feature	Decision tree	88.78%
2	Kamal Moro, et al.	--	Neural- network technique	85.33%
3	Prachi et al.	Principal Component Analysis (PCA)	Hopfield Neural classifier	93.25%
4	Prachi Mukherji and Priti Rege	shape features	fuzzy logic	86.4%
5	Hetal R. Thaker, Dr.C.K.Kumbharana	structural feature	Decision Table	-
6	Apurva A. Desai	hybrid feature	Support Vector machine (SVM)	86.66%
7	Dileep Kumar Patel,et al.	Discrete wavelet transform (DWT)	multi-resolution and Euclidean Distance Metric	90%

III. DATA SET

There is no data set available for Gujarati handwritten character so that we can work on it. So for that we collect data set from 15-20 people so that we have a different writing style that include variation which will be help full for training phase as well in testing phase also. Collecting data set from more and more different people is more help full forget better accuracy.

IV. CONCLUSION

This paper reviewed the advancement of feature extraction in the field of Handwritten Character Recognition (HCR). Here in this paper many feature extraction technique are described like structural features, statistical features and hybrid feature which include multiple features extraction technique. Here many techniques are being discussed on feature classification like decision tree, neural network, fuzzy logic, SVM which also acquired quite satisfactory success ratio, decision tree also provide good success ratio but it does not respond with all the characters set. Here many feature extraction technique also describe like structural features, statistical features and hybrid feature which include multiple features extraction technique. For in future we will expand our work on Gujarati HCR mainly on features extraction part.

REFERENCES

[1] Apurva A. Desai, Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space, CSIT (January 2015) 2(4):235–241 DOI 10.1007/s40012-014-0059-z

[2] O. D. Trier, A. K. Jain and T. Text, Feature Extraction Methods for Character Recognition- A Survey, Pattern Recognition, Vol. 29, No. 4, pp. 641-662, 1996.

[3] Hetal R. Thaker , Dr. C. K. Kumbharana, Analysis of structural features and classification of Gujarati consonants for offline character recognition , International Journal of Scientific and Research Publications, Volume 4, Issue 8,2014.

[4] Hetal R. Thaker , Dr. C. K. Kumbharana, Structural Feature Extraction to recognize some of the Offline Isolated Handwritten Gujarati Characters using Decision Tree Classifier, International Journal of Computer Applications, Volume 99 – No.15, August 2014

[5] S. S. Magare, Y. K. Gedam, D. S. Randhave, Prof. R. R. Deshmukh. Character Recognition of Gujarati and Devanagari Script: A Review, International Journal of Engineering Research & Technology, Vol.3 Issue 1, January - 2014.

[6] Kamal Moro, Mohammed Fakir, Belaid Bouikhalene, Rachid El Yachi, Bader Dinne El Kessab, New

Approach of feature extraction method based on the raw form and his skeleton for Gujarati handwritten digits using neural network classifier, University of Galati fascicle III, 2014, VOL. 37, NO. 1.

[7] Prachi Solanki,Malay Bhatt, Printed Gujarati Script OCR using Hopfield, International Journal of Computer Applications, Volume 69– No.13, May 2013.

[8] Hetal R. Thaker, Dr. C. K. Kumbharana, Analysis of structural features and classification of Gujarati consonants for offline character recognition, International Journal of Scientific and Research Publications, Volume 4, Issue 8, August 2014.

[9] Dr. Mamta Baheti, Invariant Moments Approach for Gujarati Numerals, International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume-2, Issue-2, February 2015

[10] Prachi Mukherji, Priti Rege, “Shape Feature and Fuzzy Logic Based Offline Devnagari Handwritten Optical Character Recognition”, Journal of Pattern Recognition Research 4 (2009) 52-68, 2009.

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

[12] Mamta Maloo, Dr. K.V. Kale, Gujarati Script Recognition: A Review, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): 1694-0814

[13] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, Handwritten Character Recognition Using Multi- resolution Technique and Euclidean Distance Metric, Journal of Signal and Information Processing, 2012, 3, 208-214, May 2012