iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IX **Month of publication:** September 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IX, Sep 2018- Available at www.ijraset.com

A Survey of Gujarati Handwritten Character Recognition Techniques

Arpit A. Jain[1], Harshal A. Arolkar[2]
Assistant Professor[1], Associate Professor[2] , GLS University

Abstract: ***OCR termed as Optical Character Recognition, is a technique to convert mechanically or electronically an image, photo or scanned document of a handwritten or printed text into machine encoded text. HCR termed as Handwritten Character Recognition, is a form of OCR that is specifically designed to recognize the handwritten text. OCR and HCR nowadays are used extensively for information entry from printed or handwritten data records. In this paper we have done a survey on Gujarati Handwritten Character Recognition techniques.***

Keywords: OCR, Optical Character Recognition, HCR, Handwritten Character Recognition, Image Processing, Gujarati HCR, Gujarati Handwritten Character Recognition.

I. INTRODUCTION

Characters of any language are created using two types of mechanism namely; Digital and Handwritten Format. The digital characters are created with the help of a computer. The handwritten characters are the one's that are written by person. Handwritten characters can further be classified into two categories: Offline and Online. The offline characters are written using any normal pen; while online characters are created using an optical pen or stylus on an electronic device. Figure 1 and Figure 2 shows the sample of offline and online characters.

Figure 1: Offline Characters

Figure 2: Online Characters

Languages like English, French, and Spanish have alphabets and vowels. Vowels of such languages do not reshape the characters so it is easy to create an OCR for such languages. Whereas, Devanagari and other Indian regional languages have matras and they reform the character, so it is difficult to recognize the offline handwritten characters. English Character Recognition techniques have been studied extensively in the last two decades and it has been observed that almost all characters are easily identified with high accuracy rate. Indian regional languages OCR and HCR on the other hand are still emerging. The character recognition accuracy is still a concern in them.

II. CHARACTERS IN GUJARATI LANGUAGE

The Gujarati script was adapted from the Devanagari script. Devanagari has 11 vowels and 33 consonants. Besides the consonants and the vowels, other constituent symbols in Devanagari are set of vowel modifiers called matra (placed to the left, right, above, or bottom of a character or conjunct), pure-consonant (also called half- letters) which when combined with other consonants yield conjuncts.

The difference between Gujarati and Devanagari script is the Devanagari words and characters have a upper line called Shiroekha (a header line) while Gujarati words and characters does not have Shiroekha. Gujarati language has a vast character set. Gujarati alphabet utilizes overall 75 distinct legitimate and recognized shapes, which mainly includes 59 Characters and 16 diacritics. Fifty-nine characters are divided into 2 compound consonants and 34 singular consonants also known as ornamented sounds as shown in figure 3 where 23 characters have vertical line, 13 vowels (pure sounds) as shown in figure 4 and 10 numerical digits shown in figure 5. Sixteen diacritics are divided into 13 vowels and 3 other characters [3, 5, 20]. The consonants can be combined with the vowels and can form compound characters known as Conjunct Consonants as shown in figure 6. A word can be formed by combining the basic character(s), which may be combined with vowel(s).

Gujarati Consonant											
ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ	ટ	ઠ	ડ	ઢ
ka	kha	ga	gha	cha	chha	ja	za	ta	tha	da	dha
ણ	ત	થ	દ	ધ	ન	પ	ફ	બ	ભ	મ	ચ
aNa	ta	tha	da	dha	na	pa	fa	ba	bha	ma	Ya
ર	લ	વ	સ	શ	ષ	હ	ળ	ક્ષ	જ્ઞ		
ra	la	va	sa	sha	shha	ha	ala	ksha	gna		

Figure 3: A group of Consonants [20]

Gujarati Vowel						
અ	આ(ા)	ઇ (િ)	ઈ(ી)	ઉ (ુ)	ઊ(ૂ)	ઋ
a	aa	e	ee	u	oo	ri
એ (ે)	એ (ૈ)	ઓ (ો)	ઔ (ૌ)	અં	અઃ (:)	
a	ai	o	au	am	ah	

Figure 4: Vowels [20]

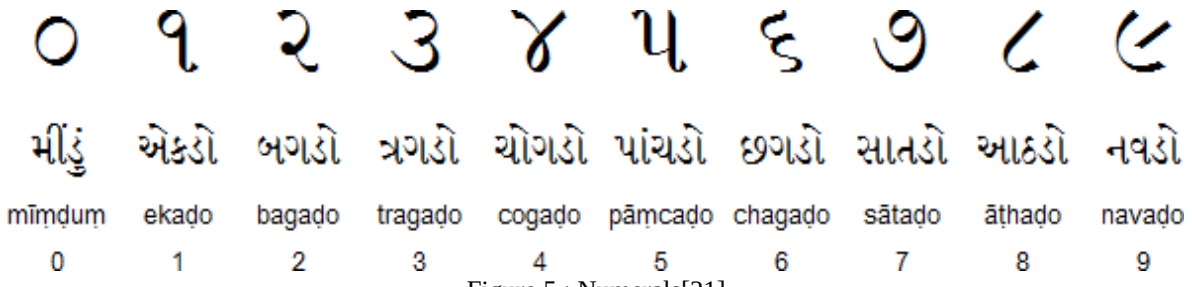


Figure 5 : Numerals[21]

Figure 6: A selection of conjunct consonants [21]

The character set of Gujarati is almost double than the character set of English language. Recognizing the characters in Gujarati at time is little bit confusing, as they have similar looks for example numeral five (૫) and alphabet 'Pa' (પ), numeral two (૨) and alphabet 'Ra' (ર) can be easily miss interpreted.

Further more the conjuct consonants creates a character by combining one and half character. This adds to another problem of identifying additional permutation and combination of an existing character. These are the few characteristics of Gujarati script which can be considered as a reason for the slow progress in development of Gujarati character recognition.

III. LITERATURE SURVEY

A large amount of literature is available for the recognition of Handwritten English, Japanese, Chinese, Arabian characters; whereas comparatively a small amount of work has been reported for the recognition of Indian scripts [8, 9, 17]. The handwritten Indian scripts vary extensively on the basis of consonants, vowels, script wise representation and conjunctional appearance. Due to this differences developing an accurate handwritten characters recognizer becomes a huge challenge.

In this section we have reviewed few of the recent papers that have been published in the domain of Gujarati Character Recognition.

1) Vishal A. Naik and Apurva A. Desai [14] proposed an algorithm for online handwritten Gujarati character recognition using hybrid features. They compared Support Vector Machine (SVM), K-Nearest Neighbor and Multi-layer perceptron (MLP) classifiers. For that they have collected training set of around 3000 samples. For the testing of their data set they have used SVM with Radial basis function (RBF) kernel and SVM with linear kernel. As there is difference in accuracy and execution time of both the kernels. They have received accuracy of 91.63% and 0.063 seconds of average execution time per stroke using SVM with RBF kernel and 90.63% and 0.056 seconds of average execution time per stroke using SVM with Linear Kernel. Their proposed system is finally tested by 100 users by providing input.

2) Milind Kumar Audichya, Jatinderkumar R. Saini[1] presented a study on Recognize printed gujarati characters using Tesseract OCR. In this paper they represented the working of teserract. They used built-in trained test data provided by Tesseract OCR to recognize characters from the digitally typed images. They have applied their testing on various font styles, types and sizes, and found satisfying results also. They conclude that the Teserract is not providing satisfactory result while working with complex characters and those characters which looks similar.

3) Modi M., et.al [11] presented a survey on Guajarati Character identification. In this paper, they listed various techniques for performing steps of character recognition like Global and Adaptive thresholding for Thresholding, Median Filter, Linear Regression Analysis, Fourier Transform based method, Edge based connected component approach for skew detection, Hough Transform, Region growing algorithm for segmentation, Template matching, zoning, Zenrike moments for Feature Extraction. They discussed about the pre-processing, Segmentation, Feature Extraction, Classification and Recognition and Post Processing part which are the necessary steps to be followed in recognition of an image. They also listed and discussed Multi-Layer Perceptron (MLP) and Collapsed Horizontal Projection (CHP) based algorithm for joint character segmentation. They talked about the non-parametric and parametric recognition which includes; KNN and Baye's method respectively.

4) Hetal R Thakkar and C. K. Kumbharana [20] worked on a subset of 5 Guajarati character namely; 'Na' (**ન**), 'Sha' (શ), 'Ga' (ગ), 'La' (લ), 'Ja' (જ). They used Decision Tree Classifier approach for classification. They split the five Gujarati Characters in terms of connected components, end component, close loop and disconnected components. They built a tree structure for the extraction of data according to the given components. The paper claims to have achieved 88.78% of average success rate for identifying the sample five characters.

5) Baheti and Kale [3] worked on Gujarati Numerals. They collected 1800 data set of numerals. They presented Hybrid Approach algorithm which uses Nearest Neighbour classifier and found that it is possible to enhance the performance of a system, if a character was divided into parts and the process of recognition is done on each part of the character. They proved recognition efficiency of 94% for specific numeral nine (૯).

6) Patel Chhaya and Desai A. Apurva [16] described the methods for identification of zone boundaries for a word and usage of zone boundaries details for segmenting the word into its subcomponents. They also stated that, "as far as handwritten OCR is concerned almost negligible work is found" [17]. They divided the word into three parts upper zone, middle zone and lower zone. The paper also introduced the concept of Mean line and Base line. Mean Line is an imaginary line that separates the upper zone and middle zone. Base line is imaginary line that separates middle zone and lower zone. The part of the text below the base line and upper the mean line is used for writing dependent vowels. The part in between mean line and base line is used for writing the consonants and independent vowels. They used a simple vertical projection profile based approach to extract upper and lower modifiers. An approach called connected component labelling was used for middle zone, it was developed to extract the subcomponents of the word. After working on the data set of 250 words they found 75% accuracy for the upper zone, 84% accuracy for the middle zone and lower zone.

7) Maloo and Kale [13] provided a path for developing recognition tools for Indian scripts. They said that there is still a scope of recognition accuracy. It provides the reasons for researchers to work for recognition of Indian scripts. They described the steps of HCR with different methods.

8) Desai [5] collected 300 samples of 300dpi with initial size of each numeral as 90x90. The author then adjusted the contrast by CLAHE i.e. Contrast Limited Adaptive Histogram Equalization algorithm considering 8x8 tiles and 0.01 as contrast enhancement constant. The boundaries were then smoothened out by median filter of 3x3 neighbourhoods. Image is then reconstructed to the size of 16x16 pixels using Nearest Neighbour Interpolation. They finally got overall performance of 81.66% which is not up to the mark. They are looking for the improvement in feature abstraction and pre-processing techniques for getting better performance.

9) Dholakia [8] attempted to use wavelet features, GRNN classifier and KNN 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 4173 samples of middle zone glyphs of initial size 32x32and 16x16 wavelet coefficients have been extracted creating the feature vector. Two sets of the randomly selected glyphs (2802 symbols) were used for training and 1371 symbols were used for testing. Two classifiers GRNN and KNN with Euclidean distance as similarity measure were used producing 97.59 and 96.71 as their respective recognition rates.

10) Jignesh Dholakia et. al [7] have presented an algorithm to identify various zones used for Gujarati printed text. In the algorithm they have proposed the use of horizontal and vertical profiles. They have identified these zones by slope of lines. The slope of line are created by upper left corner of rectangle and by the boundaries of connected components from line level. They have used 3 different document images, from which 20 lines were extracted where 19 were detected with correct zone boundary. The line where it failed was very much skewed.

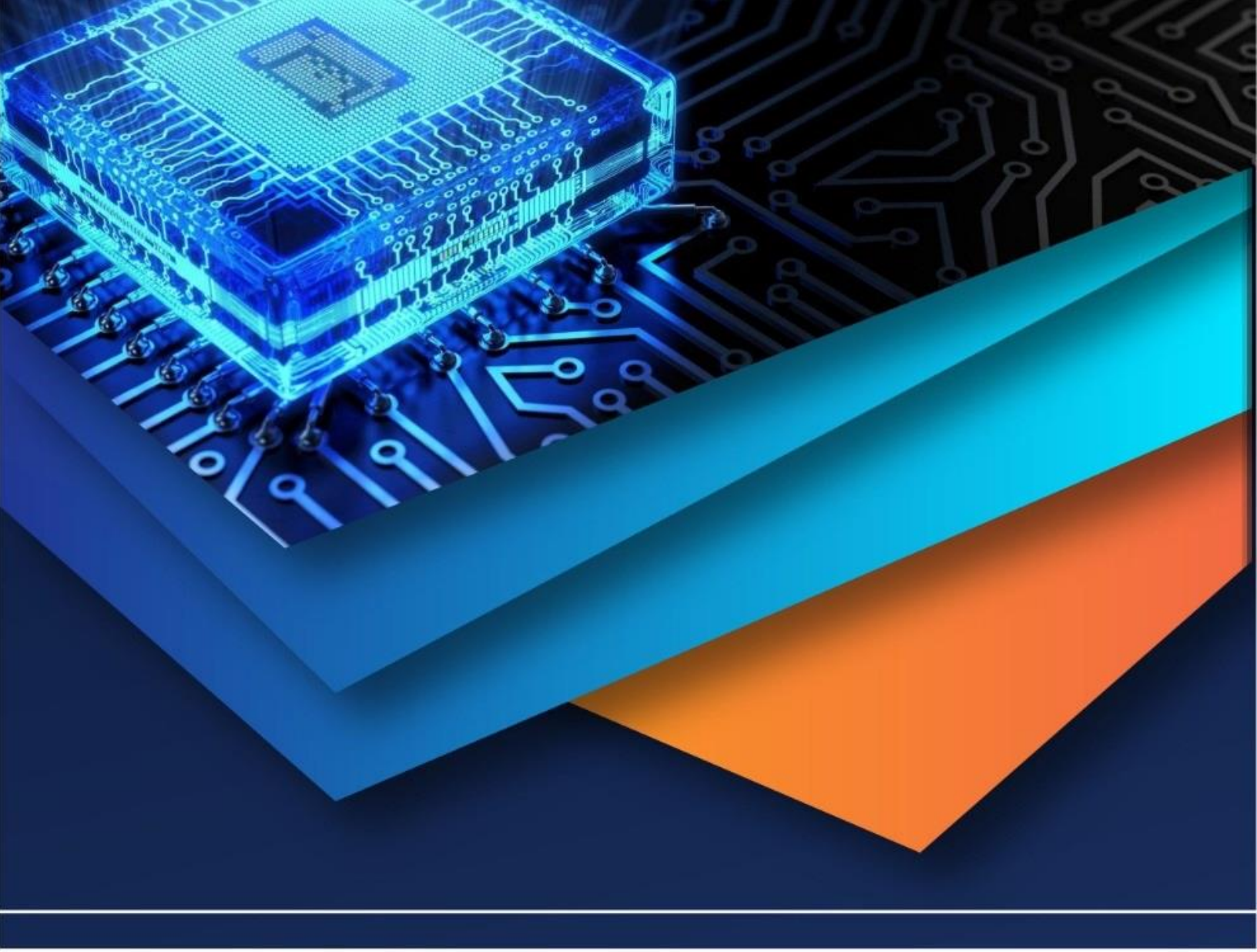
11) Antani and Agnihotri [17] created the data sets from scanned images, at 100 dpi, of printed Gujarati text as well as from various sites of internet from 15 font families. For training 5 font they created 10 samples each. The images were scaled up and then scaled down to a fixed size so that all the samples should be of same size i.e. 30x20. It does not have skew correction or noise removal features. For feature extraction the author computed both invariant moments and raw moments. Also image pixel values were used as features creating 30x20= 600 dimensional binary feature space. For classification the author used two classifiers, K-NN classifier and minimum hamming distance classifier. The best recognition rate was for 1-NN for 600 dimensional binary features space i.e. 67% 1-NN in regular moment space gave 48% while minimum distance classifier had the recognition rate of 39%. The Euclidean minimum distance classifier recognized only 41.33%.

IV. CONCLUSION

In this paper, we discussed the difference of Optical Character Recognition and Handwritten Character Recognition. The paper also gave basic knowledge and understanding of Gujarati language and its character set. To recognize a character of any language, complexity in character set of that language needs to be understood. Based on the survey of papers it can be concluded that though researchers have put efforts to build an HCR for Gujarati Character Recognition, they have not been able to achieve the desired success for entire character set. The results for individual characters, numerals or sample are promising but there still seems to be a lot of scope for improving the accuracy of recognized characters.

REFERENCES

[1] Audichya Milind & Saini R. Jatinderkumar, "A Study to Recognize Printed Gujarati Characters Using Tesseract OCR", International Journal for Research in Applied Science and Engineering Technology (IJRASET), ISSN:2321-9653, Vol. 5, Issue IX.
[2] Baheti M. J, Kale K. V., "Gujarati Numeral Recognition: Affine Invariant Moments Approach", 1st International Conference on Recent Trends in Engineering & Technology, Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering, ISSN: 2277-9477, pp.140 – 146.
[3] Baheti M. J., Kale K. V., " Recognition of Gujarati Numerals using Hybrid Approach and Neural Networks", International Conference on Recent Trends in engineering & Technology - 2013(ICRTET'2013), pp 12- 17.
[4] Chaudhari A. Shailesh and Gulati M. Ravi, "A Font and Size Independent OCR For Machine Printed Gujarati Numerals", National Journal of System and Information Technology, ISSN: 0974-3308, Vol. 3, Issue 1, pp. 70-78.
[5] Desai A. Apurva, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognition, Vol. 43, Issue 7, pp. 2582-2589.
[6] Desai A. Apurva, Patel Chhaya, "Gujarati Handwritten Character Recognition Using Hybrid Method Based On Binary Tree-Classifier And K-Nearest Neighbour," International Journal of Engineering Research & Technology,ISSN:2278-0181, Vol. 2, Issue 6.
[7] Dholakia Jignesh, Negi Atul, S Rama Mohan, "Zone Identification in the Printed Gujarati Text", Proceedings of the Eighth International Conference on Document Analysis and Recognition(ICDAR '05) ISBN:0-7695-2420-6, pp.272-276.
[8] Dholakia J, Yajnik A, Negi A, "Wavelet Feature Based Confusion Character Sets for Gujarati Script", proceeding of International Conference on Computational Intelligence and Multimedia Applications(ICCIMA), ISBN: 0-7695-3050-8, pp. 366-371.
[9] Jawahar C. V., Pavan Kumar M. N. S. S. K., Ravi Kiran S. S., "A Bilingual OCR for Hindi-Telugu Documents and its Applications", International Conference on Document Analysis and Recognition.
[10] Lehal G S, Singh C, "A Gurmukhi script recognition system" proceeding 15th International Conference on Pattern Recognition, IEEE Computer Society, Vol. 2, pp 557–560.
[11] Modi M, Macwan F., Prajapati R., "Gujarati Character Identification: A Survey", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, ISSN(online) 2321 – 2004, ISSN (print) 2321 – 5526, Vol. 2, Issue 2, pp 939- 943.
[12] Magare, S. S., Gedam. Y. K., Randhave. D. S., Deshmukh R. R., "Character Recognition of Gujarati and Devanagari Script : A Review", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 3, Issue 1, pp 3279- 3282.
[13] Maloo M, Kale K.V, "Gujarati Script Recognition: A Review", International Journal of Computer Science Issues (IJCSI), ISSN (Online): 1694-0814, Vol. 8, Issue 4, pp 480-489.
[14] Naik Vishal and Desai Apurva., "Online handwritten Gujarati character recognition using SVM, MLP, and K-NN", 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-6. 10.1109/ICCCNT.2017.8203926.
[15] Pal U, Chaudhari B.B., "Printed Devanagari Script OCR System", Vol.10, pp. 12-24.
[16] Patel Chhaya., Desai. A. Apurva, "Extraction of Characters and Modifiers from Handwritten Gujarati Words", International Journal of Computer Applications (0975 – 8887), Volume 73, Issue 3, pp 7-12.
[17] S. Antani, L. Agnihotri, "Gujarati Character Recognition", Fifth International Conference on Document Analysis and Recognition, ISBN:0-7695-0318-7, pp 418.
[18] Shah S K and Sharma A, "A Design and Implementation of Optical Character Recognition System to Recognize Gujarati Script using Template Matching", IE(I) Journal−ET ,Vol.86, pp. 44-49.
[19] Shah M Marmik, Parikh C Mehul, "A survey on Handwritten gujarati Character Recognition", International Journal for Technological Research in Engineering, ISSN(Online): 2347-4718, Vol. 2, Issue 8.
[20] Thaker H., Kumbharana. C.K., "Structural Feature Extraction to recognize some of the Offline Isolated Handwritten Gujarati Characters using Decision Tree Classifier", International Journal of Computer Applications, ISSN:0975 – 8887, Volume 99, Issue 15, pp 46-50.
[21] https://www.omniglot.com/writing/gujarati.htm



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY