Offline Handwritten Recognition of Curved Gujarati Consonants using GHCIR System

Arpit Jain
GLS University
Ahmedabad, Gujarat, India

ABSTRACT
Gujarati language has a variety of characters which makes the language diversified. The language has a total thirty-four consonants and eleven vowels that have different shapes. The characters have features like vertical lines, vertical lines with curves, only curve, curve with one loop and curve with two loops that differentiate the characters from one other. In this paper, the researcher has used a set of consonants which only has curves. Total eight (8) consonants 'ક' (k), 'ટ '(Ta), 'ડ ' (D), 'ર '(Ra), 'ળ '(Al), 'ફ '(Fa), 'ડ '(Da) and 'ઝ '(Jha) have only curve as a feature. These consonants have been taken into consideration in this research paper. For training the GHCIR system 1200 samples of each curved consonant has been collected, while the system has been tested with 100 samples of each curved consonant. The average accuracy achieved after processing for all the curved consonant is 84.25%. Further, the researcher has done the performance analysis of the Gujarati Handwritten Consonant Identification and Recognition (GHCIR) System using the Kappa coefficient. The average Kappa Coefficient justify that the system identifies all the Gujarati handwritten curved consonants with a substantial level of agreement.

Keywords
GHCIR System, Bit Combination Pattern, Kappa Coefficient, Thresholding

1. INTRODUCTION
India is a diversified country where people use multiple languages like Hindi, Gujarati, Marathi Punjabi and others for the purpose of written communication. In Gujarat, Gujarati is one of the most spoken languages which is used for the purpose of written and oral communication. Gujarati is spoken by more than fifty million people of India [12]. Gujarati language has been derived from Devanagari Script. Devanagari script is one of the oldest script of India [9]. It has been used to derive many languages like Hindi, Marathi, Punjabi and others. Gujarati language has a total 75 distinct legitimate shapes. These shapes represent various Consonants, Vowels, Numerals and Diacritics. The seventy-five distinct legitimate shapes include 59 characters and 16 diacritics. Fifty-nine characters are further divided into 34 Consonants also called as Vyanjans, 13 Pure Sounds and 10 Numerals and 2 additional rarely used characters [1, 4, 5]. The thirty-four Consonants are further divided into two Compound consonants and thirty-two Singular consonants. The consonants 'ક્ષ' (Ksa) and 'જ્ઞ' (Gna) are compound consonants whereas others are considered as compound consonants. A few Consonants of Gujarati language has vertical lines whereas others have one or two loops with a vertical line or horizontal line. The diacritics also termed as modifiers are those shapes which when used with any consonants and vowels create a different pronunciation and meaning. The Gujarati text has been available at many places in the form of old repositories and documents which need to be converted into the digitized format for better preservation. The process of converting a handwritten textual document into digitized format tends to the need of an OCR for that specific language. Gujarati OCR is available in the market whose major work is to convert only printed Gujarati text into the digitized format, but for handwritten Gujarati text, no OCR is giving promising accuracy. Handwritten character recognition is the field where researchers are working from the past many decades but the accuracy for them is not up to the mark. Document plays a major role in the recognition of a character. A document can be online and offline. A handwritten document written by using a digitizer pen is considered to be an online document where at the time of writing the classification of the character can be processed. But the document written by using a non e-device is considered as an offline document, which cannot be recognized while writing the document [7]. This type of documents is necessarily processed through a scanner and applies the steps of image processing. Figure 1 illustrates the handwritten consonants of Gujarati language in an offline document image.

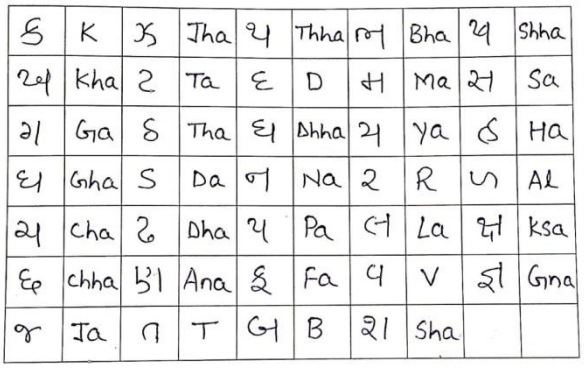

Figure 1: An offline Image of all the 34 Gujarati Consonants

2. GUJARATI HANDWRITTEN CONSONANT IDENTIFICATION AND RECOGNITION SYSTEM (GHCIRS)
GHCIR System is a six-step process towards the recognition of handwritten consonants [2]. The system has many algorithms inside it, which is incorporated at different phases. GHCIR System has divided into two phases, one is Training phase and another is Testing Phase. The architecture has been described with all the two phases and six processes. GHCIR

System has multiple processes to implement one after another i.e. Data Collection, Data digitization, Image segmentation, Bounding Rectangle, Pattern Generation and Pattern Matching [2].

The generic steps to be followed for the processing of a handwritten document image are Image Acquisition, Pre-Processing, Feature Extraction, Classification and Post-Processing which is also incorporated in GHCIR System by using multiple proposed algorithms.

Image acquisition is the first phase of image processing where the written document is transformed into digital format by considering few physical variables. This is the first phase of image processing. The data needs to be collected prior to the Image acquisition phase. Figure 2 illustrates the image for the consonant '**ક** '(k) converted into a digitized image format. The form is specifically designed for the data collection process which consists of twenty-five rows and twelve columns. There are three-hundred cells in each designed form. Thus, three-hundred samples of each consonant have been collected by using one-sample form. The size of each cell is 1cm * 1 cm. Total four forms have been used for the processing of each consonant. The consonants have been written by using multiple persons of age group 18-47 by using multiple colors and pinpoint pens. The data variation has been successfully maintained by using multi-input scenario from multiple personnel. The total numbers of collected data set for all the eight curved consonants are 9600 for the training of the GHCIR System.

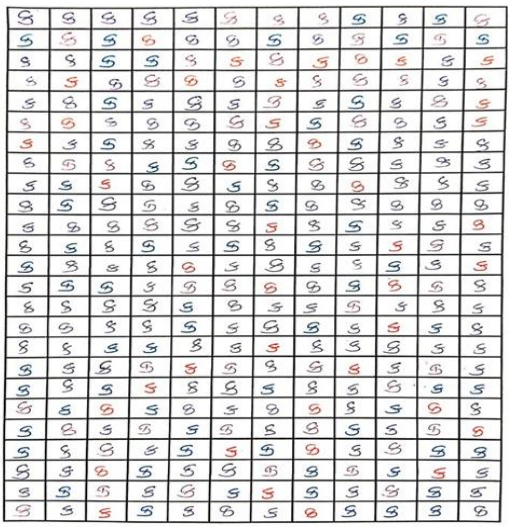

Figure 2: Sample Form for the Curved Consonant 'ક '(k)

Post Image Acquisition phase, the next is Pre-Processing, which majorly concentrates on noise removal, binarization, and edge detection, where the noise is removed from the input image. Noise can occur due to the scanner or miss-placed paper. Few consonants also have to go through by thinning process, if required. Removing such noise is necessary for smooth processing at a further level. In the pre-processing phase of the GHCIR System, the grayscale conversion has been applied where each pixel of an image is transformed to the average value of r, g, b pixels value. The thinning process has been also applied, if necessary.

The next phase is segmentation, which is the major phase for the line segmentation, word segmentation, and character segmentation. According to the content of the document the segmentation algorithm can be applied. In GHCIR System for Figure 1, only character segmentation is required, as the image already consists of 300 consonants written in appropriate cells. The segmentation algorithm results less than or equal to three-hundred individual images. Each individual image consists of one consonant image which is to be processed by using further phases of the image processing process. There is a provision of manual discard of consonants, only in case of incorrectly entered consonants in the form. The size of each consonant written in a segmented image is different. To make it all equal-sized images, apply bounding rectangle process on all the final segmented images. The bounding rectangle process converts all the different size segmented images into similar size images bounded in a rectangle. If the images are not bounded properly, there is a provision of a manual discard of the incorrect bounded images. Consolidate the final bounded images for further processing.

After segmentation and bounding rectangled process, the next phase in GHCIR system is Feature Extraction; each consonant has some unique feature. To extract the feature from the bounding rectangled image is the main feature of this phase. Transformation of an image into the binarized form and use it as features is the most abstraction mechanism. Thresholding is one of the mechanism by which the differentiation of background and foreground can be done by keeping a constant threshold value. In GHCIR System, the threshold value has taken into consideration is 127 for the separation of background and foreground of an image. The pixels of an image are now represented in the form of bits i.e. 1 and 0. One and zero is the bit structure, and a group of bits form a string which represents the transformed image. The bit combination pattern will be generated here by processing bounding rectangled image. Figure 3 represents a transformed image for the consonant '**ક** '(k) into one and zero. There are total of 8448 combination of bit patterns has been generated after applying segmentation and bounding rectangle processes for all the eight curved consonants.

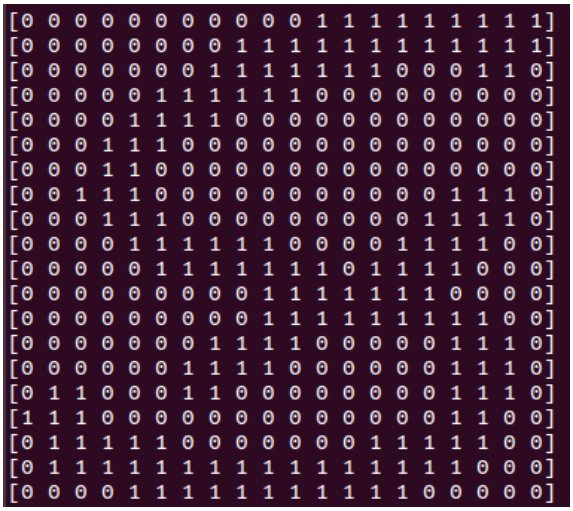

Figure 3: Transformation of 'ક '(k) in the form of 1 and 0

The next phase is Classification; classify the matched features of the tested pattern from the data repository is a major concentration of this phase. In GHCIR system, the bit combination pattern of consonants has been generated and stored in the data repository. Total numbers of generated bit

combination patterns are eight thousand four hundred forty-eight (8448) whereas; total numbers of unique bit combinations are seven thousand nine hundred ninety-five (7995). The ratio of generated bit combination pattern and unique generated bit combination patterns is listed in Table-I. The first column of Table-I represents the serial number; the second column represents the list of the curved consonant of Gujarati language; the third column is the English representation for the consonants written in the second column; the fourth column represents the number of images accepted for training the GHCIR System; and the fifth column represents the number of unique combination of generated bit patterns which is to be stored in the data repository of the GHCIR System as training dataset.

Table-I: Ratio of Accepted Images and Generated Unique Patterns

S. No	Gujarati Consonant	English Represent-ation of Gujarati Consonant	Final consonant Images Accepted For Training	Unique Patterns Generated
1.	ક	K	946	937
2.	ઝ	Jha	1101	959
3.	ટ	Ta	1095	1060
4.	ડ	Da	961	948
5.	ઇ	D	1146	1115
6.	ફ	Fa	873	708
7.	ર	Ra	1158	1112
8.	ળ	Al	1168	1156

Then the individual images for the test run has been sent to the GHCIR system, where the generated bit combination pattern is checked with each stored bit combination patterns of all the consonants and stored the matches with their accuracy in the form of key-value pair. The key-value pair consist key as the number of occurrence of matched consonant and value as the accuracy of matching. In the case of curved consonants for each test input, there are total seven thousand nine hundred ninety five (7995) comparison has been performed. The total number of comparisons are sixty three thousand nine hundred sixty (63,960) for the eight handwritten Gujarati curved consonants.

The final phase is post-processing, where the errors have been removed. Many times there is a possibility of getting the same number of occurrence for a consonant while testing. In the case of two similar occurrences of two different consonants, the post-processing phase comes into action. The consonant with the highest occurrence is considered as identified consonant but in case of the same number of occurrences, the average of accuracy for each matched pattern has been calculated and the consonant with the highest accuracy is considered as an identified consonant. Removing such errors and applying logic is the major task of the post-processing phase.

3. RESULTS AND OUTCOME

The system has been tested using one-hundred instances for each curved consonant. All these curved consonants come into the category of Joint Consonants. The GHCIR system has enabled with the training data set of seven thousand nine hundred ninety five (7995) bit combinations patterns of 1 and 0 whereas; the testing data set has a total data set of eight hundred (800) curved consonants. The results for each handwritten Gujarati curved consonants shown in Table-II, where the first column represents the serial number, the second column represents the listed curved Gujarati Consonant, the third column represents the English Representation of Gujarati Curved consonant; and the fourth column shows the total number of identified consonants out of one hundred (100).

Table-II: Achieved Accuracy for all the Eight Curved Handwritten Gujarati Consonants

S. No.	Gujarati Consonant	English Representation of Gujarati Consonant	Accuracy
1.	ક	K	96
2.	ઝ	Jha	82
3.	ટ	Ta	88
4.	ડ	Da	72
5.	ઇ	D	99
6.	ફ	Fa	39
7.	ર	Ra	98
8.	ળ	Al	100

The average accuracy for the entire eight curved consonants is 84.25%, which is a benchmark in the field of Gujarati handwritten recognition.

A graph representing the outcome is shown in Figure 4

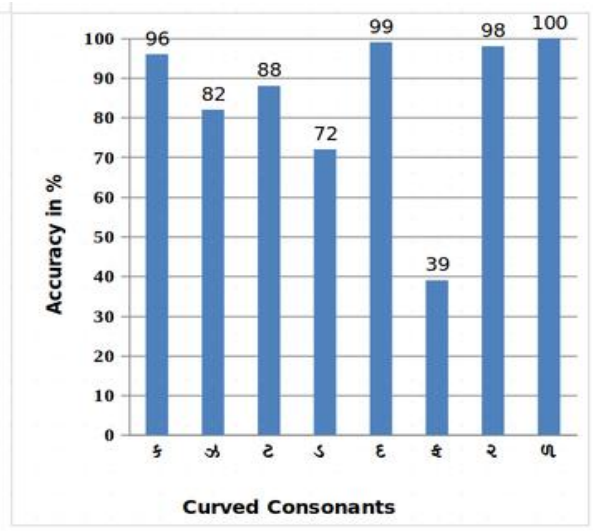

Figure 4: Identification Accuracy Chart for Curved Consonants of Gujarati Language

4. PERFORMANCE ANALYSIS OF GHCIR SYSTEM

The accuracy for the handwritten Gujarati curved consonants processed by GHCIR System is listed in the Table-II. To prove the working of GHCIR System the researchers has used kappa coefficient for the statistical analysis. The kappa coefficient is a technique which maps the inter-rater agreement. It is a statistical feature which is used to assess the quality measurement of the collected as well as identified data or a document, introduced in around 1960 in a paper published by Jacob Cohen [11]. Kappa classifies the collected and tested data set into four quadrants named as True Positive, True Negative, False Positive and False Negative [8, 12]. By analyzing the value of all the four quadrants for the eight curved consonants, Five of them comes in the range 0.81-1.00 i.e. Perfect Agreement, one of them comes in the range 0.61-0.80 i.e. Substantial Agreement, one is obtained with Moderate Agreement and one is with Fair Agreement. All the agreements display the quality of the GHCIR System. Table-III displays the calculated value of the kappa coefficient for each curved consonant. The first column represents the serial number; the second column represents the list of curved Gujarati consonants; the third column represents the English Representation of Gujarati Consonant; the fourth column represents the calculated kappa value for each curved Gujarati consonant.

Table-III: Performance Analysis by Kappa Coefficient

S. No.	Gujarati Consonant	English Representation of Gujarati Consonant	Kappa Coefficient Value
1.	ક	K	0.89
2.	ઝ	Jha	0.81
3.	ટ	Ta	0.70
4.	ડ	Da	0.58
5.	૬	D	0.96
6.	ફ	Fa	0.37
7.	ર	Ra	0.98
8.	ળ	Al	1.00

The average kappa coefficient for all the curved consonants is 0.79, which is a Substantial Agreement and agrees to the reliability of the GHCIR System.

5. CONCLUSION

In this paper, the researcher has trained and tested eight Gujarati consonants that has curved feature using a system known as GHCIR that has been developed by using python framework along with opencv2. For training nine thousand six hundred (9600) samples have been used and similarly for testing eight hundred (800) samples have been used. The results show that GHCIR System identifies the curved consonants with an average accuracy of 84.25% which as per the standard of research is an acceptable outcome. To prove the acceptance, the researcher has also applied Kappa analysis. The average result of the kappa coefficient is 0.79 and as per standards, it indicates that the consonants are identified with Substantial Agreement.

6. REFERENCES

[1] A. A. Desai, “Gujarati Handwritten Numeral OpticalCharacter Recognition Through Neural Network”, Pattern Recognition, ISSN: 0031-3203, Vol. 43, Issue 7, pp. 2582-2589, July 2010.

[2] A. A. Jain, H. A. Arolkar and C. Davda “Recognition of Offline Gujarati Handwritten Disjoint Consonants using Pattern Matching”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Vol.8 Issue-3, September 2019.

[3] A. Jain and H. Arolkar, “A Survey of Gujarati Handwritten Character Recognition Techniques”, International Journal of 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.

[4] H. Thaker and 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, ISSN: 0975 – 8887, Vol. 99, Issue 15, pp. 46-50, August 2014.

[5] M. J. Baheti and K. V. Kale, “Recognition of Gujarati Numerals using Hybrid Approach and Neural Networks”, International Journal of Computer Applications, ISSN:0975 -8887, and International Conference on Recent Trends in engineering & Technology – 2013 (ICRTET'2013), pp. 12- 17, 2013.

[6] M. J. Baheti and K. V. Kale, “Gujarati Numeral Recognition: Affine Invariant Moments Approach”, in First International Conference on Recent Trends in Engineering & Technology, Mar-2012 Special Issue of

International Journal of Electronics, Communication & Soft Computing Science & Engineering, ISSN: 2277-9477, pp. 140-146, 2012.

[7] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey", IEEE Transactions on PAMI, Vol. 22(1), pp. 63–84, 2000.

[8] S. M. Haley and J. S. Osberg, "Kappa coefficient calculation using multiple ratings per subject: a special communication". *Physical Therapy*, *69*(11), 970-974, 1989.

[9] G. Cardona and D. Jain, "The Indo-Aryan Languages", Routledge P. 115. ISBN978- 1-135-79710-2.

[10] Indo-Aryan languages, Encyclopedia Britannica Online, Retrieved 28 March 2020. https://en.wikipedia.org/wiki/Languages_of_India (On the basis of Population of 2018)

[11] Jacob Cohen Kappa Coefficient, "A coefficient of agreement for nominal scales". Educational and Psychological Measurement. 20 (1): 37–46. Retrieved on 28 March 2020. https://en.wikipedia.org/wiki/Cohen%27s_kappa#cite_note-5

[12] The Language Gulper, An Insatiable Appetite for Ancient and Modern Tongue. http://www.languagesgulper.com/ eng/Gujarati.html - retrieved on July 07, 2020.