International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Recognition of Offline Gujarati Handwritten Disjoint Consonants using Pattern Matching

Arpit A Jain, Harshal A Arolkar, Chirag S Davda

Abstract: ***Image processing is one of the most popular field nowadays. Recognition of the offline isolated handwritten characters is an area which got lot of attention within the field of Image Processing. Various techniques have been proposed in the area of online and offline handwritten character recognition (HCR). In future HCR is the key factor for the transformation of written or printed text into system understandable format. Thus, providing a boost to digitization era. Gujarati is one such language which has many challenges in creation of an accurate OCR. Researchers have achieved good accuracy in the field of online Gujarati handwritten character recognition. This paper introduces a pattern recognition system which is able to recognize isolated offline Gujarati handwritten characters with higher accuracy. Experiments have been done on sub set of (ક, ખ, ગ, ચ, છ) consonants using a total data set of 6750 handwritten consonants by different individuals. The experimental results achieved a markable contribution in the field of handwritten character recognition.***

Keywords: GHCRS Gujarati Handwritten Character Recognition System, Handwritten Character Recognition, OCR, Pattern Recognition.

I. INTRODUCTION

India is built with the complexity of multiple cultures, caste and languages. People speak various languages in the different regions of the country. Gujarati is one of the most popular language which is used for the official communication purpose in Gujarat region. Gujarati language has originated from an ancient Devanagari script. It is one of the oldest scripts that is used to derive multiple languages. In Gujarat there are variety of documents written in the Gujarati language which need to be digitized as time changes. Digitizing these texts with the use of keyboard is a tedious task. It takes undue long time and has possibility of introducing human errors. Thus, an accurate Gujarati Optical Character Recognition (OCR) system is the need of the day. OCR systems will reduce the time required for digitization, also the error introduced will reduce. OCR specially designed for recognizing handwritten documents generally termed as Handwritten character recognition (HCR). The characters can be classified as Online characters and Isolated Offline characters. “Fig. 1” shows isolated handwritten characters written by using any normal pen whereas “Fig. 2” shows the online isolated handwritten characters written by using light pen or some handheld device.

Fig. 1. Offline Characters

Fig. 2. Online Characters

Many algorithms have been defined with good enough accuracy for recognizing online characters, but the same cannot be stated for offline character recognition. Building an HCR for Indic languages is more complicated as compared to non-Indic language. The character set of most of the Indic language is complex and is twice in number then the non-indic languages. The character complexity arises with the use of matras, conjunct consonants and the shape. The complexity in Gujarati characters arise due to its curvature (ક), discrete shape (ઘ, ઙ, ઞ) and similarity of characters. Example numeral five (૫) and alphabet ‘Pa’ (પ), numeral two (૨) and alphabet ‘Ra’ (ર). The character set of Gujarati language is vast. It consists of 59 characters, and is divided into 34 singular consonants, 2 compound consonants, 13 vowels and 10 numerals [1, 3, 8].

Manuscript published on 30 September 2019
* Correspondence Author
Arpit A Jain*, Faculty of Computer Technology, GLS University, Ahmedabad, India. Email: arpit.jain@glsuniversity.ac.in
Harshal A Arolkar, Faculty of Computer Technology, GLS University, Ahmedabad, India. Email: harshalarolkar@glsuniversity.ac.in
Chirag S Davda, Student, GLS University, Ahmedabad, India. Email: chiragdavda007@gmail.com

These 34 singular consonants are also termed as ornamented sounds. The consonants can be combined with the vowels and can form compound characters known as Conjunct Consonants. A word written in Gujarati language can be combination of singular consonants, compound consonants or vowels.

II. CATEGORISATION OF GUJARATI CONSONANTS

Gujarati characters can be categorized by using multiple ways. First categorization can be based on the character that have horizontal straight line and vertical straight lines. Second categorization can be done by dividing all the consonants into two major categories i.e. joint characters and dis-joint characters. A joint character is the character which can be written without breaking the line (q, s, b). The dis-joint, character is formed by combining two different patterns. Out of 34 consonants available in the Gujarati language, 29 are joint characters, while 5 are considered as disjoint characters. The five disjoint characters in the Gujarati language are 'ગ', 'ઘ', 'લ', 'શ', 'હ'. Identification of dis-joint characters is complex when compared to the identification of joint characters.

III. ARCHITECTURE OF GHCR SYSTEM

GHCRS infers Gujarati Handwritten Character Recognition System, which is a proposed system for the identification of Gujarati handwritten consonants. "Fig. 3" shows the architecture of GHCRS. The proposed GHCRS is divided into two phases.

GHCRS is a six-step process towards the recognition of handwritten consonants. The process incorporates Data Collection, Data digitization, Image segmentation, Bounding rectangle, Pattern Generation and Pattern matching. All the processes of GHCR System are discussed below in brief.

Data Collection is one of the preliminary and required phase. The researcher has collected 1200 instances of each consonant using 4 predefined forms. The data collection form consists of tabular structure of 25 rows and 12 columns. The data has been generated by males and females in the age group of 18-25 years. They have used different colored inks and different point sizes when generating this data. A total of 6000 sample consonants have been collected.

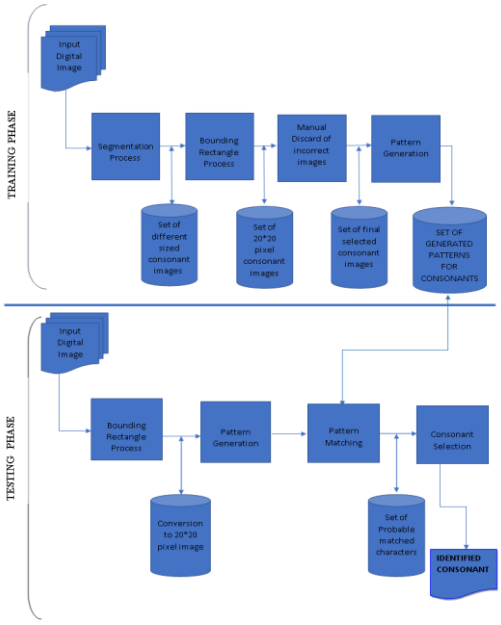

Fig. 3. Architecture of Proposed GHCR System

Data Digitization is the phase which convert the collected manual data into electronic format by using a third-party scanner at resolution 200 to 300 dpi. It is the second phase of the GHCR System for which the output is a digitized image. The image cleaning is one of the major concerns after taking input image. While scanning the predesigned filled designed from, it is possible to introduce noise or distortion in the image. To remove the noise and distortion in the input image the researchers have used third party scanner (camScanner) which works on the Guassian blur algorithm to filter out noise and edge detection algorithm. The edge detection algorithm is basically used for the cropping of an image, where only the useful and the highest perimeter part of an image can be considered for the further processing. The sheet used for the collection of data is represented in the "Fig. 4" in digitized format.

Image segmentation is another phase where the isolated images will be generated by processing the digitized form. Each digitized form consists of 300 isolated consonants and for the further processing set of individual images is required. Segmentation is the process of dividing the digitized form into multiple subparts. The output of segmentation process is set of images of given consonant in different sizes.

The images generated in segmented phase vary in size. To generate unique pattern a need to resize all images into a unique size arises. Hence each image generated in segmentation phase passes through a bounding process.

In Pattern generation phase each unique sized image is converted into 400-bits pattern. The proposed algorithm in this phase gives the multiple unique pattern for each consonant which is stored in file.

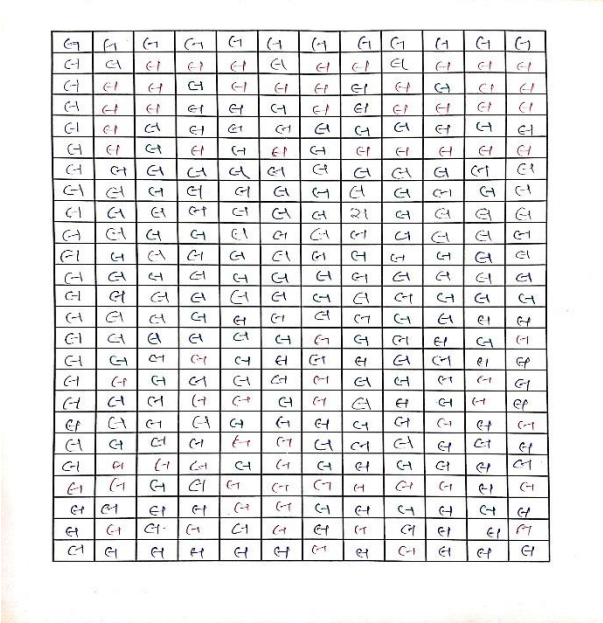

Fig. 4. Digitized A4 sized form for the dis-joint consonant 'l'

After the successful generation of patterns, the next step is to provide the test data images as an input to the GHCR System. The pattern generated by the test images is then compared with the unique patterns created during training phase.
A probability based consonant selection is then applied to generate a final consonant as outcome.

IV. RESULTS AND OUTCOME

The system has been tested using 150 instances of g, `, l, x, h consonants each. All the tested consonants belong to the dis-joint category. The result achieved by GHCR System for the five dis-joint consonants is listed in Table-I.
The first column is the serial no., the second column shows the character in Gujarati i.e. consonants of language Gujarati, the third column shows the pronunciation of character in English, the fourth column shows the number of total identified characters out of 150. Fifth column shows the achieved accuracy for all the input characters.

Table-I: Achieved Accuracy for The Gujarati Handwritten Dis-Joint Consonants

SNo.	Character	Character as pronounced in English	Total Identified	Accuracy
1	g	Ga	150	100.00
2	`	Ana	70	46.66
3	l	La	146	97.33
4	x	Sha	150	100.00
5	h	Ha	124	82.66

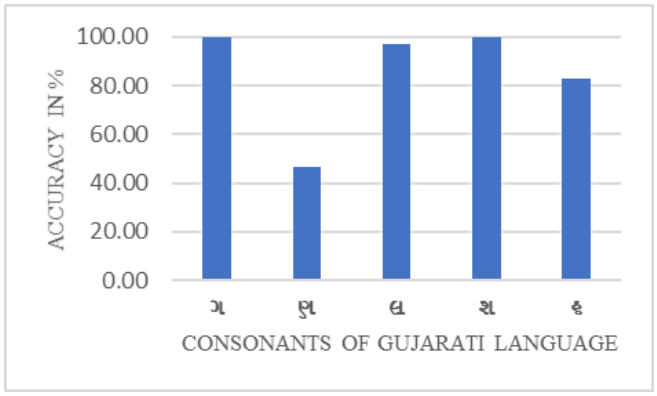

Fig. 5. Accuracy Chart of GHCRS for g, `, l, x, h

The achieved accuracy by processing the five consonants which includes dis-joint consonants, through GHCR System is 85.33%.

V. CONCLUSION

Pattern recognition is very useful for the identification of Isolated Offline Handwritten Gujarati Characters. As it is possible to generate a unique pattern for each character, which can be helpful to map the generated pattern of trained data with the generated pattern of test data. This paper proposed a system which have many algorithms inside it for the segmentation, pattern generation and pattern matching process. By applying all the algorithms together, the researcher has built a successful GHCR System. This paper is the first mile stone towards the recognition and processing of Isolated Gujarati handwritten characters. In future the author intends to test the same algorithm on entire consonant set of Gujarati language and try to achieve the acceptable accuracy. The present work seems to be a milestone in the field of OCR for Handwritten Gujarati dis-joint Characters with the achieved 85.33% accuracy of identification.

REFERENCES

1. 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.
2. 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.
3. Desai A. Apurva, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognition, Vol. 43, Issue 7, pp. 2582-2589.
4. 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.
5. Jain Arpit, Arolkar Harshal, "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.
6. 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.
7. Mehta Nikita, Doshi Jyotika, "A Review of Handwritten Character Recognition", International Journal of Computer Applications, ISSN:0975-8887, Volume 165 – No. 4, May 2017.

8. 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.

AUTHORS PROFILE

Arpit A Jain is working as an Assistant Professor at Faculty of Computer Technology, GLS University. He is having more than 09 years of teaching experience. He is pursuing Ph. D. in Image Processing. He did MCA from RGPV University and BCA from Vikram University. He has published total 2 research papers in International journals and 1 research paper in International conference.

Dr. Harshal A Arolkar is Professor and HoD, at GLS University. An ardent practitioner and teacher he possesses more than 20 years of teaching experience in Computer Science. He has published 3 books on computer science published by Wiley India and Dreamtech Press. He has also published and presented more than 30 research papers in international and national conferences and journals.

Chirag S Davda, student of MCA department, Faculty of Computer Technology, GLS university. He has completed his MCA from GLS University and BCA from M.K. Bhavnagar University.