Analysis of structural features and classification of Gujarati consonants for offline character recognition

Hetal R. Thaker[*], Dr. C. K. Kumbharana[**]

[*] Assistant Professor, Department of MCA, Atmiya Institute of Technology & Science, Rajkot, India
[**] Head, Department of computer science, Saurashtra University, Rajkot, India

Abstract- Wide range of applications and numerous other complexities involved in character recognition (CR) makes it a continuous and open area of research. Feature selection and classification plays major role in achieving higher accuracy for character recognition. In the era of digitization its compelling need to have CR system for regional script. This paper presents analysis of structural features and its classification for consonants of Gujarati script. Each character has certain characteristics which distinguishes it from other characters. Gujarati consonants are analyzed for eight such structural features and on the basis of it characters are categorized into twenty groups. Further Paper proposes decision table to classify characters based on structural features.

Index Terms- Character Recognition, Handwritten Character Recognition, Gujarati character recognition, Structural feature analysis

I. Introduction

In the world massive data are available on a paper. For preserving this data in electronic format it requires it to be digitized by scanner which will save it in an image format. Certain operations such as searching and updating is difficult if data exist in image format it requires converting image into editable form. Converting image into editable form requires certain image processing operations such as preprocessing, segmentation, feature selection, feature extraction, classification, and recognition.

Character recognition algorithm varies as diversities exist for language script and its characteristics such as 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, Devnagari script: line at top of character and matras around).

Many researchers have presented their work in the area of character recognition for English and Arabic script. Observation based on preliminary literature review indicates some work for South Indian script also, whereas very few research work is traced for character recognition in Gujarati script, which is an official language of Gujarat state, Western part of India. This paper focuses on analysis of structural feature and proposes analysis as decision table to classify offline Gujarati consonants. Paper is organized into different sections as previous work, Set of Gujarati consonants, Methodology for proposed work as structural feature selection, analysis in form of decision table for classification of Gujarati consonants.

II. Previous Work

Process of extracting unique information from binary image is called feature extraction in an area of character recognition. Feature extraction is an important step [1] [2] where it requires extracting features which helps system in deciding the character. [1] For optical character recognition methods which are used for feature extraction can be broadly classified into Global transformation and series expansion, Statistical feature, Geometrical and topological features. Geometrical and topological feature extraction is one of the popular method among researcher [2]. Character is analyzed for its constitution which includes some simple geometrical shape that includes horizontal, vertical and slanted line to complex curve i.e. C-curve, D-curve, U-curve and certain other characteristics like close region, end point, cross point etc.

Global and local properties of character identified by structural feature examination is a key to identify characters having distortions and style variations. Global and local properties like topology and geometry shapes in character. Suen et. al. have proposed many features in their work. [3].

Heutte et al. [4] has identified some structural features which include number of vertical and horizontal lines, intersections between the character and straight lines, holes position, end points, presence of loops number of intersections and junctions, number of loops.

For recognizing handwritten numeral several structural features extracted by Lee et. al [5]. Feature includes number of central, left and right cavities, location of each central cavity, crossing, and number of crossing with principal and secondary axes, pixel distribution.

In work presented by Amin et. al. [6] to recognize Arabic text, some structural features extracted are number of sub words, number and position of complimentary characters, number of loops in each peak, width and height of each pick.

Based on structural feature [7] letters are determined. Structural feature selected for extraction are loop, line etc. Further post-processing is carried out by comparing output with dictionary word to aid accuracy.

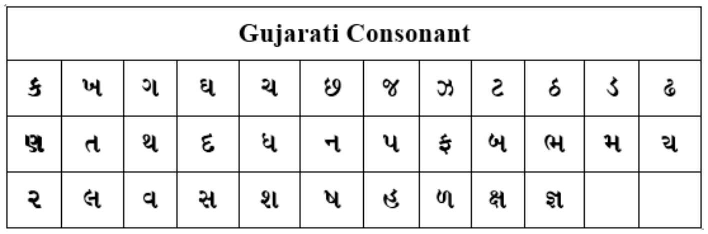
To recognize printed text of any size and font Kahan et.al.[8] have proposed set of structural feature i.e. number of holes, position and location of holes, crossing, concavities, end points and bounding box.

In an effort to recognize multi-font printed characters Rocha et. al.[9] have extracted convex arcs, singular points and their relationship as structural features.

Global features such as handwriting size, spacing between words, spacing between lines, arrangement of words, margin patterns, baseline patterns, line quality is identified for recognition of handwritten text. [10].

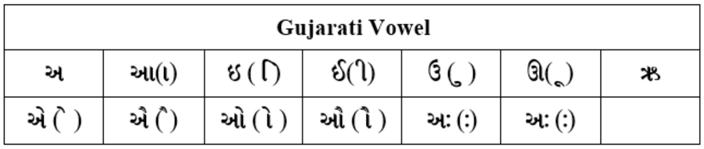
III. CHARACTERISTICS OF GUJARATI SCRIPT

Gujarati is an official language of western part of India. It has 34 consonants which are also known as ‘Vyanjans’ figure (1) and 13 vowels figure (2) called ‘Swar’ as below. This paper studies a feature for Gujarati consonants.



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 Consonants



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

Figure 2 Gujarati Vowels

IV. ANALYSIS OF STRUCTURAL FEATURE AND CATEGORIZATION

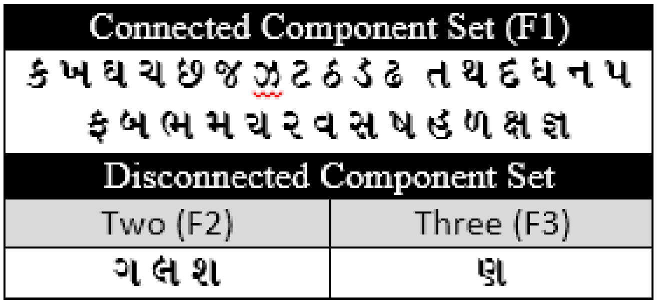
To extract structural feature it requires carefully analyzing shape of a shape and studying characteristics of shape. For proposed approach for classification following features are analyzed for Gujarati handwritten characters.

Structural features have many advantages such as font independent, size independent, works well even with shape distortion. Eight Structural features are selected as a base for categorizing consonants into twenty groups. F1, F2,…..F22 represents this groups.

A. Connected and Disconnected Components

Connected component indicates that structures that makes up the character is not broken i.e. all the lines and contours are connected.

In contradiction to that disconnected component indicates that character is broken into number of substructure and that group of these substructure makes a character. There can be one, two or three sub structures in other words disconnected components. Based on this structural feature three groups are categorized as shown in figure(3).

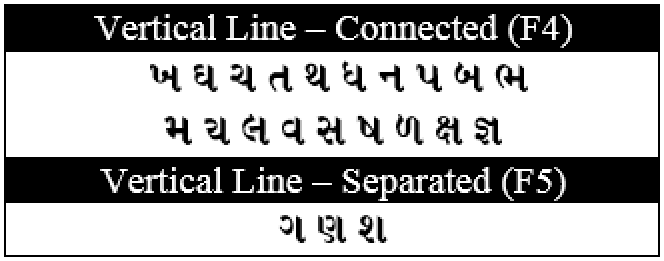


Connected Component Set (F1)	
ક ખ ઘ ચ છ જ ઝ ટ ઠ ડ ઢ ત થ દ ધ ન પ ફ બ ભ મ ય ર વ સ ષ હ ળ ક્ષ જ્ઞ	
Disconnected Component Set	
Two (F2)	Three (F3)
ગ લ શ	ણ

Figure 3 : Set of characters having connected and disconnected component components

B. Vertical Line

Character contains vertical line. Figure (4) represents two groups which are divided based on whether vertical line is a separate or connected.

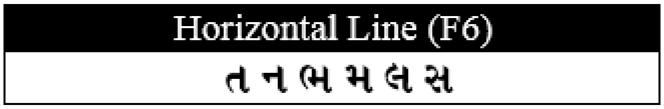


Vertical Line – Connected (F4)
ખ ઘ ચ ત થ ધ ન પ બ ભ મ ય લ વ સ ષ ળ ક્ષ જ્ઞ
Vertical Line – Separated (F5)
ગ ણ શ

Figure 4 : Set of characters having connected and separate vertical line

C. Horizontal Line

Character comprise of horizontal line. Figure (5) represents one group of such characters.

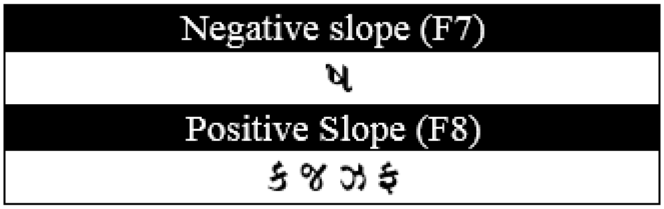


Horizontal Line (F6)
ત ન ભ મ લ સ

Figure 5 Set of characters having horizontal line

D. Diagonal Line

Diagonal line available in character where based on slope characters having diagonal line is divided into two categories i.e. character containing positive slope and negative slope or in another words right slant and left slant (figure 6).

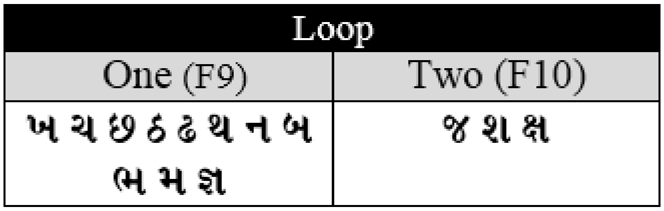


Negative slope (F7)
ષ
Positive Slope (F8)
ક જ ઝ ફ

Figure 6 Set of characters having positive and negative slope line

E. Close Region (Loop)

Loop designates any close region in character based on number of close region in character it is further categorized into two groups (figure 7).

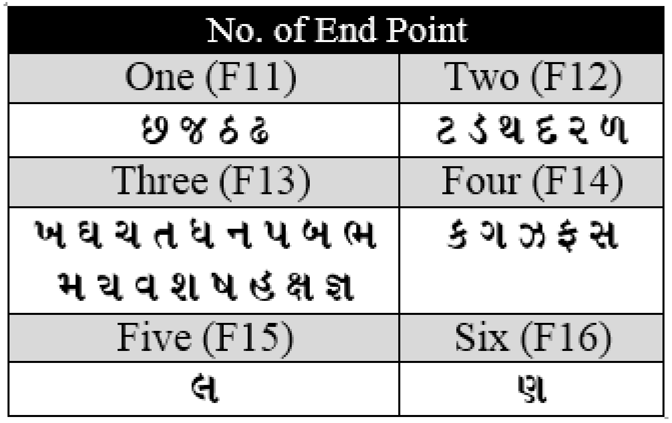


Loop	
One (F9)	Two (F10)
ખ ચ છ ઠ ઢ થ ન બ ભ મ જ્ઞ	જ શ ક્ષ

Figure 7 Set of characters having close loop

F. End Point

End point defines beginning and ending mark of that character. End points of every subcomponents of character having disconnected components are taken into account for categorization. Figure (8) represents six categories based on variation in end points.

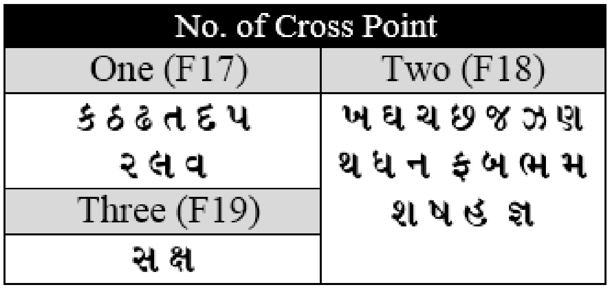


No. of End Point	
One (F11)	Two (F12)
છ જ ઠ ઢ	ટ ડ થ દ ર ળ
Three (F13)	Four (F14)
ખ ઘ ચ ત ધ ન પ બ ભ મ ય વ શ ષ હ ક્ષ જ્ઞ	ક ગ ઝ ફ સ
Five (F15)	Six (F16)
લ	ણ

Figure 8 Set of characters based on variation in end points

G. Cross Point

Cross point designates junction point or intermediate point where structure coincides. Figure (9) represents three groups categorized according to number of cross point found in characters.

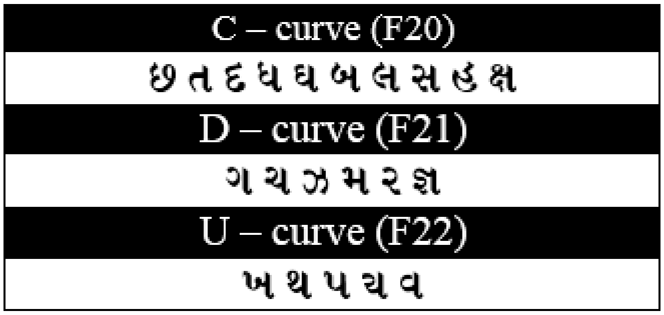


No. of Cross Point	
One (F17)	Two (F18)
ક ઠ ઢ ત દ પ ર લ વ	ખ ઘ ચ છ જ ઝ ણ થ ધ ન ફ બ ભ મ શ ષ હ જ્ઞ
Three (F19)	
સ ક્ષ	

Figure 9 Set of characters based on variation in cross (junction) points

H. C-curve, D-curve, U-curve

C – Curve designates the shape having curvature of English alphabet capital 'C'. Similarly 'D' and 'U' curve designates curve exist in English character capital 'D' and 'U' respectively. Some of the Gujarati characters contains 'C' and 'D' curve both so that categorization can be done into three groups as shown in figure (10).



C – curve (F20)
છ ત દ ધ ઘ બ લ સ હ ક્ષ
D – curve (F21)
ગ ચ ઝ મ ર જ્ઞ
U – curve (F22)
ખ થ પ ય વ

Figure 10 Set of characters having C-curve, D-curve and U-curve

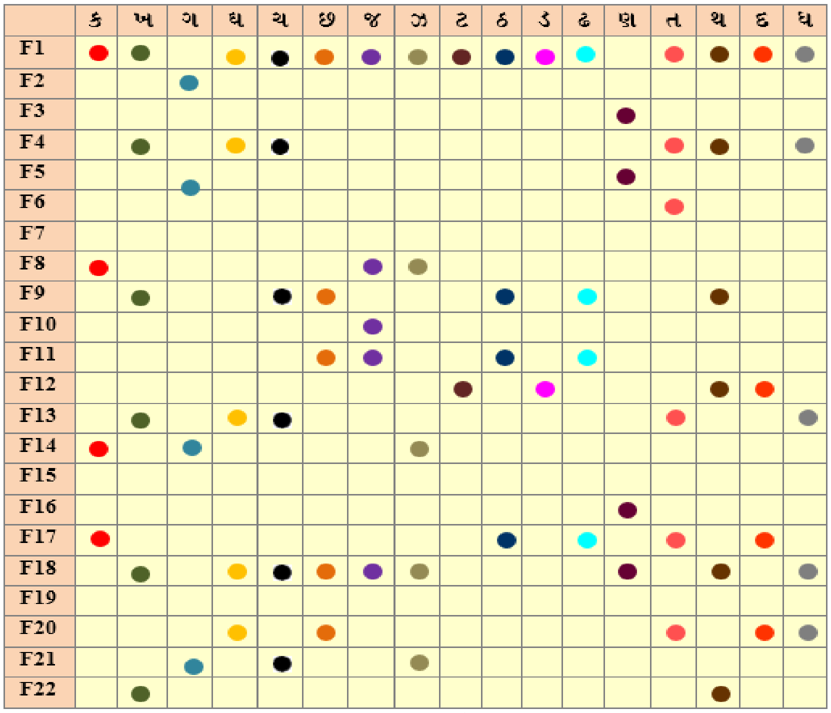
V. CLASSIFICATION OF GUJARATI CONSONANTS

Table 1 represents set of features and categorization of groups used to classify Gujarati consonants as presented in Figure 11(a) and (b) based on Table 1.

Connected / Disconnected component	
F1	Connected components
F2	Two Disconnected components
F3	Three Disconnected components
Vertical Line	
F4	Connected vertical line
F5	Disconnected vertical line
Horizontal Line	
F6	Horizontal Line
Slope Line	
F7	Negative slope line (left slanted)
F8	Positive slope line (right slanted)
Close loop	
F9	One close loop
F10	Two close loop
End point	
F11	One end point
F12	Two end point
F13	Three end point
F14	Four end point
F15	Five end point
F16	Seven end point
Cross point	
F17	One cross point
F18	Two cross point
F19	Three cross point
Type of curve	
F20	C – curve
F21	D – curve
F22	U - curve

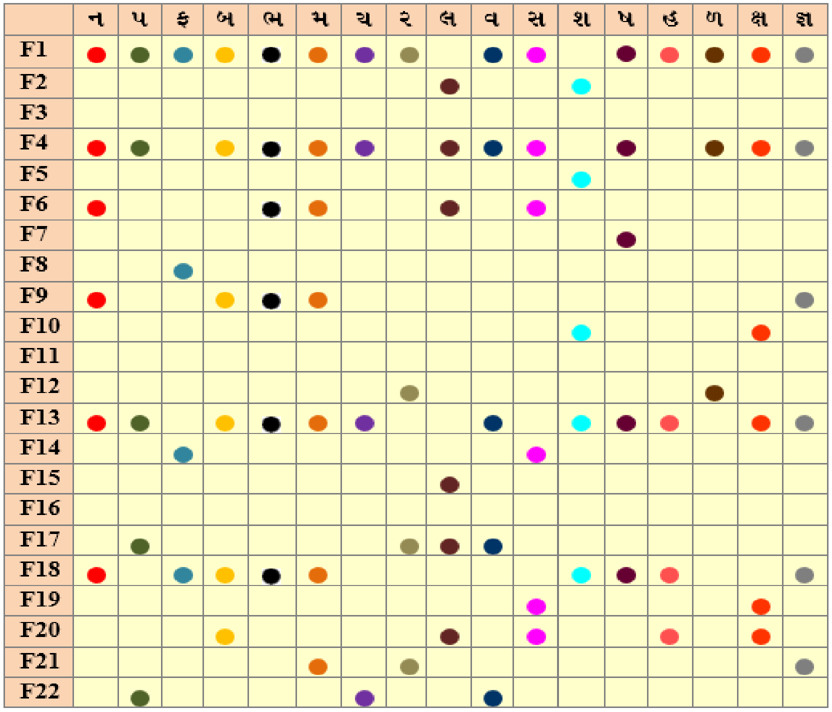
Table 1 Categorization based on feature sets

Figure 11(a) and (b) represents classification of Gujarati characters in form of table. In this table column represents Gujarati consonants and rows represents features. Intersection of row and column i.e. cell having dot mark indicates presence of that feature into that consonant. To obtain an information such as feature1 exist in how many consonants and that which consonants are they one has to trace a row. Whereas column needs to be traced to infer information about set of features for one consonant. In character recognition given an isolated consonant as input twenty one features will be examined for this image, presence of certain group of features marks a particular consonant which distinguishes it from other set of features.



	ક	ખ	ગ	ઘ	ચ	છ	જ	ઝ	ટ	ઠ	ડ	ઢ	ણ	ત	થ	દ	ધ
F1	●	●		●	●	●	●	●	●	●	●	●		●	●	●	●
F2			●														
F3													●				
F4		●		●	●									●	●		●
F5			●										●				
F6														●			
F7																	
F8	●						●	●									
F9		●			●	●				●		●			●		
F10							●										
F11						●	●			●		●					
F12									●		●				●	●	
F13		●		●	●									●			●
F14	●		●					●									
F15																	
F16													●				
F17	●									●		●		●		●	
F18		●		●	●	●	●	●					●		●		●
F19																	
F20				●		●								●		●	●
F21			●		●			●									
F22		●													●		

(a)



	ન	પ	ફ	બ	ભ	મ	ય	ર	લ	વ	સ	શ	ષ	હ	ળ	ક્ષ	જ્ઞ
F1	●	●	●	●	●	●	●	●		●	●		●	●	●	●	●
F2									●			●					
F3																	
F4	●	●		●	●	●	●		●	●	●		●		●	●	●
F5												●					
F6	●				●	●			●		●						
F7													●				
F8			●														
F9	●			●	●	●											●
F10												●				●	
F11																	
F12								●							●		
F13	●	●		●	●	●	●			●		●	●	●		●	●
F14			●								●						
F15									●								
F16																	
F17		●						●	●	●							
F18	●		●	●	●	●						●	●	●			●
F19											●					●	
F20				●					●		●			●		●	
F21						●		●									●
F22		●					●			●							

(b)

Figure 11 Classification of Gujarati Consonants (a) from '*ka*' to '*dha* '
(b) From '*na*' to '*gna*' based on structural feature s

VI. CONCLUSION

Structural features are useful in classification and recognition of characters irrespective of font or size. So as proposed paper used this method for classification of Gujarati consonants. Analyzing various structural features from Gujarati consonants presented in this paper it can be concluded that each consonant has unique structural feature which distinguishes it from other consonants. Classification approach presented here in form of decision table serves as a base for recognizing handwritten or printed Gujarati Consonants. As future work proposed analysis will be utilized for classifying and recognizing Gujarati characters.

REFERENCES

[1] N. a. Y.-V. F. Arica, "An Overview of Character Recognition Focused on Off-line Handwriting," IEEE Trans. On Systems, Man, and Cybernetics, vol. 31, no. 2, pp. 216-233, 2001.

[2] C. Suen, "Character Recognition by Computer and Applications in Handbook of Pattern Recognition and Image Processing," San Diego CA, ed. Young T.Y., Fu, K.S., Academic Press Inc, pp. 569-586.

[3] C. Y. Suen, M. Berthod and S. Mori, "Automatic recognition of hand printed characters- the state of the art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487, 1980

[4] L. Heutte, T. Paquet, J. V. Moreau, Y. Lecourtier and C. Olivier, "A structural/statistical feature based vector for handwritten character recognition", Pattern Recognition Letters, Vol. 19(7), pp. 629-641, 1998.

[5] L. L. Lee and N. R. Gomes, "Disconnected handwritten numeral image recognition", in the Proceedings of 4th ICDAR, pp. 467-470, 1997.

[6] A. Amin, "Recognition of printed Arabic text based on global features and decision tree learning techniques", Pattern Recognition, Vol. 33, pp. 1309-323, 2000.

[7] CEDAR. "Penman: Handwritten Text Recognition Project Description". Available Thein from http://www.cedar.buffalo.edu/Penman/description.htm.

[8] S. Kahan, T. Pavlidis and H. S. Baird, "On the recognition of printed characters of any font and size", IEEE Transaction, .Pattern Analalysis . Mach. Intell, Vol. 9, pp. 274-288, 1987.

[9] J. Rocha and T. Pavlidis, "A shape analysis model with applications to a character recognition system", IEEE Transactions on PAMI, Vol. 16(4), pp. 393-404, 1994

[10] G. Leedham and V. Pervouchine, "Validating the use of handwriting as a biometric and its forensic analysis", in the Proceedings of International Workshop on Document Analysis (IWDA), India, pp. 175-192, 2005.

AUTHORS

First Author – **Hetal R. Thaker**, Assistant Professor, Department of M.C.A, Atmiya Institute of Technology & Science, Rajkot, India, e-mail: hrt.research@gmail.com.

Second Author – Dr. C.K.Kumbharana, Head, Department of Computer Science, Saurashtra University, Rajkot, India,
e-mail: ckkumbharana@yahoo.com.

Correspondence Author – Hetal R. Thaker,
hrt.research@gmail.com, hrthaker@gmail.com.