7 RESULTS OF EXPERIMENTS, ANALYSIS AND DISCUSSION

We have approached the problem starting with developing model for offline Gujarati handwritten text recognition (OGHTR). The processes that we followed for OGHTR are preprocessing, segmentation and recognition. To recognize handwritten text from document image, segmentation is most vital stage. As our focus is towards document segmentation and recognition, we have selected minimum preprocessing steps for OGHCR system.

For segmentation and recognition phase of OGHTR different experiments are conducted and the experiments are set up with following hypothesis:

1) Segmentation process: as we have choice to follow holistic or analytical approach, based on literature survey and complexity of Gujarati writing system we decided to select analytical approach which has advantages of unlimited vocabulary. The aim of segmentation phase of OGHCR is to segment core character and diacritic marks from Gujarati text documents. The extensive experiment conducted on text dataset to evaluate segmentation approach. The result of segmentation is divided into core character segmentation and diacritic marks segmentation in following sections.

   To reduce the complexity of recognizing diacritical mark, we formulated segmentation process in such a way that diacritical marks are segmented with its position during the segmentation process which helps to recognize easily using decision tree classifier. Our much of the work is in the segmentation process.

2) Recognition process: to recognize core characters we have implemented IGHCR module which is capable of recognizing isolated Gujarati alphabet and numerals. According to literature survey there are wide varieties of features are available. The feature set which are used for printed characters will not

work for handwritten characters due to irregular shape and size of the characters. The feature set which are for handwritten English like language also not appropriate for Indian languages. So the right set of feature selection is major part of this research.

In this research we have suggested two novel features for Gujarati character recognition which are based on water reservoir and partial radial histogram.

7.1 RESULTS OF SEGMENTATION

The evaluation of segmentation result is based on visual criteria, which is through manual observation of segmentation output. The result of segmentation algorithm is divided into three parts, 1) result of line segmentation, 2) result of segmented component as core character and 3) result of diacritic marks extraction.

1) Result of line segmentation

In order to evaluate efficiency of OGHTR's segmentation method we conducted experiment on text dataset. The text dataset comprises of 107 handwritten Gujarati text documents containing 1135 lines. Pages contain text in one column and the number of text lines in document is minimum 5 to maximum 21. The text line considered to be segmented correctly if all its character correctly assigned to line. To identify it, the segmented component assigned different colours alternatively within two text lines as shown in Figure 31.

Table 7 Result of line segmentation

Number of document	Total number of lines	Correct line segmentation	Line estimation rates%
107	1135	1116	98.32%

The line segmentation result is displayed in Table 7. Out of 1135 text lines 1116 lines are correctly segmented. Our segmentation approach gives overall 98.32% line detection rate. The detection rate is defined as:

line detection rate. The detection rate is defined as:

$$Detection\ rate = \frac{The\ number\ of\ detected\ text\ lines}{The\ number\ of\ ground\ truth\ lines} \quad \text{Equation 13}$$

Equation 13

We consider the result of line segmentation for handwritten Gujarati text documents is satisfactory and encouraging. By observing error in line segmentation, it has been noticed that it occurred due to high variation in internal line skew or due to more initial spacing between page border and a text lines as shown in Figure 56 (a) and (b) respectively. Another reason for error in segmentation is occurred due to touching lines.

(a)

(b)

Figure 56 Line segmentation error

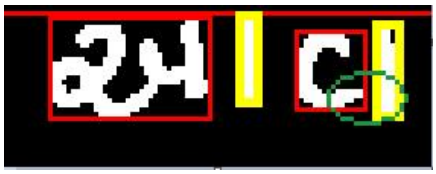
2) Result of segmented components as core characters

The result of segmentation is shown in Table 8. The same evaluation method is used to find character segmentation rate. In 107 documents 23889 ground truth characters are 23889 out of which 19683 characters are correctly segmented. The overall rate of core character segmentation is 82.39%.

Table 8 Overall character segmentation rate

Number of document	Number of characters	Correctly segmented characters	Character segmentation rates%
107	23889	19683	82.39%

The study of character segmentation error has been carried out. We found errors in character segmentation are due to following reasons:



(a)

(b)

(c)

Figure 57 Character segmentation error

i) Broken character: due to the quality of pen and paper causes broken characters. If length of broken edge is more than stroke width causes separation within the same character. This is depicted in Figure 57 (a). The broken edge (shown in green colour circle) in character "□ va ʊə" caused right part of character segmented as diacritic mark.

ii) Irregular character size: due to vide variation in character size with respect to average character width causes it to segmented as diacritic mark as shown in Figure 57 (b). The size of character "□ ṭa ţə" is smaller than average width of character segment it as diacritic mark.

iii) Joint characters: while writing writer joins two character which is not conjuncts causes identified it as single character as shown in Figure 57 (c).

3) Result of diacritic mark extraction

As described in process of segmentation, we segmented diacritic mark during the segmentation process and labelled it according to its position. The manual evaluation of diacritic mark is very tedious job as well as time consuming. Here we

have evaluated 12 documents for recognition of diacritic marks. The results are promising for diacritic mark it shows 81.07% segmentation rate shown in Table 9.

Table 9 Overall segmentation results of diacritic mark

Number of document	Number of diacritic mark	Correctly segmented diacritic mark	Inaccurate segmentation of diacritic mark	Accuracy
12	1997	1619	378	81.07%

7.2 RESULT OF IGHCR MODULE

To recognize isolated Gujarati handwritten character, neural network is used as classifier. We have conducted series of experiment to verify effectiveness of feature vector, effectiveness of removal of bottom right part. The results obtained are very promising.

7.2.1 Effectiveness of feature vector and results of IGHCR module

As described in section 6.2.3 different feature vectors are modelled to analyse the result of individual features as well as combine features. The feature vector set1 is modelled based on structural features; set2 is based on water reservoir concept and set3 based on partial radial histogram. The other feature vectors are modelled based on combining these features.

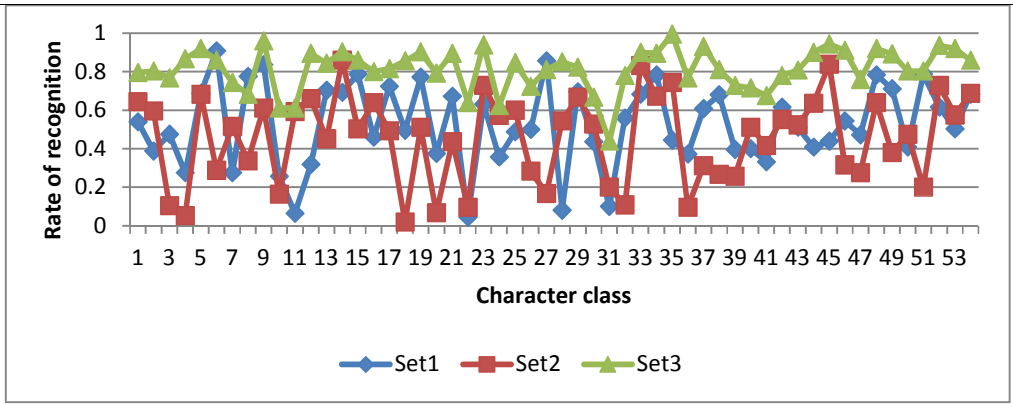


Figure 58 Results of individual feature vectors

The experiment conducted on each feature vector to analyse the effect of the features on overall recognition rate. The result of feature vectors set1, set2 and set3

is shown as chart in Figure 58. The feature vector set3 shows highest rate of recognition 81.5% amongst these three individual feature vectors. The study of chart observed that these feature vectors shows non-linear relation and hence it can be combined together to obtain better results.

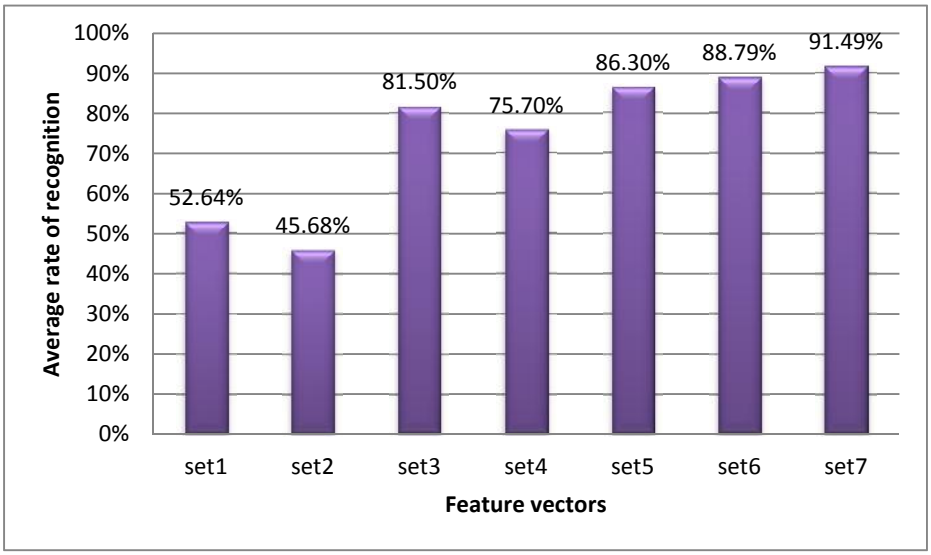


Figure 59 Performance comparisons of feature vector

The result of each of the feature vectors are shows as chart in Figure 59. Based on the results obtained it is observed that the features vector set7 shows gives highest rate of recognition for all the 54 classes that is 91.49%.

Table 10 Performance of feature vectors

Feature Vector	Rate of recognition in %			Rate of recognition greater or equal to			Character class number with recognition %	
	Min	Max	Average	70%	80%	90%	Min	Max
Set1	4.40	90.8	52.64	12	3	1	22	6
Set2	2.00	86.00	45.68	6	3	0	18	14
Set3	44.00	99.6	81.5	46	35	14	35	31
Set4	45.2	94.00	75.7	36	27	8	9	31
Set5	44.8	98.00	86.3	53	43	26	31	35
Set6	58.8	99.2	88.79	52	44	35	31	35
Set7	**69.2**	**100**	**91.49**	**53**	**50**	**38**	**31**	**35**

The comparison of results for each of the feature vector are shows in Based on the results obtained it is observed that the features vector set7 shows gives highest rate of recognition for all the 54 classes that is 91.49%. Along with the minimum, maximum and average rate of recognition, character class having more than 70%, 80% and 90% rate of recognition and character class number with minimum and maximum recognition rate.

Table 11 Performance of feature vectors

Feature Vector	Rate of recognition in %			Rate of recognition greater or equal to			Character class number with recognition %	
	Min	Max	Average	70%	80%	90%	Min	Max
Set1	4.40	90.8	52.64	12	3	1	22	6
Set2	2.00	86.00	45.68	6	3	0	18	14
Set3	44.00	99.6	81.5	46	35	14	35	31
Set4	45.2	94.00	75.7	36	27	8	9	31
Set5	44.8	98.00	86.3	53	43	26	31	35
Set6	58.8	99.2	88.79	52	44	35	31	35
Set7	**69.2**	**100**	**91.49**	**53**	**50**	**38**	**31**	**35**

The set7 feature vector is based on combination of all the extracted features. Out of all 54 character classes the recognition rate of 70.38% (38) characters are more than 90% which is encouraging. Also, 50 characters having recognition rate more than 80%.

The rate of recognition for each character class is shown as chart in Figure 59. The lowest recognition rate for character class 31 is 69.2% which is at the higher side. Recognition rate for individual characters with character class number, associated character and rate of recognition is shown in Table 12. The highest rate of recognition obtained for basic character is 99.20% for “ gña gnə”.

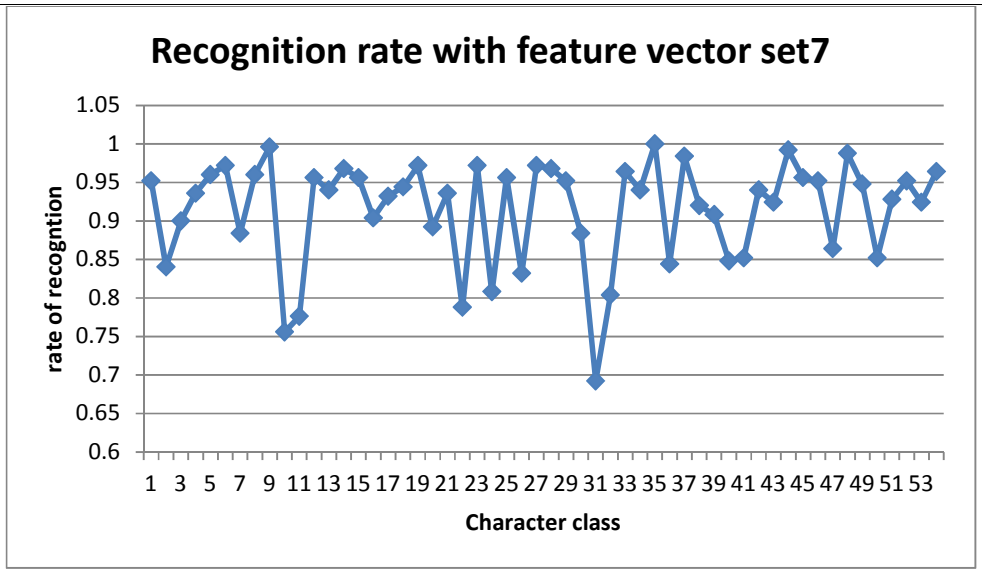


Figure 60 Recognition result with feature vector set7

Table 12 Result of characters recognition for isolated Gujarati characters

Class no.	Character	Rate of Recognition	Class no.	Character	Rate of Recognition
1	અ	**95.20%**	28	બ	**96.80%**
2	□	84.00%	29	ભ	**95.20%**
3	ઈ	**90.00%**	30	□	88.40%
4	ઉ	**93.60%**	31	□	69.20%
5	ઊ	**96.00%**	32	□	80.40%
6	ક	**97.20%**	33	લ	**96.40%**
7	□	88.40%	34	**Part1 લ**	**94.00%**
8	ગ	**96.00%**	35	**Part2 લ**	**100.00%**
9	**Part1 ગ**	**99.60%**	36	□	84.40%
10	□	75.60%	37	શ	**98.40%**
11	□	77.60%	38	**Part શ**	**92.00%**
12	છ	**95.60%**	39	સ	**90.80%**
13	જ	**94.00%**	40	□	84.80%
14	ઝ	**96.80%**	41	□	85.20%
15	ટ	**95.60%**	42	ળ	**94.00%**
16	ઠ	**90.40%**	43	ક્ષ	**92.40%**
17	□	**93.20%**	44	ત્ર	**99.20%**
18	□	**94.40%**	45	□	**95.60%**

19	□	**97.20%**	46	□	**95.20%**
20	Part1 □	89.20%	47	□	86.40%
21	□	**93.60%**	48	□	**98.80%**
22	□	78.80%	49	□	**94.80%**
23	□	**97.20%**	50	□	85.20%
24	□	80.80%	51	□	**92.80%**
25	□	**95.60%**	52	□	**95.20%**
26	□	83.20%	53	□	**92.40%**
27	□	**97.20%**	54	□	**96.40%**

The confusion matrix is studied to analyse the reasons for misclassification. The confusion matrix for each class is presented in Appendix-VI. Each entry of confusion matrix gives the count of how many times particular character is confused with other character.

Table 13 Confusion matrix analysis

Sr. No.	Character	Confused with
1.	**31 – □**	**11 – □ (13.20%), 26 – □ (6%)**
2.	**10 – □**	**24 – □ (13.20%)**
3.	**11 – □**	**31 – □ (12%)**
4.	**2 – □**	**3 – □ (14.40%)**
5.	**24 – □**	**10 – □ (13.60%)**
6.	**22 – □**	**11 – □ (5.20%) /31 – □ (4.8%)**

The confusion matrix helps to analyse nature of misclassification. The summary of highest confusing character is shown in Table 13. Based on confusion matrix Gujarati character "□" is highly confused with character "□" and "□". It shows that character class 31 that is "□" is confused 13.20% times with character class 11 ("□") and 6% times with character class 26 ("□"). Similarly character class 10 is confused

13.20% times with character class 24. Note that as our earlier discussion, we are not going to resolve confusion between characters which are exactly similar in shape. Those are character class 32 and 47 and character class 26 and 50.

Other character class which are mostly confused are 11 and 31, 2 and 3, 22 and 11. This is due to these characters have minor variation in their shape. As we are dealing with handwritten characters, the variation of shape is so minor that sometimes it is hard to recognize by human too. We find character class 10, 11, 22, 24, 26 and 31 confused with each other due to minor differences in shape from the right of the character.

7.2.2 Comparison of result with recent work on Gujarati character recognition

The comparison of result for Gujarati character recognition is shown in Table 14 presented by different authors. As described in table the result of our approach giving good results with higher number of classes and with more number of samples.

Table 14 Result comparisons of isolated Gujarati characters

Sr. No.	Authors	Character set	Classifier	Dataset size	Rate of Recognition
1	Vasant, A. R., et al. [29]	Numerals	Neural network	3900	88.79
2	Desai, A. A. [34]	Numerals	Neural network	265	82%
3.	Maloo, M., & Kale, K. V. [77]	Numerals	SVM	800	90.55%
4.	Fakir, M., et al. [45]	Numerals	Neural netword	600	80.33%
5.	Sharma, A., et al. [86]	Numerals	Neural network	12000	96.37%
6.	Desai A. A., Patel C. [27]	Consonants & Numerals	Hybrid classifier – Binary tree classifier and K-NN	NA	63.1%
7.	Prasad, J. R., et al [139]	Alphabet	ANFC	NA	68%
8.	Prasad, J. R., et al [28]	Alphabet	Template matching	NA	71.66%
9.	Desai, A. [72]	Alphabet	SVM with polynomial kernel	6898	86.66%
10.	**Our approach**	**Alphabet & Numerals**	**Neural network**	**13500**	**91.49%**
11.	**Our approach**	**Numerals**	**Neural network**	**2500**	**98.48%**

We have also experimented same approach only for numeral recognition using all feature vectors. Amongst all feature vector set5 given rate of recognition for Gujarati numeral is overall 98.48% which is higher amongst numeral classification. The recognition rate of each of the numeral is shown as chart in Figure 61. The highest rate of recognition is obtained is for numeral 8 (āṭhaṛo or āanth") which is 100%. The lowest recognition rate is 96% for 0 (mīṇḍuṃ or shunya).

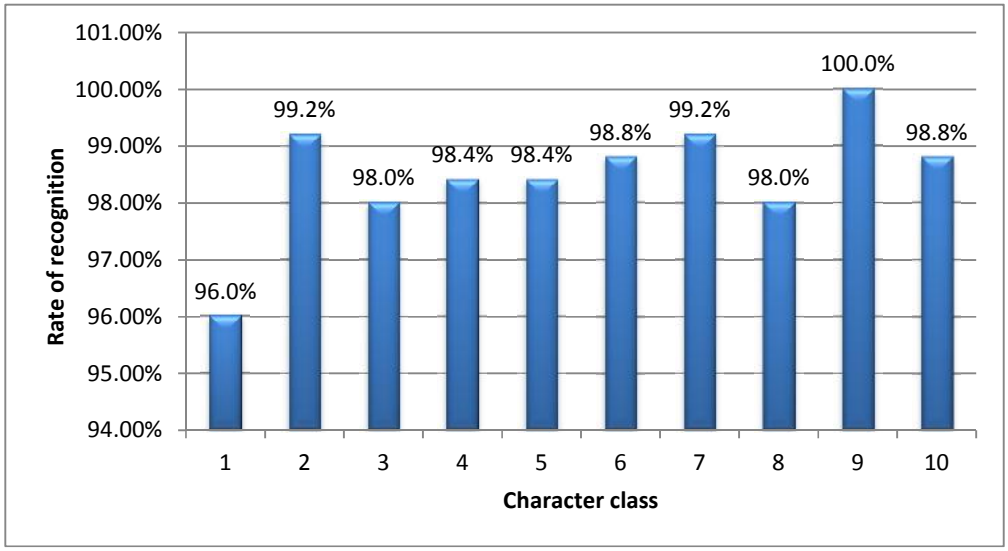


Figure 61 Rate of recognition for handwritten Gujarati numerals.

7.2.3 Effectiveness of extended line removal preprocessing step

To understand the effect of extended line removal step of preprocessing before feature extraction, we did experiment set up for it. That is features are extracted for feature vector set7 without removing extended line and with removing extended line. The result comparison is shown in Table 15 and Table 16. The overall recognition rate is increased by 1.41% and number of character class more than 90% increased by 11.11%.

Table 15 Result comparison before and after removal of extended line

Particular	Overall recognition rate	Character class more than 90% recognition rate
Before extended line removal	90.07%	32
After extended line removal	**91.49%**	**38**

We have also studied positive and negative effect of extended line removal preprocessing steps. Based on result 35 characters shows improvement in rate of recognition while, 19 characters shows decrement in rate of recognition.

Table 16 Effect on character class recognition before and after removal of extended line

Increase recognition rate for number of class	Decreases recognition rate for number of class
35	19

7.3 RESULTS OF OGHTR

In recent day few researchers started exploring research in handwritten Gujarati text recognition but as far as we know nobody yet published results of Gujarati text recognition. Overall rate of recognition is shown in Table 17 for correctly segmented character from 12 documents. The overall rate of recognition for achieved is 71.97%.

Table 17 Overall character recognition rate for text document

Number of documents	Correctly extracted core character from segmentation	Correct recognition of core characters	Rate of recognition
12	1638	1179	71.97%

The result shows that there is still need improvement in segmentation process along with recognition of more character classes.

7.4 DISCUSSION

The segmentation approach proposed in this research has been evaluated for line segmentation, character segmentation and diacritic mark segmentation. Based on experiment, results are promising.

For recognition of Gujarati character we have implemented two types of novel features, which is major contribution of this thesis and experiment conducted to evident that these features gives better results.

The overall accuracy obtained for handwritten Gujarati character recognition is 91.49% for 54 character class. The same technique is implemented for recognizing Gujarati numerals which give overall 98.48% rate of recognition. The results of characters as well as numerals are compared with recent results.