

## ABSTRACT

In the current digital era, due to very heavy usage of computers, smartphones and computing devices in all aspect of human life, it is expected that information/document has to be in e-form. So, such devices should recognize native languages to help common people to perform their daily tasks. Also, the paper is a very comfortable and feasible medium to store data, there is a great demand for the software techniques that can automatically extract, analyse and store information from the physical handwritten documents for later retrieval. But communication with the computer using this natural means of paper is not possible due to high variation of handwriting and large character set for the Indian languages like Gujarati. Handwritten text recognition for the Indian scripts is still in the need of maturity for some of the languages like Gujarati.

Recognizing handwritten character of Indian scripts are more challenging due to its large character set and nature of character shapes. In the presented work we have formulated offline handwritten text recognition system capable of recognizing isolated characters and text for Indian languages. The aim is to investigate the Handwritten text recognition for Gujarati script that can be generalized for Indian scripts. In this research, we have model the architectural solution which consisting of preprocessing, segmentation and recognition phases. Standard handwritten text document dataset is not available for the experiment. We have built dataset consist of 107 documents.

Preprocessing of image document is essential for reduction in data and for improving quality for better processing. Also, information in text document is best represented as foreground and background. We have identified preprocessing phase task as converting image to grayscale, removing noise, binarization, cropping and skew detection and correction. After the preprocessing text document image is transform into binary image which is then segmented in next phase.

Segmentation is the most challenging phase for text recognition system. In scripts like Gujarati, have diacritic marks corresponding to vowels and used in combination of consonants makes segmentation problem more challenging. Hence

the selection of appropriate segmentation strategy is very important step in construction of text recognition system. One of the major objectives of this work is to study different approaches used in text segmentation and address issues in segmentation of Gujarati handwritten text document.

Segmentation approach used in this research is based on combination approach. It uses partial projection histogram for line detection combined with RLSA to strengthen the histogram. Knowledge of constructing Gujarati Akshara is used for line tracing using parameters obtained from document after applying connected component analysis. The output of segmentation classified to core characters, and diacritic marks. The performance of line segmentation for Gujarati text document is 98.32%, character segmentation is 82.39% tested on 107 documents comprise of 1135 lines. We used structural characteristics and position for recognition of diacritic marks using tree classifier.

Isolated Gujarati handwritten character recognition module is implemented to recognise core characters using feed forward neural network. The dataset consisting of 13500 sample images of isolated Gujarati characters is used to recognize 54 character class consists of 49 basic Gujarati alphabets, numerals, and 5 character parts. In preprocessing of IGHCR, we have limited preprocessing task based on the noise in dataset. The new features namely partial radial histogram and water reservoir area methods are used in this research for character recognition. We have modelled different features vector to measure effectiveness of features. The overall highest recognition rate obtained is 91.49% for 54 classes.

# TABLE OF CONTENTS

<b>REGISTRATION DETAILS .....</b>	<b>II</b>
<b>CERTIFICATE .....</b>	<b>V</b>
<b>UNDERTAKING .....</b>	<b>VI</b>
<b>COPYRIGHT .....</b>	<b>VII</b>
<b>ACKNOWLEDGMENT .....</b>	<b>IX</b>
<b>ABSTRACT .....</b>	<b>XI</b>
<b>TABLE OF CONTENTS .....</b>	<b>XIII</b>
<b>LIST OF TABLES .....</b>	<b>XVII</b>
<b>LIST OF FIGURES .....</b>	<b>XVIII</b>
<b>LIST OF EQUATION .....</b>	<b>XXI</b>
<b>ABBREVIATION.....</b>	<b>XXII</b>
<b>1 HANDWRITTEN TEXT RECOGNITION FOR INDIAN LANGUAGES .....</b>	<b>1</b>
1.1 Introduction.....	1
1.2 OCR and its type .....	3
1.2.1 OCR: History .....	3
1.2.2 Type of OCR.....	4
1.3 General architecture for handwritten text recognition.....	6
1.4 The Gujarati script .....	9
1.5 Challenges in handwritten text recognition.....	11
1.6 Need of work .....	13
1.7 Aim and objectives.....	16
1.8 Statement of the problem.....	17

1.9 Thesis overview .....	17
<b>2 LITERATURE REVIEW FOR HANDWRITTEN TEXT RECOGNITION .....</b>	<b>20</b>
2.1 Literature review on segmentation .....	20
2.2 Literature review on features and classifiers .....	28
2.3 Review on handwritten Gujarati OCR .....	31
2.4 Discussion .....	34
<b>3 OFFLINE GUJARATI HANDWRITTEN TEXT RECOGNITION (OGHTR).....</b>	<b>35</b>
3.1 The architecture of OGHTR .....	36
3.2 Dataset .....	38
3.2.1 Gujarati handwritten text dataset.....	39
3.2.2 Isolated Gujarati character dataset .....	40
3.3 Discussion .....	43
<b>4 PREPROCESSING .....</b>	<b>45</b>
4.1 Image acquisition .....	46
4.2 Grayscale conversion .....	47
4.3 Contrast adjustment .....	48
4.4 Noise removal.....	49
4.5 Binarization.....	50
4.6 Cropping .....	52
4.7 Skew detection and correction .....	54
4.8 Discussion .....	56
<b>5 SEGMENTATION .....</b>	<b>58</b>
5.1 Character height and width estimation .....	60
5.2 Estimate tentative midlines .....	62
5.3 Tracing and labelling .....	65

5.3.1 Finding first component.....	68
5.3.2 Processing for left diacritic mark.....	69
5.3.3 Processing for core character.....	69
5.3.4 Processing for bottom diacritic mark .....	70
5.3.5 Processing for right diacritic mark.....	70
5.3.6 Processing for top diacritic marks .....	72
5.3.7 Finding next component .....	72
5.4 Discussion .....	73
<b>6 RECOGNITION.....</b>	<b>75</b>
6.1 The Recognition of diacritic mark.....	75
6.2 Recognition of core characters.....	78
6.2.1 Preprocessing .....	81
6.2.2 Feature extraction .....	83
6.2.3 Feature vector .....	91
6.2.4 Recognition engine .....	93
6.3 Text generation.....	101
6.4 Discussion .....	103
<b>7 RESULTS OF EXPERIMENTS, ANALYSIS AND DISCUSSION.....</b>	<b>104</b>
7.1 Results of segmentation.....	105
7.2 Result of IGHCR module .....	108
7.2.1 Effectiveness of feature vector and results of IGHCR module.....	108
7.2.2 Comparison of result with recent work on Gujarati character recognition .....	113
7.2.3 Effectiveness of extended line removal preprocessing step .....	114
7.3 Results of OGHTR.....	115

7.4 Discussion .....	116
<b>8 CONCLUSION .....</b>	<b>117</b>
<b>BIBLIOGRAPHY .....</b>	<b>119</b>
<b>APPENDICES .....</b>	<b>135</b>
APPENDIX I: Consonants in Gujarati character set.....	135
APPENDIX II: Vowels in Gujarati character set.....	136
APPENDIX III: Digits in Gujarati character set .....	137
APPENDIX IV: Diacritics in Gujarati.....	138
APPENDIX V: Example of conjuncts in Gujarati.....	139
APPENDIX VI: Confusion matrix for handwritten Gujarati character recognition	
140	
<b>LIST OF PUBLICATIONS .....</b>	<b>144</b>

## LIST OF TABLES

Table 1 Sample size of Gujarati isolated characters.....	43
Table 2 Diacritic mark with respect to its position.....	75
Table 4 Structural analysis of diacritic marks .....	76
Table 6 Sample entry of features end points and its location .....	85
Table 7 Reservoir area for Gujarati characters .....	90
Table 8 Feature vector model .....	93
Table 9 Result of line segmentation .....	105
Table 10 Overall character segmentation rate .....	106
Table 11 Overall segmentation results of diacritic mark.....	108
Table 12 Performance of feature vectors.....	109
Table 13 Performance of feature vectors.....	110
Table 14 Result of characters recognition for isolated Gujarati characters.....	111
Table 15 Confusion matrix analysis .....	112
Table 16 Result comparisons of isolated Gujarati characters .....	113
Table 17 Result comparison before and after removal of extended line .....	115
Table 18 Effect on character class recognition before and after removal of extended line.....	115
Table 19 Overall character recognition rate for text document.....	115

## LIST OF FIGURES

Figure 1 Classification of OCR .....	5
Figure 2 General architecture of HTR.....	7
Figure 3 The parts of Gujarati “Akshara” .....	11
Figure 4 OCR development by TDIL for Indian scripts.....	16
Figure 5 Handwritten printed text document .....	21
Figure 6 Horizontal projection .....	21
Figure 7 Binary image after RLSA .....	22
Figure 8 Offline Gujarati Handwritten Text Recognition System (OGHTR) .....	36
Figure 9 Handwritten Gujarati text document image .....	40
Figure 10 Isolated Handwritten Gujarati Characters .....	41
Figure 11 Sample datasheet.....	42
Figure 12 Preprocessing steps in OGHTR.....	46
Figure 13 A scanned document image .....	47
Figure 14 Text document converted to grayscale.....	48
Figure 15 The process of Median filter.....	50
Figure 16 Binarized handwritten document .....	51
Figure 17 Binary cropped document .....	54
Figure 18 Document image before and after skew correction.....	56
Figure 19 Block diagram of segmentation process .....	59
Figure 20 Effect of erosion with vertical line structuring element .....	60
Figure 21 Document image after horizontal and vertical run-length .....	63
Figure 22 Document image divided into 4 strips .....	64
Figure 23 (A) Horizontal Projection histogram of 1 <sup>st</sup> strip (B) Tentative line detection .....	64
Figure 24 Joint right matra િ (d rgha-ajju) with characters “વ va və”, “ર ra rə”, “બ ba bə”, and “ન na nə”.....	67
Figure 25 Character બ (ba bə) with joint <i>dīrgha-ajju</i> “બ”.....	71



Figure 26 Character ં (na nā) with joint <i>dīrgha-ajju</i> “લ” .....	71
Figure 27 Water reservoir for “લ” .....	71
Figure 28 Water reservoir for “લ” .....	71
Figure 29 Character “બ (ba bā)” after removing <i>dīrgha-ajju</i> .....	72
Figure 30 Character “લ (na nā)” after removing <i>dīrgha-ajju</i> .....	72
Figure 31 Output of segmentation process .....	73
Figure 32 Tree classifier for diacritic mark recognition .....	77
Figure 33 The model for isolated Gujarati handwritten character recognition (IGHCR) .....	79
Figure 34 Preprocessing steps of IGHCR .....	81
Figure 35 Character image after the binarization for Gujarati “ક ka kā”, “જ ja dઝઞ”, “મ ma mā” & “ર ra rā” .....	82
Figure 36 Character image after the thinning process for Gujarati “ક ka kā”, “જ ja dઝઞ”, “મ ma mā” & “ર ra rā” .....	82
Figure 37 Gujarati characters “અ a ā” “દ da dā” “પ pa pā” and “મ ma mā” with extended line .....	83
Figure 38 removal of extended line from Gujarati characters “અ a ā”, “દ da dā”, “પ pa pā” and “મ ma mā” .....	83
Figure 39 Structuring element use for detecting end points .....	85
Figure 41 End points of Gujarati character “ક ka kā” and “ર ra rā” .....	85
Figure 42 Structuring element used for branch point detection .....	86
Figure 43 Branch points in Gujarati character “ મ ma mā” .....	86
Figure 44 Loop in character “ મ ma mā” .....	87

Figure 45 a) Structuring element for horizontal line detection b) Structuring element for vertical line detection .....	87
Figure 46 Top reservoir for character “૫ પા પઠ” .....	89
Figure 47 Radial histogram .....	91
Figure 48 Partial radial histogram for character image “૫ પા પઠ” from top-left corner .....	91
Figure 49 Feature vector selection process .....	92
Figure 50 Human nervous system .....	94
Figure 51 Schematic diagram of formal neuron .....	95
Figure 52 Cyclic architecture .....	97
Figure 53 Acyclic architecture .....	97
Figure 54 A typical feedforward neural network .....	98
Figure 55 Training in neural network using supervise learning .....	99
Figure 56 Two layer feed forward neural network .....	100
Figure 57 Line segmentation error .....	106
Figure 58 Character segmentation error .....	107
Figure 59 Results of individual feature vectors.....	108
Figure 60 Performance comparisons of feature vector .....	109
Figure 61 Recognition result with feature vector set7 .....	111
Figure 62 Rate of recognition for handwritten Gujarati numerals. ....	114

## LIST OF EQUATION

Equation 1.....	20
Equation 2.....	48
Equation 3.....	51
Equation 4.....	54
Equation 5.....	66
Equation 6.....	66
Equation 7.....	87
Equation 8.....	95
Equation 9.....	95
Equation 10.....	96
Equation 11.....	96
Equation 12.....	96
Equation 13.....	105

# ABBREVIATION

<b>Abbreviation</b>	<b>Full form</b>
ACH	Average Character Height
ACW	Average Character Width
ANFC	Adaptive Neuro-Fuzzy Classifier
ANN	Artificial Neural Network
CC	Connected Component
CEDAR	Centre of Excellence for Document Analysis and Recognition
CENPARMI	The Centre for Pattern Recognition and Machine Intelligence
CLAHE	Contrast Limited Adaptive Histogram Equalization
DCT	Discrete Cosine Transform
DeitY	Department of Electronics And Information Technology
DPI	Dots Per Inch
EL	Endpoints and its Location
HCR	Handwritten Character Recognition
HTR	Handwritten Text Recognition
HWR	Height to Width Ratio
ICDAR	International Conference in Document Analysis and Recognition
IGHCR	Isolated Gujarati Character Recognition
JPEG	Joint Photographic Experts Group
k-NN	K-Nearest Neighbours
LH	Length of Horizontal Line
LV	Length of Vertical Line
MLP	Multilayer Perceptron
MNIST	Modified National Institute of Standards and Technology
NB	Number of Branch Points

NIST	National Institute of Standards And Technology
NL	Number of Loops
NLP	Natural Language Processing
OCR	Optical Character Recognition
OGHTR	Offline Gujarati Handwritten Text Recognition
PCA	Principal Component Analysis
PCR	Printed Character Recognition
PNG	Portable Network Graphics
PRH	Partial Radial Histogram
RGB	Red Green Blue
RLSA	Run Length Smoothing Algorithm
SE	Structuring Element
SVM	Support Vector Machine
SW	Stroke Width
WRA	Water Reservoir Area