



Contents

1	Introduction and Motivation	1
1.1	Introduction	1
1.2	History Of OCR	3
1.3	Image Acquisition	4
1.4	Preprocessing	11
1.5	OCR for Indian Scripts : State of the Art	24
1.6	Gujarati Script from OCR Perspective	25
1.7	Motivation	26
1.8	Organization of Thesis	27
2	Document Segmentation : A Mathematical Approach	29
2.1	Document Layout Analysis	29
2.2	Line Segmentation	32
2.3	Word Segmentation	33
2.4	Connected Component Extraction	35
2.5	Depth of Segmentation	40
2.6	Conclusions	42
3	Zone Boundary Identification : Slope based approach	43
3.1	Mathematical Preliminaries	43
3.2	Zone Boundary Identification : State-of-the-Art	44
3.3	Proposed Approach	45
3.4	Conclusions	51
4	Zone Boundary Identification : Use of Dynamic Program- ming	53
4.1	Mathematical Preliminaries	53
4.2	The Need	60
4.3	Proposed Approach	61
4.4	Conclusions	66
5	Feature Extraction	67
5.1	Introduction	67
5.2	Mathematical Preliminaries	68
5.3	Image Scaling	69
5.4	Features	70

6	Classifier Design	75
6.1	Introduction	75
6.2	Mathematical Preliminaries	76
6.3	Classifier	78
6.4	Experiment Setup	81
6.5	Results and Discussion	82
6.6	Conclusions	84
7	Finite State Machine based Text Generation	85
7.1	Mathematical Preliminaries	85
7.2	Text Generation	88
7.3	Special Cases	90
7.4	Conclusions	91
8	Results and Discussion	92
8.1	Preliminaries	92
8.2	OCR Testing	93
8.3	Results and Discussion	95
8.4	Conclusions and Future Work	98
	Publications	99
	Unicode Chart for Gujarati	101
	Bibliography	102
	Index	106