Abstract

This report discuss the methodology and implementation of two significant module segmentation & recognition part of Optical Character recognition system for scanned image. This methodology consists of three steps: At first step a segmentation approach is used in order to detect text lines, words and characters. In this step the most decisive part was to segment the characters whose regions are overlapped. We have solved this problem with a new approach we called curve segmentation. In second step these segmented characters are then used for recognition to determine the expected sequence of characters. We extracted some features for Bangla characters then examine a simple pattern-recognition system designed and simulated using KNN.

Acknowledgments

Throughout this thesis, we have received support from many people whom we wish to acknowledge here. First and foremost, we are deeply indebted to our principal supervisor Dr. Muhammed Zafar Iqbal for his advice and for being a ceaseless motivator during the last 6 month. He has always encouraged us to keep an open mind and to push the envelope at all times. We would also like to show our gratitude to our co-supervisor Md. Sabir Ismail & Md. Saiful Islam for sharing their pearls of wisdom with us during the course of this research. And we are so grateful to our supervisor & co-supervisors for arranging special meeting in every week and give their valuable time to us, which have boosted our workflow.

TABLE OF CONTENTS

Page

ABSTRACT I

ACKNOWLEDGEMENT I

TABLE OF CONTENTS II

LIST OF TABLES V

LIST OF FIGURES VI

1. INTRODUCTION 1

1.1 What is an OCR 1

1.2 Necessity of Bangla OCR 1

2. OUR PREVIOUS WORK 3

2.1 Basic Steps of Bangla OCR 3

2.2 Segmentation 4

2.2.1 Straight Line Segmentation 5

2.2.2 Curve Line Segmentation 7

3. SEGMENTED CHARACTER EXTRACTION 10

3.1 Finding Connected Components 10

4. FEATURE EXTREACTION & RECOGNITION 14

4.1 Our Observation and Analysis 14

4.2 Number of pixel in Horizontal and Vertical Histogram......................................16

4.3 Distance from different side of image...18

4.4 Skeleton of Characters 19

4.5 Corner of Character Image..20

4.6 Sub-Zoning Approach..21

5. POST PROCESSING...25

5.1 Correction of the writing sequence 25

5.2 Word Correction & Sentence Correction 29

6. RESULT ANALYSIS AND DISCUSSIONS 30

6.1 Error Rate of Recognition part...30

6.2 Compare with previous work...31

6.3 Time Analysis 34

7. CONCLUSION...36

7.1 Conclusion...36

7.2 Future Work 36

REFERENCES 37

List of Tables

		Page
Table 1	Error characters	30
Table 2	Number of Error For Different Scanned Image	31
Table 3	Time to segmentation & recognition for Different Number Of Words	34

List of Figures

		Page
Figure 2.1(a)	Steps of an OCR	4
Figure 2.2.1(a)	Straight Line segmentation	5
Figure 2.2.1(b)	Problem Of ‘ণ’ has been solved	6
Figure 2.2.1(c)	Problem of splitting গ , ণ are solved with our New Approach	6
Figure 2.2.1(d)	Failure case of straight line segmentation	7
Figure 2.2.2(a)	Coinciding problem of characters and modifiers	8
Figure 2.2.2(b)	Curve Segmentation	9
Figure 3.1(a)	Unexpected noise as a result of segmentation on matra line.	10
Figure 3.1(b)	Extracting Segmented Character with connected component	11
Figure 3.1(c)	After Extraction of each component	11
Figure 3.1(d)	After Extraction of each component	12

Figure 3.1(e)	Difference of Two different method of finding connected component	12
Figure 4.1(a)	Characters with same pattern in left side area	14
Figure 4.1(b)	Characters with almost similar outer shape	16
Figure 4.2(a)	The case where Histogram method fails to recognize correctly	17
Figure 4.2(b)	Similar character where each row and column contains approximately the same number of black pixels.	
Figure 4.2(c)	Histogram method fails to recognize mirrored image exactly	17
Figure 4.3(a)	Distance from Border Method unable to access the middle part of any character, so middle part remain unreachable.	18
Figure 4.4(a)	After making test character ‘য’ thinner and thicker	20
Figure 4.5(a)	Character where corner detection fails	21
Figure 4.6(a)	Illustration of counting black pixel in every subzone	22
Figure 4.6(b)	Ambiguous case of ‘**ব**’ and ‘**য**’	22
Figure 4.6(c)	Shifting Problem if upper ‘**ল**’ is shifted to right 3 bits then the two image would be exactly same.	24
Figure 5.1(a)	Sequence of Recognition	25
Figure 5.1(b)	Sequence of Recognized Characters	26
Figure 5.1(c)	Sequence of Recognized Characters	27
Figure 5.1(d)	Disconnected components	27
Figure 5.1(e)	Before and after processing ‘আ–কার’ and ‘দাঁড়ি’	28
Figure 6.2(a)	Previous Output for an ideal image	32
Figure 6.2(b)	Present Output for an ideal image	32

Figure 6.2(c) Previous Output for a Scanned image 33

Figure 6.2(d) Present Output for a Scanned image 33

Figure 6.2(e) Comparison between previous and present error rate 34

Figure 6.3(a) Number of words VS Time taken for segmentation & recognition Graph 35