Universiti Teknologi MARA

THE DEVELOPMENT OF AUTOMATED SEGMENTATION FOR THE FTMSK OFFICIAL LETTERS IN XML

Muhammad Muhaimin Bin Mohd Isa

Thesis submitted in fulfillment of the requirements for
Bachelor of Science (Hons) Information Technology
Faculty of Information Technology And
Quantitative Science

April 2005

DECLARATION

I certify that this thesis and the research to which it refers are the product of my own work and that any ideas or quotation from the work of other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline

APRIL, 2005

MUHD MUHAIMIN MOHD ISA
2002610078

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

LIST OF TABLES viii

LIST OF FIGURES ix

ABSTRACT x

CHAPTER 1 INTRODUCTION

1.1 Background 1

1.2 Problem Statements 2

1.3 Objectives of the Research 3

1.4 Scope of the Research 3

1.5 Significance of the Research 3

CHAPTER 2 LITERATURE REVIEW

2.1 The Document Structure and the Retrieval 5

2.2 Document Segmentation 6

2.3 The Algorithm Techniques 7

2.3.1 Searching 7

2.3.2 String Processing 8

2.3.3 Brute-Force String Matching 8

2.3.4 The Advantages and Disadvantages of using Brute-Force String Matching 9

2.4 The XML Evaluation 9

2.5 Formal Letters 12

2.6 Automated Segmentation from Digital Images 13

2.6.1 Experimental Results 14

CHAPTER 3 RESEARCH APPROACH AND METHODOLOGY

3.1 Research Methodology 15

3.2 Project Definition 15

3.3 The Prototype Architecture 16
3.3.1 The Input 16
3.3.2 The System 16
3.3.3 The System Output 17
3.4 Data Collection 17
3.4.1 The Interview 17
3.4.2 The Letters 18
3.5 Data Analysis 19
3.5.1 Letter Segmentation 19
3.6 Project testing 20
3.7 Project Implementation 20
3.8 Project Documentation 20
3.8.1 The End User Documentation 21

CHAPTER 4 PROJECT CONSTRUCTION

4.1 The Prototype Development 22
4.1.1 The User View Algorithm 23
4.1.2 The System View Algorithm 24
4.2 Label Testing Result 25
4.3 The Output Sample 26
4.4 The System Documentation 29

CHAPTER 5 RESULTS AND ANALYSIS

5.1 System testing Observation 30
5.2 Experiment Implementation 30
5.3 Result Analysis 35

CHAPTER 6 CONCLUSION AND RECOMMENDATION

6.1 Conclusion 36
6.2 Recommendation 36

ABSTRACT

The letter document has their own format, which consists of many parts. In order to process the document, the project has developed a prototype to allow the existence of content based document. This is important to divide the document into smaller, recognized labels that are intensive and flexible for managing, editing, and extracting. The target of this thesis is to apply the standard of official letter for the system, as well as to develop the algorithm which will segment the letter documents, and convert to XML documents. The main software used is Visual Basic 6.0. The project was estimated by evaluating the similarity of the contents in letter document and the output of XML document. The prototype is tested by using approach of manual checking. About 20 samples of letters have been tested to verify the prototype efficiency. The result of the experiment shows that 30% of the output can not be viewed due to the runtime errors. The findings of the research include the interviews and surveys appointed, the study of related topics and problems from the internet, books, and other information medium. The data was collected from the FTMSK administration department, and the testing was done manually using researcher's own workstations and hardware.