

WORD EXTRACTION

SITI MUNIRAH BINTI ABDUL KUDUS

This thesis is submitted as partial fulfillment of the requirements for the award of the Bachelor of Electrical Engineering (Hons.) (Electronics)

Faculty of Electrical & Electronics Engineering
Universiti Malaysia Pahang

ABSTRACT

In content-based image processing system, describing and extracting word is a key question. This project is an approach to extracting word characters from images using image processing technique of MATLAB software, which in this system developed contain two phases. Those phases are preprocessing phase, where system used binarization and smoothing technique, while in second phase which is feature extraction phase, system used morphological technique. The word extraction eliminates noise and the distortions on the same document to remain the important characters. The result generated by this system is 87.5% satisfactory with the ability to read the valueable words and documents. Thus, the system and method is effective and feasible.

ABSTRAK

Dalam rangkaian sistem pemrosesan imej, penjelasan dan pengekstrakan perkataan adalah menjadi persoalan utama. Projek ini adalah suatu pendekatan untuk mengekstrak perkataan daripada sesuatu dokumen berbentuk imej dengan menggunakan teknik pemrosesan imej, MATLAB, di mana sistem yang dibina ini mengandungi dua fasa. Fasa yang pertama adalah fasa pra-pemrosesan, di mana sistem menggunakan teknik 'binari' dan teknik 'pelembutan', manakala dalam fasa kedua iaitu fasa pengekstrakan hadapan, sistem menggunakan teknik 'morphologi'. Pengekstrakan perkataan ini membuang perkara-perkara yang tidak penting dalam sesuatu dokumen dan mengekalkan dokumen-dokumen penting di dalamnya. Hasil yang diperolehi oleh sistem ini adalah 87.5% memuaskan, dengan kebolehan membaca perkataan dan dokumen penting. Dengan itu, sistem dan teknik ini adalah efektif.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	TITLE	i
	DECLARATION	ii
	DEDICATION	iii
	ACKNOWLEDGEMENT	iv
	ABSTRAK	v
	ABSTRACT	vi
	TABLES OF CONTENTS	vii
	LIST OF TABLES	x
	LIST OF FIGURES	xi
	LIST OF APPENDICES	xii
1	INTRODUCTION	1
	1.1. Background	1
	1.2. Problem Statement	2
	1.3. Objectives of the Project	5
	1.4. Project Scope	5
	1.5 Thesis Outline	6

2	LITERATURE REVIEW	7
	2.1 Background	7
	2.2 Concept of Word Extraction	7
	2.3 Previous Work on Word Extraction	9
	2.3.1 Text Extraction	10
	2.3.2 Word Extraction	12
	2.3.3 Word Character Extraction	12
	2.4 Image Processing	18
	2.4.1 Elements of Image Processing	16
	2.4.2 Level of Image Analysis	20
	2.5 Image Processing Technique	25
	2.5.1 Binarization Technique	26
	2.5.2 Smoothing Technique	27
	2.5.3 Contour Extraction	29
	2.6 Differentiation between word characters	31
	2.6.1 English / Latin word characters	31
	2.6.2 Arabic word characters	32
	2.6.3 Chinese word characters	34
	2.6.4 Tamil word characters	34
3	METHODOLOGY	37
	3.1 Background	37
	3.2 Flow of Methodology	38
	3.3 Preprocessing Phase	40
	3.3.1 Binarization Technique	41
	3.3.2 Smoothing Technique	41
	3.4 Feature Extraction Phase	41
	3.4.1 Edge Detection	42
	3.4.2 Morphology	42

4	RESULT AND DISCUSSION	46
	4.1 Introduction	46
	4.2 Result	46
	4.3 Result Analysis	50
	4.3.1 Image & Analysis of Result	50
	4.3.2 Problem and Suggestion	56
5	CONCLUSION	58
	5.1 Conclusion	58
	5.2 Future Recommendation	58
	REFERENCES	60
	APPENDIXES	
	APPENDIX A –FIGURE OF G.U.I	65
	APPENDIX B – MATLAB CODING	68

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Summirization on Previous Work	25
2.2	Summarization on preprocessing techniques	32
2.3	Summarization on the differences characters	39
4.1	Accuracy rate of result calculation	41

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
1.1	Souvula's	3
1.2	Otsu's	4
1.3	Niblack's	5
2.1	Irregular pyramid construction process	12
2.2	Result on Pyramid Model	12
2.3	Result on Zhe Wang using Voronoi Diagram	14
2.4	Result on 3D neighborhood graph model	15
2.5	Block diagram of Image Analysis	22
2.6	Structuring Elements	25
2.7	Otsu's technique on global threshold	29
2.8	Recognizing using smoothing technique	30
2.9	Edge preserving smoothing filter	31
2.10	Alphabets characters	33
2.11	Arabic character shapes	34
2.12	Rules for finding candidate cutting points	34
2.13	Estimation Arabic baseline based	35
2.14	Chinese character shape	36
2.15	Tamil symbols for word recognition	37
3.1	General architecture of system	39
3.2	Flow chart of word extraction	41
3.3	Flow chart of method	45
3.4	Some samples of document image acquired	46
4.1	Original images and results	51
4.2	Original image sample	52
4.3	Reduce RGB Color Image	53

4.4	Applying binarization	54
4.5	Smoothing by using Averaging Filter	55
4.6	Morphological technique using erode	56

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	FIGURE OF G.U.I	62
B	MATLAB CODING	65

CHAPTER 1

INTRODUCTION

1.1 Background

Extract text information has grown rapidly as fast as technology increased, especially during the past years. As technology increased, most of the works need documentation. The document contains the words and images. When the documents have been stored for long time, they become too olds, which causes the document unclear, and difficult to read the content, especially when the important and valuable document needed to keep for long time.

This situation is faced during artifact recognizable, map interpretation, news articles search from microfilms, historical handwritten document image with uneven background and referencing system for digitized manuscripts. Also, the important word or key words of the domain are used in various applications such as the indexing, plan, text summarization, and automatic abstract generation and so on. Because of the various uses of the key words, word extraction has been investigated extensively. Thus, this problem attracted the researches interested, where they know the ability of text to provide the strong effect on description of the image content, the convenience of distinguishing them from color document and other image features and foreground.

A document image contains the text blocks including the characters which colors are darker than the colors of local background or foreground. Moreover, this kind of extraction also help to recognize and save the important information such as artifact recognizable, map interpretation, news articles search from microfilms, historical handwritten document image with uneven background and referencing system for digitized manuscripts. The Content-based image retrieval, OCR, page segmentation, license plate location, address block location, and compression, are some examples based on the text information extraction from various types of color and images.

1.2 Problem Statement

In previous work of word extraction results unsatisfied, blurring and cannot read clearly of the needed important information. The problems occur on retain characters and words during extraction. In particular, it also occur on making the extract word for content aggregation efforts and, preserving text quality. Those below images are some of the previous work result.

Followings are the key points of problem statement that have been stated and elaborated above:

- i - Retain characters and words during extraction

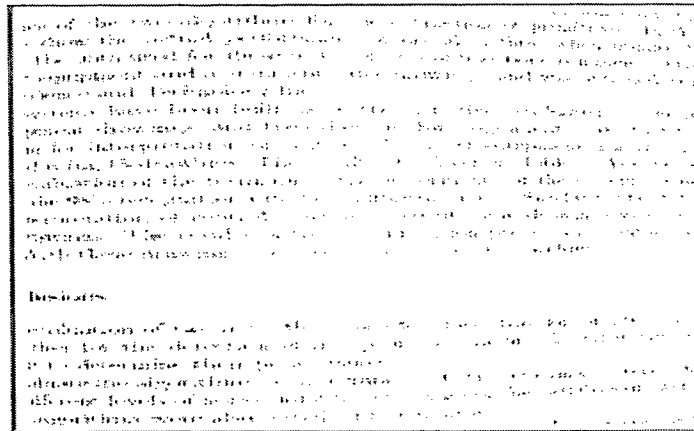


Figure 1.1: Sauvola's

Maintaining text and formatting data also allows person and organizations to build fully-searchable archives while at the same time ensuring that the conversion does not degrade the appearance or legibility of the document. This ensures that text and formatting data are maintained when converting document formats.

ii - Extract word for Content Aggregation Efforts

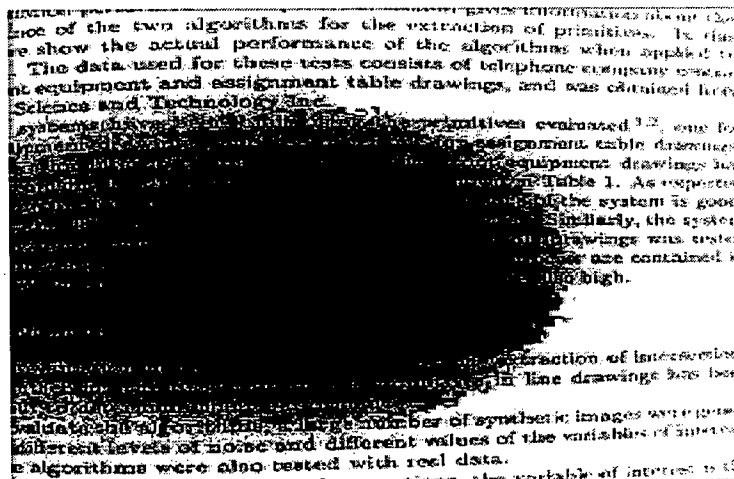


Figure 1.2: Otsu's

To create a data stream that can be processed by content aggregation tools to directly import the information into a database or repository. This data is then available to be re-purposed for publishing, archiving or searching. By using this system will speed up the data population efforts while reducing the risk of errors that commonly occur when relying on a manual data entry processed.

iii - Preserve Text Quality

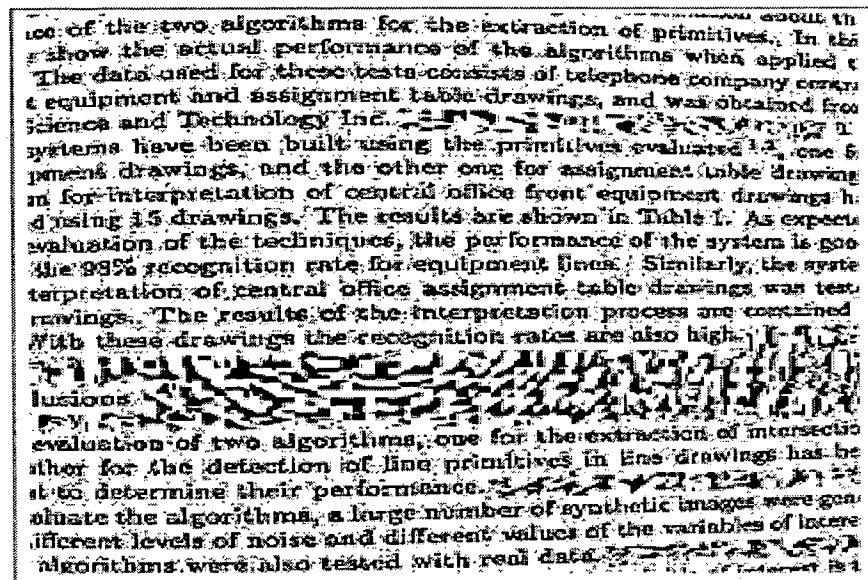


Figure 1.3: Niblack's

When there is rasterizing an image that includes the text elements, it is important to ensure image integrity and quality. If text is rendered to a lossy format, such as JPEG, letters can become misshapen impacting the legibility of the document. This problem is exaggerated as the document is continually opened, annotated and saved, potentially impacting the document's quality over time can ensure that you maintain legible text throughout the documents lifespan.

1.3 Objective of the Project

The objective is to develop a system for word extraction to make things easier for extracting the information or document needed, implemented by image processing toolbox of MATLAB.

1.4 Project Scope

The scope of this project is the document image contains text blocks constituted of characters which character is darker than foreground.

Assumption 1: Foreground

Black represented a character (or noise), while white represented the foreground.

- Word → 1 (dark)
- Non-word → 0 (light)

Assumption 2: Difference color and contour on document image.

The task of word extraction is an easily done by selecting appropriate preprocessing techniques which separating the noise in the same word and then eliminate those noises. For this purpose, I define two characteristic features, which are the differences of color, and the contour that occurred before. I would rely on them to examine each characteristic and then judge whether I should select it to perform character merging or not. So, the main point to develop this project is all about the coding and characteristic feature that selected.

The selection of black color of characteristic can never been guaranteed to show within the binary. The problem in such an approach will escalate when the extraction task goes beyond the character level and if there's exists irregular text alignment and orientation.

1.5 Thesis Outline

This thesis is composed of five chapters covering introduction, literature review, methodology, analysis and result; and the last chapter is a conclusion and recommendation in future work.

Chapter 1 explains the background of the project, problem statements, objective and also the scopes of project. Image Processing Color model, image processing algorithm and Word Extraction are the main essential in this project.

Chapter 2 focused on the literature review for those three parts that has been explained in Chapter 1. All the journals and books that are related to this project are used as a reference to guide and help completing this project. Each of this part is explain based on this finding and studies.

Chapter 3 explains and discuss about the methodology that has been used in order to complete this project. In this chapter explained about two phase used which are the preprocessing phase and processing phase of Image processing. The discussion will be focusing on how algorithm in each phase works.

Chapter 4 gives a detail result and analysis on the results obtained with accuracy rate of system, and algorithm using Image Processing of MATLAB.

Chapter 5 discussed the conclusion of development of this project. This chapter also discusses the recommendation for this system for future development or modification.

CHAPTER 2

LITERATURE REVIEW

2.1 Background

This chapter focused on the literature review for each component on this project. All the components are described in details based on the finding during the completion of this project. In this chapter, I will discuss about the concept of word extraction, previous work on word extraction, and the techniques involved in word extraction.

2.2 Concept of Word Extraction

The concept is typically associated with a corresponding representation in a language or citation needed such as a word [1]. The concept refers to the commonly agreed that any domain specific term which describes the part of the domain is called a concept. The main aim of the concept extraction subtask is to

identify the maximum possible list of concepts of the domain. The source of information known about the domain is a set of unstructured documents. With the perception of the above definition for the concept, most of the key words in the domain text are valid concepts of the domain. Though there exists an extensive literature on concept extraction, three different techniques which are closely related to our proposed approach are presented in this dissertation. These techniques were based on the contextual patterns, syntactic pattern, and co-occurrence heuristics

The concept of extraction comes out from the idea of concept mining. The concept mining is an activity that results in extraction of concepts from artifacts. For the purposes of concept mining however, these ambiguities tend to be less important than they are with machine translation, for in large documents the ambiguities tend to even out, much as is the case with text mining [2].

2.3 Previous Work on Word Extraction

Word extraction, which sometimes referred to word characters extraction, which refers generally to the process of deriving and recognizing the important information from image document. The important information is typically recognized through different color between word characters and foreground color; where the color of word character is darker than foreground color.

There are various techniques used in text extraction from document image, such as Voronoi diagram, non-pyramid and pyramid techniques. These methods focused on page segmentation by using the features of entire document page. They are not suitable for word segmentation, because the character size and inter-characters gap are different from word to word, even within the same document page. Furthermore, there exist different orientations of words in a document page. Unlike page segmentation, for the issue of word segmentation,

the local information including the descriptive features of the image elements and the relative positions should be taken into account.

2.3.1 Text Extraction

Many works on text extraction from document images have been reported previously using various techniques. It begins with text extraction where pyramid model used to extract text strings based on connected component generation [13]. This method is using irregular pyramid structure. The uniqueness of this algorithm is its inclusion of strategic background information in the analysis. The main advantage is their capability of detecting text in low resolution images. The pyramid-model efficiently and quickly identified and located words or phrases placed in an image file.

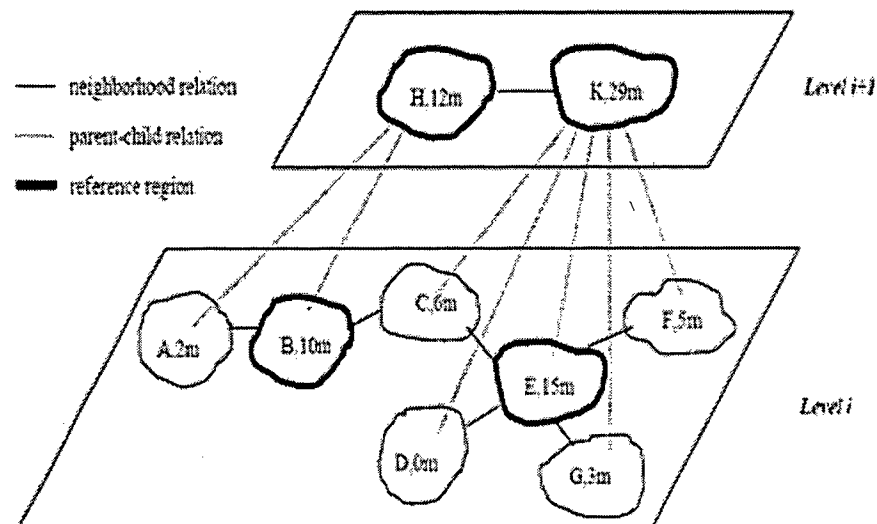


Figure 2.1: Irregular pyramid construction process

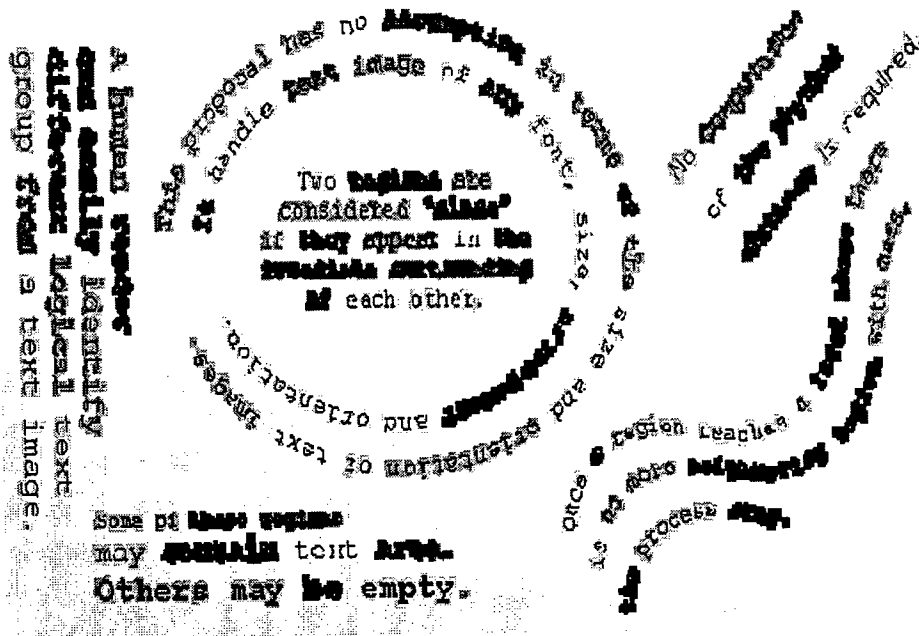


Figure 2.2: Result on Pyramid Model

However, the efficiency is decreased when characters are required to be grouped into words. This method is fast, but it results an unclear text document and required to be grouped.

Sobottka [3] proposed an approach to automatically extract text from colored books and journal covers. While Shinghal [4] conduct an experiment investigation of text-recognition by four methods. The methods of classification are without contextual information, Raviv's recursive Bayes algorithm, the modified Viterbi algorithm, and a proposed heuristic approximation to Raviv's algorithm. However, the technique of Viterbi Algorithm greatly depends on the likelihood and transition probabilities between English characters, the more contextual information those probabilities could extract the better the performance of the Viterbi algorithm.

2.3.2 Word Extraction

Continued by Zhe Wang where a diagram, namely Voronoi Diagram which this method focused on page segmentation by using the features of entire document page [12].

However, the techniques of Voronoi diagram are not suitable for word segmentation, because the character size and inter-character gap are different from word to word, even within the same document page.

So long as no
of the banks Shang-hae

3. ANALYSIS OF THE SAFETY REQUIREMENTS OF SAFETY-CRITICAL SYSTEMS

The basic concern in this paper is with the analysis of requirements and not with their elicitation (the process of acquisition of the relevant information from the user). We deal with techniques that can be used to reduce (or eliminate) the possibility of the occurrence of hazards due to faults introduced during the requirements analysis.

Figure2.3: Result on Zhe Wang using Voronoi Diagram

2.3.3 Word Character Extraction

While, the word character extraction is introduced by Fletcher, using 3D neighbourhood graph model [5]. Fletcher described a method that uses

information from each connected component in a mixed text-graphic document [6]. This method can group words in inclined lines, intersecting lines, and even curved lines.

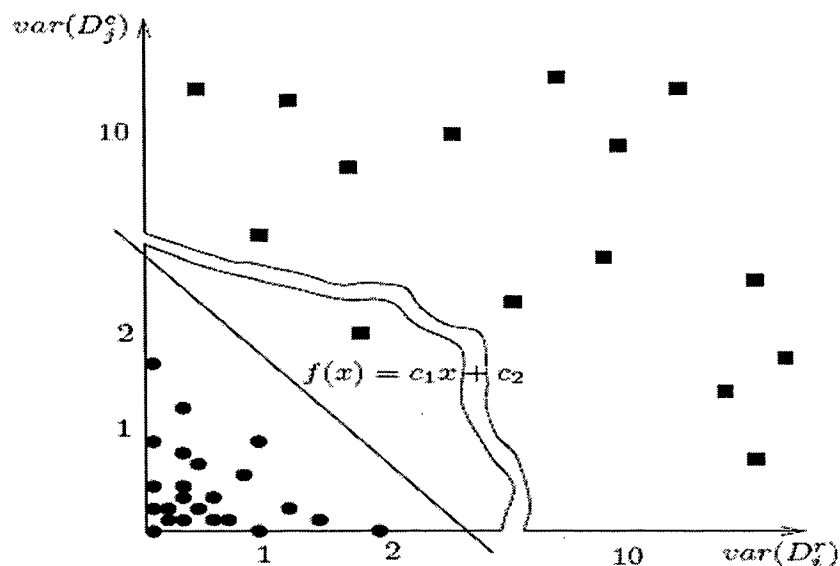


Fig. 4. The classification for characters and pictures: a circle denotes a character and a rectangle denotes a picture

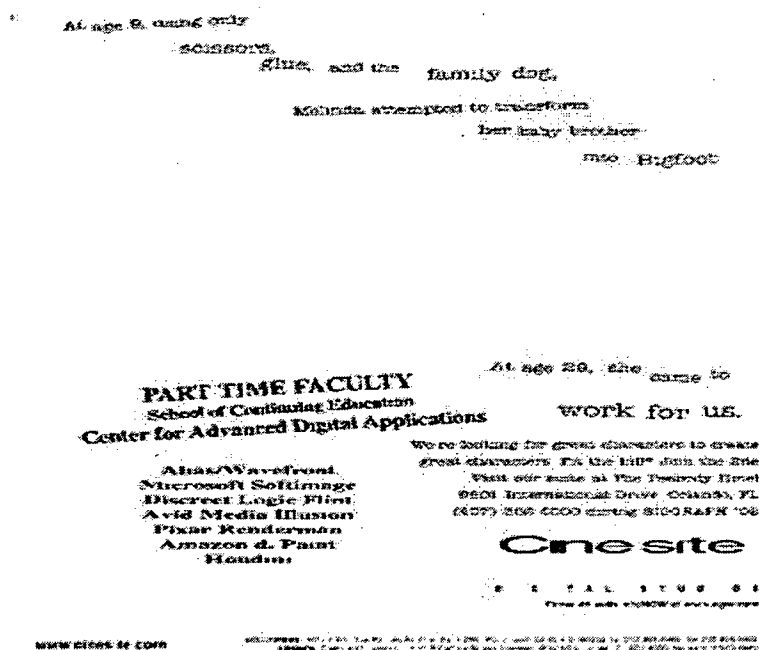
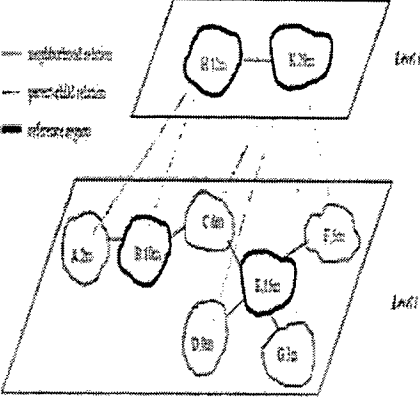
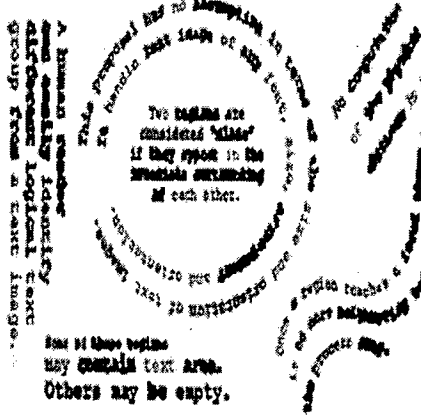
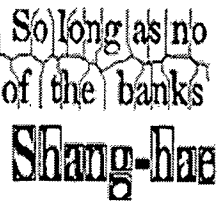


Figure2.4.: Result on 3D neighborhood graph model

Furthermore, there exist different orientations of words in a document page. Unlike page segmentation, for the issue of word segmentation, the local

information including the descriptive features of the image element and the relative positions should be taken into account. This technique extracts word character from images using image processing technique, which includes smoothing, sharpening, and contour tracing. The word extraction eliminates noise and the distortions on the same document to remain the important characters.

Evolution		Method	Result
text extraction	Tan	<p>pyramid model</p>  <p>Figure 5 - Image pyramid extraction process.</p>	 <p>- unclear, required to be grouped</p>
word extraction	Zhe Wang	<p>Voronoi Diagram</p> 	<p>3. ANALYSIS OF THE SAFETY REQUIREMENTS OF SAFETY-CRITICAL SYSTEMS</p> <p>The basic concern in this paper is with the analysis of the requirements and not with their elicitation (the process of acquisition of the relevant information from the user). We deal with techniques that can be used to reduce or eliminate the possibility of the occurrence of errors due to faults introduced during the requirements analysis.</p> <p>-focus on page segmentation -different character size&inter-character gap</p>

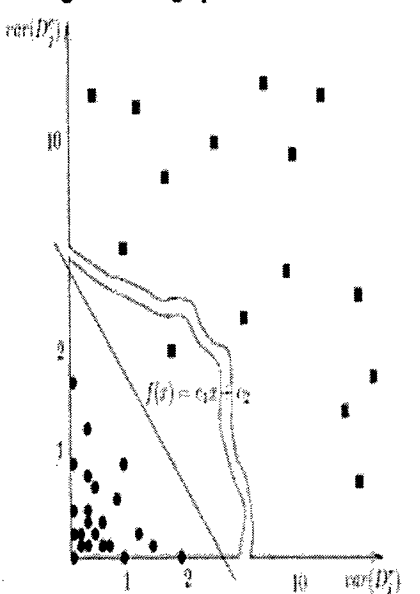
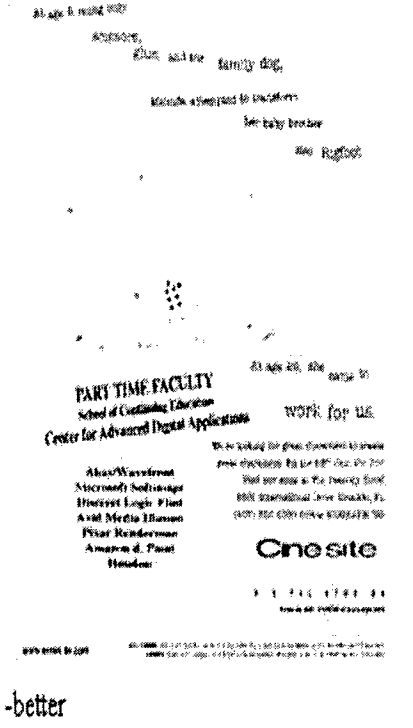
<p>word character extraction</p>	<p>Fletcher</p>	<p>3D neighborhood graph model</p>  <p>Fig. 4. The classification for characters and pictures: a circle denotes a character and a rectangle denotes a picture</p>	 <p>-better</p>
----------------------------------	-----------------	---	--

Table 2.1: summarization on previous work