RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

CHAIVATNA SUMETPHONG

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2011**

Thesis
entitled

RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

.....
Mr. Chaivatna Sumetphong
Candidate

.....
Assoc. Prof. Supachai Tangwongsan, Ph.D.
Major advisor

.....
Assoc. Prof. Damras Wongsawang, Ph.D.
Co-advisor

.....
Asst. Prof. Sukanya Phongsuphap, Ph.D.
Co-advisor

.....
Prof. Banchong Mahaisavariya, M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

.....
Assoc. Prof. Supachai Tangwongsan, Ph.D.
Program Director
Doctor of Philosophy Program
in Computer Science
Faculty of Information and Communication Technology
Mahidol University

Thesis
entitled

RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Doctor of Philosophy (Computer Science)
on
October 7, 2011

.....
Mr. Chaivatna Sumetphong
Candidate

.....
Prof. Prabhas Chongstitvatana,
Ph.D.
Chair

.....
Assoc. Prof. Supachai Tangwongsan,
Ph.D.
Member

.....
Asst. Prof. Sukanya Phongsuphap,
Ph.D.
Member

.....
Assoc. Prof. Damras Wongsawang,
Ph.D.
Member

.....
Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

.....
Assoc. Prof. Jarernsri L. Mitrpanont,
Ph.D.
Dean
Faculty of Information and
Communication Technology
Mahidol University

ACKNOWLEDGEMENTS

First and foremost, I would like to express my utmost gratitude to my advisor, Assoc. Prof. Dr. Supachai Tangwongsan. His advice and research directions kept me on track and allowed me to focus on my area of research. Over the period of six years, it was very difficult to be motivated towards completing my research works. With his encouragement, support and patience, I was able to continue and complete this epic jouney.

I would like to thank my co-advisors Assoc. Prof. Dr. Damras Wongsawong and Asst. Prof. Dr. Sukanya Phongsuphap for their valuable comments in this study. I am indebted to Prof. Dr. Prabhas Chongstitvatana for his time and involvement.

I acknowledge the scholarship from Mahidol University and I would like to thank the university for the generous financial support throughout my study.

A special thanks to Dr. Ananta Srisupab for his weekly discussions that helped to move the research forward. He was one of the very few PhD. candidate colleagues who were always willing to spend some time with me brainstorming for solutions.

I would like to express my sincere gratitude to my parents and grandparents for giving me life, understanding, love, for educating me, and giving me this opportunity with unconditional support.

Last but not least, I would like to credit my wife Viyada for standing by me through the rough times and sharing my dream of completing a post-graduate study. Her support and love kept me moving forward amidst difficult family, health and financial times. Her dedication towards taking care of our family and three lovely children Phillip, Mikey and Minnie over these rough years cleared the way for me to complete my studies.

Chaivatna Sumetphong

RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

CHAIVATNA SUMETPHONG 4838800 ITCS/D

Ph.D. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE: SUPACHAI TANGWONGSAN, Ph.D., DAMRAS WONGSAWANG, Ph.D., SUKANYA PHONGSUPHAP, Ph.D.

ABSTRACT

In this research, we focus on solving the problem of recognizing broken characters that are found abundantly in historical documents. Most OCR systems are designed to correctly recognize finely printed documents in both Thai and English scripts. When tested with degraded documents, the accuracy of these systems drops drastically. One of the most important problems that decreases the accuracy of OCR systems is the broken characters found in the document image. Although humans can read broken characters without difficulty, it is far more complicated for computers to recognize them.

We propose a strategy that aims to find the optimal set-partition of the pieces of the broken characters based on a probability functions model, thus yielding a sequence of characters. To counter the inevitability of recognition errors and the need to group the characters as words, we employ an N-grams Graph generated from a dictionary. To demonstrate the generality of the method, we performed experiments on a Thai historical document, an English historical document and a Thai fax document. The results obtained are very promising and this research could be extremely useful for researchers engaged in recognizing historical documents degraded due to the presence of broken characters.

To find optimal solutions for set-partition problems, the obvious method would be to adopt the brute-force enumerative approach. However, this is drastically slow when the number of elements in the set increases. We propose a Partition-Growing Algorithm that is capable of generating optimal set-partitions of the broken characters in a smaller timeframe by using heuristics based on characteristics of probability functions. The algorithm is further extended towards solving other problem domains that require finding an optimal set-partition. We demonstrate the applicability of the algorithm to other set partitioning problems (SPP) that might be linear or non-linear in nature by modeling these problems based on probability functions. The experiments performed show that the algorithm has a potential to be adapted to general problems that require partitioning a set optimally.

KEY WORDS : HISTORICAL DOCUMENTS / BROKEN CHARACTERS / OPTIMAL SET-PARTITIONING / N-GRAMS GRAPH / PARTITION-GROWING ALGORITHM

83 pages

การรู้จำตัวอักษรขาดในเอกสารทางประวัติศาสตร์และการแก้ปัญหาการแบ่งเซททางคณิตศาสตร์
RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

ชัยวัฒน์ สุเมธพงศ์ 4838800 ITCS/D

ปร.ด. (วิทยาศาสตร์คอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: ศุภชัย ตั้งวงศ์ศานต์, Ph.D.; ดำรัส วงศ์สว่าง, Ph.D.; สุกัญญา พงษ์สุภาพ, Ph.D.

บทคัดย่อ

ในงานวิจัยนี้ เราเน้นการแก้ปัญหาของการรู้จำตัวอักษรขาด (Broken Characters Recognition) ที่ปรากฏเป็นจำนวนมากในเอกสารทางประวัติศาสตร์ ปัจจุบันระบบการรู้จำตัวอักษร (OCR) ส่วนใหญ่จะถูกออกแบบให้จัดการกับเอกสารสิ่งพิมพ์ที่มีตัวอักษรทั้งไทยและอังกฤษได้ผลเป็นอย่างดี อย่างไรก็ตามเมื่อนำมาทดสอบกับเอกสารที่มีปัญหา เช่น เก่า เหลือง ตัวอักษรขาด ประสิทธ์ผลในความแม่นยำของระบบจะลดลงอย่างมากโดย เฉพาะกรณีของตัวอักษรขาดและไม่สมบูรณ์ แม้มนุษย์เราจะสามารถอ่านตัวหนังสือที่มีตัวอักษรขาดได้ไม่ยากนัก แต่สำหรับคอมพิวเตอร์แล้วก็เป็นเรื่องที่ค่อนข้างจะยุ่งยากด้วยความไม่สมบูรณ์ของตัวอักษร

ในงานวิจัยนี้ เราจะได้นำเสนอวิธีการแบ่งเซทเหมาะที่สุด (Optimal Set-Partition) ในการจัดการกับตัวอักษรขาดที่เกิดขึ้นในเอกสารโดยทั่วไป เพื่อให้ได้ผลลัพธ์ที่ดีด้วยแบบของฟังก์ชันความน่าจะเป็น (Probability Functions) นอกจากนี้เพื่อให้ได้ผลลัพธ์ที่ดีขึ้นในเรื่องความถูกต้องแม่นยำ เราจะตรวจคำศัพท์ของผลลัพธ์กับคำที่มีในพจนานุกรมด้วยการสร้างเอ็นแกรมกราฟ (N-grams Graph) เมื่อได้ทำการทดสอบประสิทธ์ผลของระบบงานด้วยเอกสารทางประวัติศาสตร์ฉบับภาษาไทย ฉบับภาษาอังกฤษ และฉบับโทรสารภาษาไทย พบว่าความแม่นยำอยู่ในเกณฑ์ที่น่าพอใจ จึ่งมีความเชื่อว่าผลงานนี้จะเป็นประโยชน์ต่องานการเก็บรักษาเอกสารทางประวัติศาสตร์ ของไทยและของต่างประเทศที่มีตัวอักษรขาดปรากฏโดยทั่วไป

การศึกษาหาผลลัพธ์ที่เหมาะที่สุดสำหรับปัญหาของการแบ่งเซทนั้น วิธีการที่ง่ายและตรงที่สุดเห็นจะเป็นการแยกประเมินทีละส่วนจากหัวจดท้ายจนหมดรายการ แต่วิธีนี้ต้องใช้เวลามากและไม่เหมาะสมในกรณีที่ขนาดเซทเพิ่มมากขึ้นซึ่งจะทำให้ได้คำตอบที่ช้ามากจนไม่สามารถนำไปใช้ได้ในทางปฏิบัติ ในงานวิจัยนี้เราจึงได้นำเสนออัลกอริทัมใหม่ด้วยการขยายการแบ่งส่วน (Partition-Growing) ที่สามารถแบ่งเซทได้อย่างดีที่สุดของตัวอักษรขาดภายในช่วงเวลาอันสั้น โดยอาศัยคุณลักษณะเฉพาะของฟังก์ชันความน่าจะเป็นเป็นตัวหลักในการแก้ปัญหา นอกจากนี้เรายังนำอัลกอริทัมนี้ไปแก้ปัญหาอื่นๆที่เกี่ยวกับการแบ่งเซท (SPP) ทั้งที่เป็นลักษณะเชิงเส้นและไม่เชิงเส้น ด้วยแบบจำลองตัวระบบของฟังก์ชันความน่าจะเป็น ผลการทดลองพบว่าอัลกอริทัมนี้ทำงานอย่างได้ผลสำหรับปัญหาของการแบ่งเซทที่ดีที่สุดจริง

83 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

Most commercial Thai optical recognition (OCR) systems are designed for well-formed, modern business documents. Recognizing degraded Historical documents with low-quality printing or type-writing is more challenging, due to the high occurrence of broken characters. The main aim of this work is to study the problem of broken characters and propose a solution to recognize the broken characters by rejoining or grouping the pieces accurately. Solving this problem would pave the way towards ultimate goal of the complete automation of the processo for restoration of Historical documents in the future.

Most OCR-related works require that the document image must be converted to a binary image before the main processing can be performed. Due to this binarization process, a large number of text characters on a Thai historical page show some level of breakdown into multiple clusters of pixels. This breakdown, along with some of the noise pieces, poses a big challenge towards recognition of the characters in the original text image. Other aspects that may lead to the broken characters include low-resolution image scanning, poor quality printing, variations in background color, document age, etc. These broken pieces need to be grouped correctly for the classifier to produce accurate recognition results. The complexity of the problem increases with the existence of noise pieces intermingled with the broken pieces.

An example of a zoomed image segment extracted from a Thai Historical document is shown in Figure 1.1. Note the abundance of broken pieces that originally belonged to the original text characters.

1.1 Research Motivation

Historical documents are national treasures. Thailand has 800 years of rich history, most of which have been documented over the years using different recording

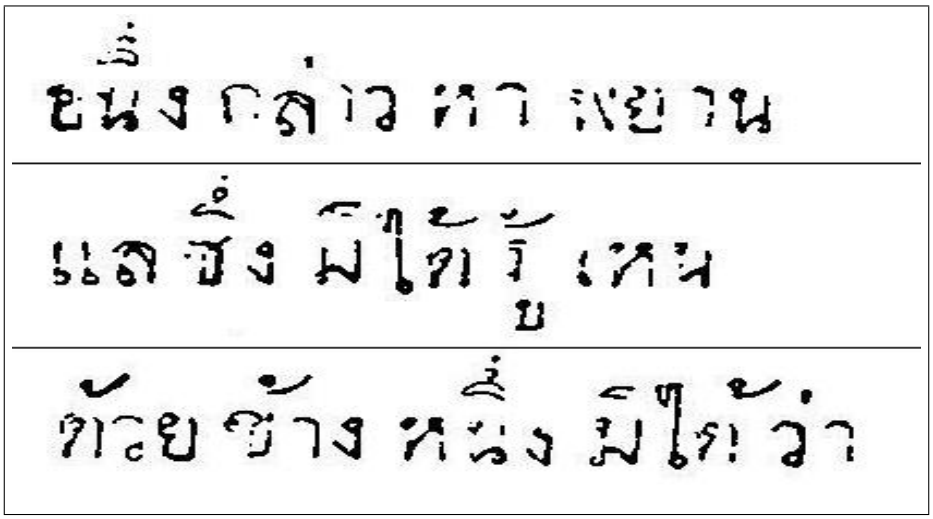


Figure 1.1: Examples of broken characters found in Historical Documents

media such as stone, palm leaf and paper. The Thai Historical documents are stored and preserved in a few organizations, mainly the Thai National Library, the Thai National Museum and some other renowned institutes. Public access to these documents is limited to on-site only and that too for short durations. In addition, searching for specific information can be painstakingly slow as the contents of these documents are not electronically transcribed.

Thai Historical documents are of major interests to historians, politicians, anthropologists and lawyers. These people need to research into the past documents as a normal part of their current day-to-day activities. Their works can be highly facilitated if these historical documents are made available for view on the internet either as restored digital images or XML documents.

Some efforts towards generating digital images of a few main historical documents have been initiated by the Thai National library. However the quality of these images highly depends on the condition of the original document and equipments used to capture the images. This procedure requires a large amount of manual effort which cannot be replicated. Moreover, searching for particular information in document images is almost impossible. OCR techniques can be employed towards digitizing the Thai Historical documents and the resulting text can be converted to document images of higher quality and XML documents that will allow people to find the specific information in history.

Different countries such as USA, Korea, China, Egypt, Greece and India are working towards digitizing their historical archives. Over the last 20 years, research into OCR techniques has advanced to a point where high accuracy in automatic transcription of these documents can be achieved. Different languages require specific techniques that are suited to particular nature of the language. Efforts towards digitally restoring Thai Historical documents have been minimal and targeted to specific issues related to the restoration process.

We believe that any effort towards creating an OCR system for degraded documents must address many areas such as noise-removal, document segmentation, recognition of broken and merged characters, automated correction of any classification errors and document reconstruction. Such an endeavor would require a large amount of effort and hence we focus mainly on broken character recognition and automated correction.

1.2 Problem Statement

As stated earlier, we seek to tackle the accurate recognition of the broken characters problem that is abundantly found in Thai Historical Documents. The input to this problem is a binarized line image L. Given the nature of Thai language where there are no inter-word gaps, we do not seek to segment words as would be the case for languages like English. Instead, we segment the line image into multiple phrase images based on the gaps that are prominently found scattered in a Thai line image.

In our problem domain, we shall focus on working with a phrase image containing a set of broken pieces X. A simple example of a phrase image containing three Thai characters that can be viewed as as set containing seven broken pieces is shown in Figure 1.2. Generally, a phrase image may contain up to 15-20 pieces. We need to group these broken characters correctly such that it closely represents character patterns of the language and hence can be accurately classified.

Given a line image L containing broken pieces X, find the best sequence of words W^* in the domain language that best explains these pieces.

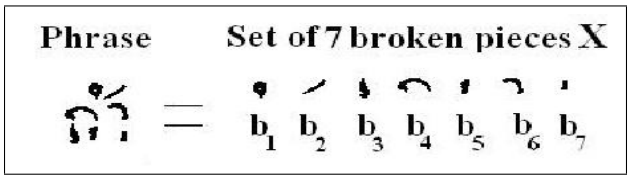


Figure 1.2: An phrase image viewed as a set of broken pieces

To simplify this general problem, we divide our approach into two specific problems.

Problem 1 *Given a set of broken pieces X, find best sequence of characters C^* that best resembles the topological arrangement of the broken pieces.*

Problem 2 *Given a series of best character sequences $C_1^*, C_2^*, ...$ derived from all the phrases belonging to a the document line L, correct any inherent classification errors and generate the best sequence of words W^* that is readable in the domain language.*

Symbolically, we summarize our problem in Equation 1.1.

$$\begin{aligned} & X_1, X_2, ... && = F_1(L) \\ \text{Problem 1}: \quad & C_i^* && = F_2(X_i) \\ \text{Problem 2}: \quad & W^* && = F_3(C_1^*, C_2^*, ...) \end{aligned} \tag{1.1}$$

We propose a solution to Problem 1 based on finding the optimal set-partition π^* of X using a probability function model and the optimal partition can then be mapped to obtain the best sequence of characters C^*. To ensure that the set-partition obtained is optimal, the obvious brute-force enumerative method is required to evaluate all partitions. However, this approach is drastically slow as the number of partitions grows exponentially as the number of broken pieces increase. To counter this performance issue, we propose a novel Partition-Growing Algorithm that allows us to obtain an optimal partition faster by evaluating only a fraction of all partitions. We then demonstrate the applicability of this algorithm to other problems such as Steel Mill Slab Design, Wedding Planner and Task Allocation.

We employ the N-grams Graph to solve Problem 2 by designing a set of error correction operators and word generation operators based on probability functions and implementing the A^* search methodology to generate the best sequence of words W^* given sequences of characters C_1^*, C_2^*....

1.3 Research Objectives

In this section, we present a list of main objectives that we have achieved during this research work.

- Survey the works related to restoring Historical documents conducted by different organizations, projects and researchers around the world.
- Study the different areas of work that may be linked towards restoration of Thai Historical documents such as image processing, pattern recognition and natural language processing.
- Tackle the main challenge of accurately recognizing broken characters that are found on Thai Historical documents.
- Investigate different OCR techniques available that may be applicable to restoring Thai Historical documents.

1.4 Scope of Research and Research Contributions

In this section, we present our scope of research work and how they contribute towards the advancement of research in the area of Thai Historical document Restoration.

- Formulate a model for solving the recognition of broken characters found in Thai Historical documents and thus solves one of the main challenges in the framework.
- Formulate a model for automated error correction of classification errors that are inherent in such OCR applications so as to improve the accuracy of the generated text and reduce the amount of manual correction required.
- Develop a prototype that will help towards understanding of the different aspects the Historical document restoration. Further development in the future can be continued based on this prototype.
- Perform experiments to demonstrate the generality of the model towards other document scripts, languages that may contain broken characters.
- Apply the solution to other problem domains of similar nature.

1.5 Structure of Research Dissertation

The rest of this research dissertation is organized as follows. In Chapter 2, we discuss existing works related to recognition of broken characters in detail. As our method to solve the broken character problem is based on optimal set-partitioning, we also provide a short review of researches linked to Set-Partition Problem (SPP). In the final part of this chapter, we include a summary of other historical documents restoration projects being conducted in the past and current. Chapter 3 focuses on the main area of this research work - accurately recognizing broken characters found in Thai Historical Documents. Chapter 4 provides a brief summary of the implementation of the prototype developed to recognize broken characters. In Chapter 5, we extend the Partition-Growing Algorithm introduced in Chapter 3 towards solving other related problem domains on which this algorithm can be applied. Chapter 6 elaborates on the experiments performed - Recognizing broken characters in three different documents, some simulation experiments and other experiments demonstrating the applicability of the Partition Growing algorithm to some other problem domains that require generating optimal set partitions. The summary and the conclusions are presented in Chapter 7 along with some unresolved issues and suggested future research work.

CHAPTER II
RELATED WORK

2.1 Recognition of Broken Characters

In this part, we review works that are related to our main research focus: Recognition of Broken Characters. It is surprising that after more than 45 years of research, OCR systems are not close to matching human performance. Current accuracy of the machine-printed characters is easily higher than 99% which is sufficient for many real applications. However, not many works have been reported towards recognizing broken characters. Some of the researches show results that are highly dependent on the level of degradation of the document image. Researchers into this area make assumptions to allow them to deal with a smaller-scoped specific-domain problem.

Hobby [72] attempts to reconstruct the character using bitmap clustering and averaging. The approach of using enhancement filter design has been studied extensively by [15] [16]. However, these techniques are limited to cases where the distance between the broken pieces is small. In [35], a border-following method is used to reconstruct the borders and missing parts of broken handwritten digits. The use of active contours to recover broken characters is extensively elaborated in [18].

Recognition-based method attempts to recognize broken characters without actually reconstructing them. T. Pavlidis et al. [19] [20] recognizes handwritten digits by using homeomorphic sub-graph matching. Broken characters are recognized by looking for gaps between features that may be interpreted as part of a character. Yu et al. [21] presents a letter-level shape description using skeleton generated by principal curve. Although both techniques rely on different bases, their results are highly depended on segmentation module, which is responsible for grouping broken segments.

Yanikoglu [22] estimated the pitch of the text and guided the segmentation of the characters by the location of the pitch window, defined by the estimated pitch and the offset. Oguro et al. [23] proposed a three step solution for restoring faxed

documents producing gray level images. Natarajan et al. [24] used Hidden Markov Model for recognizing degraded fax documents. Cannon et al. [25] suggested a method for automatically improving the quality of degraded images in a typewritten archive. Rodriguez et al. [26] developed a cost function to segment degraded typewritten digits.

Droettboom [13] proposed a robust method for rejoining broken segments based on graph combinatorics. The algorithm begin by creating an undirected graph in which nodes represents a connected component in the image. Two nodes are connected by an edge if the borders of pieces are within a certain threshold distance. Next, all of the different ways in which the connected component can be joined are evaluated using k-nearest-neighbor classifier. The dynamic programming is then used to find an optimal combination that maximizes the mean confidence of the characters across the entire subgraph. However, the efficiency of this approach is based on training data. Figure 2.1 a) shows the original broken characters and b) shows how the author creates a graph based on positions of broken pieces with respect to each other and c) shows how heuristics are applied to remove some of the edges in the graph to create subgraphs where each subgraph is used to generate a character.

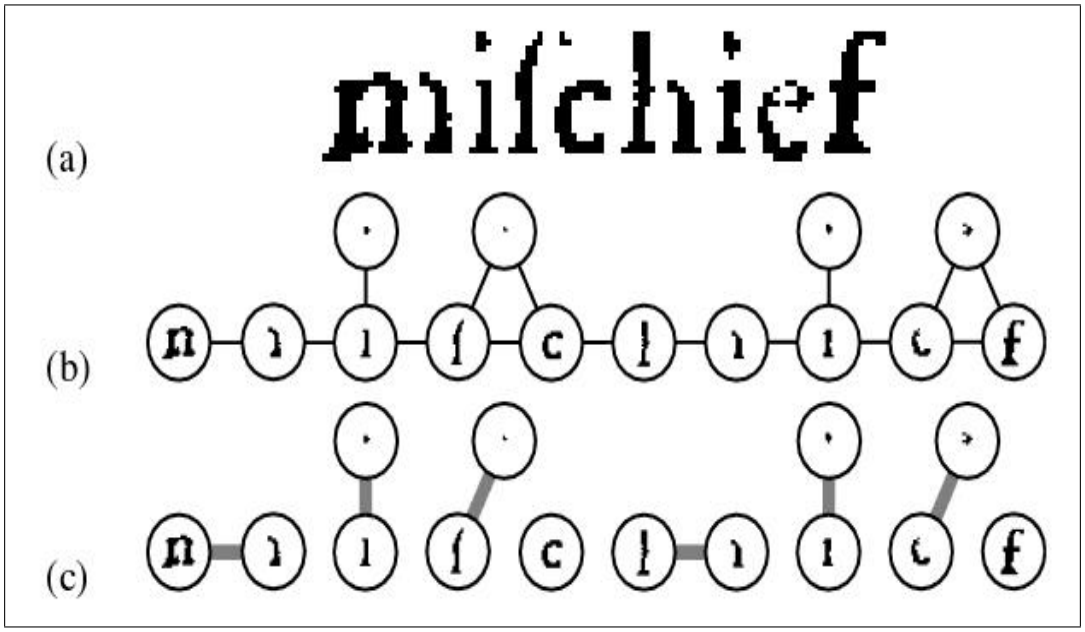


Figure 2.1: Recognizing broken Characters using Graph Combinatorics

For Thai printed document, several approaches have been proposed for Thai broken character reconstruction. Premchaiswadi et al. [27] reconstructs broken character

by using segment boundaries and distinctive features of Thai characters. Segments will be rejoined to form a character when their extended boundary overlaps each others. However, this method can detect only horizontal broken character, and is limited to a maximum of four segments leading to only sixteen pattern templates. Figure 2.2 has been extracted from the original works and included here for reference purposes.

No.	Structure	Character code	Character member	No.	Structure	Character code	Character member
1		1110	ก ค ฅ ฆ ฉ ฑ ฒ ด ต ท พ ฟ ภ ศ ส ห ฬ	9		0011	Broken character
2		1011	ข ฃ ฆ ช ซ ฌ ญ ณ น บ ป ผ ฝ ม ษ	10		1000	Broken character
3		1111	ง จ ฉ ช ซ ฎ ฏ ฐ ฌ ฝ ย ล ศ ษ ส อ ฮ	11		0010	Broken character
4		0110	Broken character	12		0100	Broken character
5		1001	เ ไ ใ	13		0001	Broken character
6		0111	ฉ ฐ ร ว อ	14		0101	Broken character
7		1101	โ	15		1010	No character
8		1100	Broken character	16		0000	No character

Figure 2.2: Character code of 16 patterns used to detect broken characters

To solve these problems, Pachiyanukul et al. [14] presents a new method for repairing vertical broken Thai character using specific characteristics of Thai character, such as position of head, the leg and the broken location of character. Although this method works well in the experiment, it still has a problem when characters are broken both vertically and horizontally. Peerawit et al. [34] employs the closing algorithm and some heuristics to group the broken pieces but the method relies solely on image processing techniques without emphasis on pattern analysis. Figure 2.3 a) shows an image containing heavily broken Thai characters and b) shows how the authors reduce the level of brokenness by closing the gaps between the pieces that are in close vicinity of each other. This allows for easier character segmentation, but leads to false merging between pieces that may belong to different characters and in some cases generated distorted thick image patterns that proved difficult to recognize.

In summary, from the above listed researches related to solving broken

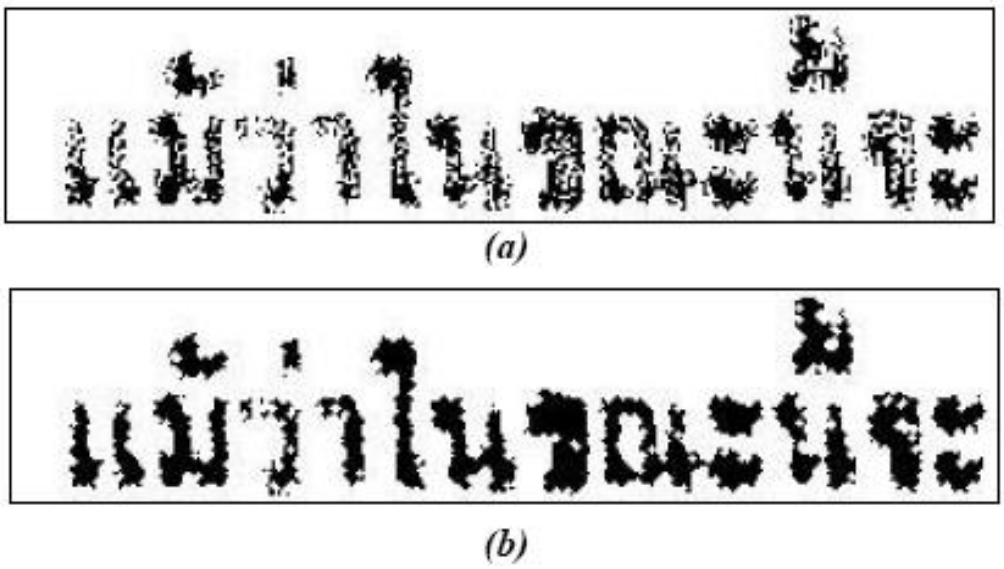


Figure 2.3: Employing Closing Algorithm to merge broken pieces

characters, we find that Droettboom [13] proposed a technique that is the most promising. Our novel method uses his approach as a basic idea but rather than looking at the pieces as nodes in a graph and relying solely on the distance between them for creating groups, we use the overall pattern and size as our basis. In addition, his method works well with English alphabets but fails when applied to multi-zone languages like Thai. Furthermore, we formulate our problem and solution mathematically based on set theory, rather than on graph metrics.

2.2 Set-Partitioning Problems

The set partitioning problem (SPP) was one of the first problems to be classified as NP-Complete by Karp [36]. Because of the importance of the SPP, a number of algorithms have been developed. These can be classified into two categories: exact algorithms which attempt to solve the SPP to optimality, and heuristic algorithms which try to find "good" solutions quickly.

Many of the researches are based on linear programming (LP) relaxation but these generally yield tight lower bounds of the optimal solution. Most researches related to SPP were directed either at developing solutions that focused on heuristics and efficient search procedures. Reviews have been extensively done with this regards by

Balas et al. [37], Joseph [38] and Boschetti et al. [40]. Looking for the optimal solution based on linear programming can pose a challenge, given a high degree of degeneracy [41]. A LP reformulation of the set partitioning problem was proposed by Diaby [39] for solving the well-known transportation-related problem. Many researches focused on proposing heuristics to solve the LP relaxation [40] [41] [42] [43] [44] [45] [87] [47]. Attempts have been made to find a good solution for SPP by imposing additional structure to the problem that is derived from real-life applications [62]. Meta-heuristic approaches have also been proposed by Alvarenga et al. [48] and Lee et al. [49]. In addition, heuristics based on reformulations have also been developed for the overall problem [54] [50] [51] [52] [53] [54] [55]. The recent literature has shown much interest in evolutionary algorithms to handle hard combinatorial problems. An example of a genetic algorithm for solving SPP can be found in [60]. Another recent approach of interest has the employment of constraint programming [61].

There have been many research efforts to develop algorithms to solve this problem to optimality. The exact procedures that have been proposed have been mostly enumerative search procedures [37] [38] [40] [44] [87] [56] [57] [58]. However, all the above procedures fall short with respect to the fundamental issue of the tractability of the problem [59]. There are two main classes of optimization algorithms for SPP - 'branch and bound' and 'branch and cut' algorithms. In this research, we employ the use of characteristics of probability functions within the branch and bound approach to generate optimal solutions of a specific set of SPP.

2.3 Historical Document Restoration

Different organizations around the world have ongoing projects to create a digital library of the valuable documents. Two main approaches are being adopted i) create a digital image library of the documents, ii) create a digital document library. The first approach is simple as it requires only manual effort of capturing the document images and cataloging them. However, this method suffers from some drawbacks such as inability to search within the image content and ever-present noise/blur inherited from the original image. The second approach is focused on eliminating these drawbacks. The document needs to be transcribed either manually or by using OCR techniques.

The **Thai National Library** has started a project to digitize some of the valuable Thai documents. However, their main effort in based on using manual capture of the image and cataloging the images. Some of these images are now available to the public for viewing online. However, the image is the original one and no restoration has been conducted on them.

In the **USA**, the World Digital Library project was initiated by the Library of Congress [110] where US historical documents have been transcribed using OCR. The Library of Congress is the research library of the United States Congress and is the oldest federal cultural institution in the United States. It is the largest library in the world by shelf space and holds the largest number of books - about 14 million. Yale Law School [116] is one of the few universities that have implemented a digital library project Avalon. The digital library consists primarily of legal documents and historical documents. The Library of Economics and Liberty is a project belonging to Liberty Fund [117]. This is a private organization dedicated to service of mankind and society.

In the **UK**, the British Library [113] is continuously investing on their project British Library: Online Gallery. The British Library (BL) is the national library of the United Kingdom. The library is one of the world's largest research libraries, holding over 150 million items in all known languages and formats: books, journals, newspapers, magazines, sound and music recordings, patents, databases, maps, stamps, drawings and much more. Its book collection is second only to the American Library of Congress. The Library's collections include around 13 million books, along with substantial additional collection of manuscripts and historical items dating back as far as 300 BC. The Institute of Historical Research of UK [114] (in collaboration with History of Parliament Trust and the University of London) has dedicated their digital library project towards restoring their historical documents British History Online project. The Institute of Historical Research (or IHR) is a British educational organization providing resources and training for historical researchers. The National Library of Wales [112] has also undergone the Welsh Journal Online project to digitize their journal archives.

In **other countries**, the Ahlul Bayt Digital Library Project (Ahlul Bayt DILP) [115], established in 1996, is a non-Profit Islamic organization; working as group of volunteers operating throughout the world. Their primary project Al-Islam.org

intends to digitize and present quality resources related to history, law, and society of the Islamic religion and its personalities with particular emphasis on the Twelver Shi'ah Islamic school of thought. Al-Islam.org is a site which also serves non-Muslims a means of introducing Islam. The Aozora Bunko is a Japanese digital library and this on-line collection encompasses several thousands of works of Japanese-language fiction and non-fiction. LacusCurtius, Marefa and Latin Library are projects to digitize ancient Rome texts and archaelogy, Arabic classisics and Latin documents respectively.

CHAPTER III
METHODOLOGY

In this chapter, we present our methodology to tackle the accurate recognition of the broken characters problem that are abundantly found in Thai Historical Documents. In our problem domain, we shall focus on working with a phrase image containing a set of broken pieces X. We need to group these broken characters correctly such that it closely represents character patterns of the language.

Let L represent a line image extracted from the document image using some line-segmentation procedure. We propose the following logic to find the best sequence of words W^* in the domain language that best explains the broken pieces found in this line image L. We describe our approach in Algorithm 1 and pictorially, we represent these functions in Figure 3.1. As can be seen, the location of big gaps and small gaps in the line image determines how the characters are later grouped together to generate words.

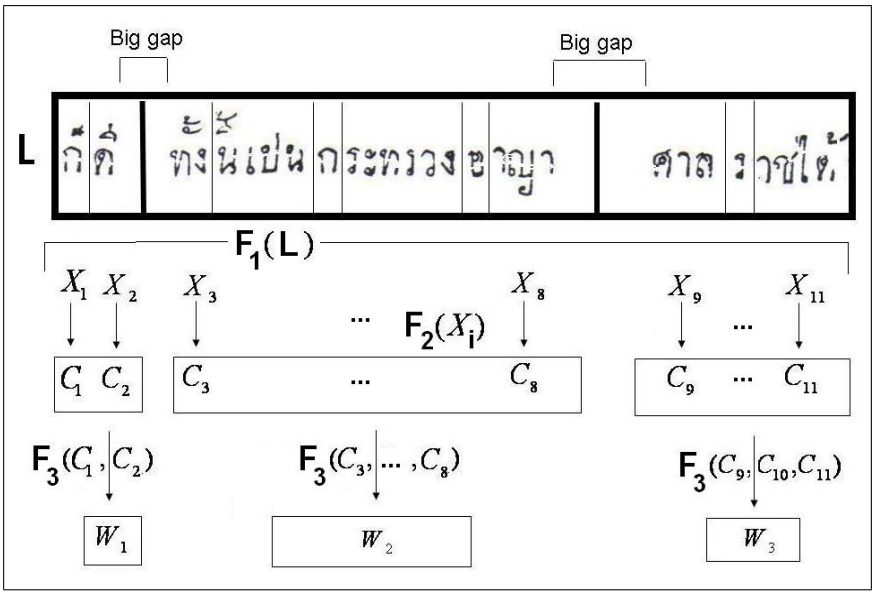


Figure 3.1: Top-level Approach to solving the Main Problem

Algorithm 1 (Recognize Broken Characters)

Input - Line image L

Output - Sequence of Words W^*

$X_1, X_2, X_3, ..., X_N$ = segment(L) // F_1 : segment
initialize $C^* = W^*$ = NULL
for i = 1 to N
  C^* = concat(C^*,recognize(X_i)) // F_2 : recognize
  if $i = N$ or big_gap_between(X_i, X_{i+1})
    W^* = concat(W^*,correct(C^*)) // F_3 : correct
    C^* = NULL
  end if
  end for
return W^*

3.1 Modelling the Broken Characters as a Set Partitioning Problem

In this section, we deal with our Problem 1, as stated in Section 1.2. We treat this problem as a set-partition problem (SPP) where an optimal partition of set X of broken characters is to be found. Let X be a set of broken pieces contained in an image segment. We wish to find the best partition π^* of X such that each subset $S \in \pi^*$ resembles closely to some character pattern c in the language script. Let $P(S|c)$ be the probability that a subset S resembles character c. Thus, we look to

Maximize $\frac{1}{|\pi|}\sum_{S\in\pi} ln\, P(S)$ $\qquad P(S) = P_\alpha(S|c^*) \times P_\beta(S|c^*) \times P_\gamma(S)$

where $P_\alpha(S|c)$ is the probability that S resembles pattern c
and $P_\beta(S|c)$ is the probability that S conforms to the sizing constraints of c
and $P_\gamma(S)$ is the probability that S is cohesive
and $c^* = \underset{c}{\operatorname{argmax}}\, P_\alpha(S|c) \times P_\beta(S|c)$
and $|\pi^*| \geq m$ (m constraints the minimum partition size).

The definition of m is approximated based on the width of the phrase image.

3.1.1 Pattern resemblance modeled as a Probability Function

To obtain a probability score based on pattern matching, a classifier can be trained and the choice of classifier can any of the classic models such as SVM, ANN, HMM or rule-based. The performance of the classifier in such images where the characters are non-uniform is largely dependent on the training data set rather than on the actual classification model. Given any image pattern containing broken pieces S, we assign $P_\alpha(S|c)$ is the 0-1 normalized classification score $g_c(S)$ for character pattern c obtained from the classifier.

$$P_\alpha(S|c) = g_c(S) \tag{3.1}$$

Because Thai is a zonal language, our model is based on a set of classifiers, one for each zone of the language. Let z_s be the derived zone of subset S, thus prompting an adaptation of the classificationn function as $g_c(S, z_s)$

$$P_\alpha(S|c) = g_c(S, z_s) \tag{3.2}$$

3.1.2 Sizing Conformity modeled as a Probability Function

To define $P_\beta(S|c)$, we compute the statistical values of each character pattern in the training set. At point of testing, the mean values are adjusted to allow for the difference between the resolution of the training document and testing document.

- Mean height and width of each character c : $\mu_{[H,c]}, \mu_{[W,c]}$ respectively
- Standard deviation height and width of each character c : $\sigma_{[H,c]}, \sigma_{[W,c]}$ respectively
- Maximum height and width of character for each c : $U_{[H,c]}, U_{[W,c]}$ respectively
- Minimum height and width of character for each c : $L_{[H,c]}, L_{[W,c]}$ respectively

Given any block S with a character class c with height h_s and width w_s, we apply lower/upper bound validation along with the normal distribution function $F(x, \mu, \sigma)$ to estimate the height conformity $P_H(S|c)$ and width conformity $P_W(S|c)$. For clarity, we show in Figure 3.1.2, how the function $P_H(S|c)$ would shape out if plotted, using a sample character class c where $\mu_{[H,c]}$ = 23.26, $\sigma_{[H,c]}$ = 2.56, $L_{[H,c]}$ = 20 and $U_{[H,c]}$ = 29. A similar behaviour can be assumed for $P_W(S|c)$, however on different range of values.

$$P_\beta(S|c) = P_H(S|c) \times P_W(S|c) \tag{3.3}$$

$$P_H(S|c) = \left\{ \begin{array}{ll} 1 & \text{if } L_{[H,c]} \leq h_S \leq U_{[H,c]} \\ F(h_S, \mu_{[H,c]}, \sigma_{[H,c]}) & \text{otherwise} \end{array} \right\}$$

$$P_W(S|c) = \left\{ \begin{array}{ll} 1 & \text{if } L_{[W,c]} \leq w_S \leq U_{[W,c]} \\ F(w_S, \mu_{[W,c]}, \sigma_{[W,c]}) & \text{otherwise} \end{array} \right\}$$

$$F(x, \mu, \sigma) = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

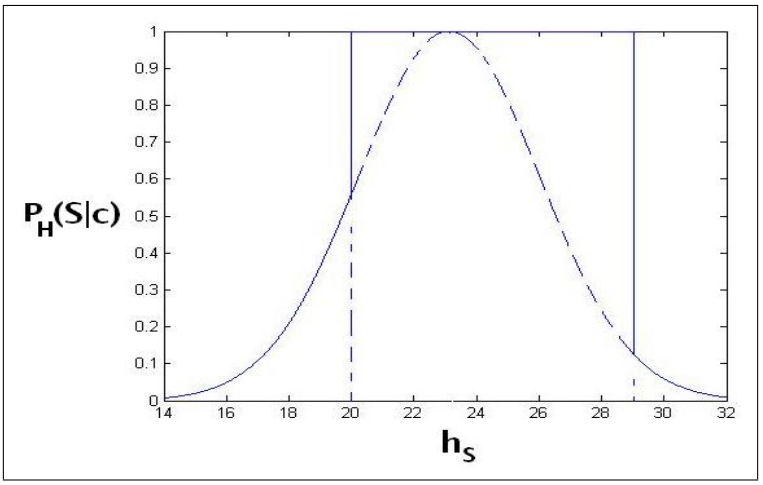


Figure 3.2: A sample of function $P_H(S|c)$

3.1.3 Cohesion modeled as a Probability Function

Given any image pattern containing broken pieces S, we assign $P_\gamma(S)$ as the probability score based on how close the pieces in the group are to each other and how far they are from the pieces of X but outside S. Let $d(x_i, x_j)$ be the distance function between broken piece x_i and x_j. We generalize the function to find the distance between any piece x_i and a set of pieces S as $d(x_i, S) = \min_{\forall x_j \in S} d(x_i, x_j)$. We estimate the separation of pieces in S as $f_1(S)$ and the separation between pieces of S with respect to $(X - S)$ as $f_2(S)$. For clarity, we show in Figure 3.1.3, the gaps that represents $f_1(X)$ and $f_2(X)$, for a sample X and S.

$$P_\gamma(S) = \left\{ \begin{array}{ll} 1 & \text{if } X = S \text{ or } |S| = 1 \text{ or } f_2(S) \geq f_1(S) \\ \frac{f_2(S)}{f_1(S)} & \text{otherwise} \end{array} \right\} \tag{3.4}$$

$$f_1(S) = \max_{\forall x_i \in S} d(x_i, S - \{x_i\})$$

$$f_2(S) = \min_{\forall x_i \in (X-S)} d(x_i, S)$$

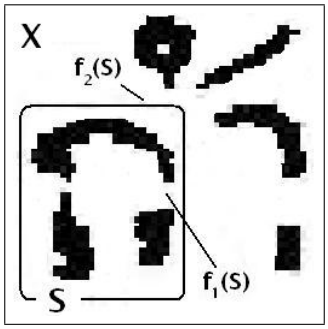


Figure 3.3: A sample of $f_1(S)$ and $f_2(S)$ functions used for computing cohesion

3.2 Partition Growing Algorithm

A simple approach to finding an optimal partition π^* is to apply the brute-force enumerative method. This would definitely yield an optimal solution but it would be very time-consuming when $|X|$ is large. We present an algorithm that effectively generates an optimal solution but is much more efficient. Our approach is based on probability functions and its mathematical implication is discussed in the next section.

3.2.1 Mathematical Discussions

We provide some fundamental definitions and some remarks on which the partition-growing algorithm is based upon.

Definition 1 *A **partition** π of X is a finite collection of non-empty subsets of X such that $(\bigcup_{S\in\pi} S) = X$ and $\forall S, T \in \pi, S \neq T, S \cap T = \emptyset$.*

Definition 2 *We define ρ, a **qualified sub-grown partition** of X, as a partition of Y where $Y \subset X$.* We shall denote a set of qualified sub-grown partitions as ρ.

Definition 3 *Let ρ be a qualified sub-grown partition with cardinality u. We define the i^{th} **comparison coefficient** denoted as $\omega_i(S, \rho)$ and its computation is shown in Eq. 3.5 where $i \geq 0$.*

$$\omega_i(S, \rho) = \frac{u \times ln\, P(\rho) + i \times ln\, P(S)}{u + i} \tag{3.5}$$

Remark 1 *Let ρ be a qualified sub-grown partition with cardinality u. Let S_1 and S_2 be blocks of X. If $P(S_1) \geq P(S_2)$, then $\omega_i(S_1, \rho) \geq \omega_i(S_2, \rho)$ for all $i \geq 1$.*

Remark 2 *Let ρ be a qualified sub-grown partition and S_k be a block being processed at iteration k. Let π^* be the best partition of X found thus far. Then, the new best partition is $\rho \cup \{S_k\}$ if the following conditions hold*

1. $(\bigcup_{S \in \rho} S) = (X - S_k)$ *and*
2. $P(\rho \cup \{S_k\}) > P(\pi^*)$

Remark 3 *Let S_k be a block being processed at iteration k. Let π^* be the best partition found thus far. Suppose that $P(S_k) \leq P(\pi^*)$. Then, the sub-grown qualified partition $\rho = \{S_k\}$ cannot be grown to an optimal partition π with $P(\pi) > P(\pi^*)$.*

Remark 4 *Let ρ be a qualified sub-grown partition with $u = |\rho|$ and S_k be block in $\mathbf{S}$ being processed at iteration k. Let π^* be the best p-partition found thus far. Suppose $ln\, P(\pi^*) > \omega_i(S_k, \rho)$, then, ρ can no longer be grown to an optimal partition π with $P(\pi) > P(\pi^*)$, where $i = p - u$.*

3.2.2 The Algorithm

Based on the mathematical discussions presemted above, here we propose a **partition-growing** algorithm that iteratively generates qualified sub-grown partitions until the optimal partition is found. Our general algorithm deals with the 3 variants discussed in Section 5.1. Here, we present the algorithm with its sub-functions and then briefly explain the algorithm to grow an optimal partition, given a set of elements and some runtime parameters.

In the algorithm, steps 1-4 generate, eliminate, evaluate blocks $\mathbf{S}$ and arrange them as a poset that will be used to grow the partitions. Step 5 initializes the default optimal partition as empty and initial sub-grown partition to contain the 1^{st} block of $\mathbf{S}$. The algorithm will terminate if no more sub-grown partitions remain or no more blocks in $\mathbf{S}$ remain to be considered (Step 5). After each outer iteration, a new set of sub-grown

function partition-growing(X**,** C**,** m**,** τ**) returns** π^*

X is set of elements, C is the set of classes,
m is minimum cardinality of π^*, τ is minimum score of any S of π^*

1. $\mathbf{S}$:= $\wp(X)$ - $\{\emptyset\}$
2. Eliminate all S from $\mathbf{S}$ that satisfies condition $|S| > |X| - m + 1$
3. Derive P(S) for all $S \in \mathbf{S}$ and eliminate all S that satisfies condition $P(S) < \tau$
4. Arrange $\mathbf{S}$ such that $S_1, S_2, ..., S_k, ...$ is based on descending value of $P(S)$.
5. $\pi^* := \emptyset$, $\rho_1 := \{\{S_1\}\}$, $k := 1$
6. while $(\rho_k \neq \emptyset) \wedge (k \leq |\mathbf{S}|)$
7. $\quad \rho_{k+1} := \emptyset$
8. $\quad$ forall $\rho \in \rho_k$
9. $\quad\quad u := |\rho|, t := max[m - u, 1]$
10. $\quad\quad \omega_1 := (ln\, P(\rho) \times u + ln\, P(S_k))/(u + 1)$
11. $\quad\quad \omega_t := (ln\, P(\rho) \times u + ln\, P(S_k) \times t)/(u + t)$
12. $\quad\quad$ if $((\bigcup \rho \cup S_k) = X) \wedge ((\bigcup \rho \cap S_k) = \phi)$ then
13. $\quad\quad\quad$ if $(\omega_1 > ln\, P(\pi^*)) \wedge (|\rho| + 1 \geq m)$ then $\pi^* := \rho \cup \{S_k\}$
14. $\quad\quad$ else
15. $\quad\quad\quad$ if $((\bigcup \rho \cap S_k) = \phi) \wedge (\omega_t > ln\, P(\pi^*))$ then $\rho_{k+1} := \rho_{k+1} \cup \{\rho \cup \{S_k\}\}$
16. $\quad\quad$ if $(\omega_t > ln\, P(\pi^*))$ then $\rho_{k+1} := \rho_{k+1} \cup \rho$
17. $\quad$ end forall
18. $\quad$ if $(P(S_k) > P(\pi^*))$ then $\rho_{k+1} := \rho_{k+1} \cup \{\{S_k\}\}$
19. $\quad$ increment k
20. end while

end function

partitions ρ_{k+1} are generated. In the inner iteration (Steps 8-17), we attempt to further the sub-grown partitions from the previous iteration using a new block S_k. Steps 9-11 initialize the comparison coefficients needed to compare scores based on Definition 3. In Steps 12-13, we check if a new complete partition is found, and compare it with the best partition π^* found thus far, and if it is better we replace it. For newly generated sub-grown partitions (Step 15) or old sub-grown partitions (Step 16), we only keep them for further iterations if they have potential to beat π^*. Finally, a newly grown partition is created containing only subset S_k (Step 18) if its individual score is better than π^*. Pictorially, we show an example in Figure 3.2.2 for X containing 4 elements and how the algorithm finds the optimal partition π^* based on the subsets S and thier scores $P(S)$.

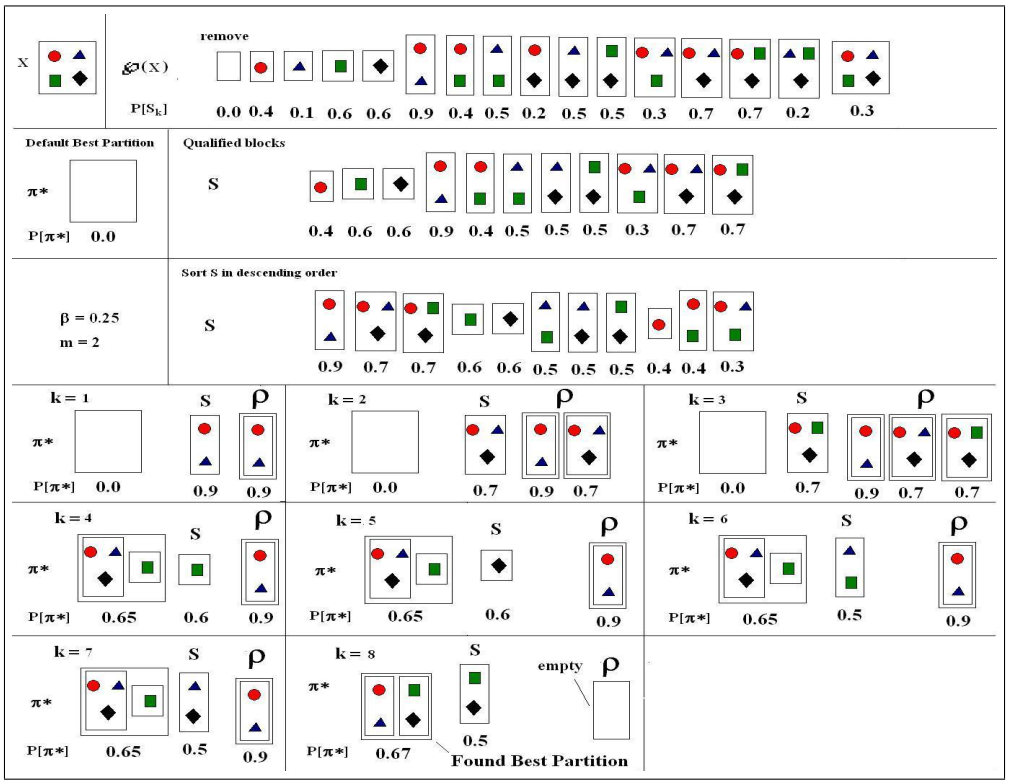


Figure 3.4: How the Partition Growing Algorithm works

3.3 Error Correction and Word Generation using N-grams Graph

In this section, we shall deal with Problem 2 as stated in Section 1.2. We focus on how to correct classification errors and generate words from a sequence of characters (obtained from the previous stage). After concatenating best sequences C of the phrases, we apply a dictionary look-up method in combination with n-grams to correct errors and generate words.

Given C as a sequence of characters each tagged with a probability $P(c_i)$ = $P(S_i|c_i^*)$ as obtained in Section 3.1, and a graph-structured dictionary, we need to find an optimal path V^* that best aligns C as formulated Equation 3.6. Such a path automatically transcribes to a sequence of words W^*.

$$W^* \equiv V^* = \underset{V}{\operatorname{argmax}}\, P(V|C) \tag{3.6}$$

3.3.1 Dictionary represented as N-grams Graph

Let each node v_i of a graph contain a character from the language and the weight of a directed edge from node v_i and v_j be the conditional probability $P(v_j|v_i)$.

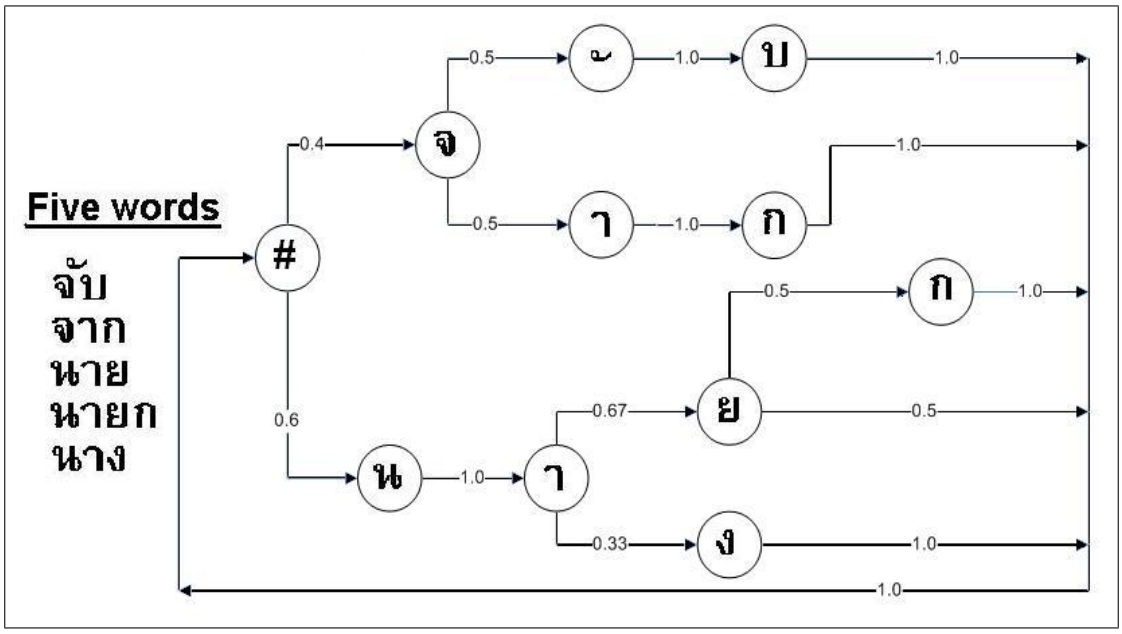


Figure 3.5: Example of N-grams Graph for a 5-word Thai Dictionary

We shall define a unique node # to represent the special start-of-word (and end-of-word) node by which we shall aim to find the best path that starts and ends at #. This path may consist of more than two visits to #, signifying more than one word is generated.

Let $N(v)$ denote the number of different non-cyclic paths in the graph from v to $\#$. The weight on the directed edge from node v_i to node v_j have been computed using the Equation 3.7. An example of a graph generated from 5 words is shown in Figure 3.5.

$$P(v_j|v_i) = \frac{N(v_j)}{N(v_i)} \tag{3.7}$$

3.3.2 Searching the N-grams graph

To generate the best path, we utilize the well-known A^* search. A path-cost from initial state to state Q is given by function $g(Q)$ and a heuristic estimate of the distance from current state Q to the goal state is given by $h(Q)$. These functions along with the actual search evaluation function $e(Q)$ are described in equation 3.8.

$$\begin{aligned} g(Q) &= P(V_Q|C_Q) \\ h(Q) &= \textstyle\prod_{i=|C_Q|+1}^{|C|}(P(c_i) \times P_\chi) \\ e(Q) &= g(Q) \times h(Q) \end{aligned} \tag{3.8}$$

where Q is the current state, C is the full sequence of characters being parsed, V_Q is the sub-path generated in state Q, C_Q is the sub-sequence of characters parsed in Q and P_χ is the mean of $P(v_2|v_1)$ over the entire n-gram graph.

We recursively denote $v_1.V$ as a path starting with node v_1 followed by sub-path V and denote the smallest eligible path as $\{\#\}$. We also recursively denote $c_1.C$ as a sequence starting with character c_1 followed by a subsequence C and denote a sequence of character of zero length as NULL. Let $P(v_2|v_1)$ be the edge transition probability from node v_1 to v_2 and let $P(c)$ be the probability that character c was correctly classified, see Section 3.1. Let $c(v)$ denote the character stored in graph node v. To facilitate the search, we define operators that allow us to generate paths in the graph.

3.3.3 Word Generation Operators

In a perfect world, where C needs no correction, the following recursive model enables the generation of paths V based on which we can estimate $P(V|C)$. As in any recursive model, there must a defined base case and we deem the generation process complete if the path generated ends at the # node and the sequence of the characters being process is exhausted.

- Base case : $P(\{\#\}|NULL) = 1$.
- Forward moving operator (Normal case)
  Pre-Condition: $V \neq \{\#\}$ and $C \neq$ NULL and $c(v_2) = c_1$

$$P(V|C) = P(v_1.v_2.V'|c_1.C') = P(c_1) \times P(v_2|v_1) \times P(v_2.V'|C') \tag{3.9}$$

- Cycle completion operator (End-of-Word case)
  Pre-condition: $V \neq \{\#\}$ and $v_2 = \#$

$$P(V|C) = P(v_1.v_2.V'|C) = P(v_2|v_1) \times P(v_2.V'|C) \tag{3.10}$$

3.3.4 Error Correction Operators

We need to address the obvious situation that errors may exist in character sequence C. We apply the conventional correction operations namely - insertion, deletion

and substitution. We pre-compute the following probabilities and use them to define the model for the error correction operators.

1. P_I is the probability that a character was almost completely blurred out on the document and hence requires insertion. $P_I = 1-(1-P_1)(1-P_2)$ where P_1 is the probability that a character is completely blurred out and P_2 is the probability that two character in the original image is recognized wrongly as a single character.

2. P_D is the probability that a group of some noise pieces was identified as a character and hence requires deletion. $P_D = 1-(1-P_3)(1-P_4)$ where P_3 is the probability that a large noise piece is recognized wrongly as a character and P_4 is the probabilitty that a broken character image is recognized wrongly as two characters.

3. P_T is the probability that a character was misclassified as another character and hence requires substitution. $P_T = 1-(1-P_5)(1-P_2)(1-P_4)$ where P_5 is the probability that a pattern is misclassified by the trained classifer and P_2, P_4 are defined as above.

The values for $P_i(1 \leq i \leq 5)$ are appromated based on the quality of the document being recognized.

- Insertion operator

  Pre-condition: $V \neq \{\#\}$ and ($C =$ NULL or $c(v_2) \neq c_1$)

$$P(V|C) = P(v_1.v_2.V'|C) = P_I \times P(v_2|v_1) \times P(v_2.V'|C) \tag{3.11}$$

- Deletion operator

  Pre-condition: $C \neq$ NULL and ($V = \{\#\}$ or $c(v_2) \neq c_1$)

$$P(V|C) = P(v_1.v_2.V'|c_1.C') = P_D \times (1 - P(c_1)) \times P(V|C') \tag{3.12}$$

The insertion and deletion operators are straightforward. However, given the dictionary is large, a substitution at node # may lead to a very large search space. To solve this performance related issue, we utilize the confusion sets to enhance the performance of

substitutions. For each character c in the language, we create a corresponding confusion set ψ_c based on classifier training results. We then define $P_\psi(c, c_i)$ as the probability that c can be confused as c_i. These values for a particular confusion are normalized such that $\sum_{i=1}^{|\psi_c|} P_\psi(c, c_i) = 1 - P_\xi$, where P_ξ is the pre-computed constant probability that a character was misclassified as another character outside its confusion set. We also limit substitutions to characters in the same zones only. Let z_c denote the zone of character c.

- Substitution operator

  Pre-condition: $V \neq \{\#\}$ and $C \neq$ NULL and $c_1 \neq c(v_2)$ and $z_{c_1} = z_{c(v_2)}$

$$P(V|C) = P(v_1.v_2.V'|c_1.C') = P_T \times (1 - P(c_1)) \times P(v_2|v_1) \times P(v_2.V'|C')$$

$$P_T = \left\{ \begin{array}{ll} P_\psi(c_1, c(v_2)) & \text{if } c(v_2) \in \psi_{c_1} \\ P_\xi & \text{if } v_1 \neq \# \text{ and } c(v_2) \notin \psi_{c_1} \\ 0 & \text{otherwise} \end{array} \right\} \tag{3.13}$$

CHAPTER IV
IMPLEMENTATION

All works researched and experimented are implemented in MATLAB 7. All images captured and produced are in JPEG format. The pages of Thai Historical document - Kod Mai Tra Sam Duang were captured as images of resolution - 150 x 150 using a standard scanner. Some of the pages that were skewed during scanning process were de-skewed manually using standard image processing tools. We present the methods employed in our implementation for each stage of the entire process. For more related information, refer to Appendix D.

4.1 Training the Classifiers

To train our classifiers, we first extracted 100 samples for each character in the Thai script. Certain characters are scarcely found in the document, so we used what could be found and generated more samples by slightly perbutating them. All character images are resized to 24x24. We employed Daubechies Wavelet 4 (Level 1) for our feature extraction using the MATLAB function **wavedec2**. This generated a large feature vector, so we applied the standard PCA method (MATLAB function **pcaconv**) to reduce the feature space with 0.80 explained. These reduced feature vectors were then normalized to values between 0 and 1.

Feed-forward Neural Networks are trained for each of the zones of the Thai language characters. Each network has a 3-layered structure - Input, Hidden and Output. The size of input layer is determined by the feature vector size, the size of the output layer is determined by the number of character classes in that zone. The hidden layer nodes is kept very high = $k\times$ output layer size and we set $k = 3$. We used the MATLAB functions **newff and train** to generated the trained networks with maximum epoch = 10000 and minimum gradient set at 1e-6. The initial weights of the network edges are randomized by the system.

4.2 Pre-processing

The colored images captured as .jpg files were converted to grayscale using the standard MATLAB function **rgb2gray** and optionally enhanced so that the grayscale covers the entire range 0-255. Next, to binarize our images, we coded the **Isodata** method [69] that is adequate in computing thresholds of the document pages that were original black and white in nature with noise evenly spread out.

4.3 Segmentation

As the main input of our works is a line segment image, we needed a method that could segment the document into lines. Using the standard histogram technique to segment this document proved inadequate. Thai language contains three zones and characters in lower zone of i^{th} line may overlap the upper zone characters of $(i+1)^{th}$ line. Moreover, given the hand-printed nature of these documents, there were unpredictable inclination angles between lines. To overcome the mentioned issues, we executed an approximate segmentation using a simple histogram technique and then employed the **piece-wise segmentation** method [104] to accurately generate line images.

4.4 Recognizing the Broken Characters

The main works of this research - Recognizing Broken Characters and Error Correction are described in Chapter 3 and are completely implemented using MATLAB. The main steps taken include

1. For each line segmented from the original document image, estimate the upper line and base line of the middle zone characters using the histogram method.

2. Employ the MATLAB function **imdilate** to close some of the tiny gaps between pieces in the three different zones. We ensure that pieces across zones are not merged. This step is taken to ensure that the number broken pieces are reduced and tiny gaps are ignored.

3. We detect the small and big vertical gaps in the line image using the histogram method. These gaps are used to generate phrase images and sets of broken pieces.

4. For each set X, we generate the power set and for each subset S we derive the zone. Based on pattern and the zone, we obtain a class and its classification score using the probability model. We then execute the Partition Growing algorithm treating the problem as Variant I, to generate a best sequence of characters for each X. The results are then concatenated to generate sentences (using the large vertical gaps to denote end of sentence).

5. The characters in each sentence is passed on the N-grams graph to correct any errors and generate words.

4.5 Performance Enhancement

Although the Partition-Growing algorithm is able to yield an optimal solution, its performance is still in the order of exponential time and hence when presented with a large set X, the algorithm might be drastically slow. In this section, we present an approach to reduce X to an acceptable size before applying the partition-growing algorithm.

function set-reduction(X**,** m**,** L**,** G**) returns** π^*
X is set of broken pieces, m is min. cardinality of π^*
L is minimum acceptable cardinality of X
G is largest acceptable distance between two pieces of X

1. if $|X| \leq L$ then π^* := **partition-growing(**X**,**m**)**
2. else
3. d := $\min_{x_i, x_j \in X}$ distance(x_i, x_j) $\quad i \neq j$
4. if $d \leq G$ then
5. X' := $X - \{x_i, x_j\} \cup \{group(x_i, x_j)\}$
6. π^* := **set-reduction(**X'**,**m**,**L**,**G**)**
7. else
8. π^* := **set-reduction(**X**,**m**,**$|X|$**,**G**)**

end function

Let L be the acceptable size of X whereby running the partition-growing algorithm is acceptable fast. Let G be the largest acceptable distance between two pieces in X such that we can presume that they belong to the same character. The actual definitions of L and G determines the trade-off between accuracy and speed, where

small values of L or large values of G will quicken the process but may not yield accurate results and vice versa.

We present a recursive **set-reduction** algorithm that first reduces X to an acceptable size before calling the partition-growing function. The main focus is on grouping 2 closest pieces in X to reduce the size of X by 1. This is repeated until the size of X becomes acceptably small. However, if the 2 closest pieces are too far apart when compared to G, then the set reduction process terminates and partition-growing is handed a set that is larger than L.

CHAPTER V
SOLVING OTHER APPLICATIONS WITH PARTITION GROWING ALGORITHM

In this chapter, we discuss a generalization of the Partition Growing algorithm to allow for solving other set partition problems of a specific nature.

5.1 Set Partitioning Problem Modeled using Probability Functions

Given a set of elements X and a set of classes C, we wish to find an optimal p-partition π^* of X that maps each block S in π^* to a class in C, such that $p \geq m$ where m is pre-defined based on some problem parameters. Based on a set of qualities $\alpha, \beta, \gamma, ...$, we determine to which class c^* the block S is best mapped to. This problem is formulated mathematically as

$$\textbf{Maximize } \frac{1}{|\pi|}\sum_{S\in\pi} ln\, P(S) \qquad |\pi| > m$$
$$\text{where } P(S) = P_\alpha(S|c^*) \times P_\beta(S|c^*) \times P_\gamma(S|c^*) \times ...$$
$$\text{and } c^* = \underset{c}{\operatorname{argmax}}\, P_\alpha(S|c) \times P_\beta(S|c) \times P_\gamma(S|c) \times ...$$

The actual definition of $P_\alpha(S|c), P_\beta(S|c), P_\gamma(S|c)...$ depends on the specific problem. They may be linear or non-linear functions that evaluates some quality of block S and assigns a probability value in the range [0..1]. We shall assume that $P(\emptyset)$ is 0.

We shall only deal with only a specific sub-class of set-partitioning problems that satisfies the following properties:

- The optimality of π^* depends directly on the probability that each individual block $S \in \pi^*$ can be mapped optimally to some class in C and each block also exhibits a high level of grouping quality.
- $P(S_i)$ can be computed independent of any $P(S_j)$ where $i \neq j$.
- $P(S)$ does not directly depend on any individual element of block S but rather on the collective qualities of its members.

There are 3 variants of the set-partitioning problem that we shall be dealing with. The problem variant must be pre-determined before the algorithm can be effectively applied. For clarity, we pictorially describe these problem variants in Figure 5.1.

1. There exists a non-injective and non-surjective mapping between blocks S and classes C (Many-to-One Mapping).

2. There exists an injective and non-surjective mapping between blocks S and classes C (One-to-One Mapping).

3. There exists an injective and surjective mapping between blocks between S and classes C (One-to-One Correspondence).

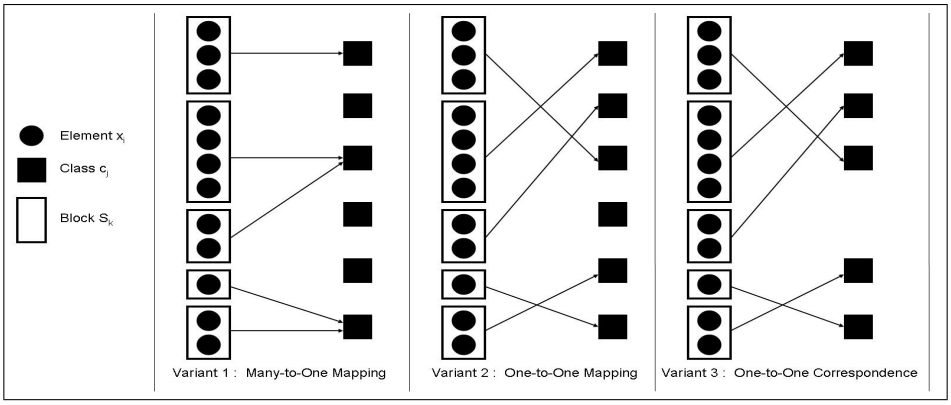


Figure 5.1: Different Set Partitioning Problem Variants

We now propose an extension of the **partition-growing** algorithm presented earlier in Section 3.2 that generates qualified sub-grown partitions until the optimal partition is found. Our general algorithm deals with the 3 variants discussed in Section 5.1. Here, we present the algorithm with its sub-functions and then briefly explain the algorithm to grow an optimal partition, given a set of elements and some runtime parameters.

Because our solution is designed to handle 3 problem variants, the sub-functions **found-solution** and **can-grow** have been separately designed to distinguish the logic needed to handle them based on the input parameter v that denotes the problem variant.

Defining m is quite straightforward as we shall see in Section 6.2, whereas defining τ can be a more tricky issue. Underestimating m could lead to slow convergence

function partition-growing-extended(X, C, v, m, τ) **returns** π^*

X is set of elements, C is the set of classes, v is the problem variant,
m is minimum cardinality of π^*, τ is minimum score of any S of π^*

1. if $v = 1$ then $\mathbf{S}$:= $\wp(X)$ - $\{\emptyset\}$ else $\mathbf{S}$:= ($\wp(X)$ - $\{\emptyset\}$) $\times$ C
2. Eliminate all S from $\mathbf{S}$ that satisfies condition $|S| > |X| - m + 1$
3. Derive P(S) for all $S \in \mathbf{S}$ and eliminate all S that satisfies condition $P(S) < \tau$
4. Arrange $\mathbf{S}$ such that $S_1, S_2, ..., S_k, ...$ is based on descending value of $P(S)$.
5. $\pi^* := \emptyset$, ρ_1 := $\{\{S_1\}\}$, $k := 1$
6. while ($\rho_k \neq \emptyset$) $\wedge$ ($k \leq |\mathbf{S}|$)
7. $\quad \rho_{k+1} := \emptyset$
8. $\quad$ forall $\rho \in \rho_k$
9. $\quad\quad u := |\rho|$, $t := max[m - u, 1]$
10. $\quad\quad \omega_1 := (ln\, P(\rho) \times u + ln\, P(S_k))/(u + 1)$
11. $\quad\quad \omega_t := (ln\, P(\rho) \times u + ln\, P(S_k) \times t)/(u + t)$
12. $\quad\quad$ if **found-solution**(ρ, S_k, X, C, v) then
13. $\quad\quad\quad$ if ($\omega_1 > ln\, P(\pi^*)$) then $\pi^* := \rho \cup \{S_k\}$
14. $\quad\quad$ else
15. $\quad\quad\quad$ if **can-grow**(ρ, S_k, v) $\wedge$($\omega_t > ln\, P(\pi^*)$) then ρ_{k+1} :=$\rho_{k+1} \cup \{\rho \cup \{S_k\}\}$
16. $\quad\quad$ if ($\omega_t > ln\, P(\pi^*)$) then ρ_{k+1} :=$\rho_{k+1} \cup \rho$
17. $\quad$ end forall
18. $\quad$ if ($P(S_k) > P(\pi^*)$) then ρ_{k+1} :=$\rho_{k+1} \cup \{\{S_k\}\}$
19. $\quad$ increment k
20. end while

end function

function found-solution(ρ, S, X, C, v) **returns boolean**

1. $\pi = \rho \cup S$
2. if ($\bigcup \pi = X$) $\wedge$ (($\bigcup \rho \cap S) = \emptyset$)
3. $\quad$ if $v = 1$ return TRUE
4. $\quad$ if $v = 2$ return $|\bigcup_{S_i \in \pi} class(S_i)| = |\pi|$
5. $\quad$ if $v = 3$ return $|\bigcup_{S_i \in \pi} class(S_i)| = |C|$
6. else return FALSE

end function

function can-grow(ρ, S, v) **returns boolean**

1. $\rho' = \rho \cup S$
2. if (($\bigcup \rho \cap S) = \emptyset$)
3. $\quad$ if $v = 1$ return TRUE
4. $\quad$ else return $|\bigcup_{S_i \in \rho'} class(S_i)| = |\rho'|$
5. else return FALSE

end function

of the algorithm whereas an underestimation of τ will lead higher requirement of internal memory. It must be stressed here that overestimations of m or τ might lead to sub-optimal solutions or no solution being found.

5.2 Remodeling some other Set Partitioning Problems

We have chosen three set partitioning problems (SPP) (one for each problem variant defined in section 5.1) to demonstrate the applicability of our methodology towards solving different set partitioning problems.

5.2.1 Steel Mill Slab Design - Variant 1

Steel is produced by casting molten iron into slabs. A steel mill can produce a finite number of slab capacities. An order has two properties: a size and a color that corresponds to the route required through the mill. Given n input orders, the problem is to assign the orders to slabs such that the total wastage is minimized. This assignment is subject to two constraints -

- Capacity constraints - The total size of the orders assigned to a slab cannot exceed the largest slab capacity.
- Color constraints - Each slab can contain at most 2 colors (max. 2 routes).

Recently, researchers [64] [65] have extended the problem to look for solutions that involve bigger slabs. We shall also deal with this extension in our model.

Given a set of orders X with assigned weights W and color L, and a set of slab capacities C, we wish to find the optimal p-partition π $(p \geq m)$ that

Maximize $\frac{1}{|\pi|}\sum_{S\in\pi} ln\, P(S)$ $\qquad P(S) = P_\alpha(S|c^*) \times P_\beta(S) \times P_\gamma(c^*)$

where $P_\alpha(S|c)$ is the probability that orders S will fully utilize a slab capacity c

and $P_\beta(S)$ is the probability that orders S involves minimal colors

and $P_\gamma(c)$ is the probability that capacity c is a relatively large slab

and $c^* = \underset{c}{\operatorname{argmax}}\, P_\alpha(S|c) \times P_\gamma(c)$

and m is the minimum slabs required, set at $\frac{1}{max(C)}\sum W(X)$

To define $P_\alpha(S|c)$, we measure how well the orders in S fits into capacity c without

exceeding it. The function promotes usage of bigger slabs in case the total wastage $\neq 0$.

$$P_\alpha(S|c) = \left\{ \begin{array}{ll} \frac{1}{c}\sum_{X_i \in S} W(X_i) & \text{if } c \geq \sum_{X_i \in S} W(X_i) \\ \\ 0 & \text{otherwise} \end{array} \right\} \tag{5.1}$$

To define $P_\beta(S)$, we apply a small penalty to S having 2 colors and reject any that have more than 2 colors. Let $|L(S)|$ be the number of unique colors for the orders in S. Note that $P_\beta(S)$ can be computed for S independent of any choice of c. We predefine P_ϵ as the probability that 1-color slab is preferred over 2-color slab.

$$P_\beta(S) = \left\{ \begin{array}{ll} 1 - P_\epsilon \times (|L(S)| - 1) & \text{if } |L(S)| \leq 2 \\ \\ 0 & \text{otherwise} \end{array} \right\} \tag{5.2}$$

To define $P_\gamma(c)$, we apply a linearly proportional penalty to smaller slabs. This function is required to distinguish solutions with bigger slab from small slabs, especially when the total wastage = 0. We predefine P_ψ as the probability that bigger slabs are preferred over smaller slabs.

$$P_\gamma(c) = 1 - P_\psi \times (max(C) - c) \tag{5.3}$$

5.2.2 Wedding Planner - Variant 2

Given a set of wedding attendees, a wedding planner must come up with a seating plan to maximize the happiness of all the guests. The happiness of a table can be determined by the individual happiness of the guests at each table. There is a set of tables with pre-defined sizes and thus we wish to minimize the number of tables needed.

Initial problems as treated in [61] [63] only focused on happiness based on undirected pair-wise relationship between guests by using the alphabets as the distance measure for simplicity. We adopt a similar approach by estimating happiness based on the ages of the guest and their gender. Grouping guests with same age group makes them happier and a good mixture of the genders at each table adds further to the guests' happiness.

Given a set of guests X with ages A and gender G (-1 as female, 1 as male), a set of tables with size C, we wish to find the optimal p-partition π $(p \geq m)$ that

Maximize $\frac{1}{|\pi|}\sum_{S\in\pi} ln\, P(S)$ $\qquad P(S) = P_\alpha(S|c^*) \times P_\beta(S) \times P_\gamma(S)$

where $P_\alpha(S|c)$ is the probability that guests S will fully utilize a table c

and $P_\beta(S)$ is the probability that guests S are age-wise cohesive

and $P_\gamma(S)$ is the probability that guests S exhibits a good mixture of gender

and $c^* = \underset{c}{\operatorname{argmax}}\, P_\alpha(S|c)$

and m is the minimum number of tables, set at $\frac{1}{max(C)}|X|$

and there exists a injective and non-surjective mapping between S and C

To define $P_\alpha(S|c)$, we seek to penalize solutions that leave empty seats on the tables. Let θ denote the occupancy of a table and computed as $\frac{|S|}{c}$. Let ϵ be the minimum occupancy of a table that is acceptable. Occupancy less then ϵ is allowed but not preferred.

$$P_\alpha(S|c) = \left\{ \begin{array}{lll} sin(\frac{\theta}{2}\pi) & \text{if } \theta \geq \epsilon & \pi \textit{ is in radians} \\ \\ sinh^2\theta & \text{otherwise} & \end{array} \right\} \tag{5.4}$$

To define $P_\beta(S)$, we pair-wise accumulate excess age gap between all pairs of guests in S by defining a sub-function $d(S)$. Excess age gap is defined based on how much cohesion the planner requires for each table. Let a be the threshold age gap. To distinguish between cardinality and absolute function, we use $||x||$ to denote absolute value of x.

$$d(S) = \sum_{X_i\in S}\sum_{X_j\in S} \left\{ \begin{array}{ll} 0 & \text{if } ||A_i - A_j|| \leq a \\ \\ ||A_i - A_j|| - a & \text{otherwise} \end{array} \right\} \tag{5.5}$$

$$P_\beta(S) = 1 - \frac{d(S)}{a \times |S|} \tag{5.6}$$

To define $P_\gamma(S)$, we compute the difference in count of the males and females on the table. Large variance leads to lesser happiness score.

$$P_\gamma(S) = 1 - \frac{||\sum_{X_i\in S} G_i||}{\sum_{X_i\in S} ||G_i||} \tag{5.7}$$

5.2.3 Task Allocation - Variant 3

Given a set of tasks to be performed in a workshop with a set of workers on the roster, the workshop foreman needs to assign the tasks to the workers based on their

skill levels. This is a set partitioning problem [63] where the tasks must be partitioned to multiple subsets with each subset assigned to a worker. The main objective to find the best match based on the required skills of each task and the skills of the worker the task is assigned to, in order to **achieve overall quality output** at the workshop. Moreover, the tasks must be **completed in the smallest timeframe** without compromising on the quality of output. We assume that skill levels are normalized to be in the range 0 to 5.

Given a set of tasks X, each requiring skills Z and effort D, a set of workers C with skills Y we wish to find the optimal p-partition π $(p \geq m)$ that

Maximize $\frac{1}{|\pi|}\sum_{S\in\pi} ln\, P(S) \qquad P(S) = P_\alpha(S|c^*) \times P_\beta(S)$

where $P_\alpha(S|c)$ is the probability worker c is qualified to complete tasks S

and $P_\beta(S)$ is the probability that tasks S will be completed within time-frame

and $c^* = \underset{c}{\operatorname{argmax}}\, P_\alpha(S|c)$

and $m = |C|$ (each worker is assigned exactly 1 subset)

and there exists a injective and surjective mapping between S and C

To define $P_\alpha(S|c)$, we compare the skill set Z_i required for each task X_i assigned to a worker c against his skill set Y. Let A be the universal set of skills used in the workshop. We use a sigmoid function to compute the probability that a worker is the best person for the task. We heavily penalize workers that are under-qualified for the task and slightly penalize workers over-qualified for task assigned. Let ν be a small penalty for assigning a task to a worker with higher than needed skills.

$$P_\alpha(S|c_j) = \prod_{i=1}^{|S|}\prod_{k=1}^{|A|} \left\{ \begin{array}{ll} (1 + e^{Z_{ik}-Y_{jk}})^{-1} & \text{if } Y_{jk} < Z_{ik} \\ \\ 1 - \nu \times (Z_{ik} - Y_{jk})^2 & \text{otherwise} \end{array} \right\} \tag{5.8}$$

To define $P_\beta(S)$, we employ the Gaussian function. Let h_s be the duration of time needed to complete all tasks in subset S by a worker. We estimate expected completion time for all tasks as $\mu = \frac{1}{|C|}\sum D$ and σ is the standard deviation computed from D.

$$P_\beta(S) = e^{-\frac{(h_S-\mu)^2}{2\sigma^2}} \tag{5.9}$$

CHAPTER VI
EXPERIMENT RESULTS

In this chapter, we present the experiments performed related to our research on the Partition Growth Algorithm. The experiments are aimed to -

- *Validate the effectiveness of the solution towards solving the Broken Characters Recognition problem.* In Section 6.1, we test the algorithm on three different documents contain broken characters. In addition, one of the documents is in English to show the generality of the solution towards handling more than one language.

- *Demonstrate applicability of the algorithm towards our Set Partitioning Problems.* In Section 6.2, we apply the algorithm towards solving some of the existing problems that are linear and non-linear in nature. We choose problems that exhibit different characters in terms of injective and surjective mapping.

- *Analyze performance of the algorithm.* In Section 6.3, we perform simulation experiments to measure the various aspects of the algorithm. We also check the reliability of the algorithm towards generating an optimal solution.

6.1 Broken Characters Recognition

To test our methodology, we conducted 3 main experiments. In the first experiment, we focused on recognizing broken characters in Thai historical document using our proposed method. In the second experiment, we tested our methodology on an English-based Historical document to show that our method can be adapted to handle other languages as well. In the third experiment, we tested our approach on a Thai facsimile document that contains broken characters.

6.1.1 Testing our Methodology on Thai Historical Document

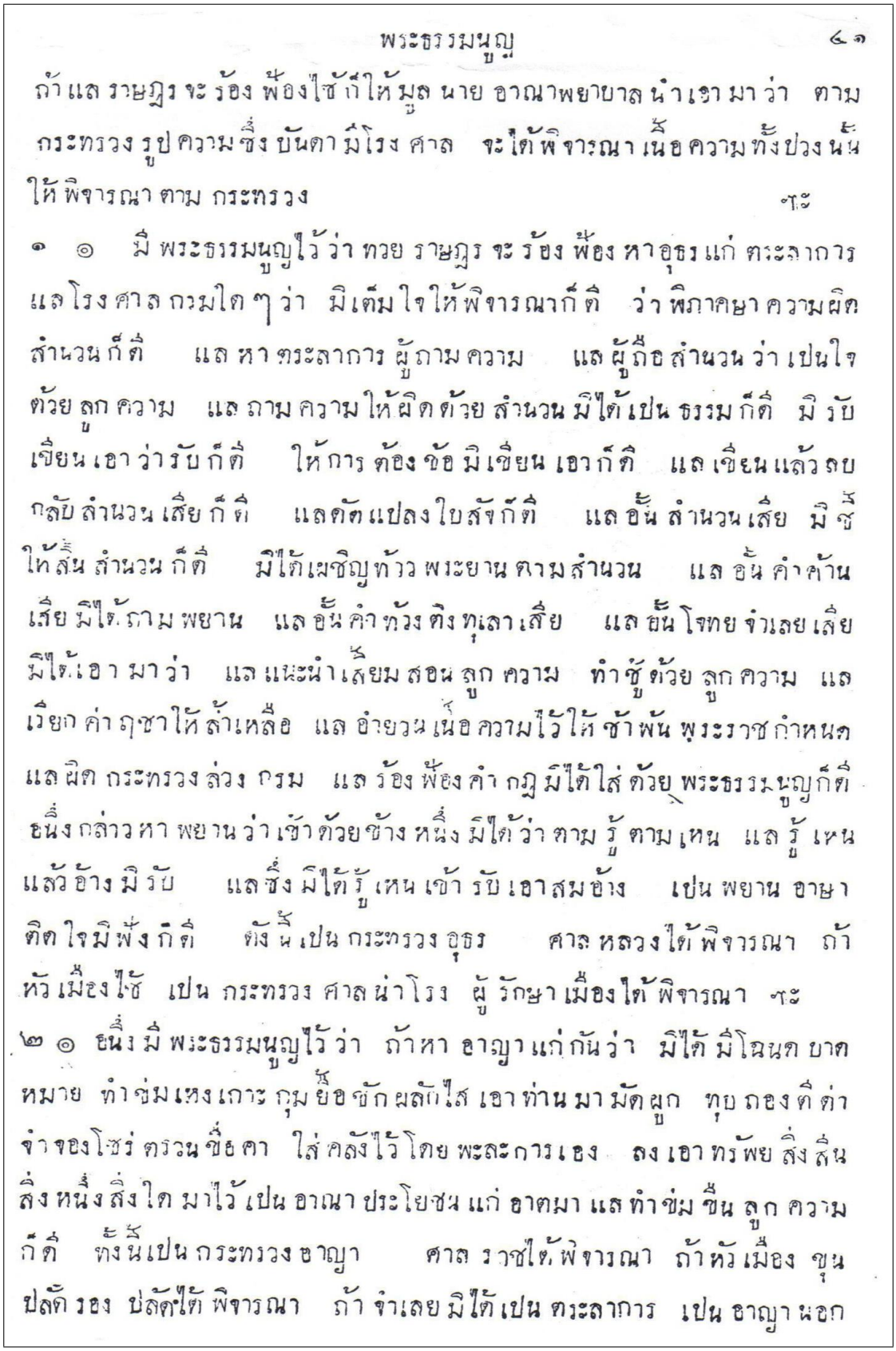
To conduct our experiments, we chose a Thai Law Historical Document "Kod Mai Tra Sam Duang" dated back to 15th century, which we obtained from the Thai National Library. Our training data contains 100 samples of each character, each instance differing to an extent in terms of pattern, stroke thickness, level of breakdown, level of blurring etc.

For testing, we chose another 10 pages from the same historical document. We implemented the Partition-growing algorithm with m=1 and τ=0.25 to recognize the broken characters on these pages. On obtaining the sequence of characters, we employed the error correction strategy based on N-grams graph and showed how it can be used to generate Thai words from a sequence of Thai characters.

A sample input image is shown in Figure 6.1, which was pre-processed and segmented automatically before applying the recognition and correction processes. Zoomed images from this document can be seen in Figure 1.1 found in Chapter 1. The recognition results after applying the recognition strategy without applying any error correction are shown in Figure 6.2. After applying error correction, the results is shown in Figure 6.3 with the misclassified characters highlighted. We adopt Levenshtein edit distance method to measure error rate and we obtain a character error rate of 6.93% and on further application of the error correction process, we are able to reduce the error rate to 4.02%. For reference, we show the detailed results in Table 6.1.

6.1.2 Testing our Methodology on Faxed Thai Document

To demonstrate that our proposed methodology is also applicable to modern documents that contain broken characters created due to other distortion sources, we chose to fax a Thai document. A zoomed segment of the received faxed image is shown in Figure 6.4, where almost all the characters are broken to a certain degree. This original document contains 15 pages from which we extracted 2 samples of Thai characters and the unbroken patterns were used to train the classifier. We then tested the solution on the 15 pages and obtained a error rate of only 6.38% without applying error correction. After applying a modern-dictionary based on the n-grams graph method, we were able to reduce the error rate to 2.79%. The original clean image of a page is shown in Figure



พระธรรมนูญ ๔๑

ถ้าแล ราษฎร ทะร้อง พ้องใช้ก็ให้มูล นาย อาณาพยาบาล นำเอา มา ว่า ตาม กระทรวง รูปความซึ่ง บันดา มีโรง ศาล ทะได้พิจารณา เนื้อความทั้งปวงนั้น ให้พิจารณา ตาม กระทรวง ฯะ

๑ ๏ มี พระธรรมนูญไว้ ว่า ทวย ราษฎร ทะ ร้อง พ้อง หา อุธร แก่ ตระลาการ แลโรง ศาล กรมใด ๆ ว่า มิเต็มใจให้พิจารณาก็ดี ว่าพิภาคษา ความผิด สำนวนก็ดี แล หา ตระลาการ ผู้ถาม ความ แล ผู้ถือ สำนวน ว่าเปนใจ ด้วย ลูกความ แล ถาม ความให้ผิดด้วย สำนวน มิได้เปน ธรรม ก็ดี มิ รับ เขียน เอา ว่ารับก็ดี ให้การ ต้อง ข้อ มิเขียน เอาก็ดี แล เขียน แล้ว ลบ กลับ สำนวน เสีย ก็ดี แลตัดแปลง ใบสัจก็ดี แล อั้น สำนวน เสีย มิ ชี้ ให้สิ้น สำนวน ก็ดี มิได้เผชิญ ท้าว พระยาน ตาม สำนวน แล อั้น คำค้าน เสีย มิได้ถาม พยาน แล อั้น คำ ท้วง ติง ทุเลา เสีย แล อั้น โจทย ขำเลย เสีย มิได้เอา มา ว่า แล แนะนำ เสียม สอน ลูกความ ทำชู้ด้วย ลูกความ แล เวียก ค่า ฤชา ให้ล้ำเหลือ แล อำยวน เนื้อความไว้ ให้ช้าพ้น พระราช กำหนด แล ผิด กระทรวง ล่วง กรม แล ร้อง พ้อง คำ กฎ มิได้ใส่ ด้วย พระธรรมนูญก็ดี อนึ่ง กล่าว หา พยาน ว่า เข้าด้วย ข้าง หนึ่ง มิได้ว่า ตาม รู้ ตาม เห็น แล รู้ เห็น แล้ว อ้าง มิ รับ แล ซึ่ง มิได้รู้เห็น เข้า รับ เอา สม อ้าง เปน พยาน อาษา ติด ใจ มิ พึ่ง ก็ดี ดังนี้ เปน กระทรวง อุธร ศาล หลวง ได้พิจารณา ถ้า หัวเมืองใช้ เปน กระทรวง ศาล น่าโรง ผู้รักษา เมือง ได้พิจารณา ฯะ

๒ ๏ อนึ่ง มี พระธรรมนูญไว้ ว่า ถ้าหา อาญา แก่กันว่า มิได้ มิโฉนด บาด หมาย ทำชิ่มเหง เกาะ กุม ย้อ ชัก ผลัก ใส เอา ท่าน มา มัดผูก ทุบ ถอง ตี ต่า จำ ของโซ่ ตรวน ขื่อคา ใส่คลังไว้ โดย พละการ เอง ลง เอา ทรัพย สิ่งสิ้น สิ่งหนึ่งสิ่งใด มาไว้ เปน อาณา ประโยชน แก่ อาตมา แล ทำชิ่ม ขืน ลูกความ ก็ดี ดังนี้เปน กระทรวง อาญา ศาล ราชได้พิจารณา ถ้าหัวเมือง ขุน ปลัด รอง ปลัดได้ พิจารณา ถ้า จำเลย มิได้เปน ตระลาการ เปน อาญา นอก

Figure 6.1: Original page from Thai Historical Document

พระธรรมนูญ ๔๑

ถ้าแลราษฎรจะร้อง ฟ้องไซ้ก็ให้มูล นาย อาณาพยาบาลนำเอามาว่า ตาม กระทรวงรูปความซึ่งบันดามี่โรง ศาล จะได้พิจารณาเนื้อความทั้งปวงนั้ม ให้พิจารณาฑม กระทรวง ฯะ

๑ ◉ มี พระธรรมนูญไว้ว่าทวย ราษฎร จะร้ปง พ้องหาอุธรแข่ ตะละลาการ แลโรงศาลกามใด ๆว่า มิเต็มใจิเห้พิจารณาก็ดี ว่าพิภาคษาความผิด สำนฤนก็ดี แล หาตระลาการ ผู้ถามความ แลผู้ถือสำนวนว่า เปนใ ด้วยลูกความ แลถามภวามให้ผิดด้ายสำนวนมิได้เปนธรรมก็ดี มิ รับ เขียนเอาว่ารับก้ดี ให้การ ต้องข้อ มิเขียน เอาก็ดิ แลเขียนแล้วลบ กลับลำนวนเสียก็ติ แลดัดแปลงใบสัจก็ดี แลอ้น สำนวนเสีย มิชื ิเห้สื้น สำนวนศดี มิไภ้เนชิญท ๆวพงะยานตามสำนวน แล อั้น คำค้าน เสียมิไภ้ถามพยาน แลอั้นคำท้วงติงทุเลาเไย แลอั้นโจทยฬวเลยเลีย มิไภ้เอา มาว่า แลแนะนำเสี้ยมสอนลูก ความ ทำชู้ด้วยลูกความ เะล เวียก คำฤชาให้ล้ำเหลือธ ะเล ปายวนเนื้อความไว้ให้ช้าพ้นพระราชกำหนต แลผิด กระทรวงล่วง กรม แล ร้องพ้องคำกฎม์ได้ใส่ด้วยพระธรรมนูญก็ดี อนึ่งกล่าวหา พยานว่ำเข้าด้วยข้างหนึ่งม์ได้ว่าตาม รู้ตามเหน แล รู้เหน แล้วอ้างมิรับ แลซึ่ง มิได้รู้เหน เข้า รับเอาสมอ้โง เปน พยาน อาษา ติดิเจมิพังก็ดี ดังพี้เปนกระทรวงยุธร ศาลหลวงได้พิจารณา ถ้า หัวเมืองเ็ช้ เปน กระทรวง ศาลนำโรง ผู้ รักษาเมืองได้พิจารณา ฯะ

๒ ◉ อนึ่งมีพระธรรมนูญไว้ว่า ถ้าหา อาญาแก่กันว่า มิได้มีโฉนด บาด หมาย ทำข่มเหงเกาะกุมย้อขักผลัณไส เอาท่านมามัดผูก ทุบถองตีดำ ศาจองโขร่ตรวนขี่ยคา ใส่คลังเว้โดยพะละการเอง ลงเอาทซัพยสิ่งลีน สิ่งหนึ่งสิ่งใด มาไว้เปนอาเนาประโยชน แป อาตมา แลทำข่มขืน ลูก ความ ก็ดี ทั้งนี้เปนกระทรวงอาญา ศาล ราชได้พียารณา ถ้าหัวเมือง ขุน ปลัดรอา ปลัศไศ้พิจารณา ถ้า จำเลยมิ๘๘้เปน ตระลาการ เปน อาญานอก

Figure 6.2: Results of Recognition without application of Error Correction

พระธรรมนูญ ๔๑

ถ้าแลราษฎรจะร้อง ฟ้องไซ้ก็ให้มูล นาย อาณาพยาบาลนำเอามาว่า ตาม กระทรวงรูปความซึ่งบันดามี่โรง ศาล จะได้พิจารณาเนื้อความทั้งปวงนั้น ให้พิจารณาทม กระทรวง ฯะ

๑ ◉ มี พระธรรมนูญไว้ว่าทวย ราษฎร จะร้อง ฟ้องหาอุธรแซ่ ตระลาการ แลโรงศาลกามใด ๆว่า มิเต็มใจไห้พิจารณาก็ดี ว่าพิภาคษาความผิด สำนวนก็ดี แล หาตระลาการ ผู้ถามความ แลผู้ถือสำนวนว่า เปนใบ ด้วยลูกความ แลถามะวามให้ผิดด้วยสำนวนมิได้เปนธรรมก็ดี มิ รับ เขียนเอาว่ารับก็ดี ให้การ ต้องข้อ มิเขียน เอาก็ดี แลเขียนแล้วลบ กลับสำนวนเสียก็ติ แลดัดแปลงใบสัจก็ดี แลอั้น สำนวนเสีย มิชิ ไห้สิ้น สำนวนคดี มิไก้เผชิญทาวพระยานตามสำนวน แล อั้น คำค้าน เสียมิไก้ถามพยาน แลอั้นคำท้วงติงทุเลาเกย แลอั้นโลทยาวเลยเลีย มิไก้เอา มาว่า แลแนะนำเสี้ยมสอนลูก ความ ทำชู้ด้วยลูกความ แล เรียก คำฤชาให้ล้ำเหลือ แล อ๊ายวนเนื้อความไว้ให้ช้าพ้นพระราชกำหนด แลผิด กระทรวงล่วง กรม แล ร้องฟ้องคำกฎมิได้ใส่ด้วยพระธรรมนูญก็ดี อนึ่งกล่าวหา พยานว่าเข้าด้วยข้างหนึ่งมิได้ว่าตาม รู้ตามเหน แล รู้เหน แล้วอ้างมิรับ แลซึ่ง มิได้รู้เหน เข้า รับเอาสมอ้าง เปน พยาน อาษา ติดไจมิฟังก็ดี ดังพี้เปนกระทรวงยุธร ศาลหลวงได้พิจารณา ถ้า หัวเมืองใช้ เปน กระทรวง ศาลนำโรง ผู้ รักษาเมืองได้พิจารณา ฯะ

๒ ◉ อนึ่งมีพระธรรมนูญไว้ว่า ถ้าหา อาญาแก่กันว่า มิได้มีโฉนด บาด หมาย ทำข่มเหงเกาะกุมยื้อชักผลับไส เอาท่านมามัดผูก ทุบถองตีดำ ต๋าจองโซร่ตรวนขื่อคา ใส่คลังไว้โดยพะละการเอง ลงเอาทชีพอสิ่งสิน สิ่งหนึ่งสิ่งใด มาไว้เปนอาเนาประโยชน แอ อาตมา แลทำข่มขืน ลูก ความ ก็ดี ทั้งนี้เปนกระทรวงอาญา ศาล ราชได้พิจารณา ถ้าหัวเมือง ขุน ปลัดรอม ปลัดไต้พิจารณา ถ้า จำเลยมิ□ดโปน ตระลาการ เปน อาญานอก

Figure 6.3: Results after Recognition with application of Error Correction

Page	Total Characters	Errors (without correction)	Error Rate	Errors (with correction)	Error Rate
1	1,497	96	6.41%	44	2.94%
2	1,429	104	7.28%	62	4.34%
3	1,536	115	7.49%	53	3.45%
4	1,521	128	8.42%	67	4.40%
5	1,477	89	6.03%	57	3.86%
6	1,492	98	6.57%	48	3.22%
7	1,467	102	6.95%	64	4.36%
8	1,491	91	6.10%	68	4.56%
9	1,420	105	7.39%	62	4.37%
10	1,452	97	6.68%	69	4.75%
Aggregate	14,782	1,025	**6.93%**	594	**4.02%**

Table 6.1: Results - Testing our Methodology on Thai Historical Document

6.5, the recognized results without correction is shown in Figure 6.6 and the corrected image is shown in Figure 6.7 where the errors are highlighted in boxes.

6.1.3 Testing our Methodology on English-based Historical Document

To test that our methodology is applicable to other languages, we chose to experiment on an English document. The historical document we chose is the printed version of the well-known American Declaration of Independence. The document image has been downloaded from the Library of Congress web [110] and represents the image of originally printed document dated back to July 8^{th}, 1776 as seen in Figure 6.1.3. As for slight skew (approximately 1.5 degrees) in the image, it was corrected automatically during the pre-processing of the image. The brownish nature of the original image was easily removed using color histograms analysis. To perform our experiment, we quickly trained a classifier (ANN) to recognize the 52 character patterns of the English alphabet based on the font used in this document. To demonstrate the level of brokenness in the characters of this image, the zoomed segment of the document is shown in Figure 6.9.

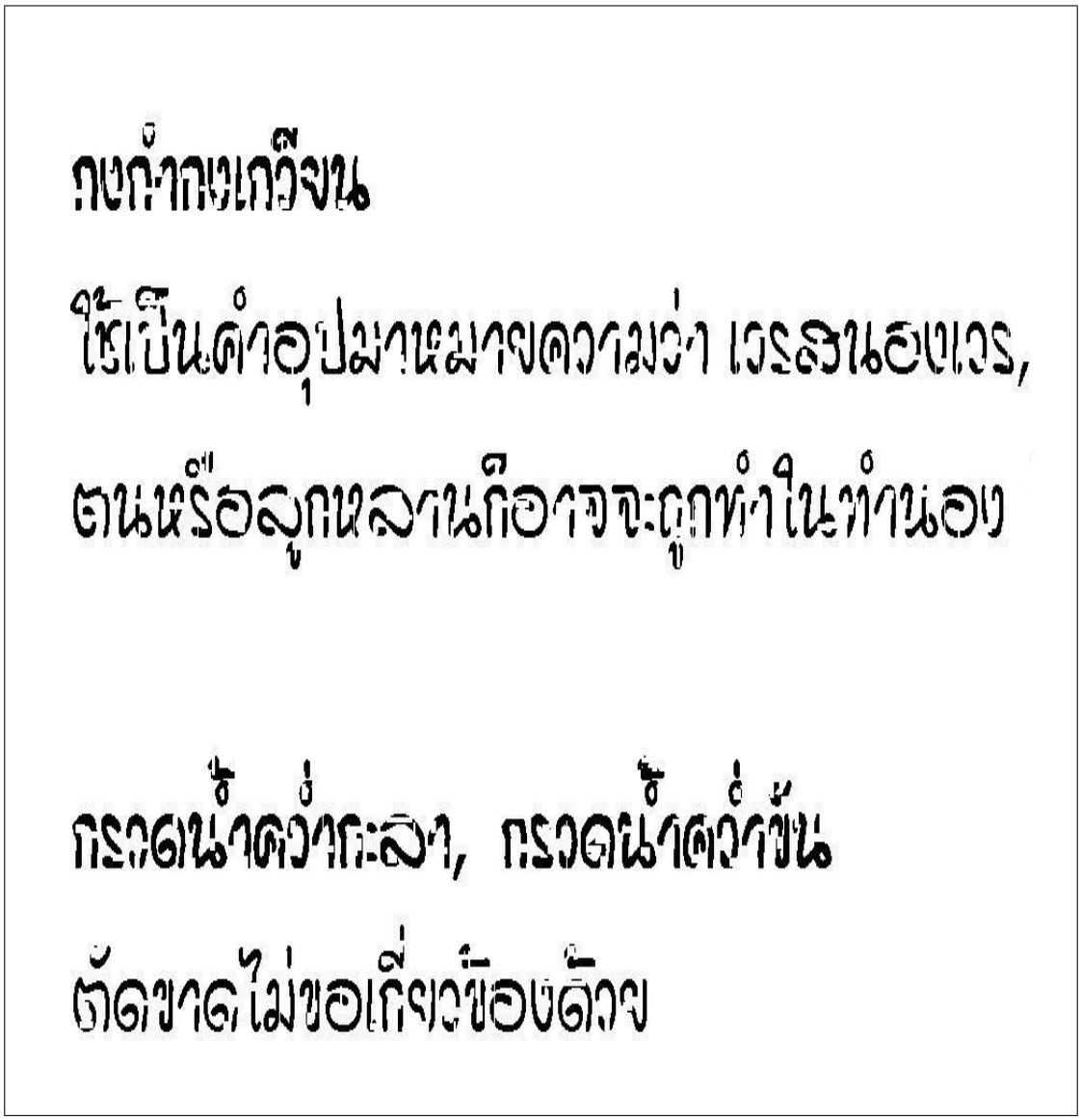


Figure 6.4: Zoomed image shown broken characters in the Faxed document

We show the results of some segments as shown in Figure 6.10 and Figure 6.12. These images show the binarized images of certain sections of the original image. The resulting text that was generated after recognition and dictionary-based correction is shown in Figure 6.11 and Figure 6.13 respectively. The boxed words denote a mis-recognized word. As can be seen, the number is acceptably low. There was a special case that needed handling such as ‘f’ in the historical document might denote a ‘f’ or a ’s’. Hence, the error correction logic was slightly adapted to allow for this confusion. For reference purposes, we show the complete image of the American Declaration of Independence used to perform this experiment in Figre 6.1.3. For our processing, we

กงกำกงเกวียน

ใช้เป็นคำอุปมาหมายความว่า เวรสนองเวร, กรรมสนองกรรม, เช่น ทำแกเขาอย่างไร ตนหรือลูกหลานก็อาจจะถูกทำในทำนองเดียวกันอย่างนั้นบ้าง เป็นกลกำกงเกวียน

กรวดน้ำคว่ำกะลา, กรวดน้ำคว่ำขัน

ตัดขาดไม่ขอเกี่ยวข้องด้วย

กระดังงาลนไฟ

หญิงที่เคยแต่งงานหรือผ่านผู้ชายมาแล้ว ย่อมรู้จักชั้นเชิงทางปรนนิบัติและเอาอกเอาใจผู้ชายได้ดีกว่าหญิงที่ยังไม่ได้แต่งงาน

กระดี่ได้น้ำ

ใช้เปรียบเทียบคนที่แสดงอาการดีอกดีใจจนตัวสั่น เช่น เขาดีใจเหมือนกระดี่ได้น้ำ

กระต่ายขาเดียว, กระต่ายสามขา

ยืนกรานไม่ยอมรับ

กระต่ายตื่นตูม

ใช้เปรียบเทียบคนที่แสดงอาการตื่นตกใจง่ายโดยไม่ทันสำรวจให้ถ่องแท้ก่อน

กระต่ายหมายจันทร์

ผู้ชายหมายปองผู้หญิงที่มีฐานะดีกว่า

กำขี้ดีกว่ากำตด

ได้บ้างดีกว่าไม่ได้อะไรเลย

Figure 6.5: Sample of faxed document image

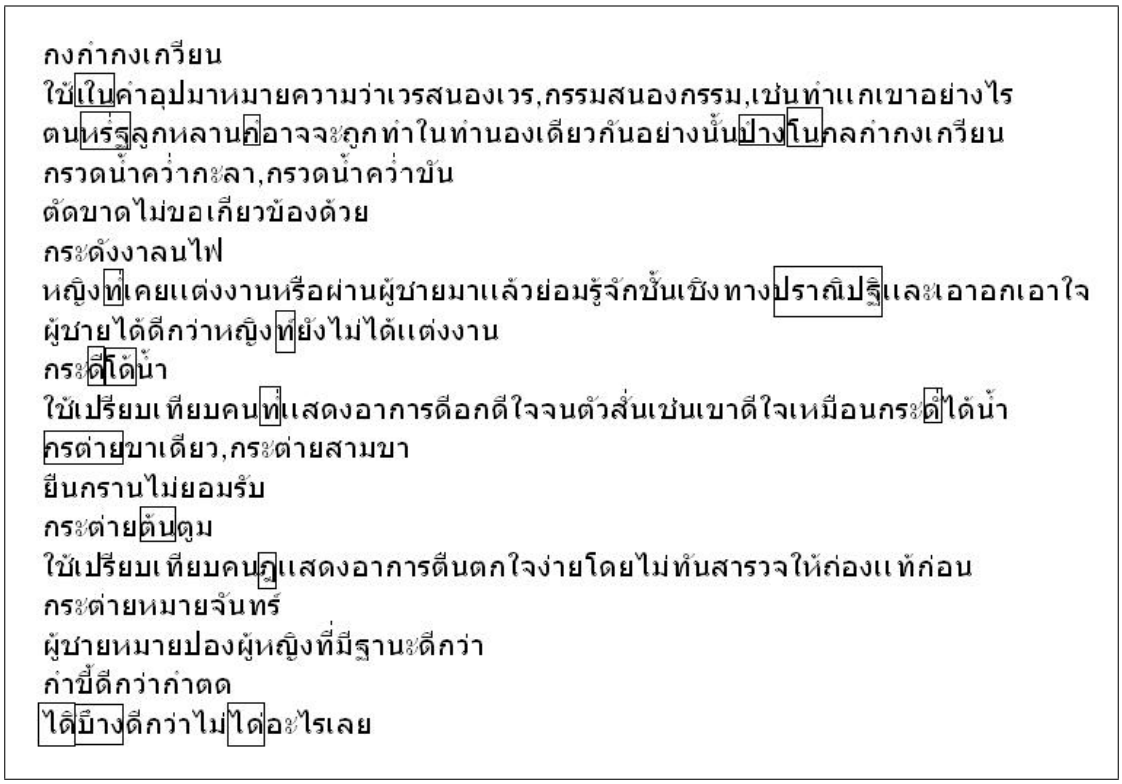

กงกำกงเกวียน
ใช้เใบคำอุปมาหมายความว่าเวรสนองเวร,กรรมสนองกรรม,เช่นทำแกเขาอย่างไร
ตนหรือลูกหลานก็อาจจะถูกทำในทำนองเดียวกันอย่างนั้นบ้างโบกลกำกงเกวียน
กรวดน้ำคว่ำกะลา,กรวดน้ำคว่ำขัน
ตัดขาดไม่ขอเกี่ยวข้องด้วย
กระดังงาลนไฟ
หญิงที่เคยแต่งงานหรือผ่านผู้ชายมาแล้วย่อมรู้จักชั้นเชิงทางปราณีปฏิเลและเอาอกเอาใจ
ผู้ชายได้ดีกว่าหญิงที่ยังไม่ได้แต่งงาน
กระดี่ได้น้ำ
ใช้เปรียบเทียบคนที่แสดงอาการดีอกดีใจจนตัวสั่นเช่นเขาดีใจเหมือนกระดี่ได้น้ำ
กรต่ายขาเดียว,กระต่ายสามขา
ยืนกรานไม่ยอมรับ
กระต่ายต้นตูม
ใช้เปรียบเทียบคนที่แสดงอาการตื่นตกใจง่ายโดยไม่ทันสารวจให้ถ่องแท้ก่อน
กระต่ายหมายจันทร์
ผู้ชายหมายปองผู้หญิงที่มีฐานะดีกว่า
ก่าขี้ดีกว่าก่าตด
ได้บ้างดีกว่าไม่ได้อะไรเลย

Figure 6.6: Recognized faxed document without applying Error Correction

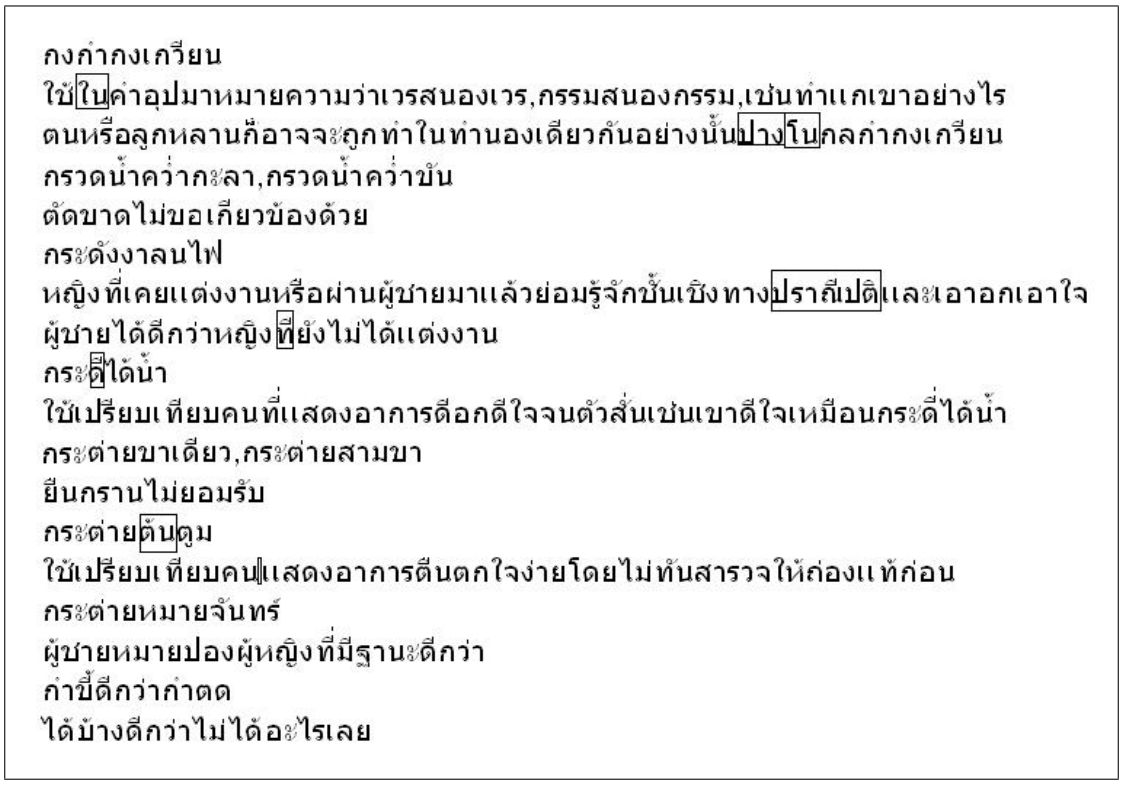

กงกำกงเกวียน
ใช้ใบคำอุปมาหมายความว่าเวรสนองเวร,กรรมสนองกรรม,เช่นทำแกเขาอย่างไร
ตนหรือลูกหลานก็อาจจะถูกทำในทำนองเดียวกันอย่างนั้นบ้างโบกลกำกงเกวียน
กรวดน้ำคว่ำกะลา,กรวดน้ำคว่ำขัน
ตัดขาดไม่ขอเกี่ยวข้องด้วย
กระดังงาลนไฟ
หญิงที่เคยแต่งงานหรือผ่านผู้ชายมาแล้วย่อมรู้จักชั้นเชิงทางปราณีปฏิเลและเอาอกเอาใจ
ผู้ชายได้ดีกว่าหญิงที่ยังไม่ได้แต่งงาน
กระดี่ได้น้ำ
ใช้เปรียบเทียบคนที่แสดงอาการดีอกดีใจจนตัวสั่นเช่นเขาดีใจเหมือนกระดี่ได้น้ำ
กระต่ายขาเดียว,กระต่ายสามขา
ยืนกรานไม่ยอมรับ
กระต่ายต้นตูม
ใช้เปรียบเทียบคนที่แสดงอาการตื่นตกใจง่ายโดยไม่ทันสารวจให้ถ่องแท้ก่อน
กระต่ายหมายจันทร์
ผู้ชายหมายปองผู้หญิงที่มีฐานะดีกว่า
ก่าขี้ดีกว่าก่าตด
ได้บ้างดีกว่าไม่ได้อะไรเลย

Figure 6.7: Recognized faxed document after applying Error Correction

manually remove all the header and footer lines where the font size is larger than the majority of the document.

6.2 Other Applications of Partition Growing Algorithm

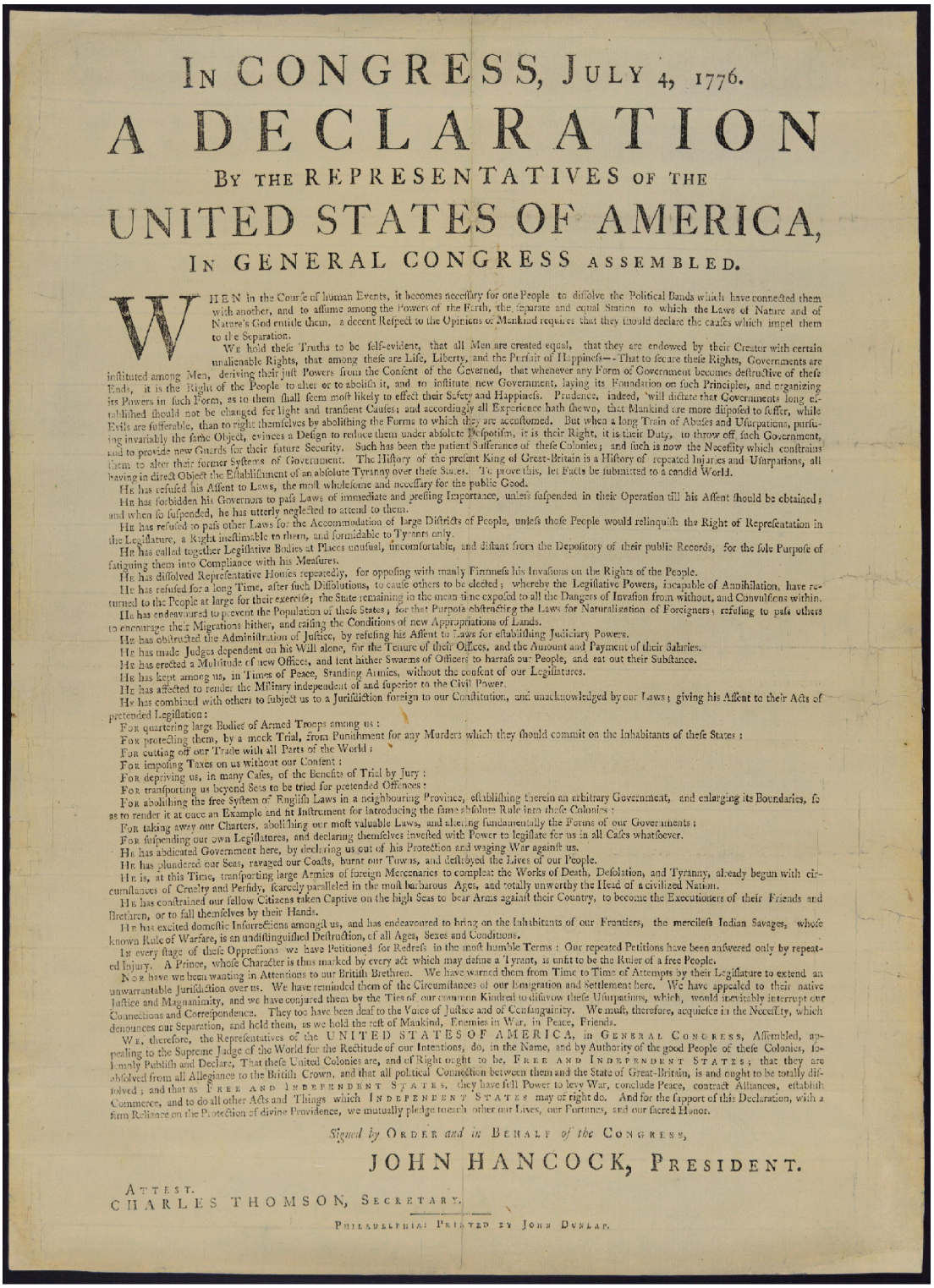
In section 5.2.1, we experiment on an old well-researched problem: Steel Mill Slab design where the objective function is a simple linear function - utilization. Then, we apply our algorithm to solve the a more complex problem in section 5.2.2 : Wedding Planner where the objective function - utilization is linear and happiness is non-linear. Finally, in section 5.2.3, we demonstrate how a complex variant 3 problem : Task Allocation can be modeled with some non-linear probability functions to generate solutions where the primary concerns are skill matching, fairness and short time-frame.

6.2.1 Experiment on Steel Mill Slab Design Problem

We conducted some experiments on the Steel Mill Slab problem discussed in Section 5.2.1 using the medium-sized samples provided in CSP Lib 38 and we present some of the results in Table 6.2.1. As can be seen, the wastages is 0 in all 3 cases. The 20 slab capacities for 3 cases is C ={12 14 17 18 19 20 23 24 25 26 27 28 29 30 32 35 39 42 43 44}. We first conduct the standard experiment without attempting to find the solution using bigger slabs. For cases 1 and 2, on applying $P_\gamma(c)$ to the model, we are able to find optimal solutions that use bigger slabs. In Figure 6.14, we also show the behaviour of function $P(S)$ with respect to the capacity c^*.

6.2.2 Experiment on Wedding Planner Problem

We conducted some experiments on the Wedding Planner Problem discussed in Section 5.2.2 and results are presented in Table 6.3. We set up the following 16 guests with following ages {22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52} and genders {1 1 1 1 -1 -1 -1 -1 1 -1 -1 1 -1 -1 -1 1} respectively. We assume there are 6 tables in the wedding hall with sizes 4 5 5 6 6 7 from which only some tables will be used. We first perform the standard experiment without considering gender of the guests. Then, we re-conduct it with gender consideration and analyze its effects. For our experiments, we set the max number of empty seats acceptable $e = 2$ and the age gap threshold $a = 6$.



IN CONGRESS, JULY 4, 1776.

A DECLARATION

BY THE REPRESENTATIVES OF THE

UNITED STATES OF AMERICA,

IN GENERAL CONGRESS ASSEMBLED.

WHEN in the Course of human Events, it becomes neceſſary for one People to diſſolve the Political Bands which have connected them with another, and to aſſume among the Powers of the Earth, the ſeparate and equal Station to which the Laws of Nature and of Nature's God entitle them, a decent Reſpect to the Opinions of Mankind requires that they ſhould declare the cauſes which impel them to the Separation.

We hold theſe Truths to be ſelf-evident, that all Men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among theſe are Life, Liberty, and the Purſuit of Happineſs--That to ſecure theſe Rights, Governments are inſtituted among Men, deriving their juſt Powers from the Conſent of the Governed, that whenever any Form of Government becomes deſtructive of theſe Ends, it is the Right of the People to alter or to aboliſh it, and to inſtitute new Government, laying its Foundation on ſuch Principles, and organizing its Powers in ſuch Form, as to them ſhall ſeem moſt likely to effect their Safety and Happineſs. Prudence, indeed, will dictate that Governments long eſtabliſhed ſhould not be changed for light and tranſient Cauſes; and accordingly all Experience hath ſhewn, that Mankind are more diſpoſed to ſuffer, while Evils are ſufferable, than to right themſelves by aboliſhing the Forms to which they are accuſtomed. But when a long Train of Abuſes and Uſurpations, purſuing invariably the ſame Object, evinces a Deſign to reduce them under abſolute Deſpotiſm, it is their Right, it is their Duty, to throw off ſuch Government, and to provide new Guards for their future Security. Such has been the patient Sufferance of theſe Colonies; and ſuch is now the Neceſſity which conſtrains them to alter their former Syſtems of Government. The Hiſtory of the preſent King of Great-Britain is a Hiſtory of repeated Injuries and Uſurpations, all having in direct Object the Eſtabliſhment of an abſolute Tyranny over theſe States. To prove this, let Facts be ſubmitted to a candid World.

He has refuſed his Aſſent to Laws, the moſt wholeſome and neceſſary for the public Good.

He has forbidden his Governors to paſs Laws of immediate and preſſing Importance, unleſs ſuſpended in their Operation till his Aſſent ſhould be obtained; and when ſo ſuſpended, he has utterly neglected to attend to them.

He has refuſed to paſs other Laws for the Accommodation of large Diſtricts of People, unleſs thoſe People would relinquiſh the Right of Repreſentation in the Legiſlature, a Right ineſtimable to them, and formidable to Tyrants only.

He has called together Legiſlative Bodies at Places unuſual, uncomfortable, and diſtant from the Depoſitory of their public Records, for the ſole Purpoſe of fatiguing them into Compliance with his Meaſures.

He has diſſolved Repreſentative Houſes repeatedly, for oppoſing with manly Firmneſs his Invaſions on the Rights of the People.

He has refuſed for a long Time, after ſuch Diſſolutions, to cauſe others to be elected; whereby the Legiſlative Powers, incapable of Annihilation, have returned to the People at large for their exerciſe; the State remaining in the mean time expoſed to all the Dangers of Invaſion from without, and Convulſions within.

He has endeavoured to prevent the Population of theſe States; for that Purpoſe obſtructing the Laws for Naturalization of Foreigners; refuſing to paſs others to encourage their Migrations hither, and raiſing the Conditions of new Appropriations of Lands.

He has obſtructed the Adminiſtration of Juſtice, by refuſing his Aſſent to Laws for eſtabliſhing Judiciary Powers.

He has made Judges dependent on his Will alone, for the Tenure of their Offices, and the Amount and Payment of their Salaries.

He has erected a Multitude of new Offices, and ſent hither Swarms of Officers to harraſs our People, and eat out their Subſtance.

He has kept among us, in Times of Peace, Standing Armies, without the conſent of our Legiſlatures.

He has affected to render the Military independent of and ſuperior to the Civil Power.

He has combined with others to ſubject us to a Juriſdiction foreign to our Conſtitution, and unacknowledged by our Laws; giving his Aſſent to their Acts of pretended Legiſlation:

For quartering large Bodies of Armed Troops among us:

For protecting them, by a mock Trial, from Puniſhment for any Murders which they ſhould commit on the Inhabitants of theſe States:

For cutting off our Trade with all Parts of the World:

For impoſing Taxes on us without our Conſent:

For depriving us, in many Caſes, of the Benefits of Trial by Jury:

For tranſporting us beyond Seas to be tried for pretended Offences:

For aboliſhing the free Syſtem of Engliſh Laws in a neighbouring Province, eſtabliſhing therein an arbitrary Government, and enlarging its Boundaries, ſo as to render it at once an Example and fit Inſtrument for introducing the ſame abſolute Rule into theſe Colonies:

For taking away our Charters, aboliſhing our moſt valuable Laws, and altering fundamentally the Forms of our Governments:

For ſuſpending our own Legiſlatures, and declaring themſelves inveſted with Power to legiſlate for us in all Caſes whatſoever.

He has abdicated Government here, by declaring us out of his Protection and waging War againſt us.

He has plundered our Seas, ravaged our Coaſts, burnt our Towns, and deſtroyed the Lives of our People.

He is, at this Time, tranſporting large Armies of foreign Mercenaries to compleat the Works of Death, Deſolation, and Tyranny, already begun with circumſtances of Cruelty and Perfidy, ſcarcely paralleled in the moſt barbarous Ages, and totally unworthy the Head of a civilized Nation.

He has conſtrained our fellow Citizens taken Captive on the high Seas to bear Arms againſt their Country, to become the Executioners of their Friends and Brethren, or to fall themſelves by their Hands.

He has excited domeſtic Inſurrections amongſt us, and has endeavoured to bring on the Inhabitants of our Frontiers, the mercileſs Indian Savages, whoſe known Rule of Warfare, is an undiſtinguiſhed Deſtruction, of all Ages, Sexes and Conditions.

In every ſtage of theſe Oppreſſions we have Petitioned for Redreſs in the moſt humble Terms: Our repeated Petitions have been anſwered only by repeated Injury. A Prince, whoſe Character is thus marked by every act which may define a Tyrant, is unfit to be the Ruler of a free People.

Nor have we been wanting in Attentions to our Britiſh Brethren. We have warned them from Time to Time of Attempts by their Legiſlature to extend an unwarrantable Juriſdiction over us. We have reminded them of the Circumſtances of our Emigration and Settlement here. We have appealed to their native Juſtice and Magnanimity, and we have conjured them by the Ties of our common Kindred to diſavow theſe Uſurpations, which, would inevitably interrupt our Connections and Correſpondence. They too have been deaf to the Voice of Juſtice and of Conſanguinity. We muſt, therefore, acquieſce in the Neceſſity, which denounces our Separation, and hold them, as we hold the reſt of Mankind, Enemies in War, in Peace, Friends.

We, therefore, the Repreſentatives of the UNITED STATES OF AMERICA, in GENERAL CONGRESS, Aſſembled, appealing to the Supreme Judge of the World for the Rectitude of our Intentions, do, in the Name, and by Authority of the good People of theſe Colonies, ſolemnly Publiſh and Declare, That theſe United Colonies are, and of Right ought to be, FREE AND INDEPENDENT STATES; that they are abſolved from all Allegiance to the Britiſh Crown, and that all political Connection between them and the State of Great-Britain, is and ought to be totally diſſolved; and that as FREE AND INDEPENDENT STATES, they have full Power to levy War, conclude Peace, contract Alliances, eſtabliſh Commerce, and to do all other Acts and Things which INDEPENDENT STATES may of right do. And for the ſupport of this Declaration, with a firm Reliance on the Protection of divine Providence, we mutually pledge to each other our Lives, our Fortunes, and our ſacred Honor.

Signed by ORDER *and in* BEHALF *of the* CONGRESS,

JOHN HANCOCK, PRESIDENT.

ATTEST.
CHARLES THOMSON, SECRETARY.

PHILADELPHIA: PRINTED BY JOHN DUNLAP.

Figure 6.8: American Declaration of Independence - Originally Printed Version

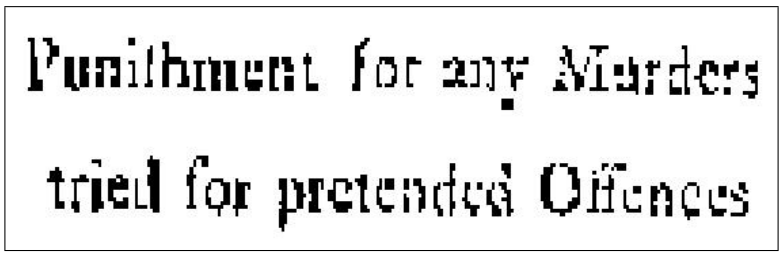


Figure 6.9: Some Broken Characters in American Declaration of Independence

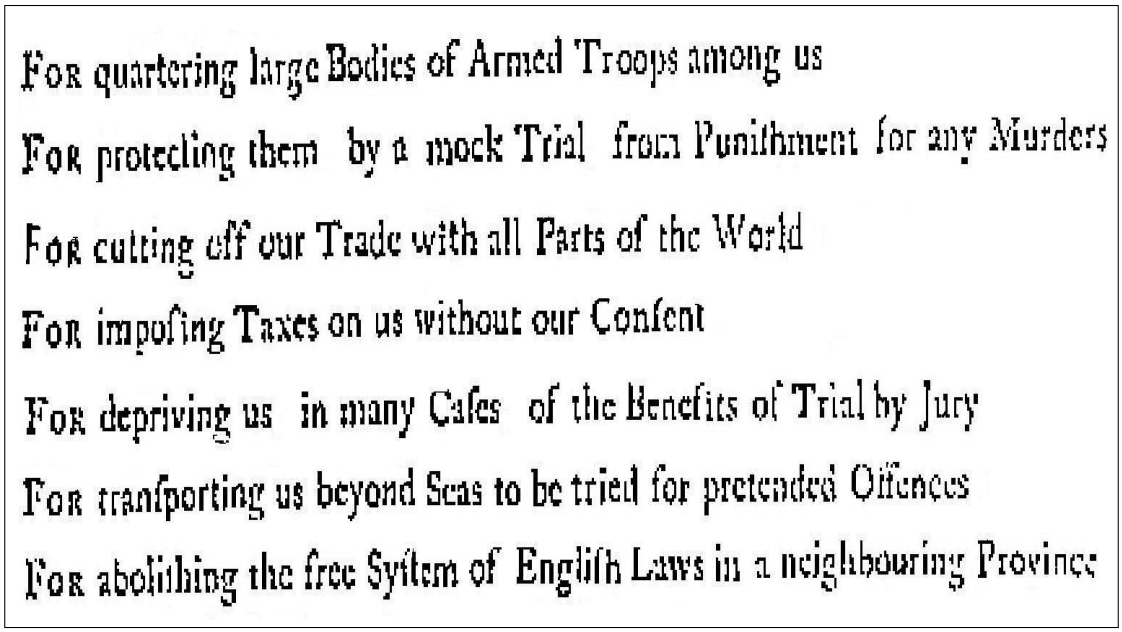


Figure 6.10: Binarized Segment I of American DOI

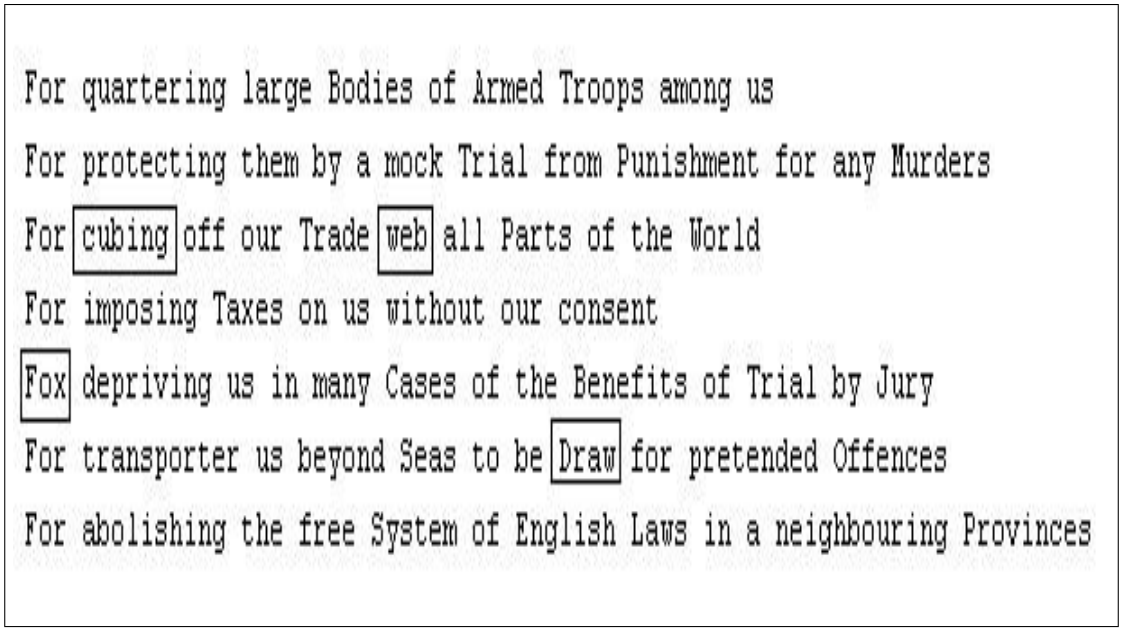


Figure 6.11: Results from OCR on Segment I of American DOI

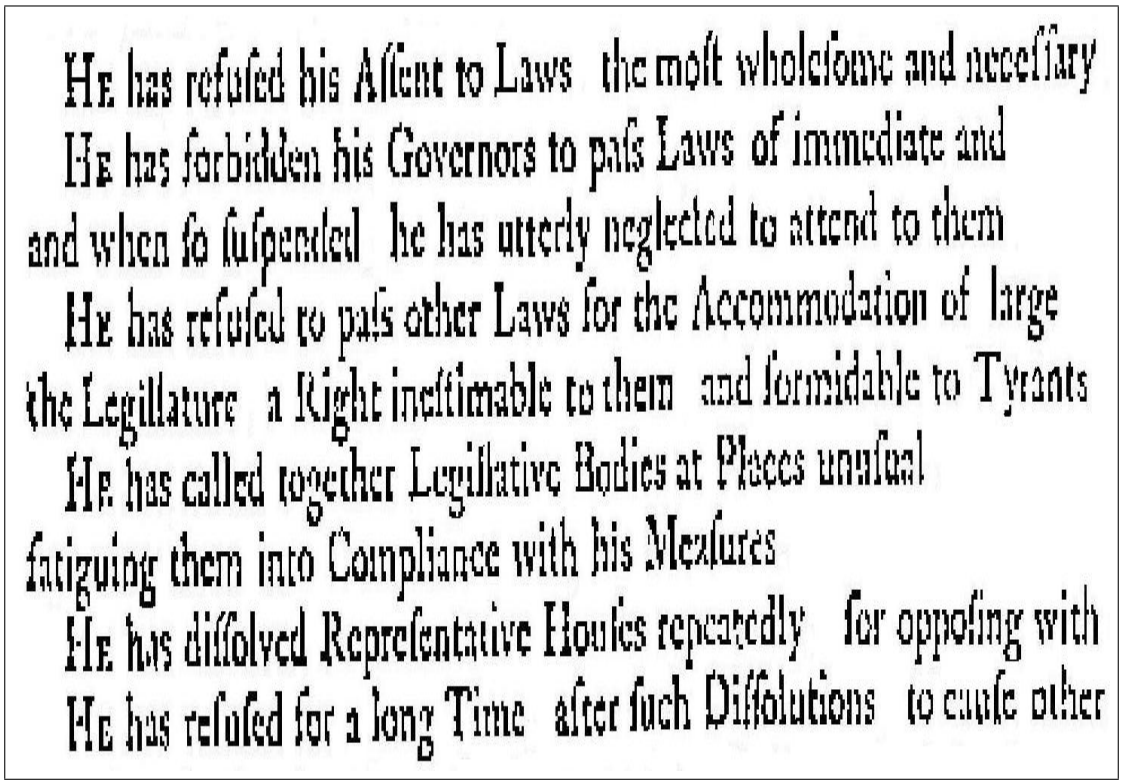


Figure 6.12: Binarized Segment II of American DOI

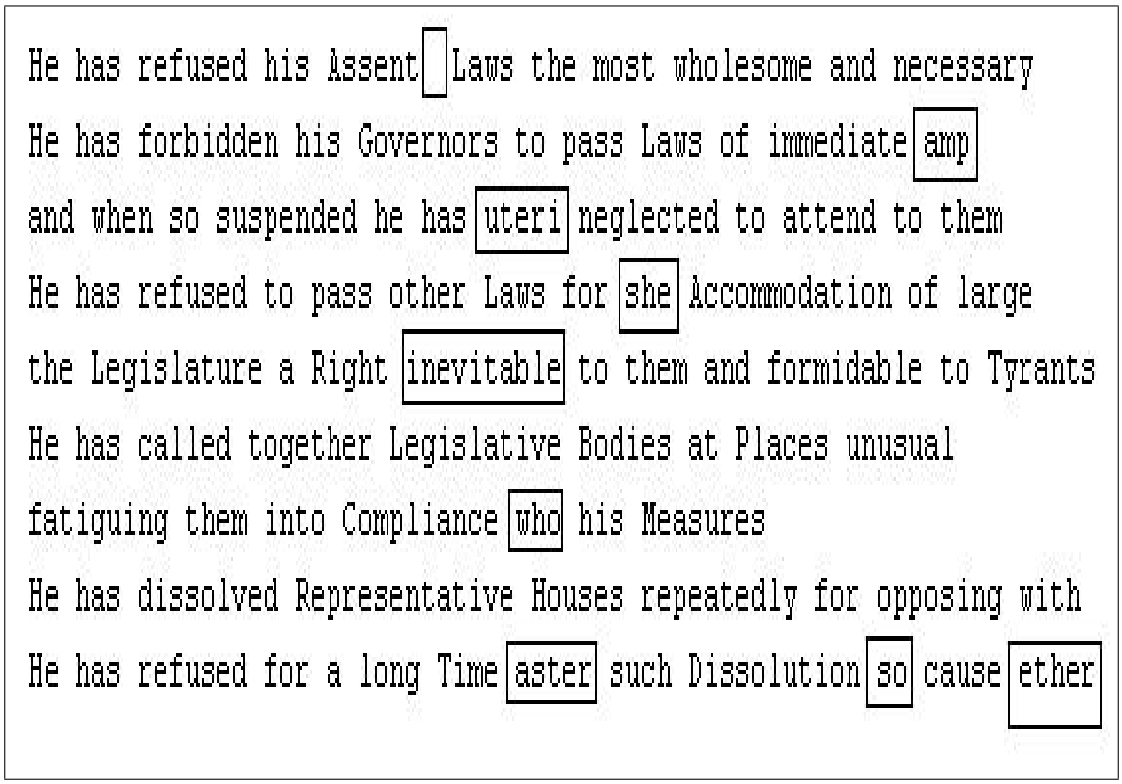


Figure 6.13: Results from OCR on Segment II of American DOI

	Problem Description	Time (Sec)	Solution without $P_\gamma(c)$	Time (Sec)	Solution with $P_\gamma(c)$
1.	X = {22 9 8 7 7 5 4 4 3 3 3 2} L = { 2 3 5 4 8 4 1 7 6 6 4 6} Number of orders = 12 Number of colors = 8	0.434	{22 7} {9 3} {7 5} {4 3 3 2} {8 4}	0.617	{22 7} {9 8} {7 5 4 3} {4 3 3 2}
2.	X = {22 9 8 7 7 5 4 4 3 3 3 2 **2**} L = { 2 3 5 4 8 4 1 7 6 6 4 6 **6**} Number of orders = 13 Number of colors = 8	1.712	{22 4} {9 8} {7 5} {4 3 3 2} {7 3 2}	2.115	{22 4} {9 8} {7 7 5 3 2} {4 3 3 2}
3.	X = {22 9 8 **8** 7 7 5 4 4 3 3 3 2 2} L = { 2 3 5 **9** 4 8 4 1 7 6 6 4 6} Number of orders = 14 Number of colors = 9	1.981	{22 7} {9 8} {7 5 3 2} {4 3 3 2} {8 4}	6.738	{22 7} {9 8} {7 5 3 2} {4 3 3 2} {8 4}

Table 6.2: Results from Experiments on Steel Mill SLab Design Problem

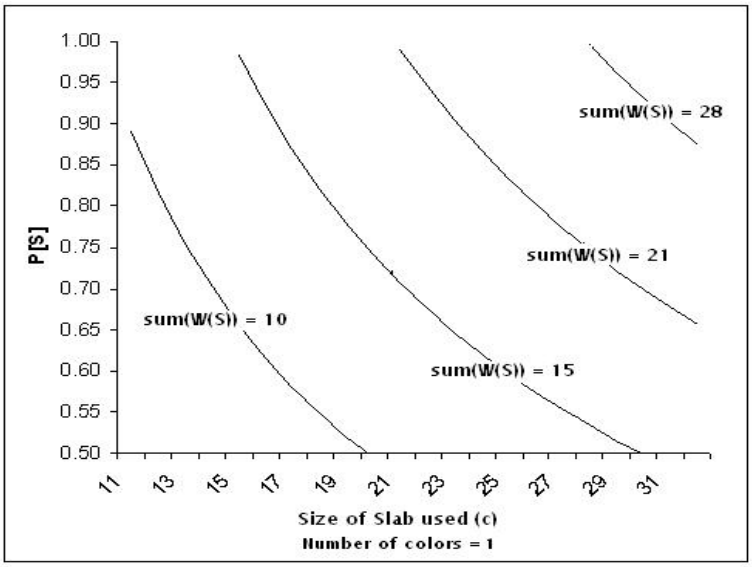


Figure 6.14: Analysis of the Steel Mill Slab Design Model

We also show the behavior of the $P(S)$ with respect to accumulative excess age $d(S)$ for different mixtures of males (M) and females (F) on table of size of 5 in Figure 6.16.

We executed an exhaustive enumeration method to determine the list of optimal partitions and used it to confirm that an optimal result is achieved for each run of the Partition Growing Method and recorded its runtime for comparative purposes as shown in Figure 6.15. For clarity, we plotted the runtime on a $\log_{10}$ scale. The performance of the Partition Growing algorithm to solve the Wedding Planner problem, depends on the number of guests and probability function employed. By incorporating the guest's Gender into the problem, the convergence of the problem is delayed to a large extent as it introduces a non-linear aspect into the problem. Another factor that negatively affects the performance of this experiment is the presence of tables of equal sizes (2 tables of size 5 and 2 tables of size 6). Given the need to match groups of subset to a particular table, the more tables of equal sizes in the problem domain the more equally optimal solutions can be found and hence the problem of symmetry still persists.

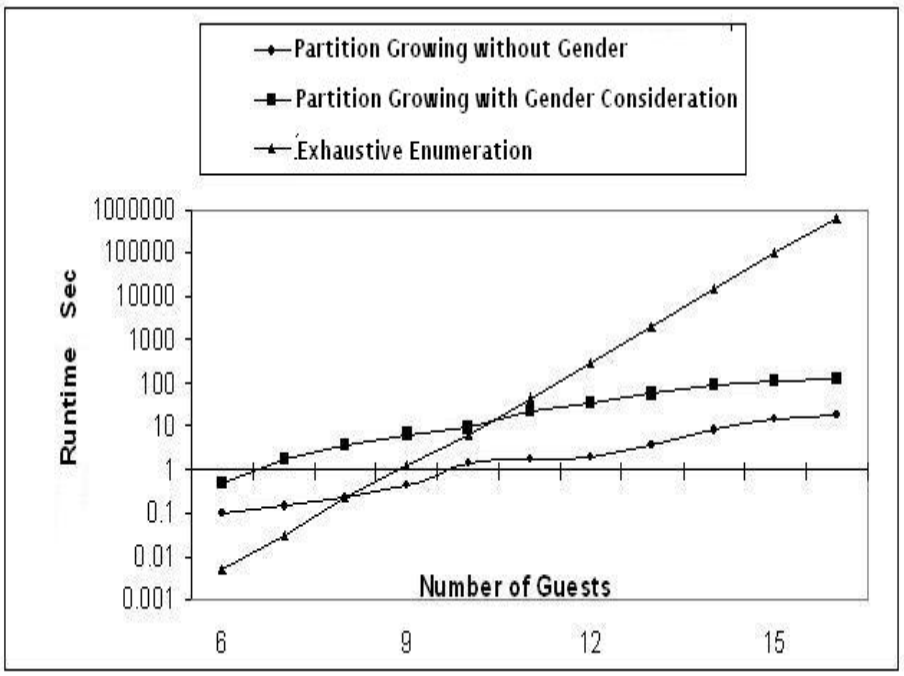


Figure 6.15: Runtime Analysis of Wedding Planner Experiment

X	Time (Sec)	Solution without Gender Mixture	Time (Sec)	Solution with Gender Mixture
1-6	0.104	{1 2 3} {5 6 7} {}{}{}{}	0.481	{3 4} {1 2 5 6} {}{}{}{}
1-7	0.143	{1 2 3} {4 5 6 7} {}{}{}{}	1.781	{3 4 7} {1 2 5 6} {}{}{}{}
1-8	0.234	{1 2 3 4} {5 6 7 8} {}{}{}{}	3.914	{1 2 5 6} {3 4 7 8} {}{}{}{}
1-9	0.422	{6 7 8 9} {1 2 3 4 5} {}{}{}{}	6.812	{8 9} {1 2 5} {3 4 6 7} {}{}{}
1-10	1.375	{1 2 3 4 5} {6 7 8 9 10} {}{}{}{}	8.868	{9 10}{1 2 5 6} {3 4 7 8} {}{}{}
1-11	1.766	{1 2 3} {4 5 6 7} {8 9 10 11}{}{}{}	23.469	{9 10 11} {1 2 5 6} {3 4 7 8}{}{}{}
1-12	1.938	{1 2 3 4} {5 6 7 8} {9 10 11 12}{}{}{}	33.609	{1 2 5 6} {3 4 7 8} {9 10 11 12}{}{}{}
1-13	3.844	{1 2 3 4} {5 6 7 8} {9 10 11 12 13}{}{}{}	58.671	{1 2 5 6} {3 4 7 8} {9 10 11 12 13}{}{}{}
1-14	8.609	{1 2 3 4} {5 6 7 8 9} {10 11 12 13 14}{}{}{}	84.547	{4 7 8 9} {1 2 3 5 6} {10 11 12 13 14} {}{}{}
1-15	14.203	{}{1 2 3 4 5} {6 7 8 9 10} {11 12 13 14 15}{}{}	108.794	{9 10 11} {1 2 5 6} {3 4 7 8} {12 13 14 15}{}{}
1-16	18.689	{1 2 3 4} {5 6 7 8} {9 10 11 12} {13 14 15 16}{}{}	124.672	{1 2 5 6} {3 4 7 8} {9 10 11 12} {13 14 15 16}{}{}

Table 6.3: Experiments on Wedding Planner Problem

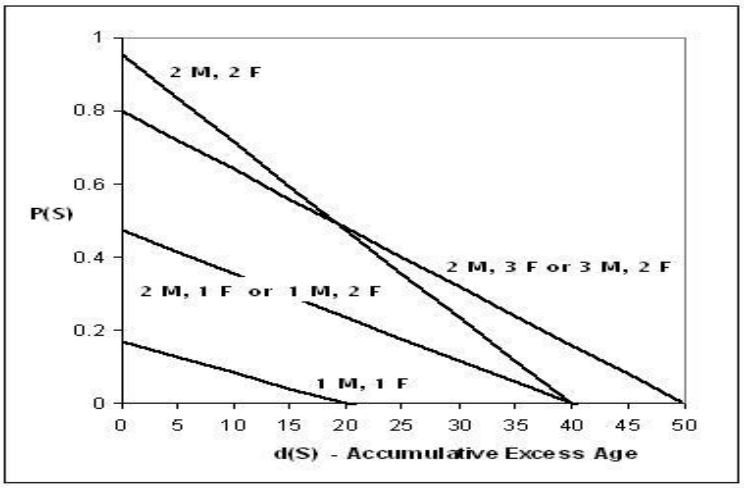


Figure 6.16: Analysis of the Wedding Planner Model

6.2.3 Experiment on Task Allocation Problem

We conducted some experiments on the Task Allocation Problem discussed in Section 5.2.3 and the results are presented in Table 6.4. For sake of simplicity and without losing any generality, we assume that there are only 2 skills required in the workshop. We set up 5 workers C with following skills Y = {(5,5) (1,5) (5,1) (2,3) (3,2)} and 15 tasks X with skills Z = {(4,4) (2,4) (4,2) (2,3) (3,2) (2,1) (0,0) (3,4) (4,3) (1,3) (3,1) (2,2) (1,2) (0,1) (1,0)} and duration D = {5 5 5 4 4 4 3 3 3 2 2 2 1 1 1} respectively. For our experiments, we set the penalty $\nu = 0.01$.

We show the behavior of the $P(S)$ with respect to timeframe of S for the 3 different levels of workers that can be assigned to S in Figure 6.17. We executed an exhaustive enumeration method to determine the list of optimal partitions and used it to confirm that an optimal result is achieved for each run of the Partition Growing Method. The performance of the algorithm depends directly on the number of tasks, the number of workers and probability function. Given that this is a problem of Variant 3, determining the optimal solution is obviously more time-consuming because in addition to simple partitioning problem, the one-to-one correspondence must be achieved too.

6.3 Simulation Experiments

In this section, we focus on analyzing the Partition Growing Algorithm. For these experiments, we only test problem variant 1 - no injective or surjective mapping

X	Time (m:s)	Solution	Qualifications Under / Over	Time Frame
1-4	0:00	No Solution		
1-5	0:01	{1}{2}{3}{4}{5}	2 / 04	5
1-6	0:03	{1}{2}{3}{4 6}{5}	2 / 06	8
1-7	0:02	{1}{2}{3}{4 6}{5 7}	2 / 11	8
1-8	0:03	{1 8}{2}{3}{4 6}{5 7}	2 / 14	8
1-9	1:34	{1 8}{2}{3 9}{4 6}{5 7}	4 / 15	8
1-10	0:04	{1 8}{2 10}{3 9}{4 6}{5 7}	4 / 17	8
1-11	0:21	{1 9}{2 8}{3 7}{4 6}{5 10 11}	4 / 18	8
1-12	1:40	{1 8}{2 6}{3 9}{4 10 12}{5 7 11}	4 / 20	9
1-13	3:12	{1 8}{2 6}{3 9}{4 10 12 13}{5 7 11}	4 / 22	9
1-14	6:55	{1 9}{2 8 13}{3 6}{4 10 12 14}{5 7 11}	4 / 26	9
1-15	17:44	{1 9 15}{2 8 13}{3 6}{4 10 12 14}{5 7 11}	4 / 31	9

Table 6.4: Experiments on Task Allocation

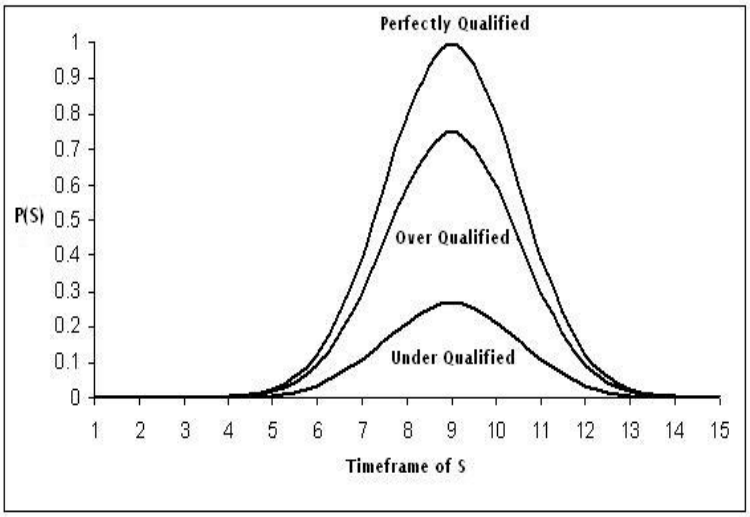


Figure 6.17: Analysis of the Task Allocation Model

between subsets S and the classes C. We used a random number generator to generate the probability values $P(S)$ over 3 different distributions - Uniform, Normal and Poisson. In total 15,000 data sets were generated, 5,000 for each distribution. Of these data sets in each distribution, there are 500 each for sizes of X ranging from 5 to 14. We used this set of data to conduct all the simulation experiments in this section.

To ensure that an optimal solution is obtainable in cases where we do not put any constraint on τ and m. We set τ to a very low value = 0.001 and $m = 2$. We first performed the brute-force search to find all possible optimal solutions. Then, we executed the Partition-Growing algorithm to generate a solution. In all 15,000 instances, the solution that was generated matched one of the optimal solutions generated by the exhaustive search. We plotted runtimes to compare the performance of our algorithm with the brute-force method in Figure. 6.18.

Given that we are looking for optimal partitions of size m, we compared the optimal solution obtained with the similar setup using the brute-force. In each case, we are able to obtain an optimal solution. To complete the analysis of impact of m, we test with $|X| = 14$ and varied m from 1 to 13, again keeping the value of τ at 0.001. The runtime is shown in Figure. 6.19, and it reflects that as m approaches $|X|$, it is faster to find the optimal solution.

Next, we divert our focus to analyzing the effect of τ on the runtime. We fix $m = 2$ for this experiment and vary τ between 0.1 and 1.0. We compute the aggregates for low, medium and high values of τ and plot average runtimes as shown in Figure 6.20. The higher ranges $0.8 \leq \tau < 1.0$ reduces the runtime dramatically, but it must be noted here that in many of the cases, the optimal solution was not obtained. For the middle range and lower ranges, the runtime is not much different, but a few of the middle range cases $0.6 \leq \tau < 0.8$, the optimal solution was not found. For lower ranges, rare cases of non-optimal solutions exist too.

Finally, we compare how the distributions of $P(S)$ effect the runtime of the algorithm. The average runtimes for different distributions is shown in Figure 6.21. For small values of $|X|$, there is no apparent different in the runtimes. As the value of $|X|$ increases, it is noticed that uniform distribution begins to show higher runtimes and the normal distribution performs best.

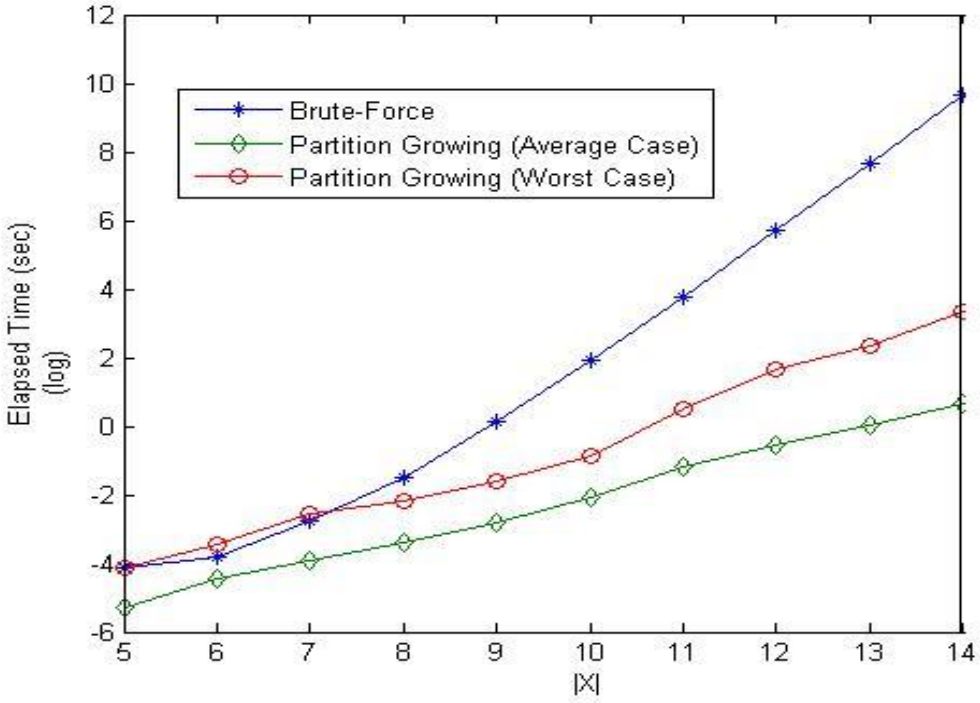


Figure 6.18: Run-time Analysis

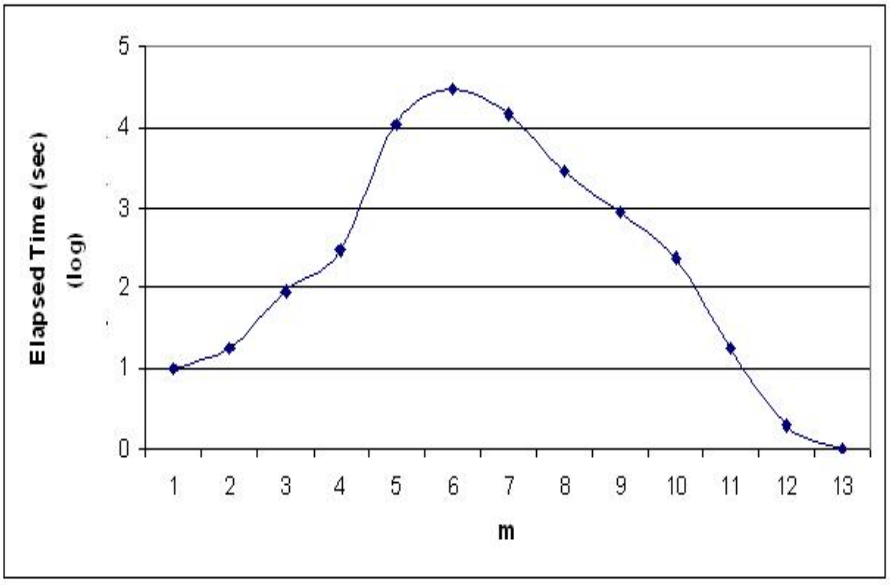


Figure 6.19: Effect of m on Runtime

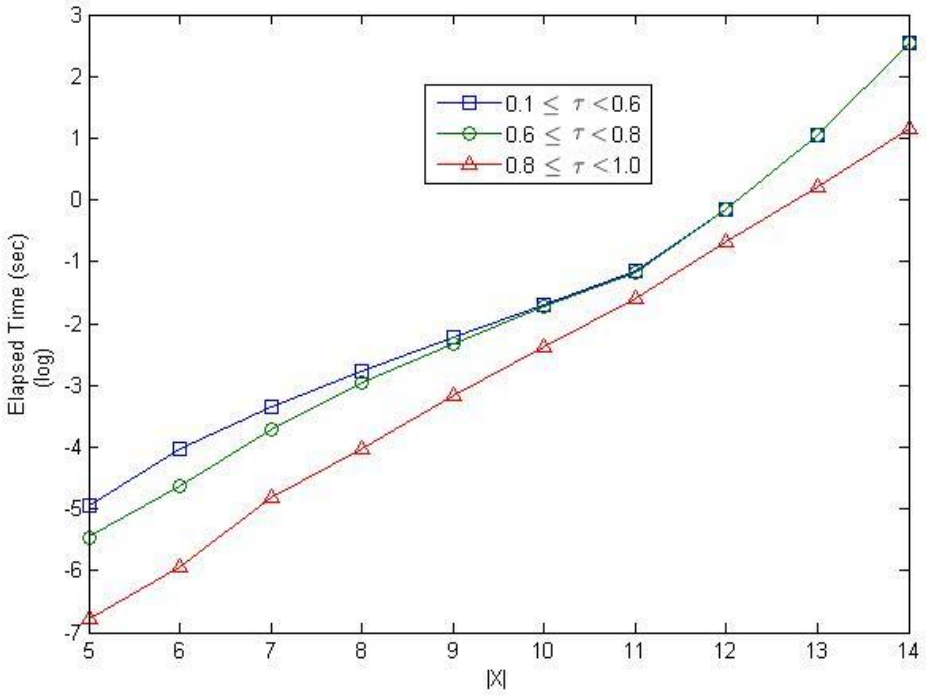


Figure 6.20: Effect of τ on Runtime

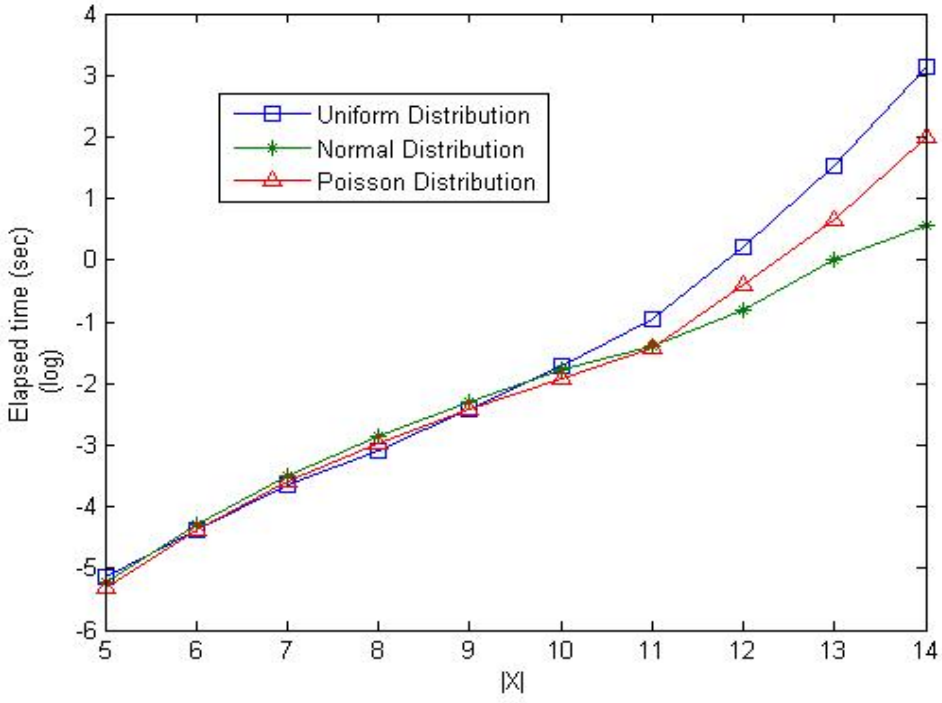


Figure 6.21: Effect of different distributions of $P(S)$

CHAPTER VII
CONCLUSION

This main focus of this research is towards solving the problem of broken characters that prominently seen on all Historical documents. We proposed a novel Partition-Growing algorithm that yields an optimal solution without having to completely enumerate all the partitions. The main idea is based on finding optimal set partitions of broken pieces and it has shown great promise towards achieving high level of accuracy of over 90%. A prototype based on the methodology to solve the broken characters problem with the automated error correction based on n-grams graphs discussed in Chapter 3 have been completed. Experiments have been performed on two different Thai documents with broken pieces and an English document to show the generality of the model. We also conducted some experiments to show that the algorithm can be applied to other problem domains of similar nature where an optimal partition needs to be found. In the next sections, we discuss some of the issues related to the solution and suggest further improvements in future works.

7.1 Unresolved Issues related to Broken Character Recognition

During performing the experiments, we encountered the following issues that have yet not been resolved due their complexity and the limited timeframe.

- Some noise pieces in close vicinity of the character pieces are treated as part of set of broken pieces X and cause a lot of distorted patterns leading to inaccurate recognition.

- Uniformly merged characters that are always coupled together in the original script has been treated as a single pattern class during the recognition process. However, the randomly merged characters due to distortions, artifacts and blurring pose a different problem altogether and have not been handled in our methodology.

These have been either manually de-merged when detected and the ones that remain undetected lead to inaccurate recognition.

- The solution proposed yields an optimal π^* for X of size less than 20 in reasonable amount of time. However, as the size of X grows, the speed of convergence drastically deteriorates as with all methods that seek optimal set partitions based on heuristic enumeration. The enhancement algorithm presented in Section 4.5 offers an improvement to the speed with a slight compromise to the optimality of the solution.

7.2 Suggested Future Work

For future work in this area, we suggest the following practical improvements to the model and other works that add more substance to the robustness and reliability of the solution.

- Enhance the model to incorporate appropriate handling of noise pieces. An unexperimented idea is given in Appendix A.

- A document to be restored is currently assumed to be single font face and single font size. Any future work can be focused towards allowing multiple font size within a single document.

- Currently, the small noise pieces do not have large impacts on the model. However, large artifacts found on historical documents such as finger prints, blobs, and pictures are currently manually removed before processing. Future works can be to incorporate the detection of such elements and either remove them permanently or temporarily during the processing.

- As part of the error correction model, we employed a dictionary to create a n-grams graph that allows us to work on a probabilistic structure. Although this effectively reduces the error rate dramatically, this can be further improved by extending it further by using some language syntax rules that will allow word-level corrections.

- Further experiments can be performed on other zonal languages like Devnagri. Although, we are confident that the model can be applied to other languages like Chinese, Japanese and Hebrew, yet no experiments have been conducted in this area and poses some extended works to future researchers.

- In our current model, the zone identification model is deterministic based on histogram method. However, a fuzzy model can be used to improve the accuracy. An un-experimented idea is given in Appendix B.

- Improve the performance of the approach in terms of speed by introducing ways to reduce the number of broken pieces within a set to an acceptable amount (< 20).

REFERENCES

[1] Sumetphong C. and Tangwongsan S. (2012). Recognizing Broken Thai characters based on Set-Partitions and N-grams Graphs, Journal of Pattern Recognition Research. Accepted for publication.

[2] Sumetphong C. and Tangwongsan S. (2012). Effectively Recognizing Broken Characters in Historical Documents - IEEE International Conference on Computer Science and Automation Engineering. Accepted for Publication.

[3] Sumetphong C. and Tangwongsan S. (2010). Recognizing broken characters in Thai Historical documents, Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering, Vol. 1, pp 99-103.

[4] Sumetphong C. and Tangwongsan S. (2008). Optical Character Recognition Techniques for Restoration of Thai Historical Documents, Proceedings of the International Conference on Computer and Electrical Engineering, pp.531-535.

[5] Pletschacher S., Hu J., Antonacupoulus A. (2009). A new Framework for Recognition of Heavily Degraded Characters in Historical Typewritten Documents Based on Semi-Supervised Clustering, Proceedings of the 10th International Conference on Document Analysis and Recognition.

[6] Jun Sun Hotta, Katsuyama Y., Naoi S. (2005). Camera based degraded text recognition using grayscale features., Proceedings of 8th International Conferent of Document Analysis and Recogition, Vol. 1 pp 182-186.

[7] Zhang P., Bui T.D., Suen C.Y. (2004). Extraction of Hybrid Complex Wavelet features for the Verification of Handwritten Numerals., 9th International workshop on Frontiers in Handwriting Recognition. pp 347-352.

[8] Tonazonni A., Vezzosi S., Bedini L. (2004). Analysis and recognition of highly degraded printed characters, International Journal on Document Analysis and Recognition, Vol 6, pp 236-247.

[9] Namane A., Arezki M., Guessoum A., Soubari E H., Meyrueis P., Bruynooghe M. (2005). Sequential neural network combination for degraded machine printed character recognition, Document Recognition and Retrieval XII vol. 5676, pp 101-110.

[10] Schenkel M., Jabri. M (1998). Low resolution, degraded document recognition using neural networks and hidden Markov models, Pattern Recognition Letters, Vol 19, pp 365-371.

[11] Kwang In Kim, Keechul Jung, Jin Hyung Kim (2003). Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, pp. 1631-1639.

[12] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran (2003). A Bilingual OCR for Hindi-Telugu Documents and its Applications, Seventh International Conference on Document Analysis and Recognition (ICDAR'03) - icdar, vol. 1, pp.408.

[13] Michael Droettboom (2003). Correcting broken characters in the recognition of historical printed documents, Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, pp 364-366.

[14] U. Pachiyanukul, W. Premchaiswadi, and N. Premchaiswadi (2002). Connection of Vertical Broken Line of Thai Printed Character, in Proc. of National Computer Science and Engineering Conference (NCSEC'02), Chonburi, Thailand, October 29-31.

[15] P. Stubberud, J. Kanai, and V. Kalluri (1995). Adaptive image restoration of text images that contain touching or broken characters. In ICDAR 95, pp. 778-781.

[16] Zheng, Q., Kanungo T. (2001). Morphological Degradation Models and Their Use in Document Image Restoration, ICIP01.

[17] Adrian P. Whichello and Hong Yan (1996). Linking Broken Character Borders With Variable Sized Masks to Improve Recognition, Pattern Recognition, vol.29, No.8, pp.1429-1435.

[18] Benedicte Allier and Hubert Emptoz (2002). Degraded character image restoration using active contours: A first approach, Proceedings of the 2002 ACM symposium on Document engineering.

[19] B. Sakoda, J. Zhou and T. Pavlidis (1993). Refinement and testing of a character recognition system based on feature extraction in grayscale space, Proc. Int. Conf. Document Analysis and Recognition, Tsukuba, Japan, pp. 464-469.

[20] Jairo Rocha and Theo Pavlidis (1995). Character Recognition Without Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.17 n.9, pp.903-909.

[21] D. Yu and H. Yan (2001). Reconstruction of broken handwritten digits based on structural morphological features, Pattern Recognition Society Journal, Vol.34, Issue 2, pp. 235-254.

[22] Berrin. A. Yanikoglu (2000). Pitch-based Segmentation and Recognition of dot-matrix text, International Journal of Document Analysis and Recognition (IJDAR), Vol 3, pp 34-39.

[23] M. Oguro, T. Akiyama and K. Oguro (1997). Faxed document image restoration using Gray level Representation, Proc. of 4th International Conference on Document Analysis and Recognition, Vol. 2, pp 679-683.

[24] Premkumar Natarajan, Issam Bazzi, Zhidong Lu, John Makhoul and Richard Schwartz (1999). Robust OCR of Degraded documents, ICDAR, pp 357-361.

[25] M. Cannon, J. Hochberg and P. Kelly (1999). Quality assessment and restoration of Typewritten document images, IJDAR, Vol 2., pp. 80-89.

[26] C. Rodriguez, J. Muguerza, M. Navarro, A. Zarate, J.I.Martin and J.M.Perez (1998). Segmentation of Low-quality Typewritten Digits, Proc. of 14th International Conference on Pattern Recognition, Vol. 2, pp 1106-1109.

[27] W. Premchaiswadi, N. Premchaiswadi, P. Limmaneewichid, and S.Narita (2002). Detection and Reconstruction Scheme for Broken Characters in Thai Character Recognition Systems, IPSJ Journal, Vol.43 No.06.

[28] Seong-Whan Lee, Dong-June Lee, and Hee-Seon Park (1996). A New Methodology for Gray-Scale Character Segmentation and Recognition, IEEE Transactions on PAMI, Vol. 18, No. 10, pp. 1045-1050.

[29] Chinmoy B. Bose and Shyh-Shiaw Kuo (1994). Connected and degraded text recognition using Hidden Markov Model, Pattern Recognition, Vol. 27, No. 10, pp. 1345-1363.

[30] S. Tsujimoto and H. Asada (1991). Resolving Ambiguity in Segmenting Touching Characters, Ist Int.Conference on Document Analysis and Recognition, Saint-Malo, France, pp. 701-709.

[31] R. G. Casey and G. Nagy (1982). Recursive Segmentation and Classification of Composite Character Patterns, Proc. 6th Int. Conf. on Pattern Recognition, Munich, Germany, 1982, pp. 1023-1026.

[32] S. Kahan, T. Pavlidis, and H. S. Baird (1987). On the recognition of printed characters of any font and size, IEEE Transactions on PAMI, Vol. 9, No. 2, pp. 274-288.

[33] Y. Lu, Machine printed character segmentation An overview (1995). Pattern Recognition, Vol. 28, No. 1, pp. 67-80.

[34] Worapoj Peerawit, Warat Yingsaeree and Asanee Kawtrakul (2004). The Utilization of Closing Algorithm and Heuristic Information for Broken

Character Segmentation, IEEE conference on Cybernatics and Intelligent Systems (CIS2004), Singapore.

[35] Whichello A.P., and Yan H. (1996). Linking Broken Character Borders With Variable Sized Masks to Improve Recognition, Pattern Recognition, vol.29, No.8, pp.1429-1435.

[36] Karp R.M. (1972). Reducibility among combinatorial problems, in R.E. Miller and J.W. Thatcher (Eds.) Complexity of Computer Computations, New York: Plenum Press, pp.85-103.

[37] Balas, E. and Padberg (1976). M., Set partitioning a survey, SIAM Review, Vol. 18, No. 4, pp.710-760.

[38] Joseph, A. (2002). A concurrent processing framework for the set partitioning problem, Computers and Operations Research, Vol. 29, No. 10, pp.1375-1391.

[39] Diaby, M. (2010). Linear programming formulation of the set partitioning problem, Int. J. Operational Research, Vol. 8, No. 4, pp.399-427.

[40] Boschetti, M.A., Mingozzi, A. and Ricciardelli, S. (2008). A dual ascent procedure for the set partitioning problem, Discrete Optimization, Vol. 5, No. 4, pp.735-747.

[41] Barahona, F. and Anbil, R. (2002). On some difficult linear programs coming from set partitioning, Discrete Applied Mathematics, Vol. 118, Nos. 1/2, pp.3-11.

[42] Barahona, F. and Anbil, R. (2000). The volume algorithm: producing primal solutions with a subgradient method, Mathematical Programing Series A, Vol. 87, No. 3, pp.385-399.

[43] Cavalcante, V.F., de Souza, C.C. and Lucena, A. (2008). Relax-and-cut algorithm for the set partitioning problem, Computers and Operations Research, Vol. 35, No. 6, pp.1963-1981.

[44] Chan, T.J. and Yano, C.A. (1992). A multiplier adjustment approach for the set partitioning problem, Operations Research, Vol. 40, Suppl., No. 1, pp.S40-S47.

[45] Conforti, M., Summa, M.D. and Zambelli, G. (2007). Minimally infeasible set-partitioning problems with balanced constraints, Mathematics of Operations Research, Vol. 32, No. 3, pp.497-507.

[46] Fisher, M.L. and Kedia, P. (1990). Optimal solution of set covering/partitioning problems using dual heuristics, Management Science, Vol. 36, No. 6, pp.674688.

[47] Lucena, A. (2005). Non delayed relax-and-cut algorithms, Annals of Operations Research, Vol. 140, No. 1, pp.375-410.

[48] Alvarenga, G.B., Mateus, G.R. and de Tomi, G. (2007). A genetic and set partitioning twophase approach for the vehicle routing problem with time windows, Computers and Operations Research, Vol. 34, No. 6, pp.1561-1584.

[49] Lee, Y.H., Kim, J.I., Kang, K.H. and Kim, K.H. (2008). A heuristic for vehicle fleet mix problem using tabu search and set partitioning, Journal of the Operational Research Society, Vol. 59, No. 6, pp.833-841.

[50] Ali, A.I. and Han, H.S. (1998). Reformulation of the set partitioning problem as a pure network with special order set constraints, Annals of Operations Research, Vol. 81, No. 0, pp.233-249.

[51] Ali, A.I. and Thiagarajan (1989). A network relaxation based enumeration algorithm for set partitioning, European Journal of Operational Research, Vol. 38, No. 1, pp.76-85.

[52] El-Darzi, E. and Mitra, G. (1992). Solution of set-covering and set-partitioning problems using assignment relaxations, Journal of the Operational Research Society, Vol. 43, No. 5, pp.483-493.

[53] El-Darzi, E. and Mitra, G. (1995). Graph theoretical relaxations of set covering and set partitioning problems, European Journal of Operational Research, Vol. 87, No. 1, pp.109-121.

[54] Lewis, M., Kochenberger, G. and Alidaee, B. (2008). A new modeling and solution approach for the set-partitioning problem, Computers and Operations Research, Vol. 35, No. 3, pp.807-813.

[55] Sherali, H.D. and Lee, Y. (1996). Tighter representations for set partitioning problems, Discrete Applied Mathematics, Vol. 68, Nos. 1/2, pp.153-167.

[56] Harche, F. and Thompson, G.L. (1994). The column subtraction algorithm: An exact method for solving weighted set covering, packing and partitioning problems, Computers and Operations Research, Vol. 21, pp.689-705.

[57] Linderoth, J.T., Lee, E.K. and Savelsbergh, M.W.P. (2001). A parallel, linear programming-based heuristic for large-scale set partitioning problems, INFORMS Journal on Computing, Vol. 13, No. 3, pp.191-209.

[58] Marsten, R.E. (1974). An algorithm for large set partitioning problems, Management Science, Vol. 20, No. 5, pp.774-787.

[59] Garey, M. and Johnson, D. (1979). Computers and Intractability: A Guide to the Theory of NP-Completeness, San Francisco: Freeman.

[60] Chu, P.C and Beasley, J.E. (1998). Constraint handling in genetic algorithms: The set partitioning problem. Journal of Heuristics, Vol 11, pp. 323-357.

[61] M O'Sullivan, C. Walker, Q.S. Lim, and S. Mitchell. (2011). Dippy - a simplified interface for advanced mixed-integer programming, Report, University of Auckland Faculty of Engineering, No. 685.

[62] Ryan, D.M and Falkner, J.C. (1988). On the integer properties of scheduling set paritioning models, European Journal of Operational Research, Vol. 35, pp.422-456.

[63] S. Mitchell, A. Kean, A.Mason, M O'Sullivan, and A. Phillips (2009). Optimization with pulp, http://www.coin-or.org/PuLP/.

[64] Alan M. Frisch and Ian Miguel and Toby Walsh (2001). Proceedings of the IJCAI-01 Workshop on Modelling and Solving Problems with Constraints, pp.39-45.

[65] CSP Libary: http://www.csplib.org, Problem 38.

[66] P. Sarkar, H. Baird, and X. Zhang. (2003). Training on severely degraded text-line images. In Proceedings of the International Conference on Document Analysis and Recognition, pages 38-43.

[67] Babu, N. Preethi, N.G. Shylaja, S.S. (2008). Degraded Document Image Enhancement Using Hybrid Thresholding and Mathematical Morphology, Computer Vision, Graphics & Image Processing, ICVGIP '08. Sixth Indian Conference, pp. 701-705.

[68] Q. Wang, T. Xia, L. Li, and C. L. Tan (2003). Document image enhancement using directional wavelet. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[69] Sezgin, M and Sankur, B (2004). Survey over Image Thresholding Techniques and Quantitative Performance Evaluation, Journal of Electronic Imaging 13(1), pp 146-165.

[70] R. Gonzalez and R.Woods (2002). Digital Image Processing. Prentice Hall.

[71] Roli Bansal, Priti Sehgal, Punam Bedi (2008). A Novel Framework for Enhancing Images Corrupted by Impulse Noise Using Type-II Fuzzy Sets, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 3, pp.266-271.

[72] John D. Hobby, Tin Kam Ho. (1997). Enhancing Degraded Document Images via Bitmap Clustering and Averaging, Proceedings of the 4th International Conference on Document Analysis and Recognition, ICDAR pp 394-400.

[73] H. Cao and V. Govindaraju. (2007). Handwritten carbon form preprocessing based on markov random field. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[74] G. Myers and K. Donaldson. (2005). Bayesian super-resolution of text in video with a text-specific bimodal prior. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 11881195.

[75] S. Geman and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence, 6(6):721741.

[76] S. Z. Li. Markov. (1995). Random Field Modeling in Computer Vision. Springer-Verlag.

[77] Y. Huang, M. S. Brown, and D. Xu (2008). A framework for reducing ink-bleed in old documents. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[78] M. D. Gupta, S. Rajaram, N. Petrovic, and T. S. Huang (2005). Restoration and recognition in a loop. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

[79] Gatos B., I. Pratikakis, S. J. Perantonis (2006). Adaptive Degraded Image Binarization, Journal of Pattern Recognition 39.

[80] Yan H. (1996). Unified Formulation of a Class of Image Thresholding Techniques, Pattern Recognition 29.

[81] Kittle J., J.Illingworth and J.Foglein. (1985). Threshold Selection Based in a Simple Image Statistic, Computer Vision, Graphics and Image Processing, vol .30, pp. 125-147.

[82] Yang J., Y. Chen, W. Hsu (1994). Adaptive Thresholding Algorithm and its Hardware Implementation, Pattern Recognition Letters. 15.

[83] Sauvola J., M. Pietikainen (2000). Adaptive Document Image Binarization, Pattern Recognition 33.

[84] Parker. J.R., C. Jennings, A.G. Salkauskas (1993). Thresholding Using an Iillumination Model, ICDAR93.

[85] Kim K., D.W. Jung, R.-H. Park (2002). Document image binarization based on Topographic Analysis Uusing a Water Flow Model, Pattern Recognition 35.

[86] Trier O.D., A.K. Jain (1995). Goal-Directed Evaluation of Binarization Methods, IEEE Trans. Pattern Anal. Mach. Intell. 17 (12).

[87] J. Fisher, S. Hinds, and D. DAmato (1990). A rule-based system for document image segmentation. 10th International Conference on Pattern Recognition, 1990. Proceedings. Vol 1 pp 567572.

[88] A. Jain and B. Yu (1998). Document representation and its application to page decomposition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(3) pp 294308.

[89] B. Kruatrachue and P. Suthaphan (2001). A fast and efficient method for document segmentation for ocr. In Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, Vol 1, pp381383.

[90] S. Imade, S. Tatsuda, and T. Wada (1993). Segmentation and classification for mixed text/image documents using neural network. In Proceedings of the Second International Conference on Document Analysis and Recognition, pp 930934.

[91] A. Jain and S. Bhattacharjee (1992). On texture in document images. In Proceedings of Computer Vision and Pattern Recognition, Vol 31, pp 677680.

[92] Y. Zheng, H. Li, and D. Doermann. (2003). Text identification in noisy document images using markov random field. In Proceedings of the Seventh International Conference on Document Analysis and Recognition.

[93] S. C. Zhu, Y. N. Wu, and D. B. Mumford. (1998). Filters random field and maximum entropy (frame): Towards a unified theory for texture modeling. International Journal of Computer Vision, 27(2) pp 120.

[94] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. (2007). Text extraction and document image segmentation using matched wavelets and mrf model. IEEE Transactions on Image Processing, 16(8).

[95] M. Tuceryan and A. K. Jain. (1998) The Handbook of Pattern Recognition and Computer Vision, chapter Texture Analysis, pages 207248. World Scientific Publishing Co.

[96] A. A. Effros and T. K. Leung. (1999). Texture synthesis by nonparametric sampling. In IEEE International Conference on Computer Vision.

[97] V. Ubeda (1993). A Parallel Thinning Algorithm Using KxK Masks. JPRAI(7) , pp. 1183-1202.

[98] Z. Hang and Y.Y. Wang. (1994). A New Parallel Thinning Methodology, IJPRAI(8), pp. 999-1011.

[99] N. J. Naccache and R. Shingha. (1984). SPTA: A proposed Algorithm for thinning binary patterns, vol. SMC-14, no.3, IEEE trans. Syst. Man and Cybern, pp. 409-418.

[100] N. J. Naccache and R. Shinghal, An investigation into the skeletonization approach of Hilditch, vol. 17(3), Pattern Recognition, pp.279-284, (1986).

[101] D. Akhter and M. M. Ali (1998). A Fast Thinning Algorithm for Bangla Characters, Proc.ICCIT pp. 132-136.

[102] I. Bar Yosef. (2005). Input sensitive thresholding for ancient Hebrew manuscript, Pattern Recognition Letters, Vol 26, pp. 1168-1173.

[103] L. Likforman-Sulem, A. Zahour, B. Taconet (2007). Text line segmentation of historical documents: a survey, International Journal of Document Analysis and Recognition. Vol 9, pp. 123-138.

[104] M. Arivazhagan, H. Srinivasan, and S. N. Srihari (2007). A statistical approach to handwritten line segmentation, Document Recognition and Retrieval XIV, Proceedings of SPIE, California, pp. 1-11.

[105] D. J. Kennard, W. A. Barrett (2006). Separating Lines of Text in Free-Form Handwritten Historical Documents, Proceedings of the Second International Conference on Document Image Analysis for Libraries, pp.12-23.

[106] A. Zahour, B. Taconet, P. Mercy, S. Ramdane (2001). Arabic hand-written text-line extraction, Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 281-285.

[107] S. Zhixin, S. Setlur, V. Govindaraju (2005). Text Extraction from Gray Scale Historical Document Images Using Adaptive Local Connectivity Map, Proceedings of the Eighth International Conference on Document Analysis and Recognition, pp. 794-798.

[108] Ju Han, Kai-Kuang Ma. (2007). Rotation-invariant and scale-invariant Gabor features for texture image retrieval, Image and Vision Computing, Vol 25, issue 9, pp 1474-1481.

[109] M. Ansani (1999). Diplomatica e diplomatisti nell'arena Digitale. Scrineum 1, pp.1-11.

[110] American Library of Congress: http://www.loc.gov/index.html

[111] Archiving Early America: http://www.earlyamerica.com/earlyamerica/freedom /doi/doi.html

[112] National Library of Wales: http://www.llgc.org.uk

[113] The British Library: http://www.bl.uk

[114] Institute of Historical Research, UK: http://www.history.ac.uk

[115] Ahlul Bayt Digital Library Project: http://www.al-islam.org

[116] Yale Law School: http://www.law.yale.edu

[117] Liberty Fund: http://www.libertyfund.org

APPENDICES

APPENDIX A
NOISE IDENTIFICATION

We present an idea to identify any noise piece(s) in the set X. We presume that any such piece is too close in the vicinity of the majority of the pieces of X that elimination by distance cannot be employed (outlier method). We employ the jack-knife method where each piece is evaluated on its contribution or non-contribution towards enhancing the score of the subset S it is a member of.

Let X be the set of broken pieces. Let $\mathbf{S} := \wp(X)$ - $\{\emptyset\}$. A piece $b \in X$ can be considered as a noise if it satisifies the following two conditions

1. $P(\{b\}) < \tau$ and
2. $\forall S \in \mathbf{S}$, if $b \in S$ and $S \neq \{b\}$, then $P(S - \{b\}) > P(S)$

A small experiment was performed on an image. Of the 184 noise pieces, 128 were correctly identified and eliminated from further processing. However, of the 6,127 non-noise pieces, 76 were deemed as noise and also eliminate causing some amount of misclassifications. Hence, if some refinements can be made on this model, then the accuracy of this noise identification can be further improved and hence improve on the entire process of broken character recognition.

APPENDIX B
FUZZY ZONE IDENTIFICATION

Given the handwritten nature of the Thai historical documents, inclination angle in the line images and the broken nature of the characters, an exact compuation of zone boundaries is impractical. We choose to implement a fuzzy model which we discuss here. For an example of the problem, see Figure B.1. We approximate the middle zone using a histogram method and define fuzzy membership functions, one for each zone.

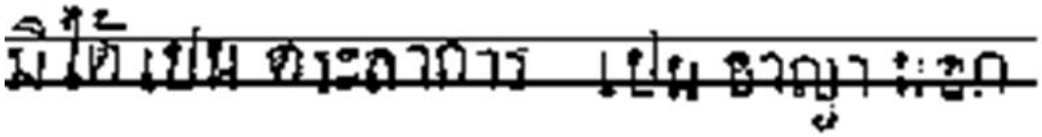


Figure B.1: Problem in definition of zone boundaries

- H : Height of the Line Image Strip
- M_T : Estimated Top Boundary of the Middle Zone
- M_B : Estimated Bottom Boundary of the Middle Zone
- $M = M_B - M_T$: Height of the Middle Zone
- $T_1 = M_T - p_1 M$ where $p_1 = 0.15$ (adjustable)
- $T_2 = M_T + p_2 M$ where $p_2 = 0.50$ (adjustable)
- $T_3 = M_B + p_1 M$
- $B_1 = M_T + p_1 M$
- $B_2 = M_T + p_2 M$
- $B_3 = M_T + p_1 M$

Let V_T and V_B be the vertical positions of the topmost and bottommost pixels of S respectively. We define the fuzzy membership functions as follows

$$f_U(S) = \left\{ \begin{array}{ll} 0 & \text{if } V_T > T_1 \\ \frac{T_1 - V_T}{T_1} & \text{otherwise} \end{array} \right\} \tag{B.1}$$

$$f_L(S) = \left\{ \begin{array}{ll} 0 & \text{if } V_T < T_2 \\ \frac{V_T - V_T}{T_3 - T_2} & \text{if } T_2 \leq V_T \leq T_3 \\ 1 & \text{otherwise} \end{array} \right\} \tag{B.2}$$

$$f_M(S) = 1 - f_U(S) - f_L(S) \tag{B.3}$$

$$g_U(S) = \left\{ \begin{array}{ll} 1 & \text{if } V_B < T_1 \\ \frac{B_2 - V_B}{B_2 - B_1} & \text{if } B_1 \leq V_B \leq B_2 \\ 0 & \text{otherwise} \end{array} \right\} \tag{B.4}$$

$$g_L(S) = \left\{ \begin{array}{ll} 0 & \text{if } V_B > B_3 \\ \frac{V_B - B_3}{H - B_3} & \text{otherwise} \end{array} \right\} \tag{B.5}$$

$$g_M(S) = 1 - g_U(S) - g_L(S) \tag{B.6}$$

The fuzzy zone values of S is [U,M,L] where

- U = mean$(f_U(S), g_U(S))$: denotes the confidence that S belongs to upper zone
- M = mean$(f_M(S), g_M(S))$: denotes the confidence the S belongs to middle zone
- L = mean$(f_L(S), g_L(S))$: denotes the confidence the S belongs to lower zone

The actual zone of S can be determined as argmax([U,M,L]). Based on this model, we are to achieve almost perfect zone identification of any given S.

APPENDIX C
AN ALTERNATIVE MODEL

In this section, we provide an alternate model to solving the problem of broken characters by obtaining the optimal set partition π^* of the broken pieces X based on linear programming.

Notations

1. X : $\{x_1,x_2,...,x_n\}$ (set of n broken pieces in set X)
2. N : $\{1, 2, .., n\}$ (set of indices for the broken pieces of set X)
3. m = 2^n - 1 (number of different non empty subsets of X)
4. $\mathbf{S}$: $\{S_1,S_2,...,S_m\}$ (set of non-empty subsets of X)
5. M = $\{1, 2, ..., m\}$ (set of indices for the non empty subsets of X)
6. u : $\{u_1,u_2,...,u_m\}$ (u_j is the score of S_j)

Let $u_j(j \in M) = -lnP(S_j)$ where $P(S_j)$ is in the continuous range 0..1. Let $\pi_j(j \in M)$ denote a binary 0/1 variable that is 1 iff S_j is in the optimal π^*. Let $A_{ij}((i, j) \in (N, M))$ be the binary 0/1 variable that is 1 iff $x_i \in S_j$. Iteratively for each $k \in N$, we solve the linear programming problem. At each iteration we increment k and update β value if the best solution found in the latest iteration is better than the previously found solution. We set the initial value of k = 2 and set initial value of $\beta = -lnP(X)$.

Minimize : $\frac{1}{k}\sum_{j\in M} u_j\pi_j$

s.t.: $\sum_{j\in M} A_{ij}\pi_j = 1 \qquad i \in N$

$\sum_{j\in M} u_j\pi_j \leq k\beta$

APPENDIX D
RESTORING DEGRADED DOCUMENTS

In this section, we present an overview of a framework for the computerized restoration of Thai Historical documents. The framework encompasses the various OCR-related steps that are required to restore these valuable documents. We organize the framework under 5 different stages as can be seen in Figure D and present a survey of OCR methods that are applicable for various stages.

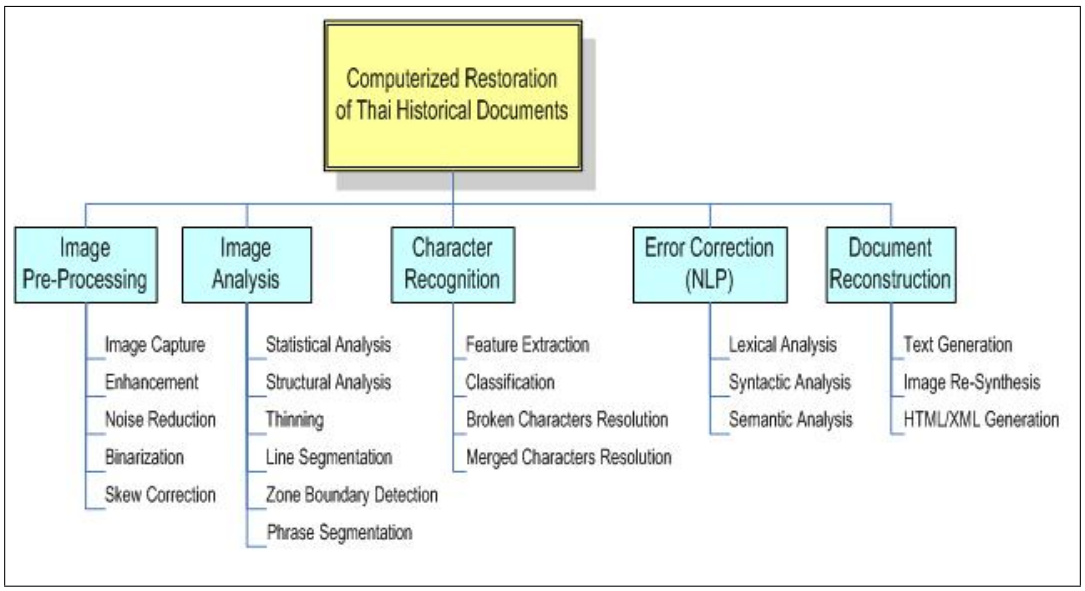


Figure D.1: Restoration Framework

Image Capture is one of the most crucial steps and is often ignored as being trivial. The end-result of restoring the historical documents is often directly related to how well it was initially captured. The resolution of the captured image has a direct impact on the performance of the later stages both in terms of accuracy and efficiency. Roli [71] proposed a fuzzy-logic based framework to **enhance images** that have been corrupted by impulse noise. Babu [67] uses a hybrid thresholding method to enhance degraded document images. Hobby [72] did some experiments on how bitmap clustering techniques can be applied to enhancing degraded images.

Images with certain known **noise models** can be restored using the traditional image restoration techniques such as Median filtering and Weiner filtering, etc. [70]. However, in practice, degradations arising from phenomena such as document aging or ink bleeding cannot be described using popular image noise models. Document processing algorithms proposed in [66] and improved upon by Huang et al. [77] works by incorporating document specific degradation models and text specific content models. Another approach combines the degradation model and the document model into a powerful MRF-based optimization framework [75] [76]. To achieve generic restoration of carbon copy documents, Cao and Govindaraju [73] used a document content model. The model consisted of a set of 55 binary patches, trained using high quality data, which is used for restoring noise and removal of rulings on the paper. Gupta et al. [78] used a patch based alphabet model to remove blurring artifacts for license plate images using a camera. [68] [74].

Some of the researchers applied local adaptive thresholding scheme which **binarizes** and enhances poor quality of the degraded document based on the specific locations of meaningful textual information [80] [81] [82] [83]. Others applied global thresholding in which a single calculated threshold value is used to classify image pixels into background classes [84] [85] [86]. Gatos [79] uses a successive application of shrink and swell filtering technique to binarize a degraded image. Yosef [102] introduces a novel thresholding technique used to binarize ancient Hebrew documents. A survey of the image thresholding methods was conducted in [69].

Many existing methods currently exist for performing text extraction for OCR and most of them are applicable to degraded documents [87] [88]. One early method for **document segmentation** into more than two region types is proposed by Imade et al [90]. The authors use different thresholding techniques and histogram features as input to a neural network in order to classify the different regions. Similarly, Kruatrachue [89] uses a modified heuristic edge following methods to efficiently segment document images for explicit OCR. Some approaches to document segmentation focus on specific features of the image or identifying word blocks to obtain segmentation results. One such approach implemented by Kumar et al. [94] uses matched wavelet filters and Fisher classifiers to estimate a three class image labeling problem: text,

background, and picture. Another approach by Zheng et al. [92] uses a set of filter and histogram features in conjunction with Fisher classifiers to identify and distinguish text regions for handwritten and machine printed text.

In one of the earlier works to examine texture in document images [91] [95], the authors examine the texture of different document images through the use of multichannel filtering based on Gabor filters in order to identify textual regions. Although many recent and sophisticated models exist to model textures in images, such as proposed by Zhu et al. [93], they fail to represent the kind of inhomogeneous texture which we find present across the different regions of a document image and remain too computationally expensive to be of any practical value. Some approaches like one proposed by Effros et al. [96] seek to get around this by using nonparametric sampling in the synthesis process based on image windows, but the methods are not yet sufficiently advanced.

The most conventional **thinning** algorithms [97] [98] and other methods were designed for English or other renowned languages. Since English characters shape is very simple and these algorithms are only dedicated for that language set, they are not applicable to Thai and other Asian-like languages that are full of different features. Some Asian researchers have adopted other techniques that yield well-thinned images for complex languages [99] [100] [101].

Most of the existing **line segmentation** methods are applicable to binary images only [103]. Valiant attempts towards gray-scale image segmentation have been proposed by Zhixin [107] and Kennard [105]. A natural choice for line segmentation of images is the project profile method [103]. The mean weakness of this approach is its sensitivity to the skew of the lines in the document. However, as often seen in historical documents (as well as in handwritten documents), text-lines can be curved in different skew angles. For that purpose, some authors [104] [106] proposed a piece-wise segmentation method that divided the document into non overlapping vertical stripes, and applied the projection to each stripe.

Pletschacher et al. [5] computes standard **statistical features** in an attempt to recognize characters in a heavily degraded English document. These features include width, height, aspect ratio, black-to-white pixels ratio etc. A popular method is to

extract Gabor features from degraded documents [108] [6] and this has been proven to be invariant towards rotation and scaling operations. Another contemporary feature extraction technique involves computing 2D wavelet transforms [7].

Different **classification models** have been proposed for OCR systems. These include Bayesian networks, Template matching, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Hidden Markov Models (HMM) etc. Artificial Neural networks have been used as the classifier in different works related to degraded OCR [8] [10]. Hopfield networks have also been used successfully in recognition of degraded documents [9]. Researchers have used Support Vector Machines to recognize degraded characters [11] [12]. A more complicated approach of using Hidden Markov Models have also been implemented by [10] [24].

A number of algorithms have been proposed in the past to recognize merged characters. Segmentation of merged characters is the major problem during recognizing characters in a degraded document. Lee et al. [28] have segmented the merged characters using projection profiles and topographic features extracted from the gray scale images. Bose and Kuo [29] used Hidden Markov Model (HMM) to recognize merged characters in degraded text. Tsujimoto and Asada [30] constructed a decision tree for resolving ambiguity in segmenting merged characters. Casey and Nagy [31] proposed a recursive algorithm for segmenting merged characters. Kahan et al. [32] have proposed a very useful double differential function to segment the merged characters. Lu [33] proposed a generalized differential technique for the purpose by mapping vertical projection on to second projection called peak-to-valley ratio.

BIOGRAPHY

NAME	Mr. Chaivatna Sumetphong
DATE OF BIRTH	6th October 1968
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Mahidol University, 1987–1991
	Bachelor of Science (Computer Science)
	Asian Institute of Technology, 1991-1992
	Master of Science (Computer Science)
	Mahidol University, 2005–2011
	Doctor of Philosophy (Computer Science)
SCHOLARSHIP	James Linen III Memorial Award, 1992
	Asian Institute of Technology
	For Computer Sceince Student
POSITION	Instructor
	Mahidol University International College
HOME ADDRESS	816/17 Charansanitwongse 3
	Thapra, Bangkokyai
	Bangkok 10600, Thailand
E-MAIL	chaivatptc@hotmail.com