A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

BUNTIDA SUVACHARAKULTON

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2010

Thesis
entitled

A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

……………….………….…..…………….
Miss. Buntida Suvacharakulton
Candidate

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

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

……………….………….…..…………….
Asst. Prof. Chomtip Pornpanomchai, 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
Master Programme in Computer Science
Faculty of Information and Communication Technology
Mahidol University

Thesis
entitled

A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Computer Science)
on
May 12, 2010

…………………..…………..…..……………
Miss. Buntida Suvacharakulton
Candidate

…………………..…………..…..……………
Asst. Prof. Panjai Tantasanawong, Ph.D.
Chair

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

……………………..…...……………...…
Asst. Prof. Chomtip Pornpanomchai, Ph.D.
Member

…………………..…………..…..……………
Asst. Prof. Sukanya Phongsuphap, 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.
Acting Dean
Faculty of Information and Communication
Technology
Mahidol University

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Assoc. Prof. Dr. Supachai Tangwongsan for the continuous support of my master degree study and research, for his patience, motivation, enthusiasm, and immense knowledge. He gave appreciate suggestion, checked and corrected the fault of this thesis.

Besides my advisor, I would like to special thank the rest of my thesis committee: Asst. Prof. Dr. Panjai Tantasanawong, Asst. Prof. Dr. Sukanya Phongsuphap and Asst. Prof. Dr. Chomtip Pornpanomchai for their encouragement, insightful comments, and helpful suggestions on this research.

I am deeply thankful to the staff of the Department of Computer Science and the University Computing Center, Mahidol University, for their helpful in providing facilities and materials for my thesis experiment. Also, I would like to thank to my friends at the graduate classes, Mahidol University, for many helpful recommendations.

Furthermore, I am grateful to Miss. Koonthida Suvacharakulton, my younger sister for her extremely cheerfulness. Thank also to my relation and my friends for their help and encouragement.

Last but not the least, my graduation would not be achieved without best wish from my parents, Mr. Veerasak and Mrs. Nukoon Suvacharakulton, who help me for everything and always give me greatest love, willpower and spiritual support throughout my life.

Buntida Suvacharakulton

A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

BUNTIDA SUVACHARAKULTON 4836577 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE : SUPACHAI TANGWONGSAN, Ph.D., SUKANYA PHONGSUPHAP, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D., PANJAI TANTASANAWONG, Ph.D.

ABSTRACT

This thesis proposes a model of optical character recognition with the technique of word prediction for bilingual documents in Thai and English. The model is hence named, BOCR-WP (Bilingual Optical Character Recognition with Word Prediction).

The BOCR-WP is an enhancement of conventional OCR with two additional and distinctive processes: language identification and word prediction. For language identification, the process attempts to distinguish which language mode those image strips should belong to, Thai or English, as a result of the identification. In word prediction, the process is actually followed by character verification after the processing of character recognition. The main idea is that instead of attempting to recognize each individual character via the conventional method, the new approach is trying to identify whole words, either in Thai or English, by using contextual analysis to predict those probable words. Then verify them to obtain the right one by template matching. Obviously, the longer the matched word is, the better the speed of recognition will be. Finally, the technique of dictionary look-up is used in order to improve the accuracy of the final answer for the whole recognition process.

A series of experiments showed that the BOCR-WP was able to classify the script modes, Thai or English, correctly with a high accuracy of 99.99% on average. This system also yielded a better performance compared to conventional OCR in terms of speed improvement with a best case of 28.85%, 22.20% on average, and a minimal improvement of 15.69% while still being able to maintain a quality of accuracy of 100% in Thai and 99% in English from a source of 141 bilingual documents or a total of 284,417 characters.

KEY WORDS: BILINGUAL OCR / N-GRAM / THAI CHARACTER RECOGNITION / WORD PREDICTION / WORD VERIFICATION

90 pages

ระบบประสิทธิภาพสูงสำหรับการรู้จำตัวพิมพ์ภาษาไทยและภาษาอังกฤษด้วยวิธีพยากรณ์คำศัพท์
A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

บุญธิดา สุวัชระกุลธร 4836577 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., สุกัญญา พงษ์สุภาพ, Ph.D., ชมทิพ พรพนมชัย, Ph.D., ปานใจ ธารทัศนวงศ์, Ph.D.

บทคัดย่อ

งานวิจัยนี้ นำเสนอแบบจำลองสำหรับการรู้จำตัวพิมพ์อักษรไทยและอังกฤษ ด้วยวิธีการพยากรณ์คำ โดยแบบจำลองที่นำเสนอนี้ มีชื่อว่า ระบบ BOCR-WP

ระบบ BOCR-WP สามารถเพิ่มประสิทธิภาพของการรู้จำตัวอักษรของระบบรู้จำทั่วไป โดยเพิ่มเติมเทคนิคพิเศษ 2 วิธีการ กล่าวคือ การระบุภาษา และการพยากรณ์คำ สำหรับการระบุภาษา นำมาใช้ในการแยกโหมดของรูปภาพตัวอักษร ว่าเป็นภาษาไทยหรืออังกฤษ ส่วนวิธีพยากรณ์คำ นำมาใช้แทนการรู้จำตัวอักษรทีละตัวของระบบการรู้จำทั่วไป ในแนวทางใหม่ โดยพยายามทำนายเซตของคำที่น่าจะเป็น ของแถบรูปภาพตัวอักษร ด้วยการวิเคราะห์เชิงบริบท และตรวจสอบคำเหล่านี้ ด้วยวิธีการเข้าคู่รูปแบบ เพื่อหาคำที่เป็นคำตอบสำหรับการรู้จำ โดยคำที่เข้าคู่กับแถบรูปภาพตัวอักษร ยิ่งมีความยาวมาก ยิ่งทำให้การรู้จำรวดเร็วยิ่งขึ้น นอกจากนี้ ยังได้นำเทคนิคการค้นหาคำในพจนานุกรม มาช่วยปรับปรุงการรู้จำ ให้มีความถูกต้องมากขึ้น

ผลการทดลองกับเอกสารสองภาษา ไทยและอังกฤษ 141 หน้า จำนวนทั้งสิ้น 284,417 ตัวอักษร แสดงให้เห็นถึงความสามารถของระบบ BOCR-WP ในการระบุภาษา ไทยหรืออังกฤษ ด้วยความแม่นยำสูง โดยมีความถูกต้องโดยเฉลี่ย เท่ากับ 99.99% นอกจากนี้ ระบบสามารถรู้จำตัวอักษรไทยและอังกฤษ ได้เร็วขึ้น 28.85% สำหรับกรณีที่ดีที่สุด 22.20% สำหรับค่าเฉลี่ย และ 15.69% สำหรับกรณีที่แย่ที่สุด ในขณะเดียวกัน ระบบยังสามารถรักษาคุณภาพของการรู้จำ ที่ความถูกต้องเฉลี่ย 100% สำหรับตัวอักษรไทย และ 99% สำหรับตัวอักษรอังกฤษ

90 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

Page

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

This thesis presents a model of optical character recognition with the technique of word prediction for bilingual documents in Thai and English. The model is hence named as BOCR-WP (Bilingual Optical Character Recognition with Word Prediction).

In the globalization era, since electronic documents are widely used for information interchange, therefore; the OCR system is a necessary tool for converting paper documents from image into text as online equivalent contents. For the past decades or even further, most of the local research works in OCR would deal with only monolingual of Thai documents, whereas in daily situation, documents would contain not only monolingual but bilingual or even multi-lingual text, such as Thai-English, Thai-English-Chinese. Therefore, we could see that it is quite a big challenge to solve the problem of character recognition in bilingual documents. For practicality and possibility, we would like to take bilingual documents with Thai and English text at this stage of work.

Furthermore, despite of the fact that there are quite a number of outstanding research studies in printed Thai or English character recognition with high accuracy and performance, the speed of recognition could still be improved if we are able to find a new approach in attempting to speed up the recognition process. If that could be achieved, it is also undoubtedly a big challenge for our research work.

In the present work, we propose the model of BOCR-WP as the solution of the aforementioned problems. The BOCR-WP system is an enhancement of the conventional OCR with two additional and distinctive processes: language identification and word prediction. For the language identification, the process attempts to distinguish which language mode those image strips should belong to, Thai or English as the result of identification. In the word prediction, the process is actually followed by character verification after processing of character recognition. The main idea is instead of attempting to recognize each individual character as the conventional method, the new approach is trying to identify whole words either in Thai or English by using contextual analysis to predict those probable words, and verify them to obtain the right one by the template matching. Obviously, the longer the matched word is, the better the system speed of recognition will be. Finally, the technique of dictionary look-up is used in order to improve the accuracy of the final answer for the whole recognition process.

Based on a series of experiments of the BOCR-WP system in the design and the experimental studies, the results are quite impressive as follows: The new system is able to classify the script mode: Thai or English correctly with 99.99% on the average. Then, this system is able to yield a correct result in character recognition with 100% in Thai and 99% in English at the end of processing. For processing time in recognition, this system is able to recognize a character in 0.52 milliseconds on the average with the technique of word prediction, compared to 0.66 milliseconds as by the conventional one. From the whole series of experiments, we found that the speed improvement would be 28.85% in the best case, 22.20% in the average, and 15.69% as the very minimal one from the source of 141 documents in bilingual of a total 284,417 characters.

The thesis is organized into 7 chapters as follows:

Chapter I: Introduction.

Chapter II: Problem Statement. It discusses the motivations leading to this research, related works, problem outline, including its objective and scope.

Chapter III: Conceptual Design. It describes a main concept to solve these research problems. For further clarity, conceptual design of language identification, word prediction, word verification and dictionary look-up are described in subsequent sections.

Chapter IV: System Design. It describes a detailed design of the BOCR-WP system. The system overview is shown in structure chart and step-by-step work flows of those processes are presented with activity diagrams and algorithms.

Chapter V: System Implementation. It describes hardware and software specification used to create the system and also those program modules to make BOCR-WP as a reality.

Chapter VI: Experimental Results. It presents a series of experimental results for evaluating the system performance and also a table of comparison with the conventional approach.

Chapter VII: Discussion and Conclusion. It discusses those issues in experimental results and makes the final conclusion. In addition, we suggest some research directions for future works.

CHAPTER II
PROBLEM STATEMENT

In this chapter, we describe the motivation of the present research work in Section 2.1, followed by literature survey of those previous works in Section 2.2. Furthermore, the problem statement is given in Section 2.3, then; research objectives and research scope are in Sections 2.4 and 2.5 respectively. Details are as follows:

2.1 Motivation

The present work is motivated by the following challenges:

1. Despite of the fact that there are quite a number of outstanding research studies in printed Thai character recognition over past two decades with high accuracy and performance, the speed of recognition could still be improved if we were able to find a new approach in attempting to speed up the recognition process. If that be successful, it is indeed a big challenge.

2. Most of the local research works deal with only monolingual of Thai documents, whereas in daily situation, documents would contain not only monolingual but also bilingual or even multi-lingual text. Therefore, it is also another big challenge to recognize bilingual documents with Thai and English text at this stage of work.

2.2 Literature Survey

In this section, those related words to the present research interest are described in 3 topics: Language identification, Thai OCRs and Multi-lingual OCRs. Details are as follows:

2.2.1 Related Works in the Language Identification

In optical character recognition (OCR), initially, a number of research studies over past half century attempt to propose high accuracy OCRs for monolingual document. Then, in recent years, research direction continues in advance to extend OCRs for multilingual documents. Some research studies attempt to enhance methodology of the language identification while the others try to extend the monolingual OCR system into multilingual one. Some details will be explained in the following.

Several approaches for language identification as seen in the past few decades were in the followings:

Juan Cheng et al. [1] proposed normalized histogram statistic approach for script identification of document analysis in 4 scripts: Chinese, Japanese, English and Russian in 2006. The average accuracy of script identification is better than 92.5% while the precision of individual script varies between 90% and 96.7%.

Then, in 2009, P.A. Vijaya et al. [2] proposed an efficient technique of language identification for Kannada, Hindi and English based on the characteristic features of top-profile and bottom-profile of individual text lines of the input document image. On the testing set of 600 text lines, the overall classification accuracy was 96.6%.

Additionally, in Thai, S. Chanda et al [3] proposed word-wise Thai and Roman script identification approach named SVM based method and simple feature obtained from structural shape, profile behavior, component overlapping information, topological properties, and water reservoir. The accuracy of script identification tested on 10,000 words was 99.62% on the average.

2.2.2 Related Works in Thai OCRs

In printed Thai character recognition (Thai OCR) for monolingual document, a number of researches proposed various approaches for enhancing the recognition performance. The early works used structural matching technique which compared topological features of unknown character image with character templates. Next, neural network approach, for instance, backpropagation neural network (BNN), ARTMAP neural network and so on were applied to the recognition system of printed

Thai characters for improving the recognition accuracy. Then fuzzy logic theory–a statistical approach–was proposed in research studies. In addition, printed Thai character recognition researches focused on structural analysis were proposed since the beginning of the current decade and genetic algorithm was used to recognize printed Thai characters in recent years. In the following, we explain some printed Thai character recognition researches as only proposed in this decade.

Firstly, in structural matching, S. Srisuk [4] proposed Hausdorff distance technique for printed Thai character recognition to measure the similarity of Thai characters in 1999. Its experimental results tested on over 60,000 isolated noised characters confirmed high recognition accuracy: 100% for noise free and a little noisy characters and above 95% for noisy characters.

In 2000, A. Kawtrakul and P. Waewsawangwong [5] proposed minimum Euclidean distance technique for classifying unknown Thai characters based on multiple features of characters. This research experimented on a thousand characters of 3 fonts. The recognition speed was 5 character images per second with 97.44% correctness.

Secondly, in statistical approach, P. Le-wan and A. Kawtrakul [6] proposed a hybrid technique: fuzzy C-means and K-means for the recognition of printed Thai characters in 2001. This system tested on small noise free and noisy characters made high recognition accuracy and noise tolerant.

In 2004, R. Foopratheepsiri et al [7] proposed fuzzy logic theory for enhancing the recognition of printed Thai character by using principle of dividing object boundary and Fast Active Contours for classification and using Wavelet Transform Modulus Curvature (WTMC) for searching outstanding features of each character image. This research experimented on Thai character recognition on various character fonts and styles based on the decision of fuzzy logic. The average recognition accuracy was over 95%.

Next, based on neural network theory, there were several kinds of neural network presented in a number of research studies. For instance, P. Phokharakul and C. Kimpan [8] proposed backpropagation neural network (BNN) to recognize handprinted Thai characters using cavity features of character in 1998. The experimental results proved the usefulness of this method with the recognition rate of 98.3% for 3,200 handprinted Thai characters.

In addition, BNN was combined with inductive logic programming (ILP) in printed Thai character recognition research by B. Kijsirikul and S. Sinthupinyo [9] in 1999. This system experimented on a few thousands characters of two fonts and seven sizes. The recognition accuracy was 94.26% on average.

In 2005, A. Thammano and P. Duangphasuk [10] proposed hierarchical cross-correlation ARTMAP neural network for recognizing printed Thai characters of no head fonts. This system experimented on small character sets of twelve no head fonts. The recognition accuracy was about 84% on average.

Then, in structural analysis, S. Mitatha and et al [11] proposed rough sets in printed Thai character recognition in 2001. This research built 3 sets of decision rules to classify unknown characters and used Voting algorithm to determine the best rule for each ones. It is 100% recognition accuracy tested on a few thousands characters.

In 2003, two stages rough sets were proposed in printed Thai character recognition by S. Mitatha and et al [12]. This research classified characters into 38 groups and 4 additional groups and recognized characters in each group using set of decision rules. The recognition accuracy tested on 2,064 training characters and 6,880 testing characters of 4 fonts and 4 sizes are 100% and 81% respectively.

In 2008, S. Tangwongsan and O. Jungthanawong [13] proposed stroke structural features and classification rules for the recognition system of printed Thai documents with over 1,000,000 characters of multi-font and multi-size. The recognition accuracy was above 99% on average and the average recognition time per character was about 36 milliseconds.

Finally, based on genetic algorithm, C. Pornpanomchai and M. Daveloh [14] proposed genetic algorithm in the recognition experiments of printed Thai characters in 2007. This research applied genetic algorithm in cellular automata pattern and through Conway’s rules of the game of life and experimented on a thousand characters. The recognition accuracy was 97.04% and the average recognition time was about 85 seconds per character.

From the researches of printed Thai character recognition mentioned above, almost researches proposed the systems to improve only recognition accuracy. Therefore, the recognition rates of these systems were above 95% while the speeds of the recognition were quite slow due to complex computation in recognition process. In addition, sizes of testing data sets were rather small except the system using stroke structural features and classification rules [13]. However, recognition performance among these researches might not be comparable due to the differences of research criteria such as the characteristics of tested input.

2.2.3 Related Works in Multilingual OCRs

In multilingual OCR, Q.Huo et al. [15] proposed minimum classification error (MCE) based character-pair modeling and negative training for improving Chinese/English OCR performance in 2003. The experimental results tested on 16,216 Chinese and English mixed characters yielded the recognition accuracy of 99.38%.

Then, R.S. Kunte and R.D.S. Samuel [16] proposed a bilingual machine-interface OCR for printed Kannada and English text by applying gabor filter based features for language identification, wavelets for feature extraction, and multilayer feed forward neural network for character recognition. The accuracy of recognition was 90.5% on the average.

Then, Kai Wang et al. [17] proposed Chinese/English split algorithm based on global information and a language identification rule by using Bayesian formula for Chinese/English mixed OCR with character level identification in 2009. In comparison with the conventional mixed OCR, accuracy of recognition increases from 98.48% to 99.13% on magazine samples and from 98.68% to 99.25% on book samples.

From research studies mentioned above, although there were various approaches proposed in printed Thai character recognition, researchers still try to modify the previous methods or to apply the new one for enhancing the recognition performance.

2.3 Problem Statement

Problem statement of this research is given as follows:

1. How to find a new approach for printed Thai character recognition which could speed up the recognition process while the recognition accuracy could still be maintained at the high level just at least as good as the previous one.

2. How to enhance the work to cover the recognition for bilingual documents in Thai and English with high performance in accuracy and speed perspectives.

2.4 Research Objectives

There are some objectives of this research as follows:

1. To study methodologies of the character recognition and analyze the advantages and disadvantages of these techniques especially in printed Thai characters.

2. To study fast techniques for searching and verifying and apply the appropriate techniques in printed Thai character recognition for speeding up the system while maintaining 100% of recognition accuracy.

3. To study a number of approaches of the language identification processing in multilingual documents in order to select those effective ones for identifying Thai and English characters.

2.5 Research Scope

The scopes of this research are described in the following.

1. We limit the work to cover only printed characters in Thai and English, but not any hand written types at all. All printed characters are listed as follows:

1.1 Thai character group

1.1.1 Forty four consonants: ก, ข, ฃ, ค, ฅ, ฆ, ง, จ, ฉ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ผ, ฝ, พ, ฟ, ภ, ม, ย, ร, ล, ว, ศ, ษ, ส, ห, ฬ, อ, and ฮ.

1.1.2 Eighteen vowels: -ะ, -า, เ-, แ-,-ิ,-ี, -ึ,-ื, -ุ, -ู, โ-, ไ-, ใ-, -ํ, -ั, -็, ฤ, and ฦ.

1.1.3 Four tonal marks: -่, -้, -๊, and -๋.

1.1.4 Three special characters: -์, ๆ, and ฯ.

1.2 English character group

1.2.1 Twenty six capital letters: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, and Z.

1.2.2 Twenty six small letters: a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, and z.

1.2.3 Ten digits: 1, 2, 3, 4, 5, 6, 7, 8, 9, and 0.

2. We also limit the work to cover only printed Thai characters of the fonts: EucrosiaUPC and FreesiaUPC, the sizes: 16 and 18, and the styles: normal and bold and English letters of the fonts: Arial and Cordia New, the sizes: 11 and 12 for Arial font and 16 and 18 for Cordia New font, and the styles: normal and bold. Moreover, those characters could be in the same A4 document image with one character size per line.

3. In addition, we will consider those documents prepared in grayscale bitmap file with noise free and no line skewing.

CHAPTER III
CONCEPTUAL DESIGN

From the problem statement mentioned in Chapter II, this chapter presents a concept of how to solve the problem by using the technique of word prediction, and name the working model as BOCR-WP (bilingual optical character recognition by word prediction) system. In Section 3.1, we show the BOCR-WP system overview. Next in Section 3.2, we show the concept of language identification process as to discriminate character scripts to their respect types before character recognition. In addition, word prediction and word verification are conceptually described in Sections 3.3 and 3.4 respectively. Finally, dictionary look-up is presented as the final step to improve the over-all accuracy in Section 3.5. Details are as follows:

3.1 Overview

The conceptual design of the BOCR-WP system is shown in Figure 3.1. The BOCR-WP system is an enhancement of the OCRs with two additional processes: language identification and word prediction followed by verification after processing of character recognition. The main idea is instead of attempting to recognize individual characters; the methodology is trying to identify whole words by using contextual analysis to predict those probable words, and verity the right one by the template matching technique. Actually, the longer the matched word is, the better the system performance will be. Finally, the technique of dictionary look-up is used in order to improve the accuracy of the final answer for the whole recognition process.

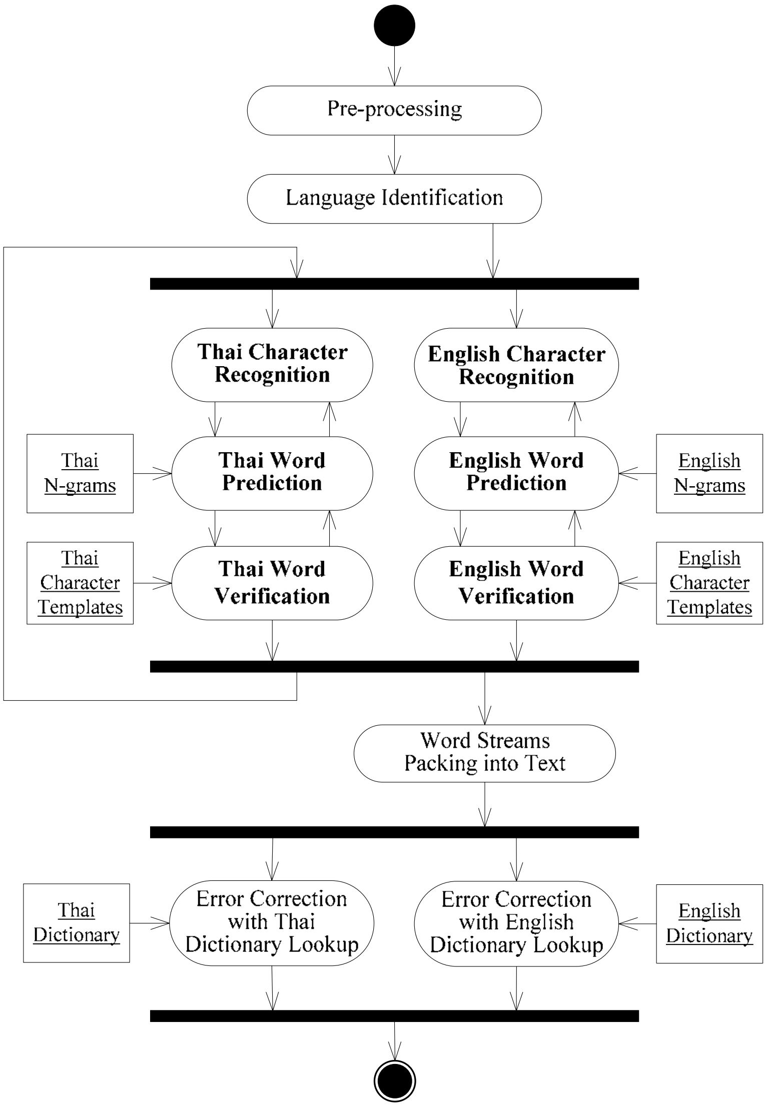
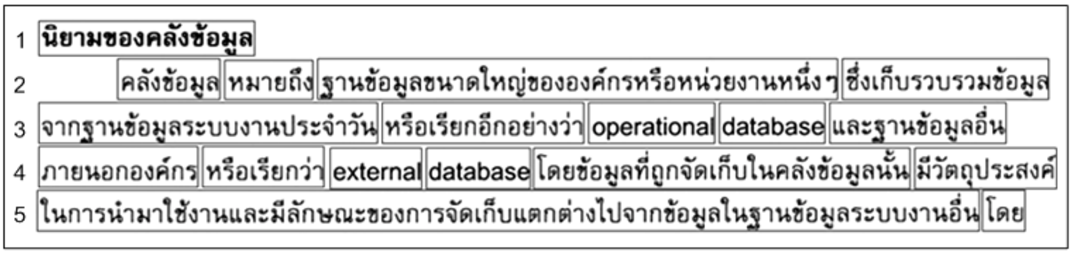

Pre-processing
Language Identification
Thai Character Recognition
English Character Recognition
Thai N-grams
Thai Word Prediction
English Word Prediction
English N-grams
Thai Character Templates
Thai Word Verification
English Word Verification
English Character Templates
Word Streams Packing into Text
Thai Dictionary
Error Correction with Thai Dictionary Lookup
Error Correction with English Dictionary Lookup
English Dictionary

Figure 3.1: Conceptual design of the BOCR-WP system

Details are described in the following Sections.

3.2 Language Identification

After the pre-processing step, the next is language identification with an aim to distinguish which character scripts of language they belong to. In this research, language identification in word level is used to identify Thai/English bilingual characters. This process is performed in iterations on block images contained in the strip images one after another. A sample of block images extracted from part of Thai and English mixed document is shown in Figure 3.2.



1 นิยามของคลังข้อมูล

2 คลังข้อมูล หมายถึง ฐานข้อมูลขนาดใหญ่ขององค์กรหรือหน่วยงานหนึ่ง ๆ ซึ่งเก็บรวบรวมข้อมูล

3 จากฐานข้อมูลระบบงานประจำวัน หรือเรียกอีกอย่างว่า operational database และฐานข้อมูลอื่น

4 ภายนอกองค์กร หรือเรียกว่า external database โดยข้อมูลที่ถูกจัดเก็บในคลังข้อมูลนั้น มีวัตถุประสงค์

5 ในการนำมาใช้งานและมีลักษณะของการจัดเก็บแตกต่างไปจากข้อมูลในฐานข้อมูลระบบงานอื่น โดย

Figure 3.2: A sample of block images in Thai and English mixed document

From Figure 3.2, the part of Thai and English mixed document comprised of 5 strip images could be divided into 18 block images. In printed Thai and English mixed document, there are 4 sets of decision rules to identify character script according to the differences of Thai and English characteristics [18] as shown in Figure 3.3.

```
Rule 1 : Number of middle level character >21
    If true then
        language mode is Thai
    Otherwise
        goto rule 2

Rule 2 : Having bottom level character(s)
    If true then
        language mode is Thai
    Otherwise
        goto rule 3

Rule 3 : Having top level character(s)
    If true then
        language mode is Thai
    Otherwise
        goto rule 4

Rule 4 : Having a character head
    If true then
        language mode is Thai
    Otherwise
        language mode is English

Note: Iteration until end of the image strips.
```

Figure 3.3: The decision rules of language identification

3.3 Word Prediction

In word prediction, a main concept is to predict a list of next probable words based on the previous word token. This process requires linguistic knowledge of the language, for this reason, a dictionary is used to retrieve all possible candidates and list them in an N-gram tree.

As to reduce the search space in the N-gram tree, we simply sort nodes of those candidates according to their probabilistic order of occurrences. Samples of 3-gram implemented in tree structure of the word token “โรง” and “bea” are shown in Figure 3.4 and 3.5.

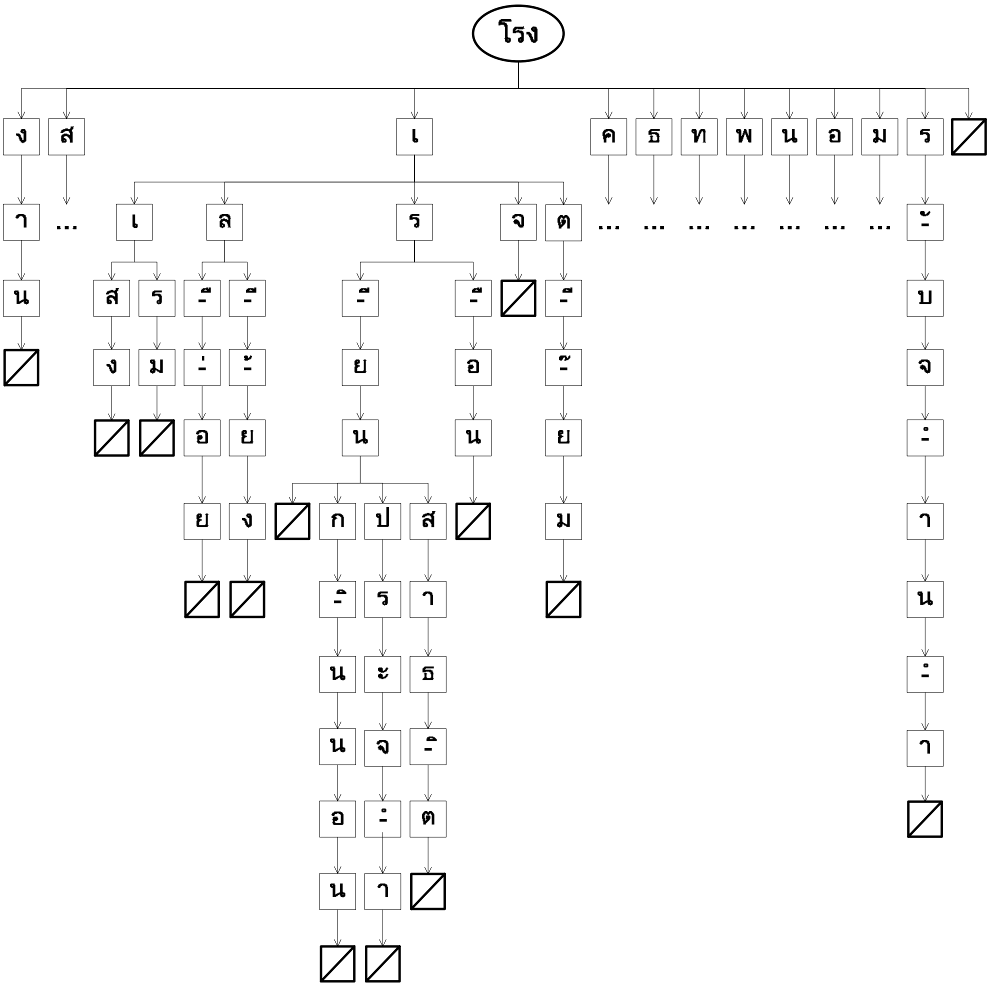


Figure 3.4: A sample of 3-gram tree of the word token “โรง”

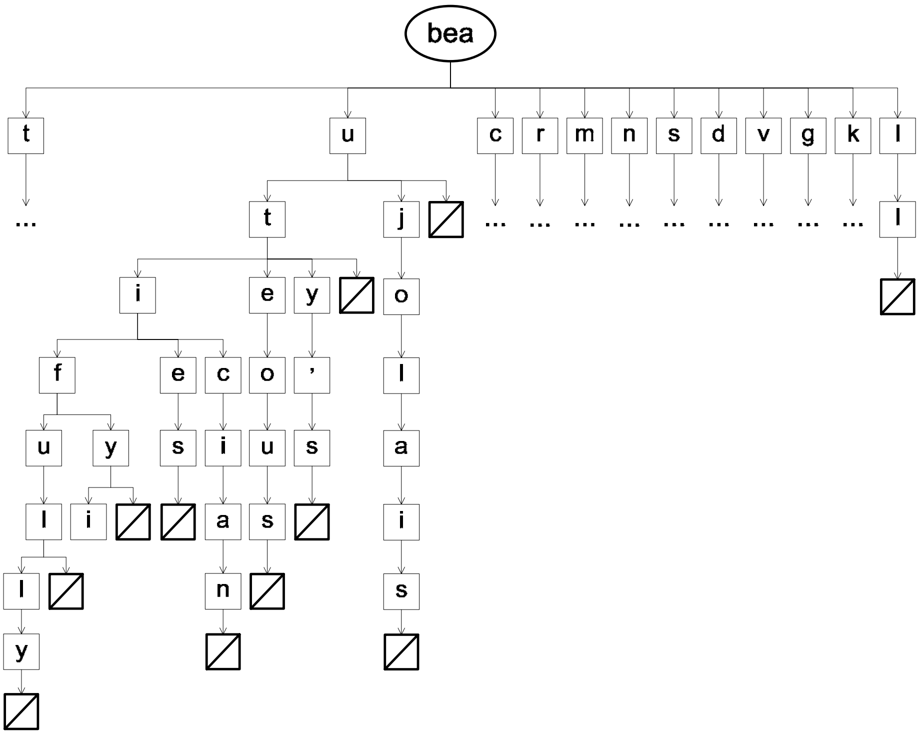


Figure 3.5: A sample of 3-gram tree of the word token “bea”

A node in the tree of Figure 3.4 and 3.5 contains one character. A terminal node represents the end of word.

From the token “โรง” in Figure 3.4, the system will retrieve all candidates of 28 predictive words as follows: โรงงาน โรงสี โรง โรงแสง โรงเลื่อย โรงแรม โรงครัว โรงคัล โรงธาร โรงธารคำนัล โรงทาน โรงทึม โรงเรียน โรงเรียนกินนอน โรงเจ โรงเตี๊ยม โรงเรือน โรงเลี้ยง โรงเรียนประจำ โรงเรียนสาธิต โรงพัก โรงพิมพ์ โรงนา โรงพยาบาล โรงสีข้าว โรงอาหาร โรงมหรสพ โรงรับจำนำ. Then the next probable characters of the previous word token could be ‘ง’ ‘ส’ ‘เ’ ‘ค’ ‘ธ’ ‘ท’ ‘พ’ ‘น’ ‘อ’ ‘ม’ ‘ร’ and ‘$’ (end of word).

As well as the token “bea” in Figure 3.5, the system will retrieve all candidates of 66 predictive words as follows: beat, beautiful, beach, bear, beauty, bears, beam, beating, bearing, beaten, beats, beans, bearings, bean, beasts, beard, beaches, beautifully, bead, beacon, beaver, beau, bearer, beagle, bearded, beatings, beatrice, beak, beaming, beauties, bearish, bear's, beaker, beards, bearable, beater, beaded, beaut, beadles, beanstalk, beady, beall, beaujolais, beautify, beatniks, beauteous, beardless, beachcomber, beaching, beakers, beadle, beautician, bearskin, beautifying, beavertail, beanpole, beast, beads, beams, beatnik, beachhead, beachwear, beastly, beauty's, beardown and beadsman. Then the next probable characters of the previous word token could be ‘t’, ‘u’, ‘c’, ‘r’, ‘m’, ‘n’, ‘s’, ‘d’, ‘v’, ‘g’, ‘k’, and ‘l’.

After the matching is successful with any one of next probable characters in the list by the verification process as elaborated in the following section, the prediction will continue to look for new candidates of predictive words.

3.4 Word Verification

Word verification is to match unknown character images of the block image with candidate of predictive words based on its features. In the BOCR-WP system, template matching technique is used in the verification step. This process is done in iteration on unknown characters in the block image by matching with candidates of the predictive words. It will not continue if there is a word match, end of the candidate list or end of the block image; otherwise, it continues to work in this step.

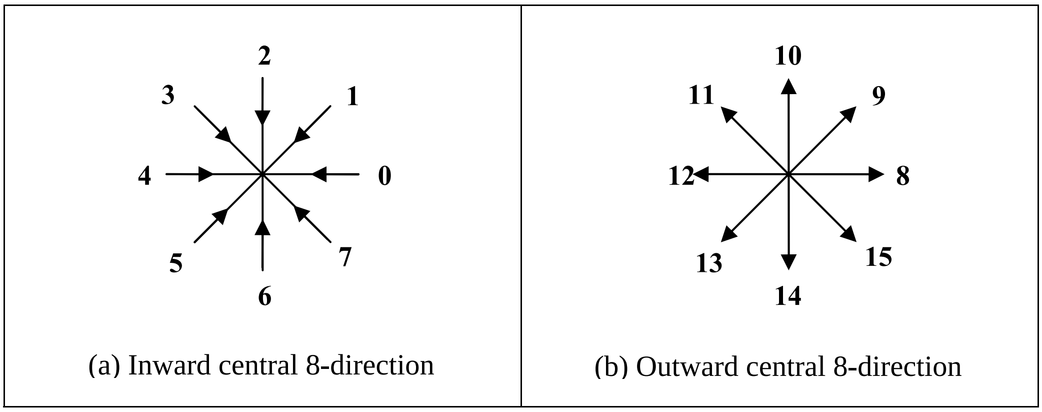
According to the concept of word verification mentioned above, matching characters require information of the character features; as a result, the templates of the whole characters are used in verification for matching the unknown character with templates of the candidate characters as presented in Section 3.4.1.

3.4.1 Template Matching Technique

In verification processes, there are three template matching approaches implemented with negative and positive matching. Firstly, in Section 3.4.1.1, negative matching is performed to reject dissimilarity of next probable character in the predictive word and the unknown character image. This process will not continue if the result is a mismatch; otherwise, it continues to take the positive matching. Next, in Sections 3.4.1.2 and 3.4.1.3, positive matching is used to verify similarity of next probable character in the predictive word and the unknown character image. Details are as follows:

3.4.1.1 Template Matching with 8-Direction Vector

In this approach, there are three character features: zone, overlapped zone and width-height ratio extracted from the unknown character image without thinning. To match these features of the unknown character image and ones of candidate character is done by a set of decision rules. In addition, 8-direction vector which comprised of 16 features in inward and outward central 8-directions is used to represent outer and inner bounds of the character with distance between central pixel and first black pixel found in all directions as shown in Figure 3.6. Then, Euclidean distance is implemented to match 8-direction vector of the unknown character image and template of the candidate one.



(a) Inward central 8-direction

(b) Outward central 8-direction

Figure 3.6 (a)-(b): Inward and outward central 8-directions

3.4.1.2 Template Matching with Black Pixel Percentages

In this approach, we extract character features: black pixel percentages from 9 equal parts of the unknown character image without thinning. Then, Euclidean distance is implemented to match these features of the unknown character image and template of the candidate one.

3.4.1.3 Template Matching with Loop Feature

In this approach, we extract character features: number of loop and position(s) of loop(s) from the unknown character image without thinning. Then a set of decision rules is implemented to match these features of the unknown character image and ones of candidate character.

3.5 Dictionary Look-up

Dictionary look-up is to search a word in dictionary with matching criteria; for instance, approximate matching, longest matching, maximum matching, and so on according to the purpose of implementation. In the BOCR-WP system, we apply dictionary look-up in order to improve recognition accuracy with longest matching search.

The main concept is to match a word of recognition result and vocabulary of dictionary with longest matching criteria, on the other hand, the word will be searched in dictionary which is sorted in alphabetical order by using binary search technique. Then, the longest matching vocabularies derived from dictionary look-up will be probable word(s) for a correct recognition. Therefore the maximum matching word from all candidates is the optimal one for correcting recognition mistake. Conceptual view of dictionary look-up process is shown in Figure 3.7.

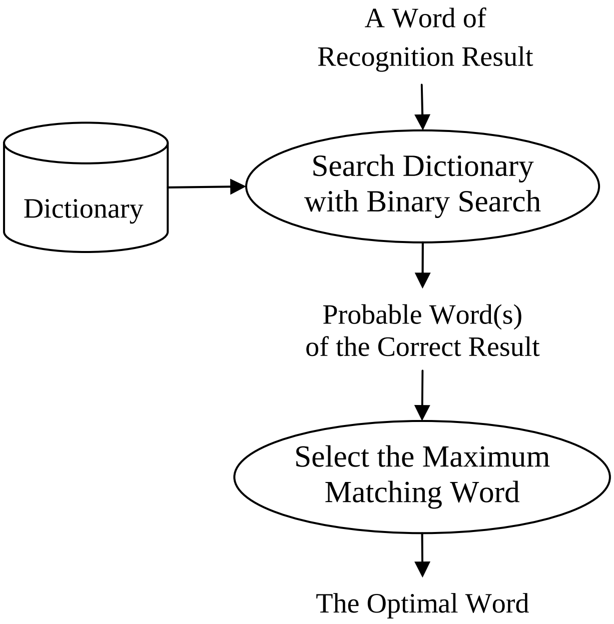


Figure 3.7: Conceptual view of dictionary look-up

CHAPTER IV
SYSTEM DESIGN

According to the conceptual design mentioned in Chapter III, this chapter explains the BOCR-WP (bilingual optical character recognition by word prediction) system in details. Firstly, in Section 4.1, we present the BOCR-WP system overview. Next, in Section 4.2, we present the process of language identification, followed by character recognition in Section 4.3. Furthermore, word prediction and word verification are meticulously described in details in Sections 4.4 and 4.5 respectively. Finally, dictionary look-up is explained in Section 4.6. Details are as follows:

4.1 System Overview

From the conceptual design mentioned in Chapter III, the BOCR-WP system comprises of six main processes as shown in structure chart as in Figure 4.1. Firstly, in pre-processing, binarization is implemented to convert a documents image in gray scale into binary image, then noise reduction is used to reduce isolated point noises and isolated holes noises in the binary image and segmentation is implemented to segment the noise free binary image into character images of strip images by using projection and loop [22] prior recognition processing in the next steps. Then, in language identification, we identify boundary of block images, followed by classifying their language modes as implemented in Section 4.2. In addition, character recognition comprises of three processes: thinning, feature extraction and classification is implemented in Section 4.3. Furthermore, in word prediction, the implementations of N-gram construction and N-gram look-up are described in Section 4.4, followed by word verification by using template matching is explained in Section 4.5. Finally, dictionary look-up is implemented to solve incorrect recognition results in both Thai and English as shown in Section 4.6. Details are as follows:

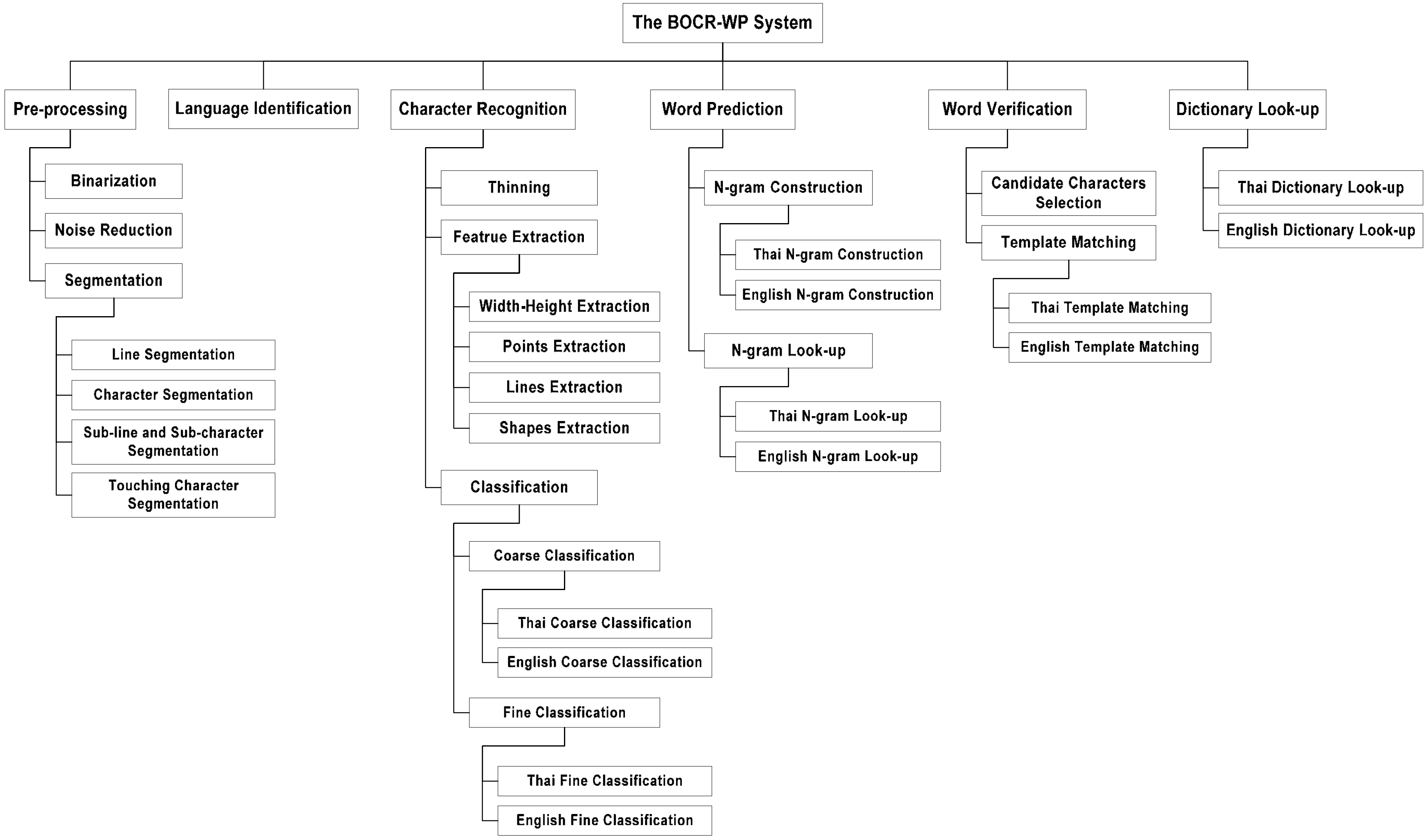


Figure 4.1: Structure chart of the BOCR-WP system

4.2 Language Identification

There are two processes in the language identification: image block detection, and language mode classification. In the process of image block detection, the system simply scans an image strip from left to right, and find those regions of white space, then each image block is formed by using the white spaces as its block boundaries. The next process is the language mode classification, in which each image block is classified to its probable language mode by using a set of decision rules. The output will be a series of image blocks with their respective language modes for further recognition processing. Figure 4.2 shows an activity diagram of the language identification, followed by Algorithms 2.1 and 2.2 as the processes for image block detection and language mode classification respectively.

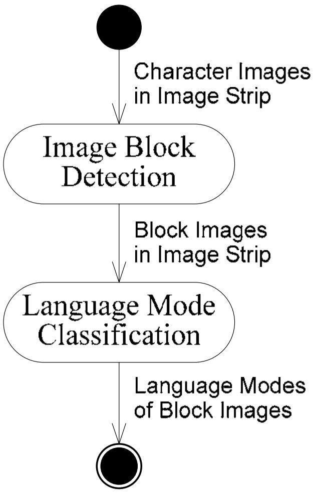


Figure 4.2: Activity diagram of language identification

Algorithm 2.1 Image Block Detection

Description: Checking the boundary of block image (end of word or phrase)

Input: Character image arrays, start and end rows, start and end columns

Output: End of block images

```
Scan all rows of the strip images from start row to end row
Begin
    Scan all columns of the character images from start column to end column
    Begin
        // Find white space widths between character images
        Set white space width of this character image to be start columns of next
        images - end column of this image
    End
    Sum all white space widths
    Set average white space width to be sum of all white space widths / numbers
    of white space
    Scan all white space widths of character images
    Begin
        // Find start and end columns of image blocks
        If this white space width > average white space width then
            Set end of this block image to be end column of this character image
    End
End
```

Algorithm 2.2 Language Mode Classification

```
Description: Classify language mode of the block image
Input:  Block image array, start and end rows, start and end columns of the
        block image
Output: Language mode of the block image
// Identify language mode of the block image
Scan all block images
Begin
    Set initial mode of this block image to be NULL
    While mode is not NULL do
    Begin
        // Rule 1: number of middle level characters > 20
        If number of middle level characters > 20 then
            Set mode of this block image to be “Thai”
        // Rule 2: having bottom level character(s): ุ, ู
        Else If number of bottom level character > 0 then
            Set mode of this block image to be “Thai”
        // Rule 3: having top level character
        Else If number of top level character > 0 then
            // Having dot: i, j
            If number of dot > 0 then
                Set mode of this block image to be “English”
            // No having dot: ั, ็, ิ, ี, ึ, ํ, ื, ้, ่, ์, ๊, ๋
            Else
                Set mode of this block image to be “Thai”
        // Rule 4:  having character head
        Else
            Set position to be this start block image
            While (position <= this end block image) and (mode is not NULL) do
            Begin
                If number of loop of this character image > 0 then
                    // Having head: ข, ฃ, ค, ฅ, ง, จ, ฉ, ช, ซ, ฌ, ญ, ฎ, ฏ, ฐ, ฑ, ฒ,
                    // ณ, ด, ต, ถ, ท, น, บ, ป, ถ, ผ, ฝ, พ, ฟ, ภ, ม, ย, ร, ฤ, ล, ฦ, ว, ศ,
                    // ษ, ส, ห, ฬ, อ, ฮ, เ, ไ, ใ, โ, ๆ, ฯ
                    If number of character head > 0 then
                        Set mode of this block image to be “Thai”
                    // No having head: A, B, D, O, P, Q, R, a, b, d, e, g, o, p, q,
                    // 0,4, 6, 8, 9
                    Else
                        Set mode of this block image to be “English”
                Else
                    // Non-middle zone character: C, E, F, G, H, I, J, K, L, M,
                    // N, S, T, U, V, W, X, Y, Z, f, h, k, l, t, y, 1, 2, 3, 5, 7
                    If this character image is not middle zone character then
                        Set mode of this block image to be “English”
                        Increment value of position
            End
        End
    End
```

4.3 Character Recognition

In character recognition process, there are three procedures: thinning, feature extraction and character recognition as shown in Figure 4.3.

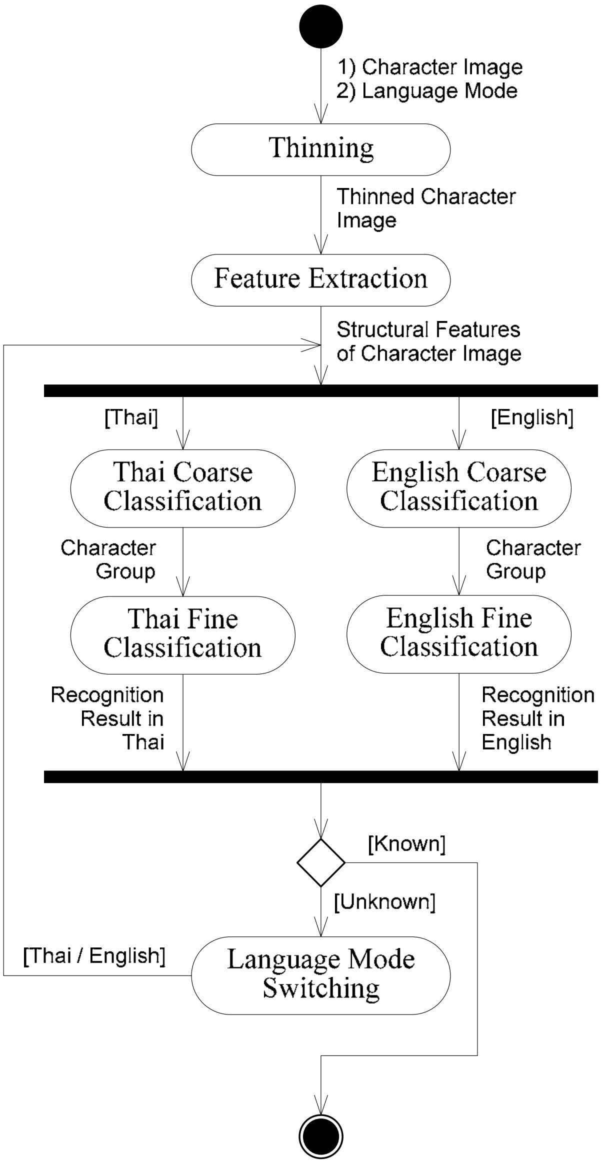


Figure 4.3: Activity diagram of character recognition

First, in thinning process, ZS algorithm using parallel processing [23] is implemented to skeletonize character image. Next in feature extraction process as presented in Section 4.3.1, structural features of character image are extracted. Finally, in character recognition process as presented in Section 4.3.2, the skeletonized character image is mapped to its probable language script. Details are as follows:

4.3.1 Feature Extraction

This process is used to extract structural character features. In the BOCR-WP system, character features are categorized into 4 groups: width-height information, points, lines and shapes. The implementation of feature extraction is shown in activity diagram in Figure 4.4.

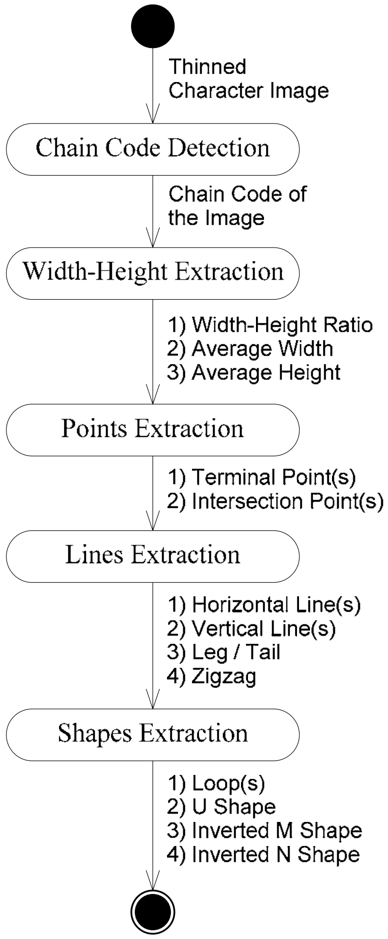


Figure 4.4: Activity diagram of feature extraction

In feature extraction, 8-direction chain code is first detected from the thinned character image to represent character structure. Next, in Section 4.3.1.1, we present the implementation of width-height information, followed by step by step of points extraction in Section 4.3.1.2. In addition, the implementation of lines and shapes extraction are shown in Section 4.3.1.3 and 4.3.1.4 respectively. Details are as follows:

4.3.1.1 Width-Height Extraction

Width-height information is used to determine type of character: narrow, normal or wide character. In this step, width-height ratio, average width and average height are extracted from the thinned character image as shown in Algorithm 3.1.

Algorithm 3.1 Width-height information calculation

```
Description: Calculate width-height ratio, average width and height
Input:  Start and end rows and columns of all thinned character images,
        number of thinned character images
Output: Width-height ratio, average width and height
Scan all start and end rows and columns of all thinned character images
Begin
    Set image width of this input image to be end column - start column
    Set image height of this input image to be end row - start row
    Set width-height ratio of this input image to be image width / image height
    Sum image width of input image
    Sum image height of input image
End
Set average width to be sum of image width / number of the input images
Set average height to be sum of image height / number of the input images
```

4.3.1.2 Points Extraction

This process looks up terminal and intersection points of the thinned character image according to its nearest neighbor information. The results of this step contain numbers of terminal and intersection points and positions of the terminal and intersection points according to the positioning pattern as shown in Figure 4.5. A sample of terminal and intersection points is illustrated in Figure 4.6.

0	1	2
3	4	5
6	7	8

Figure 4.5: The positioning pattern for identifying position of the character feature

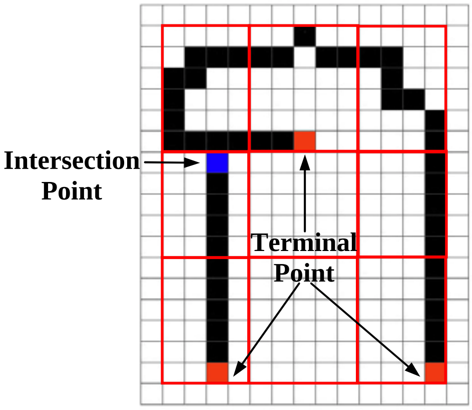


Figure 4.6: A sample of terminal and intersection points

In Figure 4.6, there are three terminal points occurred in position 1, 6 and 8 and one intersection point occurred in position 3. Details of terminal point extraction and intersection point extraction are shown in Algorithms 3.2 and 3.3 respectively.

Algorithm 3.2 Terminal point extraction

```
Description: Find terminal point(s) of the character image and define
             position(s) of the terminal point(s)
Input: Thinned character image array, start and end row, start and end column
Output: Number of terminal point and position(s) of the terminal point(s)
Scan all columns of the input mage from start column to end column
Begin
    Scan all rows of the input image from start row to end row
    Begin
        If the input pixel is black pixel then
            Count number of white to black pixels in sequence $P_1P_2P_3P_4P_5P_6$
            $P_7P_8P_1$ of pattern P mentioned in thinning process
            If number of white to black pixels is equal to 1 then
                Set this input pixel to be terminal point
                Find the position of this terminal point
    End
End
Remark: White represents 0 and black represents 1.
```

Algorithm 3.3 Intersection point extraction

```
Description:   Find intersection point(s) of the character image and define
               position(s) of the intersection point(s)
Input: Thinned character image array, start and end row, start and end column
Output: Number of intersection point and position(s) of the intersection point(s)
Scan all columns of the input image from start column to end column
Begin
    Scan all rows of the input image from start row to end row
    Begin
        If the input pixel is black pixel then
            Count number of white to black pixels in sequence $P_1P_2P_3P_4P_5P_6$
            $P_7P_8P_1$ of pattern P mentioned in thinning process
            If number of white to black pixels >= 3 then
                Set this input pixel to be intersection point
                Find the position of this intersection point
    End
End
Remark: White represents 0 and black represents 1.
```

4.3.1.3 Lines Extraction

Horizontal and vertical lines are extracted from chain code of the thinned character image. Then tail is extracted by determining the occurrences of long vertical line ended with terminal point in position 2 (top-right segment) and leg is extracted by determining the occurrences of long vertical line ended with terminal point in position 8 (bottom-right segment). In addition, zigzag is extracted from vertical projection profile of the character image. Samples of lines appeared in the characters are shown in Figure 4.7. Details are presented in Algorithms 3.4 - 3.6.

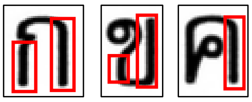
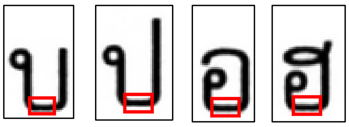
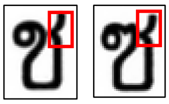
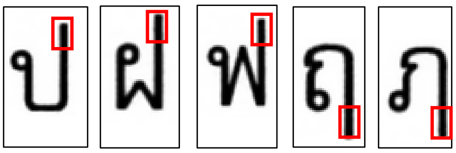
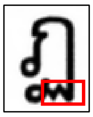
Line	Samples
Vertical Line	
Horizontal Line	
Tail	
Leg	
Zigzag	

Figure 4.7: Samples of lines appeared in the characters

Algorithm 3.4 Vertical and horizontal lines extraction

Description: Find vertical and horizontal lines in the chain code of thinned character image

Input: Chain code of the thinned character image, average width and height of the thinned character image

Output: Number of vertical and horizontal line, position(s) of the vertical and horizontal line(s), type(s) of the vertical and horizontal line(s)

```
// Find vertical and horizontal lines using chain code
Set initial number of vertical line to be zero
Set initial number of horizontal line to be zero
Scan all chain code of the input image
Begin
    Count number of adjacent 0 chain code
    Count number of adjacent 2 chain code
    Count number of adjacent 4 chain code
    Count number of adjacent 6 chain code
    If number of adjacent 2 or 6 chain code <= 0.8*average height then
        Increment number of vertical line
        Set vertical line type to be long
    Else If number of adjacent 2 or 6 chain code <= 0.6*average height then
        Increment number of vertical line
        Set vertical line type to be medium
    Else If number of adjacent 2 or 6 chain code <= 0.3*average height then
        Increment number of vertical line
        Set vertical line type to be short
    If number of adjacent 0 or 4 chain code <= 0.8*average width then
        Increment number of horizontal line
        Set horizontal line type to be long
    Else If number of adjacent 0 or 4 chain code <= 0.6*average width then
        Increment number of horizontal line
        Set horizontal line type to be medium
    Else If number of adjacent 0 or 4 chain code <= 0.3*average width then
        Increment number of horizontal line
        Set horizontal line type to be short
End
```

Algorithm 3.5 Tail and leg extraction

Description: Find leg or tail of the character image and define its type

Input: Number of terminal and intersection point, position(s) of the terminal and intersection point(s), chain code, number of vertical line and type(s) of vertical line(s) of the thinned character image

Output: Number of tail and leg, type of tail and leg

```
Scan all terminal and intersection points of the input image
Begin
    // Find tail of the character image
    If there are terminal and intersection point in position 2 then
        If there are set of adjacent 1 chain code at end parts of chain code then
            Set number of tail to be 1
        Else
```

```
            Set number of tail to be 0
        Else
            Set number of tail to be 0
        // Find leg of the character
        If there is intersection point in position 3 then
            If there is terminal point in position 2 and have long vertical line then
                Set number of leg to be 1
                Set type of leg to be 1
            Else If there is terminal point in position 8 and have long vertical line
            then
                Set number of leg to be 2
                Set type of leg to be 1
            Else
                Set number of leg to be 0
        Else
            Set number of leg to be 0
    End
```

Algorithm 3.6 Zigzag extraction

Description: Detect zigzag in the original character image

Input: Character image without thinning, start and end row, start and end column

Output: Number of zigzag

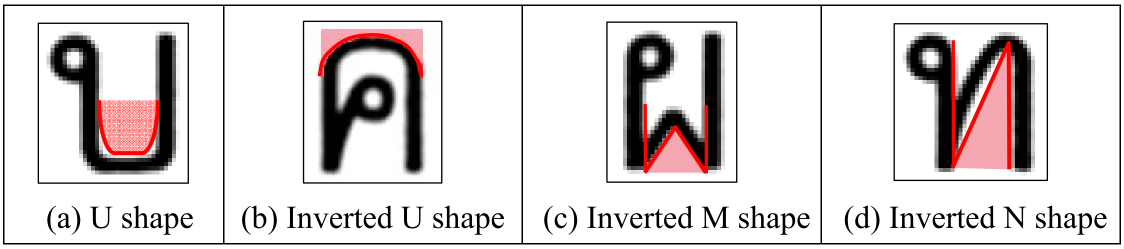
```
Scan all input pixels from start column to end column
Begin
    Scan all input pixels from start row to end row
    Begin
        If input pixel is black pixel then
            Set vertical profile to be the position of current row - start row
            Exit scanning this row
    End
End
Scan all vertical profile
```

```
Begin
    If vertical profile change in descending order and renew in ascending order
        Set number of zigzag to be 1
    Else
        Set number of zigzag to be 0
End
```

Remark: White represents 0 and black represents 1.

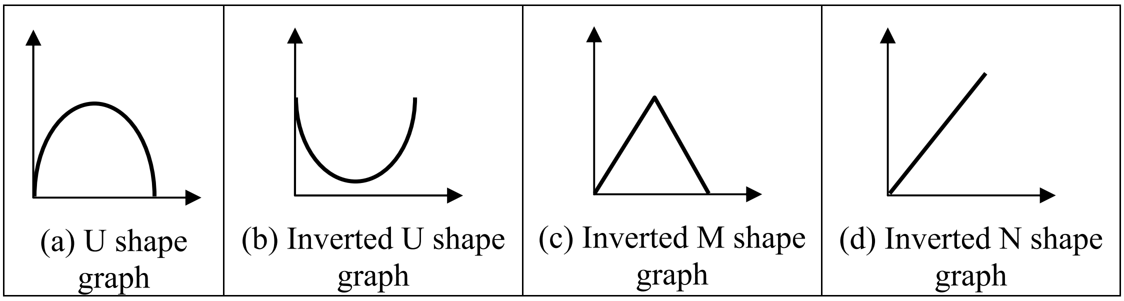
4.3.1.4 Shapes Extraction

In this step, loop(s) would be extracted from chain code of the thinned character image. In addition, vertical projection profile of the character image is used to extract U, inverted M and inverted N shapes as shown in Figure 4.8 - 4.9. The results of this step contain number of loop, position(s) of loop(s), type(s) of loop(s), number of U, inverted M and inverted N shape, position(s) of U, inverted M and inverted N shape(s), and type of U shape. Details are shown in Algorithms 3.7 - 3.11.



(a) U shape (b) Inverted U shape (c) Inverted M shape (d) Inverted N shape

Figure 4.8: Shapes and shape extraction areas



(a) U shape graph (b) Inverted U shape graph (c) Inverted M shape graph (d) Inverted N shape graph

Figure 4.9: Vertical profile graphs of shapes

Algorithm 3.7 Loop extraction

```
Description: Find loop(s) of the thinned character image from its chain code
Input: Chain code of the thinned character image array, start and end positions
Output: Number of loop, position(s) of loop(s)
Set initial number of loop to be 0
Scan all chain codes
Begin
    If start position of chain code = end position of chain code then
        Increment number of loop
        Find the position of this loop
End
```

Algorithm 3.8 Loop type extraction

```
Description: Detect clockwise or counter clockwise direction
Input: Number of loop, position(s) of loop(s), intersection points
Output: Type(s) of loop(s)
Scan all loops
Begin
    Find the position of central loop
    Scan all intersection points
    Begin
        If there is the minimum distance between intersection point and loop then
            Set connection position to be intersection position
    End
    If the position of central loop > connection position then
        Set loop type to be clockwise loop
    Else If the position of central loop < connection position then
        Set loop type to be counterclockwise loop
    Else
        Set loop type to be unknown
End
```

Algorithm 3.9 U shape extraction

```
Description:  Find U shape in the character image and define its type (U shape
              or inverted U shape)
Input: Character image array, start and end row, start and end column
Output: Number of U shape, type of U shape
// Find Inverted U shape
Set initial number of U shape to be 0
Scan all input pixels in one-third top area of character image array
Begin
    Find top vertical profile
    If top vertical profile shape like U shape
        Increment number of U shape
        Set U shape type to be 1
End
// Find U shape
Scan all input pixels in one-third bottom area of the input image
Begin
    Find inner bottom vertical profile
    If inner bottom vertical profile shape like inverted U shape
        Increment number of U shape
        Set U shape type to be 2
End
```

Algorithm 3.10 Inverted M shape extraction

```
Description: Find inverted M shape in the character image
Input: Character image array, start and end row, start and end column
Output: Number of inverted M shape
Scan all input pixels from left to right and bottom to top
Begin
    Find bottom vertical profile
    If bottom vertical profile shape like inverted M shape
        Set number of M shape to be 1
End
```

Algorithm 3.11 Inverted N shape extraction

```
Description: Find inverted N shape in the character image
Input: Character image array, start and end row, start and end column
Output: Number of Inverted N shape
Scan all input pixels from left to right and bottom to top
Begin
    Find bottom vertical profile
    If bottom vertical profile shape like inverted N shape
        Set number of N shape to be 1
End
```

4.3.2 Classification

In this process, two steps classification: coarse and fine classification are used to recognize the unknown character image based on its features derived from the feature extraction step as follows:

4.3.2.1 Coarse Classification

This step is used to classify the unknown character image into a character cluster by using set of decision rules. In the BOCR-WP, there are two sets of decision rules defined for Thai and English coarse classification as shown in Section 4.3.2.1.1 and 4.3.2.1.2 respectively.

4.3.2.1.1 Thai Coarse Classification

In Thai coarse classification, we define a set of decision rules to classify 68 Thai characters into 18 groups according to global features as shown in Table 4.1.

Table 4.1: The clusters of Thai characters

Group	Global Features	Characters
1	Hen's beak	ก, ฎ, ฏ, ถ, ภ, ฤ, ฦ
2	Narrow U shape	ข, ฃ, ช, ซ
3	Inverted U shape	ค, ฅ, ด, ต, ศ
4	Second loop on bottom-left zone	ฆ, ม
5	Second loop on bottom-right zone	ฉ, น
6	Roof	ล, ส
7	Broad U shape and roof	อ, ฮ
8	Broad character	ฌ, ญ, ณ, ฒ
9	Inverted N shape	ฑ, ท, ห
10	Broad U shape	บ, ป, ษ
11	Inverted M shape	ผ, ฝ, พ, ฟ, ฬ
12	Other consonants	ง, จ, ฐ, ธ, ย, ร, ว
13	Tonal mark level or upper vowel level without broad loop	-่, -้, -๊, -๋, -ํ, -ั, -็, -์
14	Upper vowel level and broad loop	-ิ, -ี, -ึ, -ื
15	Lower vowel level	-ุ , -ู
16	Right vertical line	-า, ๆ, ฯ
17	Left vertical line and bottom loop	เ-, โ-, ไ-, ใ-
18	Two loops on left zone	-ะ

4.3.2.1.2 English Coarse Classification

In English coarse classification, we define a set of decision rules to classify 62 English characters into 13 groups according to global features as shown in Table 4.2.

Table 4.2: The clusters of English characters

Group	Global Features	Characters
1	Having one top and one bottom loops and no leg	B, 8
2	Having one large loop and no leg	D, O, Q, o, 0
3	Having one top loop and two bottom legs	A, R
4	Having one top loop and one bottom leg	P, g, p, q, 4 9
5	Having one bottom loop and one top leg	b, d, 6
6	Having one loop and no leg	a, e
7	Having one three-intersection point	T, Y, y
8	Having two corners (┌) and no loop	E, F
9	Having two bottom legs and no loop	M, h, m, n
10	Having two top legs and no loop	U, V, W, u, v, w
11	Having one four-intersection point	H, K, X, k, x
12	Narrow characters	I, f, i, j, l, t, 1
13	Miscellaneous	C, G, J, L, S, Z, c, r, s, z, 2, 3, 5, 7

4.3.2.2 Fine Classification

This step is used to decide what character is by using set of decision rules of its character cluster according to its local features. If one of all decision rules in the cluster is matched, recognition result derived from this step would be one of all characters in the cluster. Otherwise, recognition result would be unknown. In the BOCR-WP, there are a numbers of sets of decision rules defined for Thai and English fine classification as shown in Section 4.3.2.2.1 and 4.3.2.2.2 respectively.

4.3.2.2.1 Thai Fine Classification

In Thai fine classification, we define 18 sets of decision rules to classify Thai characters within 18 groups according to local features of the character cluster [13].

4.3.2.2.2 English Fine Classification

In Thai fine classification, we define 13 sets of decision rules to classify Thai characters within 13 groups according to local features of the character cluster.

4.4 Word Prediction

Based on the concept of word prediction by using N-gram mentioned in Section 3.3, character based N-gram is applied to predict next probable words according to the previous word token. In implementation, there are two processes of word prediction as follows: N-gram construction and N-gram look-up as shown in Sections 4.4.1 and 4.4.2.

4.4.1 N-gram Construction

In training phase, this step is used to build lexicon N-grams of all n prefix word token occurred in dictionary for 2-gram to 5-gram. Firstly, vocabularies in dictionary are sorted in probabilistic ordering based on their frequencies of occurrences in general documents. Next, these N-grams are built from the list of lexicon words in probabilistic order, followed by storing all N-grams information in database.

In the BOCR-WP system, the implementation of N-gram construction is presented in activity diagram in Figure 4.10. Then, we show details of constructing Thai and English lexicon N-grams in Sections 4.4.1.1 and 4.4.1.2 respectively.

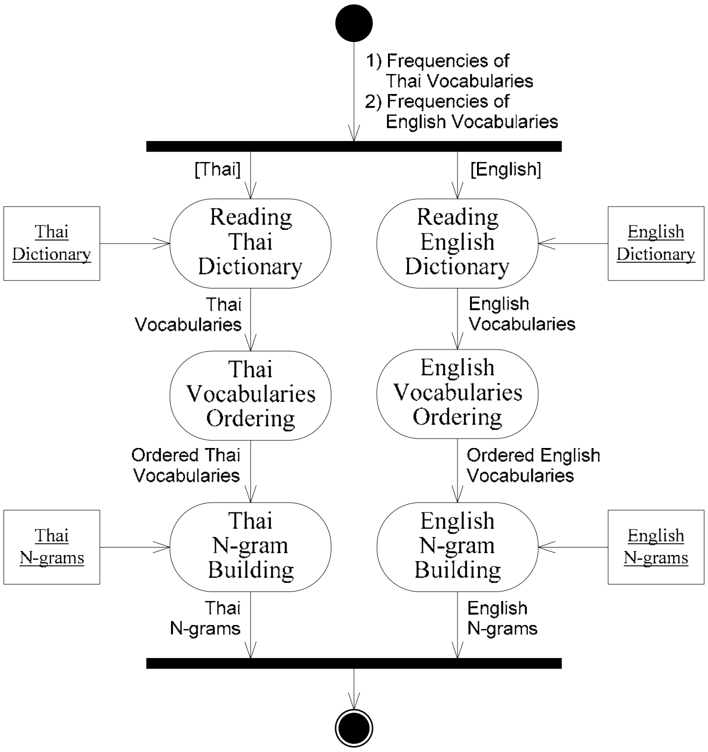


Figure 4.10 (a): Activity diagram of Thai and English N-gram construction

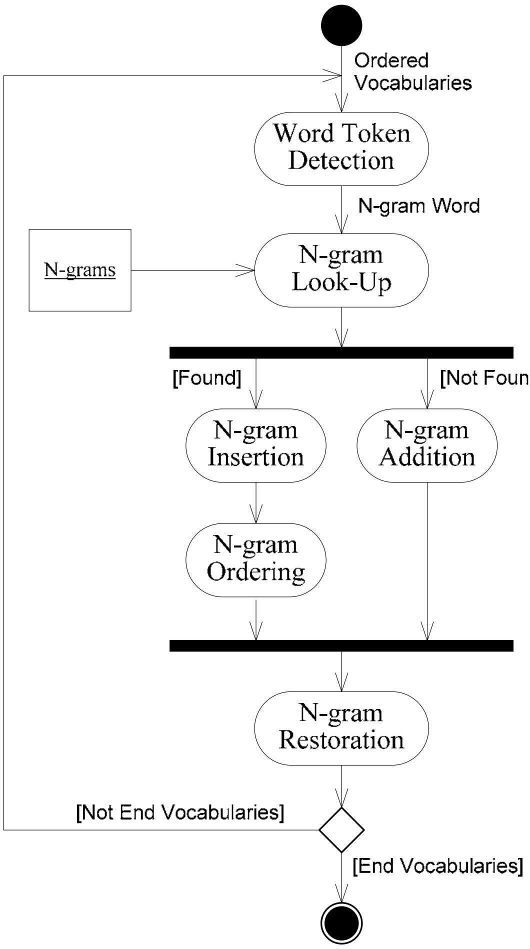


Figure 4.10 (b): Activity diagram of N-gram building

Figure 4.10 (a)-(b): Activity diagrams of N-gram construction

4.4.1.1 Thai N-gram Construction

This step is used to build Thai N-gram of 36,853 Thai vocabularies. In Thai dictionary [19], we first sorted these Thai words in probabilistic order according to their frequencies of occurrences in general documents derived from term frequencies of Thai words occurred in a large document collection.

Algorithm 4.1 Thai N-gram construction

```
    Description:   Build Thai lexicon N-grams (2 to 5 gram)
    Input: Thai dictionary vocabularies, frequencies of Thai vocabularies
    Output: Thai N-grams array and N-grams indexes for 2-gram to 5-gram
    Read Thai vocabularies from dictionary file
    Read frequencies of Thai vocabularies
    // Sort all Thai dictionary vocabularies in probabilistic order
    Sort Thai vocabularies by using quick sort
    For N = 2 to 5 do           // Start N-gram construction
    Begin
        Scan all Thai vocabularies
        Begin
            If N prefix word token of this vocabulary is not NULL then
                Search this token in N-gram array
                If this token is not exist in N-gram array then
                    Insert this token into N-gram array
                    Insert index of this vocabulary to N-gram index
                    Sort this N-gram array in alphabetical order
                Else
                    Add index of this vocabulary to N-gram index that the token
                    existed
            End
        End                     // End N-gram construction
End
```

4.4.1.2 English N-gram Construction

This step is used to build English N-gram of 40,479 English vocabularies. In English lexicon [20], we first sorted these English words in probabilistic order according to their frequencies of occurrences in general documents derived from term frequencies of English words proposed in [20].

Algorithm 4.2 English N-gram construction

```
Description:   Build English lexicon N-grams (2 to 5 gram)
Input: English dictionary vocabularies, frequencies of English vocabularies
Output: English N-grams array and N-grams indexes for 2-gram to 5-gram
Read English vocabularies from dictionary file
Read frequencies of English vocabularies
// Sort all English dictionary vocabularies in probabilistic order
Sort English vocabularies by using quick sort
For N = 2 to 5 do             // Start N-gram construction
Begin
    Scan all English words
    Begin
        If N prefix word token of this vocabulary is not NULL then
            Search this token in N-gram array
            If this token is not exist in N-gram array then
                Insert this token into N-gram array
                Insert index of this vocabulary to N-gram index
                Sort this N-gram array in alphabetical order
            Else
                Add index of this vocabulary to N-gram index that the token
                existed
        End
    End                       // End N-gram construction
End
```

4.4.2 N-gram Look-up

In testing phase, this step is used to predict probable words of a block image according to previous N word token by searching in N-gram. Firstly, we retrieve all N-grams information from database. Next, N-gram look up is used to search a word token for finding those predictive words. In this step, binary search technique is combined for effective searching in the vocabularies listed in alphabetical ordering.

In the BOCR-WP system, the implementation of Thai and English N-gram look-up is shown in Figure 4.11. Details are presented in Section 4.4.2.1 and 4.4.2.2.

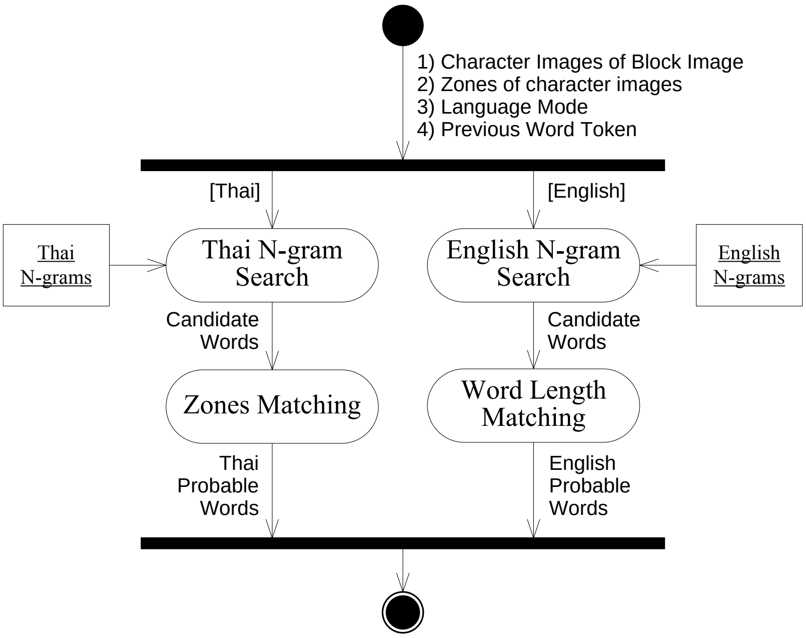


Figure 4.11: Activity diagram of N-gram look-up

4.4.2.1 Thai N-gram Look-up

According to activity diagram of N-gram look-up mentioned above, in this step, we explained how to search the previous word token in Thai N-gram in Algorithm 4.3.

Algorithm 4.3 Thai N-gram look-up

Description: Search the previous word token in Thai N-gram and retrieve the words pointed by this N-gram to be probable words

Input: Thai N-gram arrays and N-gram indexes from 2-gram to 5-gram, the previous N word token, language mode, zones and number of character images in the block image

Output: Set of probable words

```
Set N to be length of the previous word token
// Search in Thai N-gram
If language mode is Thai then
    Search this word token in Thai N-gram
    If this word token is existed in N-gram array then
        Retrieve candidates of probable words
        Scan all candidates of probable words
        Begin
            If zones of the character images = zones of this candidate word then
                Insert this candidate word into set of probable words
        End
```

4.4.2.2 English N-gram Look-up

According to activity diagram of N-gram look-up mentioned in Figure 4.11, in this step, we explained how to search the previous word token in English N-gram in Algorithm 4.4.

Algorithm 4.4 English N-gram look-up

```
Description: Search the previous word token in English N-gram and retrieve
             the words pointed by this N-gram to be probable words
Input: English N-gram arrays and N-gram indexes from 2-gram to 5-gram, the
       previous N word token, language mode, zones and number of character
       images in the block image
Output: Set of probable words
Set N to be length of the previous word token
// Search in English N-gram
If language mode is English then
    Search this word token in Thai N-gram
    If this word token is existed in N-gram array then
        Retrieve candidates of probable words
        Scan all candidates of probable words
        Begin
            If number of the character images = length of this candidate word then
                Insert this candidate word into set of probable words
        End
```

4.5 Word Verification

According to the concept of word verification mentioned in Section 3.4, template matching technique is implemented to match the unknown character image and the template of candidate characters of probable words derived from the word prediction process. This step is worked in iteration on matching character images contained in the block image and character templates of the probable words one after another until a matched word are found or end of predictive words. In this step, we show the implementation of word verification in Figure 4.12.

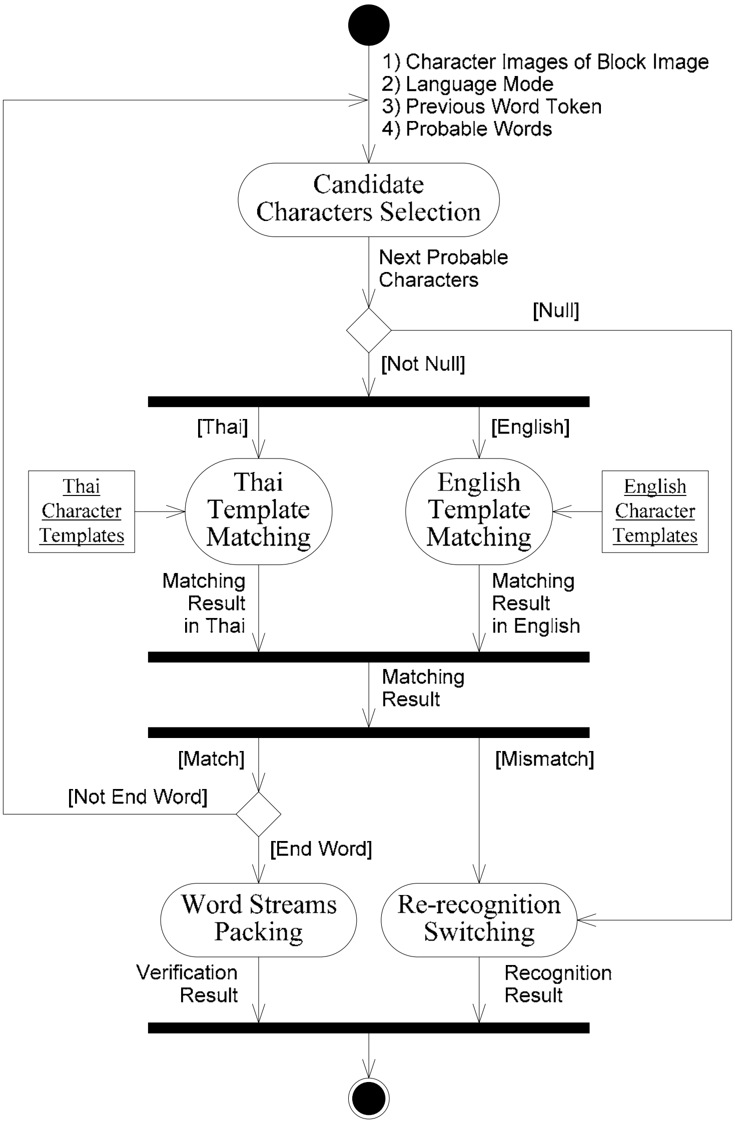


Figure 4.12: Activity diagram of word verification

Then, we present step by step of two main processes in word verification: candidate characters selection and template matching in Sections 4.5.1 and 4.5.2. Details are as follows:

4.5.1 Candidate Characters Selection

This process selects candidates of the next unknown character image from predictive words derived from the word prediction. In this step, set of next probable characters is retrieved prior matching in the next step as shown in Algorithm 5.1.

Algorithm 5.1 Candidate character selection

```
Description: Retrieve next probable characters of probable words at the same
             unknown character position
Input: Set of Probable words, position of the character image
Output: Candidates of next probable character
Set current position to be the position of unknown character image
Scan all probable words
Begin
    If position of the character image <= length of this probable word then
        Set candidate character to be a character at current position of this
        probable word
    Else
        Remove this probable word from probable word set.
End
```

4.5.2 Template Matching

According to the concept of template matching mentioned in Section 3.4.1, this step will match the unknown character image and templates of next probable characters in iteration by using 3 sets of character features: 8-direction vector, black pixel percentages and loop information. In the BOCR-WP system, there are 2 procedures: Thai template matching and English template matching as presented in Sections 4.5.2.1 and 4.5.2.2.

4.5.2.1 Thai Template Matching

This step is used to match Thai character image with template of next probable characters by using three template matching approaches mentioned in Section 3.4.1. In addition, Thai template matching would match with additional character features, for example, zigzag, tail, leg, and so on, by sets of decision rules. The implementation of Thai template matching is shown in Figure 4.13.

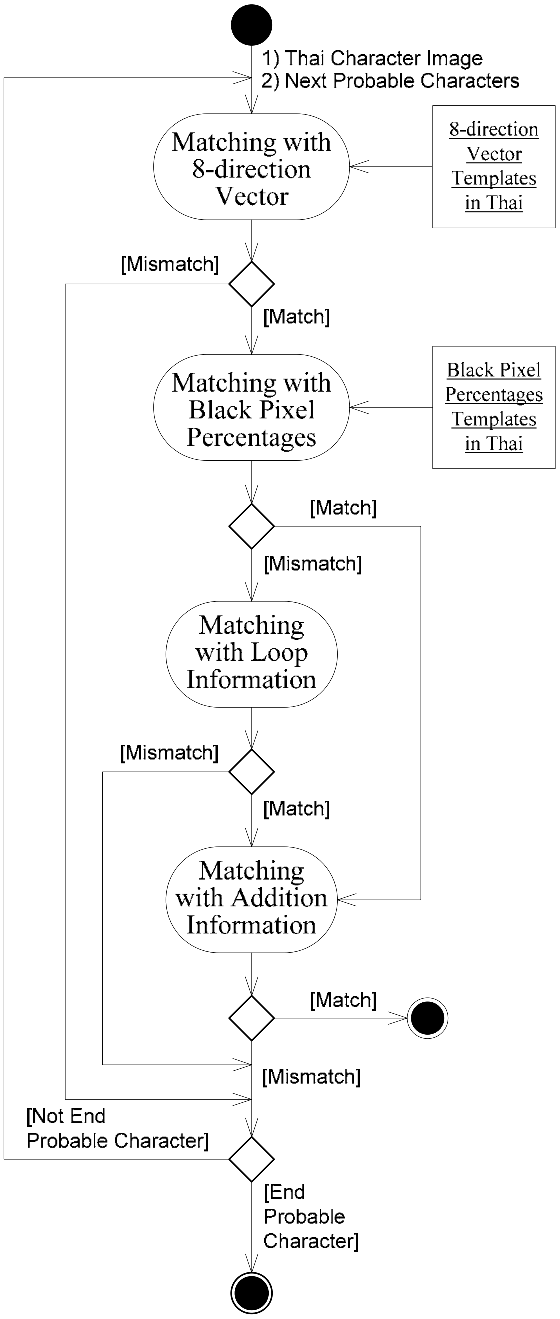

1) Thai Character Image
2) Next Probable Characters
Matching with 8-direction Vector
8-direction Vector Templates in Thai
[Mismatch]
[Match]
Matching with Black Pixel Percentages
Black Pixel Percentages Templates in Thai
[Match]
[Mismatch]
Matching with Loop Information
[Mismatch]
[Match]
Matching with Addition Information
[Match]
[Mismatch]
[Not End Probable Character]
[End Probable Character]

Figure 4.13: Activity diagram of Thai template matching

4.5.2.2 English Template Matching

This step is used to match English character image with template of next probable characters. The implementation of English template matching is shown in Figure 4.14.

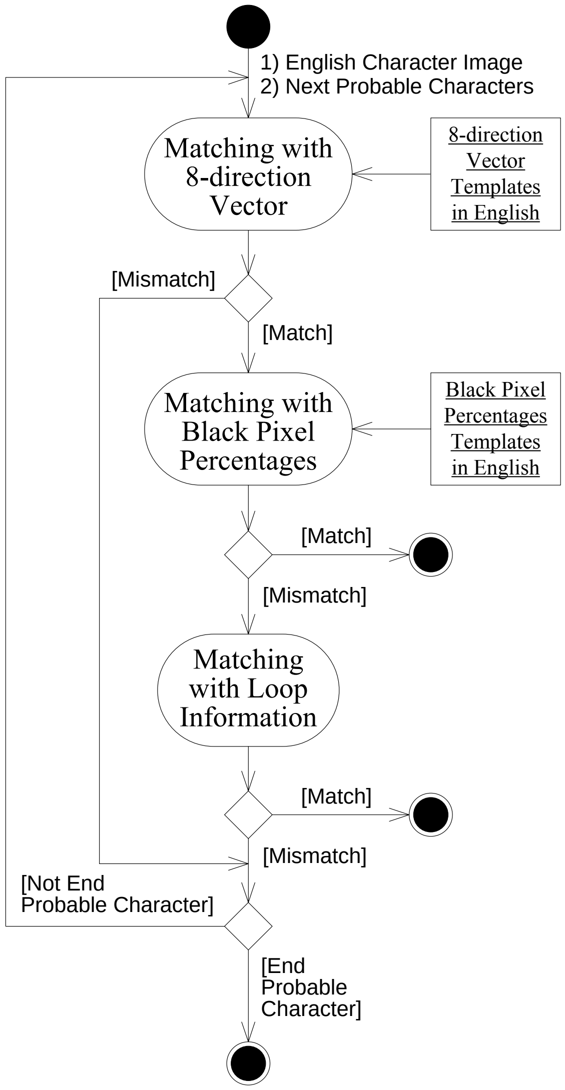

1) English Character Image
2) Next Probable Characters
Matching with 8-direction Vector
8-direction Vector Templates in English
[Mismatch]
[Match]
Matching with Black Pixel Percentages
Black Pixel Percentages Templates in English
[Match]
[Mismatch]
Matching with Loop Information
[Match]
[Mismatch]
[Not End Probable Character]
[End Probable Character]

Figure 4.14: Activity diagram of English template matching

Details of template matching are shown in Algorithms 5.2 - 5.6.

Algorithm 5.2 Eight-direction vector extraction

Description: Calculate 16 features vector of the original character image

Input: Character image array without thinning

Output: 16 features vector of inward and outward central 8-direction vector

```
// Inward central 8-direction vector
Scan all 8 directions from border of the character image to its central pixel
Begin
    If the current input pixel is black pixel then
        Calculate the distance between this pixel and the central pixel
        Set the vector value to be the distance value
    Else If no black pixel in this direction then
        Set the vector value to be infinity
End
// Outward central 8-direction vector
Scan all 8 directions from central pixel of the character image to its border
Begin
    If the input pixel is black pixel then
        Calculate the distance between this pixel and the central pixel
        Set the vector value to be the distance value
    Else If no black pixel in this direction then
        Set the vector value to be infinity
End
```

Remark: White represents 0 and black represents 1.

Algorithm 5.3 Black pixel percentages extraction

Description: Calculate black pixel percentages of the original character image

Input: Character image array without thinning

Output: black pixel percentages array in 9 segments

```
Scan all segments of the character image array
Begin
    Set initialize number of pixel to be zero
```

```
        Set initialize number of black pixel to be zero
        Scan all pixels of the character image array in this segment
        Begin
            Increment number of pixel
            If this pixel is black pixel then
                Increment number of black pixel
        End
        Set black pixel percentage of this segment = (number of black pixel /
        number of pixel) x 100
    End
```

Remark: White represents 0 and black represents 1.

Algorithm 5.4 Loops extraction without thinning

Description: Find loop(s) of the original character image and define the position(s) of the loop(s)

Input: Character image array without thinning

Output: number of loop, position(s) of loop(s)

```
Set initial number of loop to be 0
Scan all pixels in the character image array
Begin
    // Edge Detection by using zero crossing
    Detect edge of this character image array
    Remove outer edge by setting edge pixels to be white pixels
    // Loop detection
    Detect loop(s) from inner edges of this character image array
    If an inner edges is a loop then
        Increment number of loop
        Find the position of this loop
End
```

Remark: White represents 0 and black represents 1.

Algorithm 5.5 Euclidean distance measure

```
Description: Calculate Euclidean distance between feature values of the
             character image and the template of next probable character
Input: Feature values of the character image array and the template of next
       probable character
Output: Euclidean distance
Set distance_sum to be zero
Scan all of feature values of the character image array
Begin
    Calculate distance of this feature value of image and template
    Calculate sum square of this distance
    Increment distance_sum with sum square of this distance
End
Set Eu_distance to be square root of distance_sum
Return Eu_distance
```

Algorithm 5.6 Template matching

```
Description: Match features of the unknown character image and templates of
             next probable characters
Input: 16 features vector of inward and outward central 8-direction vector,
       width-height ratio, zone, overlapped zone, character image array
       without thinning, next probable characters, templates of next
       probable characters
Output: Match or mismatch
Scan all next probable characters
Begin
    // (1) Matching with 8-direction vector
    Match zone, overlapped zone and width-height ratio of the character
    image array and this probable character
    If match result is true then
        // Find 16 features vector of the character image
        Call eight-direction vector extraction               // Algorithm 4.4
        // Calculate Euclidean distance of 16 features vector
```

```
            Call Euclidean distance measure with 16 features values of the
            image and the template
            If Eu_distance <= threshold of this template then      // Match (1)
                // (2) Matching with black pixel percentages
                Call black pixel percentages extraction            // Algorithm 4.5
                // Calculate Euclidean distance of black pixel percentages
                Call Euclidean distance measure with black pixel percentages of
                the image and the template
                If Eu_distance <= threshold of this template then  // Match (2)
                    // (3) Matching with loop information
                    Call loops extraction without thinning         // Algorithm 4.6
                    // Calculate Euclidean distance of black pixel percentages
                    Call Euclidean distance measure with black pixel percentages
                    of the image and the template
                    If Eu_distance <= threshold of this template then // Match (3)
                        Set match result to be true
                Else                                               // Mismatch (2)
                    Set match result to be false
            Else                                                   // Mismatch (1)
                Set match result to be false
        Else
            Set match result to be false
End
```

4.6 Dictionary Look-up

According to the concept of dictionary look-up mentioned in Chapter III, there are two procedures in this regard: Thai dictionary look-up and English dictionary look-up implemented in Sections 4.6.1 and 4.6.2 respectively. Details are as follows:

4.6.1 Thai Dictionary Look-up

Due to a number of recognition errors in Thai documents occurred in character recognition and word verification, dictionary look-up is used to solve incorrect recognition results. According to the fact that no end word boundary between Thai words of a phrase or sentence, in dictionary look-up process, we apply a word segmentation tool named SWATH [24] to match Thai words in a block image with longest matching and maximal matching algorithms. Details are shown in activity diagram in Figure 4.15.

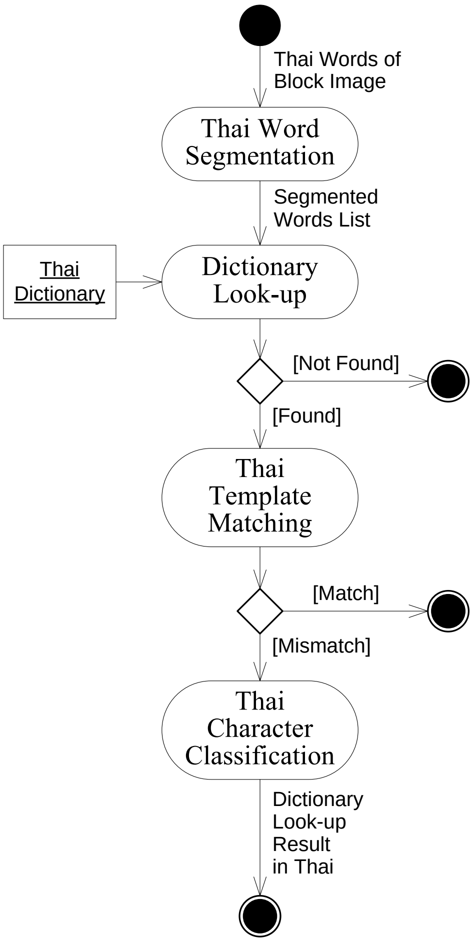


Figure 4.15: Activity diagram of Thai dictionary look-up

4.6.2 English Dictionary Look-up

Due to a number of recognition errors in English documents occurred in character recognition and word verification, dictionary look-up is used to solve two types of recognition problems: unknown results of English connected characters and incorrect results of English similar characters. Details are shown in activity diagram in Figure 4.16.

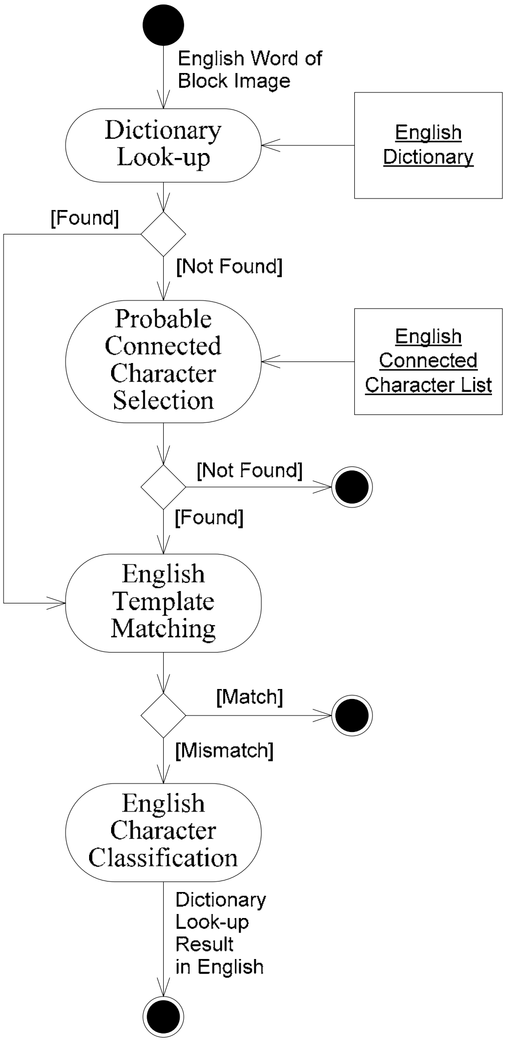

English Word of Block Image
Dictionary Look-up
English Dictionary
[Found]
[Not Found]
Probable Connected Character Selection
English Connected Character List
[Not Found]
[Found]
English Template Matching
[Match]
[Mismatch]
English Character Classification
Dictionary Look-up Result in English

Figure 4.16: Activity diagram of English dictionary look-up

In Figure 4.16, English connected character list derived from training phase contains 20 connected characters: ‘ca’, ‘rm’, ‘ff’, ‘ft’, ‘fy’, ‘rf’, ‘rt’, ‘rv’, ‘rw’, ‘ry’, ‘tf’, ‘tt’, ‘tw’, ‘ty’, ‘vy’, ‘wt’, ‘yw’, ‘yt’, ‘rty’, and ‘ryt’. In probable connected character selection, we define a set of decision rules to classify 20 connected characters into 6 groups according to structural features as shown in Figure 4.17.

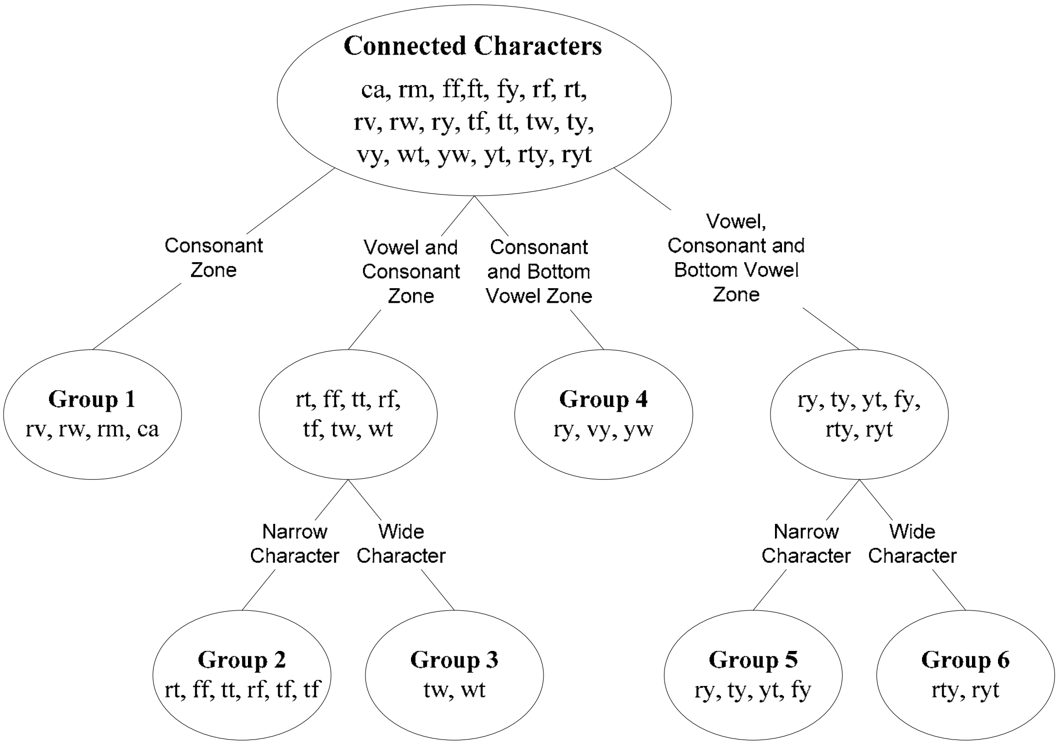


Figure 4.17: Probable connected characters

CHAPTER V
SYSTEM IMPLEMENTATION

In this chapter, we describe the system environment of BOCR-WP in terms of hardware and software specification in Section 5.1. In addition, module structure of the system implementation in Delphi source codes is presented in Section 5.2. Details are as follows:

5.1 System Environment

The computer used for the BOCR-WP development is described in the following.

1. Hardware

1.1 Laptop hardware: Intel(R) Core 2 Duo CPU 2.2 GHz, Memory 2 GB, HDD 250 GB

1.2 Scanner: Epson All in One Scanner CX3700 resolution 600 x 1200 dpi

2. Software

2.1 Operating System: Microsoft Windows Vista Home Premium service pack 2

2.2 Programming Language: Borland Delphi 7.0

5.2 Module Structure

To develop the BOCR-WP system, there are 6 main modules: preprocessing, language identification, character recognition, word prediction, word verification, and dictionary look-up. The module structure of Delphi source codes is shown in Figure 5.1.

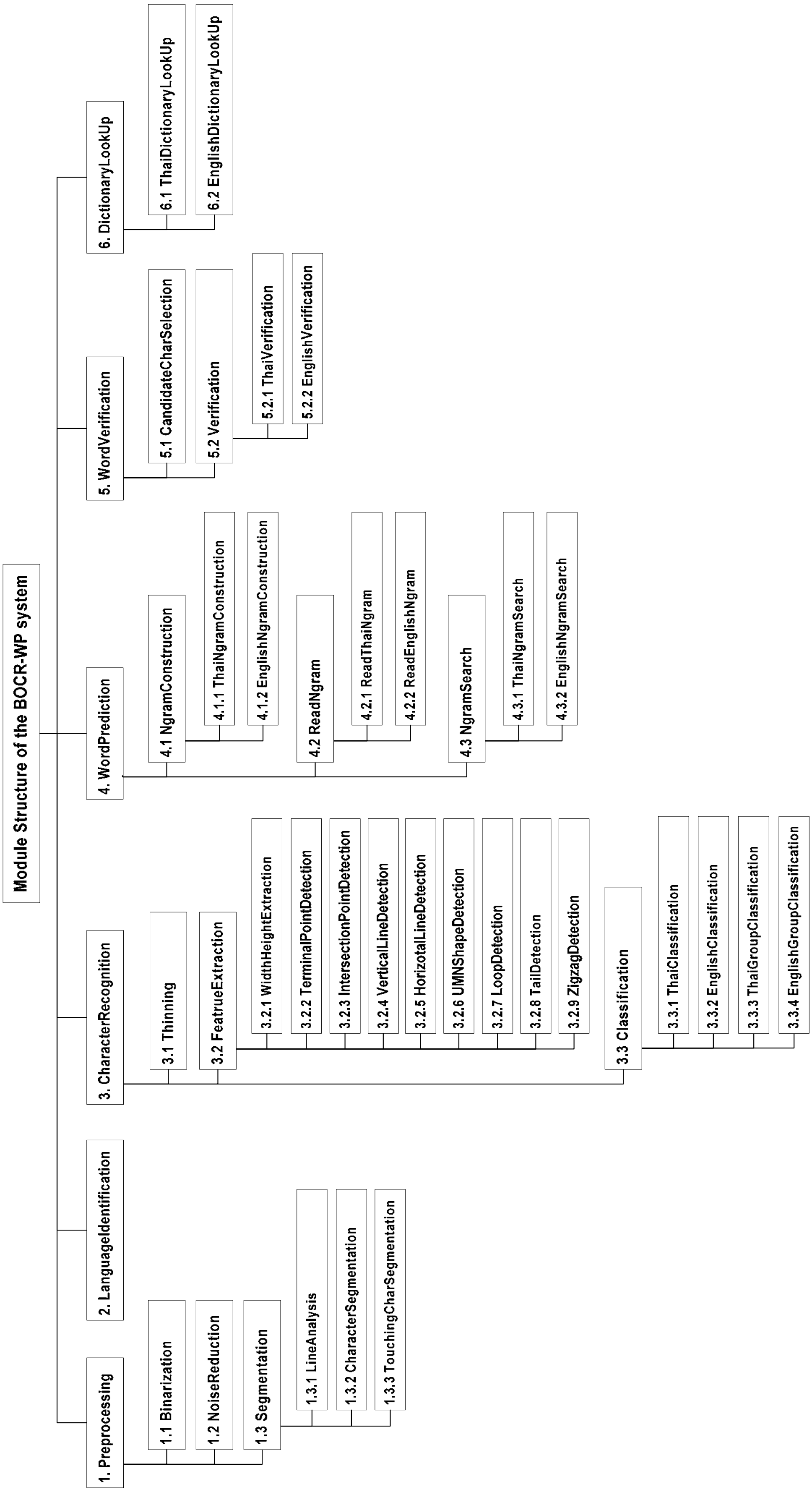


Figure 5.1: Module Structure of the BOCR-WP system

Details are as follows:

1. Preprocessing is a procedure to increase the quality of input document image derived from scanning process, meanwhile to reduce complex computation in the subsequent procedures. In pre-processing step, there are three procedures as follows:

1.1 Binarization is a procedure to convert a gray scale document image into a binary document image.

1.2 NoiseReduction is a procedure to remove isolated point noises and isolated hole noises appeared in the binary document image.

1.3 Segmentation is a procedure to prepare the sequential character images of image strips for processing in the next step. In this process, there are four procedures as follows:

1.3.1 LineAnalysis is a procedure to detect three reference lines: top line, central line and bottom line, and four zones: top vowel zone, vowel zone, consonant zone and bottom vowel zone.

1.3.2 CharacterSegmentation is a procedure to separate the document image in to strip images and each strip image is separated into a sequence of character images.

1.3.3 TouchingCharSegmentationProc is a procedure to segment touching character image into single character images.

2. LanguageIdentification is a procedure to identify the language modes of a block image contained the sequence of character images of word or phrase.

3. CharacterRecognition is a procedure to recognize the character images in iterations. In recognition step, there are there procedures as follows:

3.1 Thinning is a procedure to thin a character image into a thinned character image which remains its skeleton.

3.2 FeatureExtraction is a procedure to extract features of the character from its thinned character image.

3.2.1 WidthHeightExtraction is a procedure to calculate width-height ratio, average width and average height of the character image and the thinned character image.

3.2.2 TerminalPointDetection is a procedure to find terminal point(s) and terminal position(s) in the thinned character image.

3.2.3 IntersectionPointDetection is a procedure to find intersection point(s) and intersection position(s) in the thinned character image.

3.2.4 VerticalLineDetection is a procedure to find vertical line(s) and vertical line position(s) in the thinned character image.

3.2.5 HorizontalLineDetection is a procedure to find horizontal line(s) and horizontal line position(s) in the thinned character image.

3.2.6 UMNShapeDetection is a procedure to find U shape, inverted M shape and inverted N shape in the thinned character image.

3.2.7 LoopDetection is a procedure to find loop and loop position(s) in the thinned character image.

3.2.8 TailDetection is a procedure to find tail or leg in the thinned character image. This step is used in fine classification of some character groups.

3.2.9 ZigzagDetection is a procedure to find zigzag in the character image without thinning. This step is used in fine classification of some character groups.

3.3 Classification is a procedure to classify the character in the thinned character image into a character group and identify what the character is by using sets of decision rules. This procedure is used for both Thai and English by taking two steps classification as follows:

3.3.1 ThaiClassification is a procedure to classify Thai character image into a character group in which features of character image matched one of decision rules.

3.3.2 EnglishClassification is a procedure to classify English character image into a character group in which features of character image matched one of decision rules.

3.3.3 ThaiGroupClassification is a procedure to identify Thai character in the character image by matching with decision rules of its group. In this step, there are 18 group classifications for Thai.

3.3.4 EnglishGroupClassification is a procedure to identify English character in the character image by matching with decision rules of its group. In this step, there are 13 group classifications for English.

4. WordPrediction is a procedure to predict next probable words according to the previous word token. In this process, there are three procedures as follows:

4.1 NgramConstruction is a procedure to build lexicon N-grams and to store these N-grams in the database. This step is done once in the training phase. This process is used for both Thai and English as follows:

4.1.1 ThaiNgramConstruction is a procedure to build Thai lexicon N-grams and to store these N-grams in the database.

4.1.2 EnglishNgramConstruction is a procedure to build English lexicon N-grams and to store these N-grams in the database.

4.2 ReadNgram is a procedure to retrieve N-gram information from the database. This step is implemented in the beginning of testing phase. This process is used for both Thai and English as follows:

4.2.1 ReadThaiNgram is a procedure to retrieve Thai N-gram information from the database.

4.2.2 ReadEnglishNgram is a procedure to retrieve English N-gram information from the database.

4.3 NgramSearch is a procedure to predict next probable words according to the previous word token. This process is used for both Thai and English as follows:

4.3.1 ThaiNgramSearch is a procedure to predict next probable words of Thai character images according to the previous word token.

4.3.2 EnglishNgramSearch is a procedure to predict next probable words of English character images according to the previous word token.

5. WordVerification is a procedure to verify the character images of the block image by matching with the next probable words derived from NgramSearch. In the BOCR-WP system, there are two word verification procedures as follows:

5.1 CandidateCharSelection is a procedure to retrieve candidates of next probable characters from the predictive words at the same position as the current unknown character image.

5.2 Verification is a procedure to verify character images contained in a block image one after another by matching with templates of candidate characters of predictive words until matched word(s) is found or end of the candidate characters list.

5.2.1 ThaiVerification is a procedure to verify Thai character images by matching with templates of candidate characters.

5.2.2 EnglishVerification is a procedure to verify English character images by matching with templates of candidate characters.

6. DictionaryLookUp is a procedure to solve recognition mistakes by matching a word result of the block image and dictionary vocabularies with maximum matching search.

6.1 ThaiDictionaryLookUp is a procedure to match Thai word(s) of the block image and Thai vocabularies.

6.2 EnglishDictionaryLookUp is a procedure to match English word of the block image and English vocabularies.

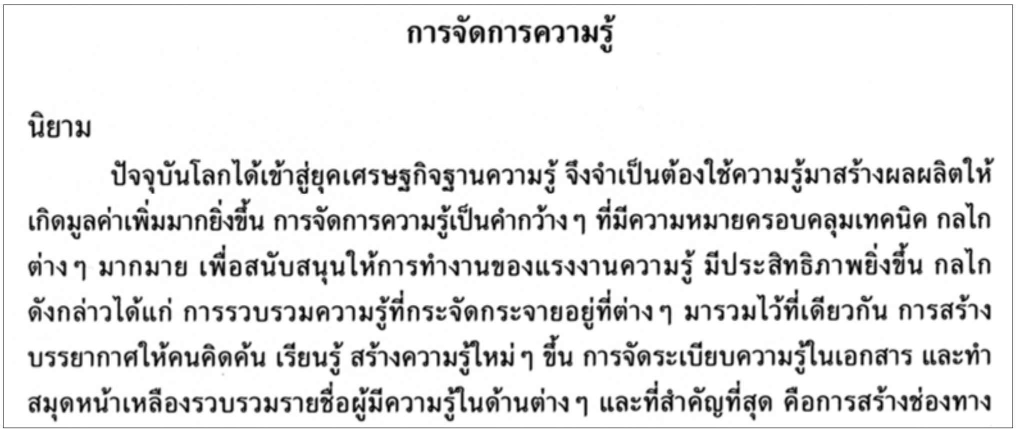
CHAPTER VI
EXPERIMENTAL RESULTS

According to the design and implementation of the BOCR-WP system as proposed in Chapters III and IV, this chapter will present a series of experimental results for evaluating the system performance of this work and also a comparison with the conventional approach. In this series of experiment, the testing data contained over 280,000 characters in 141 pages of Thai/English documents which are divided into 3 sets: Thai document set, English document set, and Thai/English mixed document set. Details are as follows:

1. Thai document set

This set consists of 53 pages of text with 117,347 characters, the fonts of EucrosiaUPC and FreesiaUPC, the sizes of 16 and 18 and the styles of normal and bold. In this group, there are 5 sets of Thai documents as follows:

Thai document set I contains 5 document images of 8,632 characters, the font of EucrosiaUPC, the sizes of 16 and 18 and the styles of normal and bold. A sample of scanned document images in EucrosiaUPC font is shown in Figure 6.1.



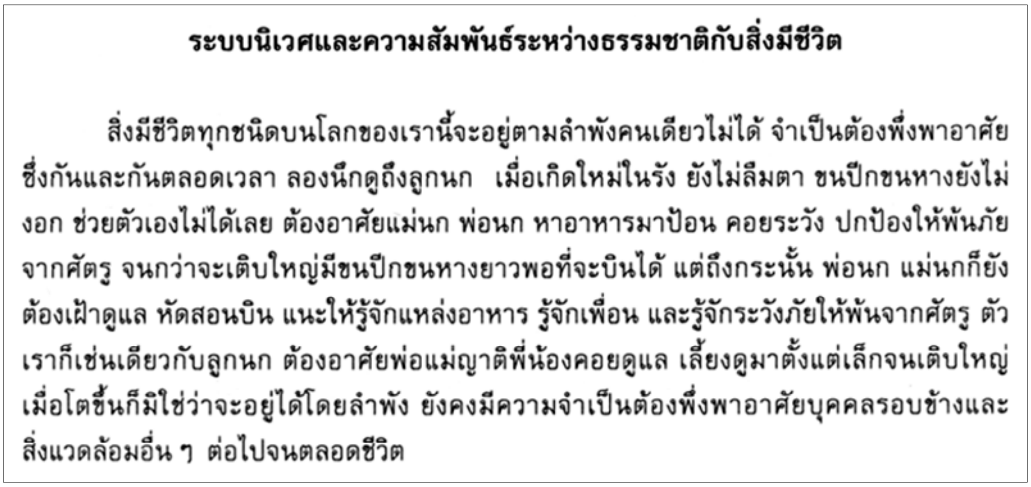
การจัดการความรู้

นิยาม

ปัจจุบันโลกได้เข้าสู่ยุคเศรษฐกิจฐานความรู้ จึงจำเป็นต้องใช้ความรู้มาสร้างผลผลิตให้เกิดมูลค่าเพิ่มมากยิ่งขึ้น การจัดการความรู้เป็นคำกว้าง ๆ ที่มีความหมายครอบคลุมเทคนิค กลไกต่าง ๆ มากมาย เพื่อสนับสนุนให้การทำงานของแรงงานความรู้ มีประสิทธิภาพยิ่งขึ้น กลไกดังกล่าวได้แก่ การรวบรวมความรู้ที่กระจัดกระจายอยู่ที่ต่าง ๆ มารวมไว้ที่เดียวกัน การสร้างบรรยากาศให้คนคิดค้น เรียนรู้ สร้างความรู้ใหม่ ๆ ขึ้น การจัดระเบียบความรู้ในเอกสาร และทำสมุดหน้าเหลืองรวบรวมรายชื่อผู้มีความรู้ในด้านต่าง ๆ และที่สำคัญที่สุด คือการสร้างช่องทาง

Figure 6.1: A sample of scanned document images in EucrosiaUPC font

Thai document set II contains 5 document images of 10,854 characters, the font of FreesiaUPC, the sizes of 16 and 18 and the styles of normal and bold. A sample of scanned document images in FreesiaUPC font is shown in Figure 6.2.

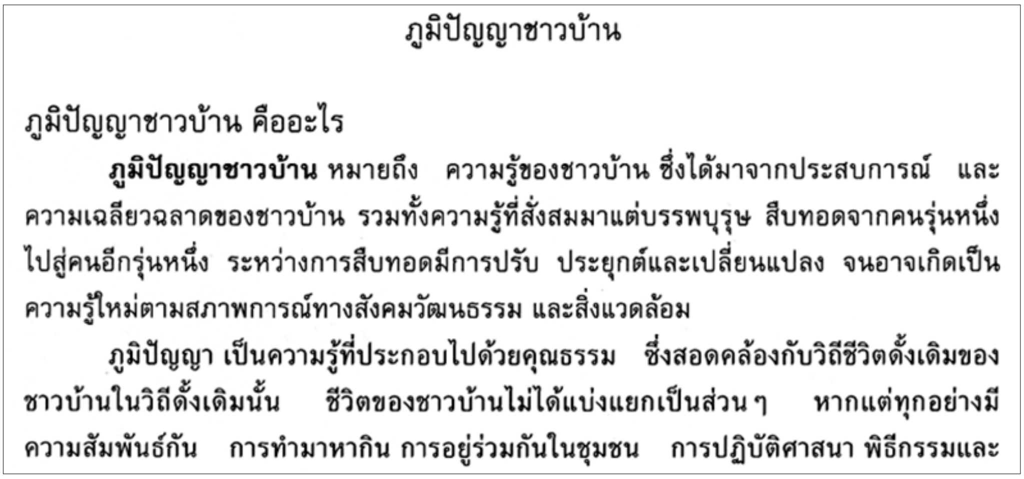


ระบบนิเวศและความสัมพันธ์ระหว่างธรรมชาติกับสิ่งมีชีวิต

สิ่งมีชีวิตทุกชนิดบนโลกของเรานี้จะอยู่ตามลำพังคนเดียวไม่ได้ จำเป็นต้องพึ่งพาอาศัยซึ่งกันและกันตลอดเวลา ลองนึกดูถึงลูกนก เมื่อเกิดใหม่ในรัง ยังไม่ลืมตา ขนปีกขนหางยังไม่งอก ช่วยตัวเองไม่ได้เลย ต้องอาศัยแม่นก พ่อนก หาอาหารมาป้อน คอยระวัง ปกป้องให้พ้นภัยจากศัตรู จนกว่าจะเติบใหญ่มีขนปีกขนหางยาวพอที่จะบินได้ แต่ถึงกระนั้น พ่อนก แม่นกก็ยังต้องเฝ้าดูแล หัดสอนบิน แนะให้รู้จักแหล่งอาหาร รู้จักเพื่อน และรู้จักระวังภัยให้พ้นจากศัตรู ตัวเราก็เช่นเดียวกับลูกนก ต้องอาศัยพ่อแม่ญาติพี่น้องคอยดูแล เลี้ยงดูมาตั้งแต่เล็กจนเติบใหญ่ เมื่อโตขึ้นก็มิใช่ว่าจะอยู่ได้โดยลำพัง ยังคงมีความจำเป็นต้องพึ่งพาอาศัยบุคคลรอบข้างและสิ่งแวดล้อมอื่น ๆ ต่อไปจนตลอดชีวิต

Figure 6.2: A sample of scanned document images in FreesiaUPC font

Thai document set III contains 3 document images of 6,349 characters, the fonts of EucrosiaUPC and FreesiaUPC, the sizes of 16 and 18 and the styles of normal and bold. A sample of scanned document images in Thai mixed fonts is shown in Figure 6.3.



ภูมิปัญญาชาวบ้าน

ภูมิปัญญาชาวบ้าน คืออะไร

ภูมิปัญญาชาวบ้าน หมายถึง ความรู้ของชาวบ้าน ซึ่งได้มาจากประสบการณ์ และความเฉลียวฉลาดของชาวบ้าน รวมทั้งความรู้ที่สั่งสมมาแต่บรรพบุรุษ สืบทอดจากคนรุ่นหนึ่งไปสู่คนอีกรุ่นหนึ่ง ระหว่างการสืบทอดมีการปรับ ประยุกต์และเปลี่ยนแปลง จนอาจเกิดเป็นความรู้ใหม่ตามสภาพการณ์ทางสังคมวัฒนธรรม และสิ่งแวดล้อม

ภูมิปัญญา เป็นความรู้ที่ประกอบไปด้วยคุณธรรม ซึ่งสอดคล้องกับวิถีชีวิตดั้งเดิมของชาวบ้านในวิถีดั้งเดิมนั้น ชีวิตของชาวบ้านไม่ได้แบ่งแยกเป็นส่วน ๆ หากแต่ทุกอย่างมีความสัมพันธ์กัน การทำมาหากิน การอยู่ร่วมกันในชุมชน การปฏิบัติศาสนา พิธีกรรมและ

Figure 6.3: A sample of scanned document images in Thai mixed fonts

In addition, the large document sets: Thai document set IV contains 20 document images of 43,465 characters and Thai document set V contains 20 document images of 48,047 characters, the fonts of EucrosiaUPC and FreesiaUPC, the sizes of 16 and 18 and the styles of normal and bold. The details of documents in Thai document set are outperformed in Table 6.1.

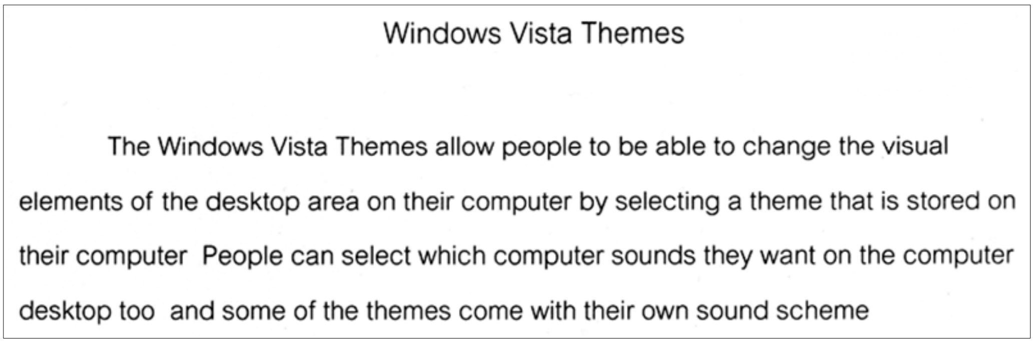
Table 6.1: The testing documents in Thai document set.

Document	Font	Character No.			Character No. (%)	
		Thai	English	All	Thai	English
Thai Document Set I	EucrosiaUPC	8,632	0	8,632	100.00	0.00
Thai Document Set II	FreesiaUPC	10,854	0	10,854	100.00	0.00
Thai Document Set III	Mixed Fonts	6,349	0	6,349	100.00	0.00
Thai Document Set IV	Mixed Fonts	43,465	0	43,465	100.00	0.00
Thai Document Set V	Mixed Fonts	48,047	0	48,047	100.00	0.00
All Thai Document Sets		**117,347**	**0**	**117,347**	**100.00**	**0.00**

2. English document set

This second set consists of 53 pages of text with 96,260 characters, the fonts of Arial and Cordia New, the sizes of 11 and 12 for Arial, and 16 and 18 for Cordia New and the styles of normal and bold. In this group, 5 sets of English documents are as follows:

English document set I contains 5 document images of 9,252 characters, the font of Cordia New, the sizes of 16 and 18 and the styles of normal and bold. A sample of scanned document images in Cordia New font is shown in Figure 6.4.

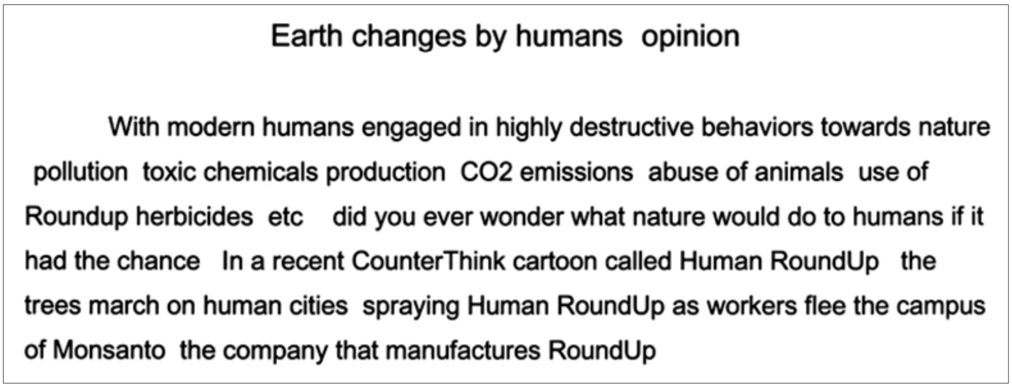


Windows Vista Themes

The Windows Vista Themes allow people to be able to change the visual elements of the desktop area on their computer by selecting a theme that is stored on their computer People can select which computer sounds they want on the computer desktop too and some of the themes come with their own sound scheme

Figure 6.4: A sample of scanned document images in Cordia New font

English document set II contains 5 document images of 9,084 characters, the font of Arial, the sizes of 11 and 12 and the styles of normal and bold. A sample of scanned document images in Arial font is shown in Figure 6.5.

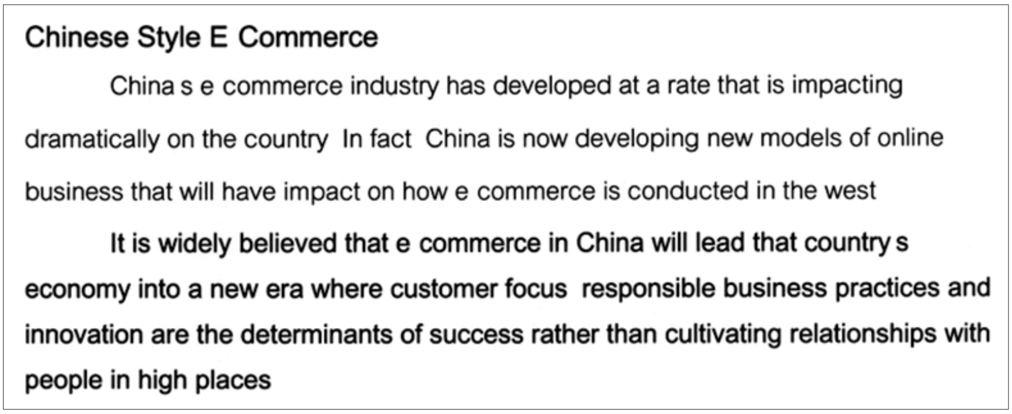


Earth changes by humans opinion

With modern humans engaged in highly destructive behaviors towards nature pollution toxic chemicals production CO2 emissions abuse of animals use of Roundup herbicides etc did you ever wonder what nature would do to humans if it had the chance In a recent CounterThink cartoon called Human RoundUp the trees march on human cities spraying Human RoundUp as workers flee the campus of Monsanto the company that manufactures RoundUp

Figure 6.5: A sample of scanned document images in Arial font

English document set III contains 3 document images of 5,472 characters, the fonts of Arial and Cordia New, the sizes of 11 and 12 for Arial, and 16 and 18 for Cordia New, and the styles of normal and bold. A sample of scanned document images in English mixed fonts is shown in Figure 6.6.



Chinese Style E Commerce

China s e commerce industry has developed at a rate that is impacting dramatically on the country In fact China is now developing new models of online business that will have impact on how e commerce is conducted in the west

It is widely believed that e commerce in China will lead that country s economy into a new era where customer focus responsible business practices and innovation are the determinants of success rather than cultivating relationships with people in high places

Figure 6.6: A sample of scanned document images in English mixed fonts

In addition, the large document sets: English document set IV contains 20 document images of 34,969 characters and English document set V contains 20 document images of 37,483 characters, the fonts of Arial and Cordia New, the sizes of

11 and 12 for Arial, and 16 and 18 for Cordia New, and the styles of normal and bold. The details of documents in English document set are outperformed in Table 6.2.

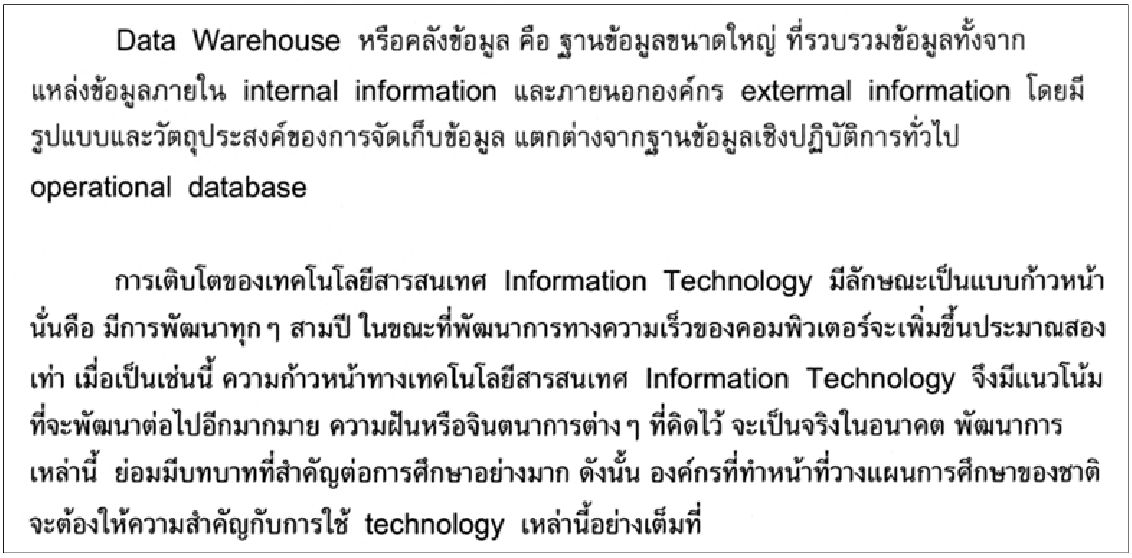
Table 6.2: The testing documents in English document set.

Document	Font	Character No.			Character No. (%)	
		Thai	English	All	Thai	English
English Document Set I	Cordia New	0	9,084	9,084	0.00	100.00
English Document Set II	Arial	0	9,252	9,252	0.00	100.00
English Document Set III	Mixed Fonts	0	5,472	5,472	0.00	100.00
English Document Set IV	Mixed Fonts	0	34,969	34,969	0.00	100.00
English Document Set V	Mixed Fonts	0	37,483	37,483	0.00	100.00
All English Document Sets		**0**	**96,260**	**96,260**	**0.00**	**100.00**

3. Thai and English mixed document set

This third set consists of 35 pages of text with 70,810 characters, the fonts of EucrosiaUPC and FreesiaUPC for Thai characters, and Arial and Cordia New for English characters, the sizes of 16 and 18 for EucrosiaUPC, FreesiaUPC and Cordia New, and 11 and 12 for Arial and the styles of normal and bold. In this group, there are 5 sets of Thai and English mixed bilingual documents as follows:

Thai and English mixed document set I contains 5 document images of 10,478 characters, the fonts of EucrosiaUPC and FreesiaUPC for Thai, and Arial for English, the sizes of 11 and 12 for Arial font, and 16 and 18 for the others, and the styles of normal and bold. A sample of scanned document images in Thai mixed fonts and Arial font for English is shown in Figure 6.7.

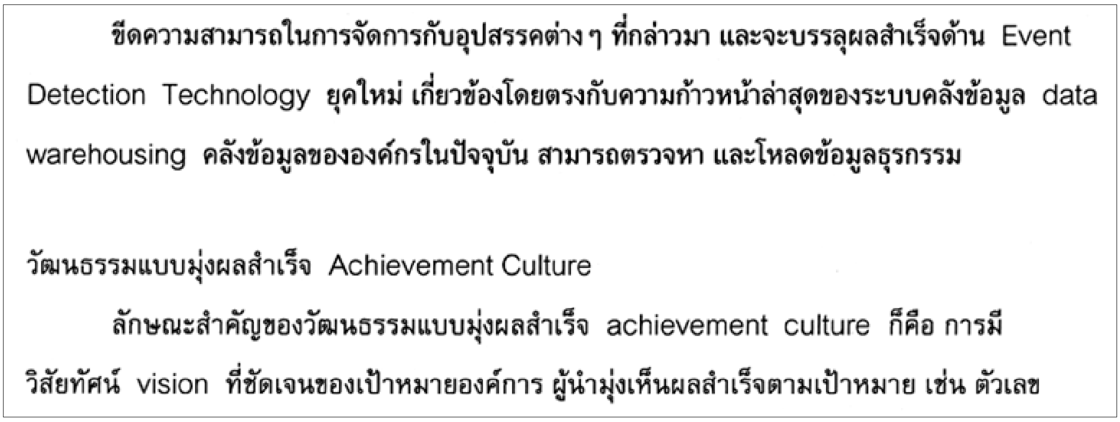


Data Warehouse หรือคลังข้อมูล คือ ฐานข้อมูลขนาดใหญ่ ที่รวบรวมข้อมูลทั้งจากแหล่งข้อมูลภายใน internal information และภายนอกองค์กร extermal information โดยมีรูปแบบและวัตถุประสงค์ของการจัดเก็บข้อมูล แตกต่างจากฐานข้อมูลเชิงปฏิบัติการทั่วไป operational database

การเติบโตของเทคโนโลยีสารสนเทศ Information Technology มีลักษณะเป็นแบบก้าวหน้านั่นคือ มีการพัฒนาทุก ๆ สามปี ในขณะที่พัฒนาการทางความเร็วของคอมพิวเตอร์จะเพิ่มขึ้นประมาณสองเท่า เมื่อเป็นเช่นนี้ ความก้าวหน้าทางเทคโนโลยีสารสนเทศ Information Technology จึงมีแนวโน้มที่จะพัฒนาต่อไปอีกมากมาย ความฝันหรือจินตนาการต่าง ๆ ที่คิดไว้ จะเป็นจริงในอนาคต พัฒนาการเหล่านี้ ย่อมมีบทบาทที่สำคัญต่อการศึกษาอย่างมาก ดังนั้น องค์กรที่ทำหน้าที่วางแผนการศึกษาของชาติจะต้องให้ความสำคัญกับการใช้ technology เหล่านี้อย่างเต็มที่

Figure 6.7: A sample of scanned document images in Thai mixed fonts and Arial font for English

Thai and English mixed document set II contains 5 document images of 9,773 characters, the fonts of EucrosiaUPC and FreesiaUPC for Thai, and Cordia New for English, the sizes of 16 and 18, and the styles of normal and bold. A sample of scanned document images in Thai mixed fonts and Cordia New font for English is shown in Figure 6.8.



ขีดความสามารถในการจัดการกับอุปสรรคต่าง ๆ ที่กล่าวมา และจะบรรลุผลสำเร็จด้าน Event Detection Technology ยุคใหม่ เกี่ยวข้องโดยตรงกับความก้าวหน้าล่าสุดของระบบคลังข้อมูล data warehousing คลังข้อมูลขององค์กรในปัจจุบัน สามารถตรวจหา และโหลดข้อมูลธุรกรรม

วัฒนธรรมแบบมุ่งผลสำเร็จ Achievement Culture

ลักษณะสำคัญของวัฒนธรรมแบบมุ่งผลสำเร็จ achievement culture ก็คือ การมีวิสัยทัศน์ vision ที่ชัดเจนของเป้าหมายองค์การ ผู้นำมุ่งเห็นผลสำเร็จตามเป้าหมาย เช่น ตัวเลข

Figure 6.8: A sample of scanned document images in Thai mixed fonts and Cordia New font for English

Thai and English mixed document set III contains 5 document images of 10,197 characters, the fonts of EucrosiaUPC and FreesiaUPC for Thai, and Cordia New and Arial for English, the sizes of 11 and 12 for Arial font, and 16 and 18 for the others, and the styles of normal and bold.

In addition, the large document sets: Thai and English mixed document set IV contains 10 document images of 19,156 characters and Thai and English mixed document set V contains 10 document images of 19,220 characters, the fonts of EucrosiaUPC and FreesiaUPC for Thai, and Cordia New and Arial for English, the sizes of 11 and 12 for Arial font, and 16 and 18 for the others, and the styles of normal and bold. The details of documents in Thai/English mixed document set are outperformed in Table 6.3.

Table 6.3: The testing documents in Thai and English mixed document set.

Document	Font	Character No.			Character No. (%)	
		Thai	English	All	Thai	English
Mixed Modes Document Set I	Thai Mixed Fonts and Arial	9,239	1,239	10,478	88.18	11.82
Mixed Modes Document Set II	Thai Mixed Fonts and Cordia New	8,431	1,342	9,773	86.27	13.73
Mixed Modes Document Set III	Thai and English Mixed Fonts	9,127	1,070	10,197	89.51	10.49
Mixed Modes Document Set IV	Thai and English Mixed Fonts	19,329	1,791	19,156	100.90	9.35
Mixed Modes Document Set V	Thai and English Mixed Fonts	17,147	2,073	19,220	89.21	10.79
All Mixed Modes Document Sets		63,273	7,515	68,824	91.93	10.92

6.1 Results of the Language Identification Process

The experiment in this section attempts to identify the language modes from the three sets of documents as mentioned previously.

An example of language modes for a part of Thai and English mixed document derived from the language identification process is illustrated in Figure 6.9.

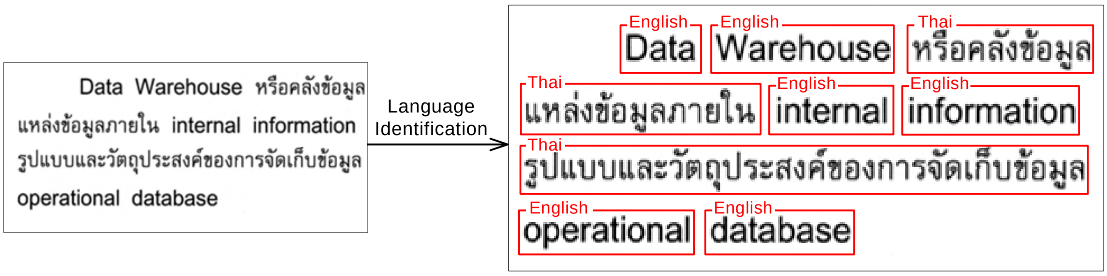


Figure 6.9: An example of language modes for
a part of Thai/English mixed document

The results of language identification process tested on all 3 document sets: Thai, English, and Thai and English mixed document sets are given in Table 6.4. For the first set with Thai scripts only, the system is able to identify the script correctly with 100% accuracy in 4 sets, and 99.93% accuracy in a set of document. For the second set with English scripts only, the system could yield the result with 99.98% accuracy in over all 5 sets of documents. For the third set with bilingual scripts, the system is able to identify all scripts correctly without a single error, or 100% accuracy. Therefore the overall accuracy in this experiment is 99.99%.

Table 6.4: The results of language identification process.

Document	No. of Characters		No. of Correct Identified Scripts		Language Identification Accuracy (%)
	Thai	English	Thai	English	
Thai Document Set I	8,632	0	8,632	0	100.00
Thai Document Set II	10,854	0	10,846	0	99.93
Thai Document Set III	6,349	0	6,349	0	100.00
Thai Document Set IV	43,465	0	43,465	0	100.00
Thai Document Set V	48,047	0	48,047	0	100.00
All Thai Document Sets	**117,347**	**0**	**117,339**	**0**	**99.99**
English Document Set I	0	9,084	0	9,084	100.00
English Document Set II	0	9,252	0	9,252	100.00
English Document Set III	0	5,472	0	5,472	100.00
English Document Set IV	0	34,969	0	34,956	99.96
English Document Set V	0	37,483	0	37,481	99.99
All English Document Sets	**0**	**96,260**	**0**	**96,245**	**99.98**
Mixed Mode Document Set I	9,239	1,239	9,239	1,239	100.00
Mixed Mode Document Set II	8,431	1,342	8,431	1,342	100.00
Mixed Mode Document Set III	9,127	1,070	9,127	1,070	100.00
Mixed Mode Document Set IV	19,336	1,791	19,336	1,791	100.00
Mixed Mode Document Set V	17,162	2,073	17,162	2,073	100.00
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,295**	**7,515**	**100.00**
All Document Sets	**180,642**	**103,775**	**180,634**	**103,760**	**99.99**

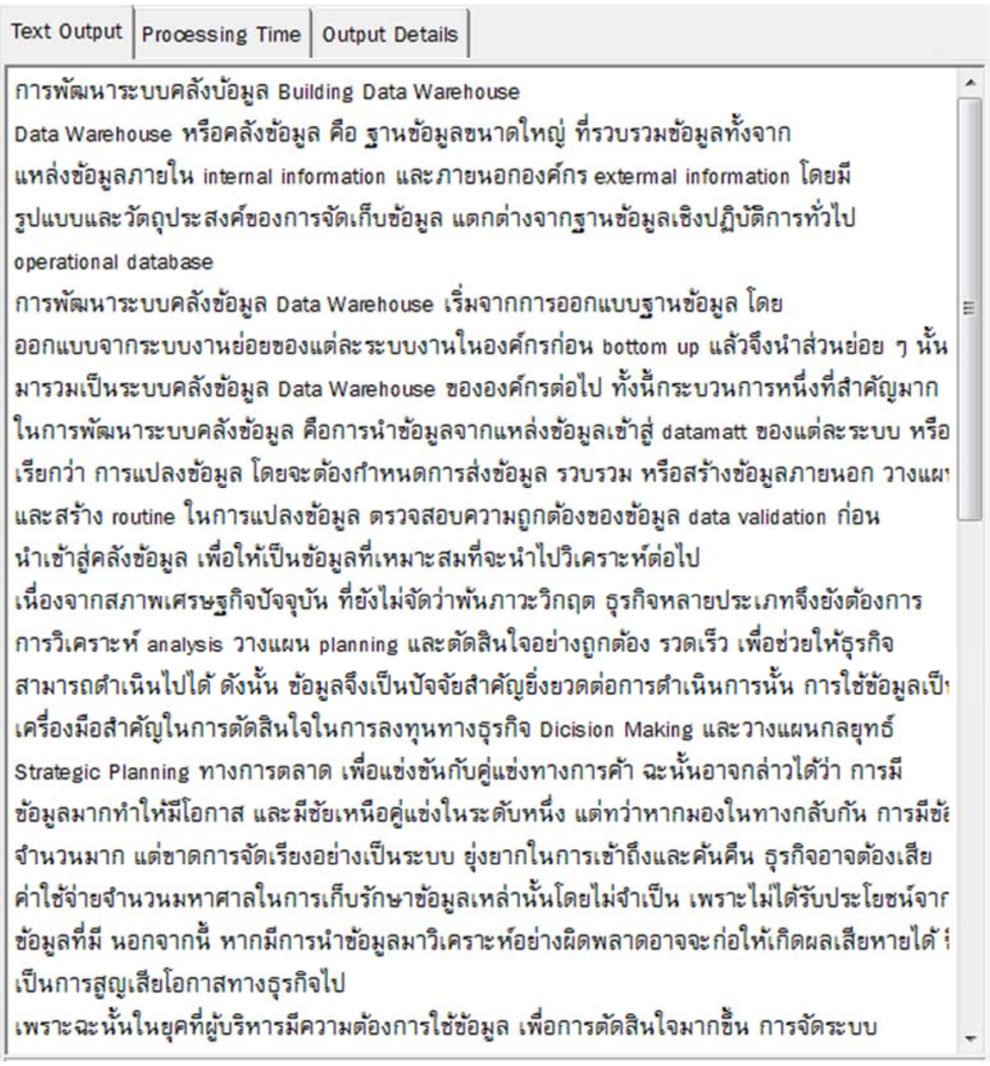
6.2 Results of the Conventional Approach

The following series of experiment attempts to recognize character images for bilingual document in Thai and English with their language scripts being identified first. Therefore, the character identified “Thai” will be recognized in the conventional Thai-OCR where-as the character identified “English” will be recognized in the conventional English-OCR.

In Section 6.2.1, we show the experimental results of the conventional approach in terms of accuracy and speed. Next, in Section 6.2.2, we show the improvements of the conventional approach by using dictionary look-up for error correction. Finally, in Section 6.3.3, we show performance evaluation of the conventional approach with dictionary look-up in comparison with the conventional approach.

6.2.1 The Accuracy and Speed of the Conventional Approach

In the experiment, character images contained in the block images with their language scripts being identified are used to test the conventional system. The character images are recognized in the recognition process one after another. A sample of recognition result is shown in Figure 6.10.



```
การพัฒนาระบบคลังบ้อมูล Building Data Warehouse
Data Warehouse หรือคลังข้อมูล คือ ฐานข้อมูลขนาดใหญ่ ที่รวบรวมข้อมูลทั้งจาก
แหล่งข้อมูลภายใน internal information และภายนอกองค์กร extermal information โดยมี
รูปแบบและวัตถุประสงค์ของการจัดเก็บข้อมูล แตกต่างจากฐานข้อมูลเชิงปฏิบัติการทั่วไป
operational database
การพัฒนาระบบคลังข้อมูล Data Warehouse เริ่มจากการออกแบบฐานข้อมูล โดย
ออกแบบจากระบบงานย่อยของแต่ละระบบงานในองค์กรก่อน bottom up แล้วจึงนำส่วนย่อย ๆ นั้น
มารวมเป็นระบบคลังข้อมูล Data Warehouse ขององค์กรต่อไป ทั้งนี้กระบวนการหนึ่งที่สำคัญมาก
ในการพัฒนาระบบคลังข้อมูล คือการนำข้อมูลจากแหล่งข้อมูลเข้าสู่ datamatt ของแต่ละระบบ หรือ
เรียกว่า การแปลงข้อมูล โดยจะต้องกำหนดการส่งข้อมูล รวบรวม หรือสร้างข้อมูลภายนอก วางแผ
และสร้าง routine ในการแปลงข้อมูล ตรวจสอบความถูกต้องของข้อมูล data validation ก่อน
นำเข้าสู่คลังข้อมูล เพื่อให้เป็นข้อมูลที่เหมาะสมที่จะนำไปวิเคราะห์ต่อไป
เนื่องจากสภาพเศรษฐกิจปัจจุบัน ที่ยังไม่จัดว่าพ้นภาวะวิกฤต ธุรกิจหลายประเภทจึงยังต้องการ
การวิเคราะห์ analysis วางแผน planning และตัดสินใจอย่างถูกต้อง รวดเร็ว เพื่อช่วยให้ธุรกิจ
สามารถดำเนินไปได้ ดังนั้น ข้อมูลจึงเป็นปัจจัยสำคัญยิ่งยวดต่อการดำเนินการนั้น การใช้ข้อมูลเป็
เครื่องมือสำคัญในการตัดสินใจในการลงทุนทางธุรกิจ Dicision Making และวางแผนกลยุทธ์
Strategic Planning ทางการตลาด เพื่อแข่งขันกับคู่แข่งทางการค้า ฉะนั้นอาจกล่าวได้ว่า การมี
ข้อมูลมากทำให้มีโอกาส และมีชัยเหนือคู่แข่งในระดับหนึ่ง แต่ทว่าหากมองในทางกลับกัน การมีข้
จำนวนมาก แต่ขาดการจัดเรียงอย่างเป็นระบบ ยุ่งยากในการเข้าถึงและค้นคืน ธุรกิจอาจต้องเสีย
ค่าใช้จ่ายจำนวนมหาศาลในการเก็บรักษาข้อมูลเหล่านั้นโดยไม่จำเป็น เพราะไม่ได้รับประโยชน์จาก
ข้อมูลที่มี นอกจากนี้ หากมีการนำข้อมูลมาวิเคราะห์อย่างผิดพลาดอาจจะก่อให้เกิดผลเสียหายได้
เป็นการสูญเสียโอกาสทางธุรกิจไป
เพราะฉะนั้นในยุคที่ผู้บริหารมีความต้องการใช้ข้อมูล เพื่อการตัดสินใจมากขึ้น การจัดระบบ
```

Figure 6.10: A sample of recognition result

In this experiment, we show performance of the conventional system in terms of accuracy and speed in Table 6.5(a)-(b). Details are as follows:

Table 6.5(a): The recognition accuracy of the conventional approach.

Document	No. of Characters		No. of Correct Recognition		Recognition Accuracy (%)
	Thai	English	Thai	English	
Thai Document Set I	8,632	0	8,632	0	100.00
Thai Document Set II	10,854	0	10,854	0	100.00
Thai Document Set III	6,349	0	6,349	0	100.00
Thai Document Set IV	43,465	0	43,465	0	100.00
Thai Document Set V	48,047	0	48,047	0	100.00
All Thai Document Sets	**117,347**	**0**	**117,347**	**0**	**100.00**
English Document Set I	0	9,084	0	8,989	98.95
English Document Set II	0	9,252	0	9,110	98.47
English Document Set III	0	5,472	0	5,363	98.01
English Document Set IV	0	34,969	0	34,376	98.30
English Document Set V	0	37,483	0	36,666	97.82
All English Document Sets	**0**	**96,260**	**0**	**94,504**	**98.18**
Mixed Mode Document Set I	9,239	1,239	9,237	1,228	99.88
Mixed Mode Document Set II	8,431	1,342	8,428	1,331	99.86
Mixed Mode Document Set III	9,127	1,070	9,127	1,051	99.81
Mixed Mode Document Set IV	19,336	1,791	19,334	1,761	99.85
Mixed Mode Document Set V	17,162	2,073	17,159	2,048	99.85
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,285**	**7,419**	**99.85**
All Document Sets	**180,642**	**103,775**	**180,632**	**101,923**	**99.35**

Table 6.5(b): The processing time of the conventional approach.

Document	No. of Characters		Total Processing Time in ms	Average Processing Time per Character in ms
	Thai	English		
Thai Document Set I	8,632	0	6,044.21	0.70
Thai Document Set II	10,854	0	7,093.13	0.65
Thai Document Set III	6,349	0	3,988.79	0.63
Thai Document Set IV	43,465	0	28,351.79	0.65
Thai Document Set V	48,047	0	31,240.16	0.65
All Thai Document Sets	**117,347**	**0**	**76,718.07**	**0.65**
English Document Set I	0	9,084	5,291.28	0.58
English Document Set II	0	9,252	6,676.20	0.72
English Document Set III	0	5,472	3,575.26	0.65
English Document Set IV	0	34,969	21,612.74	0.62
English Document Set V	0	37,483	27,572.31	0.74
All English Document Sets	**0**	**96,260**	**64,727.80**	**0.67**
Mixed Mode Document Set I	9,239	1,239	7,049.88	0.67
Mixed Mode Document Set II	8,431	1,342	6,609.36	0.68
Mixed Mode Document Set III	9,127	1,070	6,967.43	0.68
Mixed Mode Document Set IV	19,336	1,791	13,049.62	0.62
Mixed Mode Document Set V	17,162	2,073	12,875.55	0.67
All Mixed Mode Document Sets	**63,295**	**7,515**	**46,551.84**	**0.66**
All Document Sets	**180,642**	**103,775**	**187,997.71**	**0.66**

6.2.2 Improvements of the Conventional Approach with Dictionary Look-up

To improve accuracy of the conventional system, we apply dictionary look-up method for correcting some recognition errors after recognizing a strip of character images.

In this experiment, we show performance of the conventional system with dictionary look-up in terms of accuracy and speed in Table 6.6(a)-(b). Details are as follows:

Table 6.6(a): The recognition accuracy of the improvements of conventional system.

Document	No. of Characters		No. of Correct Recognition		Recognition Accuracy (%)
	Thai	English	Thai	English	
English Document Set I	0	9,084	0	9,084	100.00
English Document Set II	0	9,252	0	9,252	100.00
English Document Set III	0	5,472	0	5,472	100.00
English Document Set IV	0	34,969	0	34,949	99.94
English Document Set V	0	37,483	0	37,454	99.92
All English Document Sets	**0**	**96,260**	**0**	**96,211**	**99.95**
Mixed Mode Document Set I	9,239	1,239	9,239	1,237	99.98
Mixed Mode Document Set II	8,431	1,342	8,430	1,340	99.97
Mixed Mode Document Set III	9,127	1,070	9,127	1,070	100.00
Mixed Mode Document Set IV	19,336	1,791	19,335	1,791	100.00
Mixed Mode Document Set V	17,162	2,073	17,160	2,065	99.95
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,291**	**7,503**	**99.98**
All Document Sets	**63,295**	**103,775**	**63,291**	**103,714**	**99.96**

Table 6.6(b): The processing time of the improvements of conventional system.

Document	No. of Characters		Total Processing Time in ms	Average Processing Time per Character in ms
	Thai	English		
English Document Set I	0	9,084	5,462.73	0.60
English Document Set II	0	9,252	6,801.20	0.74
English Document Set III	0	5,472	3,659.31	0.67
English Document Set IV	0	34,969	22,025.56	0.63
English Document Set V	0	37,483	28,430.78	0.76
All English Document Sets	**0**	**96,260**	**66,379.57**	**0.69**
Mixed Mode Document Set I	9,239	1,239	26,668.01	2.55
Mixed Mode Document Set II	8,431	1,342	26,482.45	2.71
Mixed Mode Document Set III	9,127	1,070	26,565.33	2.61
Mixed Mode Document Set IV	19,336	1,791	64,349.07	3.05
Mixed Mode Document Set V	17,162	2,073	51,472.59	2.68
All Mixed Mode Document Sets	**63,295**	**7,515**	**195,537.44**	**2.76**
All Document Sets	**63,295**	**103,775**	**261,917.01**	**1.57**

6.2.3 Evaluation of the Conventional Approach with Dictionary Look-up

From the experimental results of Sections 6.2.1 and 6.2.2, we compare the performance of the two approaches as shown in Table 6.7. Details are as follows:

Table 6.7: The comparison of the conventional approach and the conventional system with dictionary look-up.

Document	No. of Characters		No. of Correct Recognition by Conventional Approach			No. of Correct Recognition by the improvements of Conventional Approach			Accuracy Improvement (%)
	Thai	English	Thai	English	Accuracy Rate (%)	Thai	English	Accuracy Rate (%)	
English Document Set I	0	9,084	0	8,989	98.95	0	9,084	100.00	1.06
English Document Set II	0	9,252	0	9,110	98.47	0	9,252	100.00	1.56
English Document Set III	0	5,472	0	5,363	98.01	0	5,472	100.00	2.03
English Document Set IV	0	34,969	0	34,376	98.30	0	34,949	99.94	1.67
English Document Set V	0	37,483	0	36,666	97.82	0	37,454	99.92	2.15
All English Document Sets	**0**	**96,260**	**0**	**94,504**	**98.18**	**0**	**96,211**	**99.95**	**1.81**
Mixed Mode Document Set I	9,239	1,239	9,237	1,228	99.88	9,239	1,237	99.98	0.11
Mixed Mode Document Set II	8,431	1,342	8,428	1,331	99.86	8,430	1,340	99.97	0.11
Mixed Mode Document Set III	9,127	1,070	9,127	1,051	99.81	9,127	1,070	100.00	0.19
Mixed Mode Document Set IV	19,336	1,791	19,334	1,761	99.85	19,335	1,791	100.00	0.15
Mixed Mode Document Set V	17,162	2,073	17,159	2,048	99.85	17,160	2,065	99.95	0.09
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,285**	**7,419**	**99.85**	**63,291**	**7,503**	**99.98**	**0.13**
All Document Sets	**63,295**	**103,775**	**63,285**	**101,923**	**99.35**	**63,291**	**103,714**	**99.96**	**1.09**

6.3 Results of the BOCR-WP System

The experiment in this part attempts to enhance the performance of the conventional approach to a new system in terms of accuracy and speed with two additional processes: word prediction and word verification after processing of character recognition for bilingual document in Thai and English. The new system is named as BOCR-WP (Bilingual Optical Character Recognition with Word Prediction). In addition, based on bilingual document, as if their language modes are first identified correctly, the characters will be recognized in the appropriate OCR-WP system according to their identified language modes. Therefore, the character identified “Thai” will be recognized in the Thai-OCR-WP where as the character identified “English” will be recognized in the English-OCR-WP.

In Section 6.3.1, the results of the BOCR-WP system in terms of performance are presented. Next, in Section 6.3.2, performance evaluation of the BOCR-WP system is presented by comparing with the conventional approach. Then, in Section 6.3.3, the improvements of the BOCR-WP system by using dictionary look-up for error correction is presented. Finally, in Section 6.3.4, performance evaluation is presented by comparing the BOCR-WP with and without the dictionary look-up.

6.3.1 System Performance of BOCR-WP

In the experiment, character images contained in the block images with their language scripts being identified are used to test the BOCR-WP system. In each block image, a few character images would be recognized first by the recognition process one after another that given a previous word token. Next, we predict the next probable words of the token and verify unknown character images by using template matching technique.

In this experiment, we show performance of the BOCR-WP system in terms of accuracy and speed in Table 6.8(a)-(b). Details are as follows:

Table 6.8(a): The recognition accuracy of the BOCR-WP system.

Document	No. of Characters		No. of Correct Recognition		Recognition Accuracy (%)
	Thai	English	Thai	English	
Thai Document Set I	8,632	0	8,632	0	100.00
Thai Document Set II	10,854	0	10,854	0	100.00
Thai Document Set III	6,349	0	6,348	0	100.00
Thai Document Set IV	43,465	0	43,465	0	100.00
Thai Document Set V	48,047	0	48,047	0	100.00
All Thai Document Sets	**117,347**	**0**	**117,346**	**0**	**100.00**
English Document Set I	0	9,084	0	9,072	99.87
English Document Set II	0	9,252	0	9,218	99.63
English Document Set III	0	5,472	0	5,454	99.67
English Document Set IV	0	34,969	0	34,909	99.83
English Document Set V	0	37,483	0	37,275	99.45
All English Document Sets	**0**	**96,260**	**0**	**95,928**	**99.66**
Mixed Mode Document Set I	9,239	1,239	9,237	1,233	99.92
Mixed Mode Document Set II	8,431	1,342	8,428	1,335	99.90
Mixed Mode Document Set III	9,127	1,070	9,127	1,063	99.93
Mixed Mode Document Set IV	19,336	1,791	19,334	1,789	99.98
Mixed Mode Document Set V	17,162	2,073	17,160	2,061	99.93
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,286**	**7,481**	**99.94**
All Document Sets	**180,642**	**103,775**	**180,632**	**103,409**	**99.87**

Table 6.8(b): The processing time of the BOCR-WP.

Document	No. of Characters		Total Processing Time in ms	Average Processing Time per Character in ms
	Thai	English		
Thai Document Set I	8,632	0	4,882.21	0.57
Thai Document Set II	10,854	0	5,641.04	0.52
Thai Document Set III	6,349	0	3,337.00	0.53
Thai Document Set IV	43,465	0	22,829.44	0.53
Thai Document Set V	48,047	0	26,339.91	0.55
All Thai Document Sets	**117,347**	**0**	**63,029.60**	**0.54**
English Document Set I	0	9,084	3,786.86	0.42
English Document Set II	0	9,252	4,749.94	0.51
English Document Set III	0	5,472	2,611.58	0.48
English Document Set IV	0	34,969	15,598.49	0.45
English Document Set V	0	37,483	19,718.21	0.53
All English Document Sets	**0**	**96,260**	**46,465.07**	**0.48**
Mixed Mode Document Set I	9,239	1,239	5,565.32	0.53
Mixed Mode Document Set II	8,431	1,342	5,319.71	0.54
Mixed Mode Document Set III	9,127	1,070	5,473.96	0.54
Mixed Mode Document Set IV	19,336	1,791	10,205.23	0.48
Mixed Mode Document Set V	17,162	2,073	10,210.82	0.53
All Mixed Mode Document Sets	**63,295**	**7,515**	**36,775.04**	**0.52**
All Document Sets	**180,642**	**103,775**	**146,269.71**	**0.51**

6.3.2 Evaluation of the BOCR-WP System

From the experimental results of Sections 6.2.1 and 6.3.1, we compare the performance of the two approaches as shown in Table 6.9(a)-(b). Details are as follows:

Table 6.9(a): The comparison of the conventional approach and the BOCR-WP system in term of accuracy.

Document	No. of Characters		No. of Correct Recognition by Conventional Approach			No. of Correct Recognition by BOCR-WP			Accuracy Improvement (%)
	Thai	English	Thai	English	Accuracy Rate (%)	Thai	English	Accuracy Rate (%)	
Thai Document Set I	8,632	0	8,632	0	100.00	8,632	0	100.00	-
Thai Document Set II	10,854	0	10,854	0	100.00	10,854	0	100.00	-
Thai Document Set III	6,349	0	6,349	0	100.00	6,348	0	100.00	-
Thai Document Set IV	43,465	0	43,465	0	100.00	43,465	0	100.00	-
Thai Document Set V	48,047	0	48,047	0	100.00	48,047	0	100.00	-
All Thai Document Sets	**117,347**	**0**	**117,347**	**0**	**100.00**	**117,346**	**0**	**100.00**	**-**
English Document Set I	0	9,084	0	8,989	98.95	0	9,072	99.87	0.92
English Document Set II	0	9,252	0	9,110	98.47	0	9,218	99.63	1.19
English Document Set III	0	5,472	0	5,363	98.01	0	5,454	99.67	1.70
English Document Set IV	0	34,969	0	34,376	98.30	0	34,909	99.83	1.55
English Document Set V	0	37,483	0	36,666	97.82	0	37,275	99.45	1.66
All English Document Sets	**0**	**96,260**	**0**	**94,504**	**98.18**	**0**	**95,928**	**99.66**	**1.51**
Mixed Mode Document Set I	9,239	1,239	9,237	1,228	99.88	9,237	1,233	99.92	0.05
Mixed Mode Document Set II	8,431	1,342	8,428	1,331	99.86	8,428	1,335	99.90	0.04
Mixed Mode Document Set III	9,127	1,070	9,127	1,051	99.81	9,127	1,063	99.93	0.12
Mixed Mode Document Set IV	19,336	1,791	19,334	1,761	99.85	19,334	1,789	99.98	0.13
Mixed Mode Document Set V	17,162	2,073	17,159	2,048	99.85	17,160	2,061	99.93	0.07
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,285**	**7,419**	**99.85**	**63,286**	**7,481**	**99.94**	**0.09**
All Document Sets	180,642	**103,775**	**180,632**	**101,923**	**99.35**	**180,632**	**103,409**	**99.87**	**0.53**

Table 6.9(b): The comparison of the conventional approach and the BOCR-WP system in term of speed.

Document	Original Characters		Average Processing Time per Character by Conventional Approach	Average Processing Time per Character by BOCR-WP	Speed Improvement (%)
	Thai	English			
Thai Document Set I	8,632	0	0.70	0.57	19.22
Thai Document Set II	10,854	0	0.65	0.52	20.47
Thai Document Set III	6,349	0	0.63	0.53	16.34
Thai Document Set IV	43,465	0	0.65	0.53	19.48
Thai Document Set V	48,047	0	0.65	0.55	15.69
All Thai Document Sets	**117,347**	**0**	**0.65**	**0.54**	**17.84**
English Document Set I	0	9,084	0.58	0.42	28.43
English Document Set II	0	9,252	0.72	0.51	28.85
English Document Set III	0	5,472	0.65	0.48	26.95
English Document Set IV	0	34,969	0.62	0.45	27.83
English Document Set V	0	37,483	0.74	0.53	28.49
All English Document Sets	**0**	**96,260**	**0.67**	**0.48**	**28.21**
Mixed Mode Document Set I	9,239	1,239	0.67	0.53	21.06
Mixed Mode Document Set II	8,431	1,342	0.68	0.54	19.51
Mixed Mode Document Set III	9,127	1,070	0.68	0.54	21.44
Mixed Mode Document Set IV	19,336	1,791	0.62	0.48	21.80
Mixed Mode Document Set V	17,162	2,073	0.67	0.53	20.70
All Mixed Mode Document Sets	**63,295**	**7,515**	**0.66**	**0.52**	**21.00**
All Document Sets	**180,642**	**103,775**	**0.66**	**0.51**	**22.20**

6.3.3 Improvement of the BOCR-WP System with Dictionary Look-up

To improve accuracy of the BOCR-WP system, we apply dictionary look-up method for correcting some recognition errors after recognizing a strip of character images.

In this experiment, we show performance of the BOCR-WP system with dictionary look-up in terms of accuracy and speed in Table 6.10(a)-(b). Details are as follows:

Table 6.10(a): The recognition accuracy of the improvements of the BOCR-WP system.

Document	No. of Characters		No. of Correct Recognition		Recognition Accuracy (%)
	Thai	English	Thai	English	
English Document Set I	0	9,084	0	9,081	99.97
English Document Set II	0	9,252	0	9,248	99.96
English Document Set III	0	5,472	0	5,471	99.98
English Document Set IV	0	34,969	0	34,964	99.99
English Document Set V	0	37,483	0	37,441	99.89
All English Document Sets	**0**	**96,260**	**0**	**96,205**	**99.94**
Mixed Mode Document Set I	9,239	1,239	9,239	1,237	99.98
Mixed Mode Document Set II	8,431	1,342	8,430	1,342	99.99
Mixed Mode Document Set III	9,127	1,070	9,127	1,070	100.00
Mixed Mode Document Set IV	19,336	1,791	19,335	1,789	99.99
Mixed Mode Document Set V	17,162	2,073	17,160	2,064	99.94
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,291**	**7,502**	**99.98**
All Document Sets	**63,295**	**103,775**	**63,291**	**103,707**	**99.96**

Table 6.10(b): The processing time of the improvements of the BOCR-WP.

Document	No. of Characters		Total Processing Time in ms	Average Processing Time per Character in ms
	Thai	English		
English Document Set I	0	9,084	3,809.50	0.42
English Document Set II	0	9,252	4,800.94	0.52
English Document Set III	0	5,472	2,659.88	0.49
English Document Set IV	0	34,969	15,714.67	0.45
English Document Set V	0	37,483	19,972.13	0.53
All English Document Sets	**0**	**96,260**	**46,957.12**	**0.49**
Mixed Mode Document Set I	9,239	1,239	24,620.92	2.35
Mixed Mode Document Set II	8,431	1,342	25,305.97	2.59
Mixed Mode Document Set III	9,127	1,070	24,773.22	2.43
Mixed Mode Document Set IV	19,336	1,791	59,521.02	2.82
Mixed Mode Document Set V	17,162	2,073	47,529.85	2.47
All Mixed Mode Document Sets	**63,295**	**7,515**	**181,750.98**	**2.57**
All Document Sets	**63,295**	**103,775**	**228,708.10**	**1.37**

6.3.4 Evaluation of the BOCR-WP System with Dictionary Look-up

From the experiments of Sections 6.3.1 and 6.3.3, we compare the performance of the two approaches as shown in Table 6.11. Details are as follows:

Table 6.11: The comparison of the BOCR-WP system and the BOCR-WP with dictionary look-up.

Document	No. of Characters		No. of Correct Recognition by BOCR-WP			No. of Correct Recognition by the improvements of BOCR-WP			Accuracy Improvement (%)
	Thai	English	Thai	English	Accuracy Rate (%)	Thai	English	Accuracy Rate (%)	
English Document Set I	0	9,084	0	9,072	99.87	0	9,081	99.97	0.10
English Document Set II	0	9,252	0	9,218	99.63	0	9,248	99.96	0.33
English Document Set III	0	5,472	0	5,454	99.67	0	5,471	99.98	0.31
English Document Set IV	0	34,969	0	34,909	99.83	0	34,964	99.99	0.16
English Document Set V	0	37,483	0	37,275	99.45	0	37,441	99.89	0.45
All English Document Sets	**0**	**96,260**	**0**	**95,928**	**99.66**	**0**	**96,205**	**99.94**	**0.29**
Mixed Mode Document Set I	9,239	1,239	9,237	1,233	99.92	9,239	1,237	99.98	0.06
Mixed Mode Document Set II	8,431	1,342	8,428	1,335	99.90	8,430	1,342	99.99	0.09
Mixed Mode Document Set III	9,127	1,070	9,127	1,063	99.93	9,127	1,070	100.00	0.07
Mixed Mode Document Set IV	19,336	1,791	19,334	1,789	99.98	19,335	1,789	99.99	0.00
Mixed Mode Document Set V	17,162	2,073	17,160	2,061	99.93	17,160	2,064	99.94	0.02
All Mixed Mode Document Sets	**63,295**	**7,515**	**63,286**	**7,481**	**99.94**	**63,291**	**7,502**	**99.98**	**0.04**
All Document Sets	**63,295**	**103,775**	**63,286**	**103,409**	**99.78**	**63,291**	**103,707**	**99.96**	**0.18**

CHAPTER VII
DISCUSSION AND CONCLUSION

This chapter will present discussion and conclusion from the results as described in Chapter VI. Details are as follows:

7.1 Discussion

In Section 7.1.1, we discuss the performance of the BOCR-WP system in terms of accuracy, processing time and storage. Next, complexity of the program is explained in Section 7.1.2. Finally, in Section 7.1.3, we present the problems found during the system implementation. Details are as follows:

7.1.1 Performance

From the experimental results mentioned in Chapter VI, the proposed BOCR-WP system could give a high accuracy performance in language identification that is 99.99% on the average. In recognition processing, the BOCR-WP system is evaluated by comparing with the results of the conventional approach. In comparison, both methods could yield the results with the utmost accuracy of 100% in Thai document and could also yield with a quite high accuracy above 98% on average for the whole document sets.

In system speed, the average processing time of the BOCR-WP system is 0.51 ms per character. The experimental studies of the BOCR-WP as in Table 6.9 could outperform its peer in terms of speed enhancement by a big margin with 22.20% on the average and 15.69% as the very minimal.

In hard disk storage, the average image storage of one A4 document is 8.29 MB per page and the program storage is 8.98 MB. In addition, the database storage of the additional information used only in the BOCR-WP system is 9.82 MB as shown in Table 7.1. Therefore, in the BOCR-WP system, the overall storage used for recognizing one page of document image is 8.29 + 8.98 + 9.82 = 27.09 MB. Obviously, an extra storage of 9.82 MB is quite negligible in today hard disk capacity and space availability.

Table 7.1: Database storage of the additional information used in the BOCR-WP system

Additional Information	Database Storage (MB)
Thai Dictionary	0.625
English Dictionary	0.751
Thai N-grams	3.630
English N-grams	4.790
Thai Character Templates	0.011
English Character Templates	0.013
All	**9.820**

7.1.2 Complexity of Program

To test the complexity of the program, we use Cyclomatic Complexity which is a standard measurement of the overall complexity of source code based loosely on the number of loops and branches that are contained within the source module. Therefore, the BOCR-WP program implemented in Delphi programming language is run on the RTS Cyclomatic Complexity for Pascal (CCP) [25] for testing the program complexity.

The result shows that there are 24 extremely complex modules and 35 highly complex modules from the overall 123 modules due to number of loops and branches. The complexity value of extremely complex modules is bigger than or equal to 55 and the complexity value of the highly complexity ones is 22 - 50. The first two highest complexity modules are feature extraction and segmentation which these complexity values are equal to 821 and 708 respectively.

7.1.3 Lessons Learnt

From the problems of the BOCR-WP implementation, there are several factors contributed to build an effective BOCR-WP system as discussed in the following:

1. *r/v* ratio, it is the speed ratio between the average recognition time per character versus the average verification time per character. Obviously, the higher is the better as to signify the gain in speed improvement; otherwise, for a low ratio of *r/v*, it is useless to implement the system at all. At the first stage of development we were able to obtain the ratio of 3.1 as the very best. Later after several refinements in the verification process, the ratio is steeply improved, and 5 is the recent figure in the experiment.

2. Predictive words, they are those probable words of occurrence. It is the main issue for designing the BOCR-WP with effectiveness, as they should be found and retrieved correctly to form a list of candidacies for later use in the verification process. If they could not be retrieved correctly, the searching for a match would be ended up in vain. Then another case happens, if the input document images contain all randomized characters, which are so unlikely to happen in practical situation. In that case, however, the BOCR-WP is not applicable at all as it is definitely not possible to find any correct predictive words in this worst case scenario.

3. Order of candidacies. The search process should be guided in order instead of a blind search. In order to shorten the search time, it is better to rank those candidacies of probable words according to their probabilistic orders of occurrence. The highest one should be listed on the top and be matched first. This simple strategy proves its effectiveness in the BOCR-WP implementation.

4. η, the upper limit number. It is the number to limit the depth of the search process, and there is no fixed rule to set this number. If the number is set too low, it will be simply a shallow search, and most of the cases, it will end up with nothing. In opposite, if the number is set too high, it will be a long march search with so many attempts to match, and of course, a sheer waste of time. Presently, η is set to 5 as the integer of *r/v*, it works quite satisfactorily in our experiments.

7.2 Conclusion

Based on the experimental results mentioned in Chapter VI, the BOCR-WP system in the design could fulfill the requirements and objectives as described in the problem statement with quite an outstanding achievement: coverage of bilingual documents in Thai and English, maintaining high accuracy in character recognition, and the last, speed improvement with significance in recognition time in both Thai and English.

7.3 Future Works

We would like to suggest the following possible directions for the future works:

1. Expand the work to various different scripts, fonts, sizes and styles instead of bilingual scripts and 4 fonts as the limit.

2. Apply other approaches in character recognition methods instead of the rule based technique.

3. Improve the work in word prediction methodology with higher level of linguistic knowledge, for instance, applying word based N-gram in the word prediction process.

REFERENCES

1. Juan Cheng et al, “Script identification of document image analysis”, Proceedings of the 1st International Conference on Innovative Computing, Information Control, 2006, pp. 178-181.
2. P.A. Vijaya and M.C. Padma, “Text line identification from a multilingual document”, Proceedings of the International Conference on Digital Image Processing, 2009, pp. 302-305.
3. S. Chanda et al, “Word-wise Thai and Roman Script Identification”, Proceedings of ACM Transactions on Asian Language Information Processing, Vol.8, No. 3, Article 11, 2009, pp. 1-21.
4. S. Srisuk, “Thai printed character recognition using the Hausdorff distance”, Proceedings of National Computer Science and Engineering Conference (NCSEC), 1999.
5. A. Kawtrakul and P. Waewsawangwong, “Multi-feature extraction for printed Thai character recognition”, Proceeedings of the 4th Symposium on Natural Language Processing (SNLP), 2000.
6. P. Le-wan and A. Kawtrakul, “Thai character recognition by using hybrid techniques: fuzzy C-means and K-means”, Master Degree Thesis, Department of Computer Exgineeering, Kasetsart University, Thailand, 2001.
7. R. Foopratheepsiri et al, “An improved accuracy method for Thai-OCR using fuzzy logic”, Proceedings of the National Computer Science and Engineering Conference, 2004.
8. P. Phokharakul and C. Kimpan, “Recognition of handprinted Thai characters using the cavity features of character based on neural network”, Proceedings of the 1998 IEEE Asia-Pacific Conference on Circuits and Systems, 1998, pp.149-152.

9. B. Kijsirikul and S. Sinthupinyo, “Approximate ILP rules by backpropagation neural network: a result on Thai character recognition”, Springer-Verlag Berin Heidelberg, 1999, pp. 162-173.

10. A. Thammano and P. Duangphasuk, “Printed Thai character recognition using hierarchical cross-correlation ARTMAP”, Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, 2005.

11. S. Mitatha et al, “Experimental results of using rough sets of printed Thai characters recognition”, Proceedings of IEEE Region 10th International Conference on Electrical and Electronic Technology, 2001, pp. 331-334.

12. S. Mitatha et al, “Some experimental results of using rough sets for printed Thai characters recognition”, International Journal of Computational Cognition, Vol. 1, 2003, pp. 109-121.

13. S. Tangwongsan and O. Jungthanawong, “A refinement of stroke structure for printed Thai character recognition”, Proceeedings of the 9th International Conference on Signal Processing (ICSP), 2008, pp. 1504-1507.

14. C. Pornpanomchai and M. Daveloh, “Printed Thai character recognition by genetic algorithm”, Proceedings of the 6th Conference on Machine Learning and Cybernetics, 2007, pp. 3354-3359.

15. Q. Huo and Z.D. Feng, “Improving Chinese/English OCR performance by using MCE-based character-pair modeling and negative training”, Proceedings of the 7th International Conference on Document Analysis and Recognition, 2003, pp. 364-368.

16. R.S. Kunte and R.D.S. Samuel, “A bilingual machine-interface OCR for printed Kannada and English text employing wavelet features”, Proceedings of the 10th International Conference on Information Technology, 2007, pp. 202-207.

17. Kai Wang et al, “High performance Chinese/English mixed OCR with character level language identification”, Proceedings of the 10th International Conference on Document Analysis and Recognition, 2009, pp. 406-410.

18. B. Krutrachue and P. Piyatrakul, “Automatic Thai and English fonts identification without character recognition”, Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 2001, pp. 603-606.

19. The Royal Institute of Thailand, “The Royal Institute Dictionary B.E. 2542”, Retrieved on 6 November 2007 from http://rirs3.royin.go.th/.

20. D.A. Balota et al, “The English Lexicon Project”, Behavior Research Methods, 2007, pp. 445-459.

21. D. Nadeem and S. Rizvi, “Character recognition using template matching”, Submitted in partial fulfillment for the award of the Bachelor of Information Technology Degree, Department of Computer Science, Jamia, Millia, Islamia, 2002.

22. S. Watcharabutsarakham, “Using projection and loop for segmentation of touching Thai typewritten”, International Symposium on Communications and Information Technology (ISCIT), 2004, pp. 504-508.

23. Jun-Sik Kwon et al, “An enhanced thinning algorithm using parallel processing”, Proceedings of the International Conference on Image Processing, 2001, Vol. 3, pp. 752-755.

24. National Electronics and Computer Technology Center (NECTEC), “Smart Word Analysis for Thai (SWATH)”, NECTEC, 1995, Retrieved on 21 January 2010 from http://hlt.nectec.or.th/products/swath.php.

25. Ryan Vanidrstine, “Run-time system cyclomatic complexity calculator for Delphi”, 2005, Retrived on 9 March 2010 from http://www.run-time-systems.com/ccp/.

BIOGRAPHY

NAME	Miss. Buntida Suvacharakulton
DATE OF BIRTH	3 June 1982
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Silpakorn University, 2001-2004 Bachelor of Science (Computer Science) Mahidol University, 2005-2009 Master of Science (Computer Science)
HOME ADDRESS	2/2 Taksin Road, Cheongneon, Muang, Rayong, 21000, Thailand
EMPLOYMENT ADDRESS	Mahidol University, Bangkok, Thailand Tel. 0-2354-4333 E-mail : itbsv@mahidol.ac.th