การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย

SEGMENTATION OF VOWEL AND TONE FROM THAI CHARACTER STRING

นายศุภกร รัตนปราการ
SUPAKORN RATTANAPRAKARN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2542
ISBN 974-622-450-6

การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย

SEGMENTATION OF VOWEL AND TONE FROM THAI CHARACTER STRING

นายศุภกร รัตนปราการ

SUPAKORN RATTANAPRAKARN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชา วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2542

ISBN 974-622-450-6

SEGMENTATION OF VOWEL AND TONE FROM THAI PRINTED CHARACTER STRING

SUPAKORN RATTANAPRAKARN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE AND

INFORMATION TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

1999

ISBN 974-622-450-6

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย
SEGMENTATION OF VOWEL AND TONE FROM THAI PRINTED CHARACTER STRING

ชื่อนักศึกษา นายศุภกร รัตนปราการ

รหัสประจำตัว 35628016

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ

อาจารย์ผู้ควบคุมวิทยานิพนธ์ ผศ.ดร.บุญธีร์ เครือตราชู

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ผศ.ดร.บุญธีร์	เครือตราชู	
รศ.ดร.วิเชียร	เปรมชัยสวัสดิ์	
รศ.ดร.ชม	กิ้มปาน	
รศ.ดร.บุญวัฒน์	อัตชู	
ดร.วรพจน์	กรีสุระเดช	

วัน/เดือน/ปี ที่สอบ 7 พฤษภาคม 2542 **เวลา** 10.30 น. **เป็นต้นไป**

สถานที่สอบ ห้องบรรยาย 234 ชั้น 2 คณะเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.มนัส สังวรศิลป์)

คณบดีบัณฑิตวิทยาลัย

วันที่ 27 **เดือน** พฤษภาคม **พ.ศ.** 2542

หัวข้อวิทยานิพนธ์	การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย
นักศึกษา	นายศุภกร รัตนปราการ
รหัสประจำตัว	35628016
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
พ.ศ.	2542
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ.ดร.บุญธีร์ เครือตราชู

บทคัดย่อ

รูปแบบตัวอักษรภาษาไทย ประกอบด้วย 4 ระดับ ได้แก่ ระดับเหนือบน, ระดับบน, ระดับกลาง และระดับล่าง หมายถึง ตัวอักษรซึ่งประกอบด้วย พยัญชนะ,สระ และวรรณยุกต์ ที่อยู่ทั้งด้านบน และด้านล่าง ในบางครั้งอาจทำให้เกิดการสัมผัส หรือซ้อนทับกันของตัวอักษร ซึ่งมีได้หลายรูปแบบ สิ่งพิมพ์หลายประเภทมีปริมาณตัวอักษรที่ติดกันในระดับบน และล่างมากกว่าที่จะเกิดการติดกันในระดับกลาง ดังนั้นการวิจัยนี้ได้มุ่งเน้นการนำเสนอส่วนของการจัดเตรียมข้อมูล ในขั้นตอนการตัดแยกตัวอักษรภาษาไทยออกจากภาพของประโยคในระดับที่นอกเหนือจากระดับกลาง โดยแบ่งเป็น 3 ขั้นตอน ได้แก่ ขั้นตอนการแยกภาพตัวอักษรด้วยวิธีการเปลี่ยนรหัสขอบ ในขั้นตอนนี้อาศัยการติดตามขอบ และทำการแทนที่ขอบด้วยเครื่องหมายที่แสดงตำแหน่งของขอบด้านซ้าย ด้านขวาและมุม เพื่อสามารถที่จะทำการคัดลอกตัวอักษรที่เหลื่อมล้ำกันได้ ขั้นตอนการวิเคราะห์การติดกันของตัวอักษรและเสนอแนวทางการตัดแยก โดยอาศัยระดับของตัวอักษรเพื่อทำการแบ่งประเภทของตัวอักษรที่ติดกัน ทำให้สามารถแบ่งแยกประเภทของตัวอักษรที่ติดกันได้ 7 ประเภท จากนั้นใช้การหาค่าอีสโตแกรมแบบต่างๆ ได้แก่ Pixel Projection, Profile Projection และ Modify Pixel Projection มาทำการวิเคราะห์ อาศัยแนวการตัดแยกที่ได้จากการวิเคราะห์ ทำการตัดแยกในขั้นตอนการตัดแยกตัวอักษรโดยใช้วิธีการหาค่า Break Cost ที่ต่ำที่สุดเพื่อกำหนดจุดตัดของตัวอักษรที่ติดกัน สรุปผลการทดสอบกับข้อมูลตัวอักษรตัวพิมพ์ จากหนังสือพิมพ์ นิตยสาร และวารสาร จำนวนตัวอักษร 8,570 ตัวอักษร สามารถวิเคราะห์ตัวอักษรที่ติดกันถูกต้อง 92 % และตัดแยกได้ถูกต้อง 87 % ด้วยวิธีการที่นำเสนอนี้ทำให้สามารถวิเคราะห์และตัดแยกภาพตัวอักษรตัวพิมพ์ภาษาไทยที่ติดกันได้โดยให้มีการเปรียบเทียบโดยการรู้จำให้น้อยที่สุด

Thesis Title	Segmentation of vowel and tone from Thai printed character string
Student	Mr. Supakorn Rattanaprakarn
Student ID.	35628016
Degree	Master of Science
Programme	Computer Science and Information Technology
Year	1999
Thesis Advisor	Asst. Prof. Dr. Boontee Kruatrachue

ABSTRACT

Thai characters consist of 4 levels: above upper, upper, middle and lower for consonants, vowels and tones. Due to these many levels of characters, there are many forms of crossing or touching of characters. Many publishing has a larger amount of touching or crossing characters in vertical than horizontal. Therefore, this research proposed the method in 'Preprocessing Process', Segmentation process of the Thai characters in the position aside from the middle level. The segmentation process can be divided into 3 steps: Image Segmentation step, Connected Characters Detection and propose segmented point step, and Character Segmentation step.

In the Image Segmentation Step, we segment the image of Thai character strings by encoding the boundary along the contour of the character strings with the left, right and corner codes. This method to solve overlapping of the characters. In the Connected Characters Detection and segmented point step, we can devide the characters into 7 types by using the character level, and using the histogram value of the consonant physical characteristics in many techniques such as Pixel Projection, Profile Projection and Modify Pixel Projection to analyze the connected characters. And use segmented point which get from analyzing. In the Character Segmentation step, we apply minimum Break Cost value to settle the segmented point of connected characters.

The experimental result, analyzing 8,570 characters from newspapers, magazines and journals, the Connected Characters Detection accuracy rate is 92% and the segmentation accuracy rate is 87%. This proposed method can analyze and segment the Thai printed connected character by using less recognition process.

กิตติกรรมประกาศ

การจัดทำวิทยานิพนธ์ในครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี เพราะได้รับความเมตตากรุณาจากท่าน อาจารย์ ผศ. ดร.บุญธีร์ เครือตราชู ซึ่งได้ให้คำปรึกษา และแนะนำผู้วิจัยตลอดมา ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่าน และกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณ คุณบุญช่วย เจ้าหน้าที่ที่ดูแลนักศึกษาที่ได้ให้ข้อมูลข่าวสารตลอดจนแนะนำวิธีการต่างๆ

ขอขอบคุณ เพื่อนๆที่คอยให้กำลังใจ และความช่วยเหลือเป็นอย่างดี เจ้าหน้าที่สารสนเทศทุกท่าน ที่ได้ให้ความช่วยเหลือทุกๆ ท่าน

และขอขอบคุณเป็นพิเศษสำหรับ คุณอุมาภรณ์ สร้อยประเสริฐ สำหรับทุกสิ่งทุกอย่าง

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ศุภกร รัตนปราการ

สารบัญ(ต่อ)

หน้า

3.2.2 การหาค่าตำแหน่งพิกัด..... 24

3.2.3 การคัดลอกตัวอักษร..... 25

บทที่ 4 การแยกสระและวรรณยุกต์ระดับบนล่างออกจากภาพประโยค..... 27

4.1 การกำหนดประเภทการติดกันของตัวอักษร..... 27

4.2 การกำหนดตัวอักษรที่อยู่ระดับกลาง..... 32

4.3 ข้อสังเกตของตัวอักษรในระดับกลาง..... 34

4.4 การวิเคราะห์การติดกันของภาพตัวอักษร..... 35

4.5 การวิเคราะห์การเหลื่อมล้ำของตัวอักษร..... 36

4.6 การตรวจสอบและแนวทางการตัดแยกตัวอักษร..... 37

4.7 การตัดแยกตัวอักษรที่ติดกัน..... 52

4.8 การเรียงรูปประโยคหลังการรู้จำ..... 54

บทที่ 5 ผลการทดสอบและปัญหาที่พบ..... 57

5.1 ผลการทดสอบความถูกต้องของการวิเคราะห์การติดกันของตัวอักษร..... 57

5.2 ผลการทดสอบความถูกต้องของการตัดแยกตัวอักษร..... 58

บทที่ 6 สรุปผลการวิจัย และข้อเสนอแนะ..... 60

6.1 สรุปผลงานวิจัย..... 60

6.2 แนวทางในการพัฒนาในอนาคต..... 61

เอกสารอ้างอิง..... 62

ภาคผนวก ก..... 63

ภาคผนวก ข..... 65

ประวัติผู้เขียน..... 75

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	I
บทคัดย่อภาษาอังกฤษ	II
กิตติกรรมประกาศ	IV
สารบัญ	V
สารบัญตาราง	VII
สารบัญภาพ	VIII
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการศึกษา	1
1.3 สมมติฐานของการศึกษา	2
1.4 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย	4
1.5 ขอบเขตของการดำเนินการวิจัย	6
1.6 วิธีที่ใช้ในการดำเนินการวิจัย	6
1.7โครงสร้างวิทยานิพนธ์	7
บทที่ 2 หลักการทำงานของทฤษฎี	10
2.1 ลักษณะตัวอักษรภาษาไทย	10
2.2 ลักษณะตัวพิมพ์ภาษาไทย	11
2.3 การหาค่า Histogram ของภาพตัวอักษร	12
2.3.1 Pixel Projection	12
2.3.2 Profile Projection	13
2.3.3 Modify Pixel Projection	13
2.4 การหาค่า Break Cost เพื่อการกำหนดจุดตัด	15
บทที่ 3 การแยกภาพตัวอักษรออกจากภาพประโยคโดยวิธีการเปลี่ยนรหัสขอบ	17
3.1 การกำหนดแนวทางและลำดับการค้นหาภาพตัวอักษรในแต่ละระดับ	17
3.2 การแยกภาพตัวอักษร	19
3.2.1 การกำหนดรหัสจุดขอบ	20

สารบัญตาราง

ตารางที่	หน้า
2.1. แสดงประเภทของตัวอักษรเดี่ยว	11
4.1 แสดงการคำนวณประเภทของตัวอักษร	29
4.2 แสดงประเภทการติดกันของตัวอักษรและแนวทางการตัดแยก	30
4.3 แสดงกลุ่มพยัญชนะและสระ ในระดับกลาง	33
4.4 แสดงตัวอักษรเดี่ยวในแต่ละประเภทของตัวอักษร	37
5.1 แสดงผลการวิเคราะห์การติดกันของสิ่งพิมพ์แต่ละประเภท	57
5.2 แสดงสาเหตุและ เปอร์เซ็นต์ความผิดพลาดในการวิเคราะห์	58
5.3 แสดงผลการทดสอบความถูกต้องของการตัดแยกตัวอักษรในสิ่งพิมพ์แต่ละประเภท	58
5.4 แสดงสาเหตุและ เปอร์เซ็นต์ความผิดพลาดในการตัดแยก	59

สารบัญรูป

รูปที่	หน้า
1.1 แสดงการเหลื่อมล้ำกันของตัวอักษรที่ได้จากการแสกน	2
1.2 แสดงการติดกันหรือสัมผัสกันของตัวอักษร	3
1.3 แสดงตัวอักษรซ้อนทับกัน	3
1.4 แสดงผังขั้นตอนการทำงานเริ่มตั้งแต่การแยกตัวอักษร	5
2.1 แสดงภาพประโยคของภาษาไทย	9
2.2 แสดงการจัดเรียงของสระในระดับบนของตัวอักษร	11
2.3 แสดงตัวอักษรที่มีความสูงถึงระดับเหนือบน	12
2.4 แสดงตัวอักษรติดกันที่ความกว้างไม่เกิน 1 ตัวอักษร	12
2.5 แสดงกราฟ Vertical Pixel Projection	13
2.6 แสดงกราฟ Horizontal Profile Projection	13
2.7 แสดงกราฟ Horizontal Modify Pixel Projection	15
2.8 แสดงนัยสำคัญของการสัมผัสกัน	15
3.1 แสดงภาพระดับของตัวอักษรและเส้นแบ่งระดับ	17
3.2 แสดงการแทรกภาพที่ได้จากการตรวจหาครั้งที่ 2	18
3.3 แสดงการแทรกภาพที่ได้จากการตรวจหาครั้งที่ 3	18
3.4 แสดงตัวอักษรที่เรียงลำดับอย่างถูกต้องในแต่ละระดับ	19
3.5 แสดงกรอบของตัวอักษร	19
3.6 แสดงการคัดลอกโดยใช้รหัสขอบเพียงอย่างเดียว	20
3.7 แสดงรหัสทิศทางของฟรีแมน	21
3.8 แสดงมุมปลายด้านบน	22
3.9 แสดงมุมปลายด้านล่าง	22
3.10 แสดงจุดปลายที่มีการเปลี่ยนรหัส	22
3.11 แสดงมุมเว้าด้านบน	23
3.12 แสดงมุมเว้าด้านล่าง	23
3.13 แสดงขอบในแนวนอน	23
3.14 แสดงรหัสขอบของตัวอักษร	24
3.15 แสดงขอบเขตของตัวอักษร	25
4.1 แสดงการแบ่งประเภทตัวอักษรตามความสูง	27

สารบัญรูป(ต่อ)

รูปที่		หน้า
4.2	แสดงการแบ่งระดับและกราฟฮีสโตแกรม	32
4.3	แสดงการหาจำนวนภูเขาของฮีสโตแกรม	33
4.4	แสดงการหาแนวเส้นตัวอักษรระดับกลาง	34
4.5	แสดงขั้นตอนการวิเคราะห์การติดกันของตัวอักษร	35
4.6	แสดงสรุปเส้นทางการวิเคราะห์และตัดแยก	36
4.7	แสดงการเหลื่อมล้ำด้านหลังและด้านหน้า	36
4.8	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 1	38
4.9	แสดงลักษณะที่ทำให้การวิเคราะห์ความกว้างผิดพลาด	38
4.10	แสดงตัวอย่างการตัดผิดพลาด	38
4.11	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 2	39
4.12	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 3	39
4.13	แสดงแนวการตัดในกลุ่ม("บ","พ","ผ")	40
4.14	แสดงการติดโดยความกว้างไม่เกิน 1 ตัวอักษร	41
4.15	แสดงการติดโดยความกว้างเกิน 1 ตัวอักษร	41
4.16	แสดงตัวอักษรกลุ่ม ("พ","บ","ผ") ที่มีการเหลื่อมล้ำ	42
4.17	แสดงการวิเคราะห์ผิดพลาดให้ผลเป็นอักษรเดี่ยว	42
4.18	แสดงการวิเคราะห์ที่ผิดทำให้การตัดผิดพลาด	43
4.19	แสดงการวิเคราะห์ผิดพลาดเนื่องจากขนาดของตัวอักษร	43
4.20	แสดงขั้นตอนในการตรวจสอบประเภท 3	44
4.21	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 4	45
4.22	แสดงภาพเปรียบเทียบตัวอักษรเดี่ยว และ “ก” ที่ติดกับสระ“อุ”	46
4.23	แสดงภาพเปรียบเทียบตัวอักษรเดี่ยว และ “ภ” ที่ติดกับสระ“อู”	46
4.24	แสดงขั้นตอนในการตรวจสอบประเภทที่ 4	47
4.25	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 5	48
4.26	แสดงขั้นตอนในการตรวจสอบประเภทที่ 5	49
4.27	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 6	50
4.28	แสดงขั้นตอนในการตรวจสอบประเภทที่ 6	51
4.29	แสดงผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 7	52

สารบัญรูป(ต่อ)

รูปที่ หน้า

4.30 แสดงการหาเส้นตัดในแนวนอน และขอบเขตของการคำนวณ..... 53

4.31 แสดงการหาเส้นตัดในแนวตั้ง และขอบเขตของการคำนวณ..... 53

4.32 แสดงลำดับของตัวอักษรที่ถูกตัดแยกแล้วในแต่ละระดับ..... 54

4.33 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 2..... 55

4.34 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 1..... 55

4.35 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 4..... 55

4.36 แสดงลำดับของตัวอักษรที่เรียงอย่างถูกต้อง..... 56

บทที่ 1
บทนำ

1.1 ความเป็นมา และความสำคัญของปัญหา

จากผลงานวิจัยที่ผ่านๆมา มีการมุ่งเน้นทางด้านการคิดค้นวิธีการรู้จำตัวอักษรให้ได้ประสิทธิภาพ และความถูกต้องของการรู้จำสูงสุดเป็นส่วนใหญ่ ไม่ค่อยได้กล่าวถึงส่วนของการเตรียมข้อมูลสำหรับการรู้จำ หรืออาจกล่าวไว้เพียงคร่าวๆ ซึ่งส่วนของการเตรียมข้อมูลนี้เป็นขั้นตอนหนึ่งของส่วนการจัดการล่วงหน้าของระบบ (Preprocessing Process) ที่มีความสำคัญไม่น้อยไปกว่าส่วนของการรู้จำ (Recognition) ในการประยุกต์ใช้งานจริงของระบบการรู้จำตัวอักษร อินพุทของระบบจะมีลักษณะเป็นเอกสารบนแผ่นกระดาษที่บรรจุข้อความเป็นบรรทัดหรือเป็นย่อหน้า เพื่อให้ระบบการรู้จำตัวอักษรทำงานได้อย่างมีประสิทธิภาพและถูกต้อง จำเป็นจะต้องมีขบวนการทำงานเข้ามาจัดการกับงานในส่วนนี้ก่อนที่จะเข้าสู่ขั้นตอนการรู้จำ นั่นคือส่วนการจัดการล่วงหน้า โดยจะทำการวิเคราะห์และระบุส่วนประกอบหน้าเอกสาร ที่ประกอบด้วยตัวอักษรเรียงต่อกันเป็นคำ เป็นข้อความ เป็นประโยค และเป็นย่อหน้า ก็จะต้องมีขบวนการทำงานที่สามารถแยกตัวอักษรให้ได้เป็นตัวอักษรเดี่ยวๆ ออกจากภาพของข้อความ (Segmentation) ก่อนส่งต่อไปให้ในขั้นตอนของการรู้จำ

ขั้นตอนการแยกตัวอักษรออกจากภาพของข้อความ (Segmentation) เป็นขั้นตอนการดึงภาพเฉพาะ 1 ตัวอักษร ออกมาจากภาพประโยคของเอกสารในแต่ละบรรทัด โดยเฉพาะตัวอักษรไทย การแยกตัวอักษรแบบอัตโนมัติยิ่งมีความซับซ้อนเพิ่มขึ้น เนื่องจากในแต่ละประโยคประกอบด้วย พยัญชนะ, สระ และวรรณยุกต์ ที่มีระดับต่างๆถึง 4 ระดับ และในขั้นตอนนี้อาจจะพบตัวอักษรที่สัมผัสหรือซ้อนทับกันทั้งในระดับเดียวกัน (ตามแนวนอน) และในระดับต่างกัน (ตามแนวตั้ง) จึงจำเป็นต้องมีกระบวนการเพื่อตัดแยกตัวอักษรที่ติดกัน ซึ่งการติดกันของตัวอักษรมักจะมีสาเหตุมาจากกลไกในการพิมพ์, ขนาด (size), สิ่งรบกวน (noise) หรือแม้แต่รูปแบบของตัวอักษร

จากผลที่ได้ทำการเก็บรวบรวมข้อมูลที่แสดงไว้ในภาคผนวก ก. พบว่า รูปแบบการติดกันของตัวอักษรในระดับบนและล่างมีหลายรูปแบบยากต่อการวิเคราะห์ และยังพบอีกว่าอัตราการติดกันของตัวอักษรในระดับบนและล่างของสิ่งพิมพ์บางประเภท เช่น นิตยสาร มีสูงถึง 90.37 % ดังนั้นผู้วิจัยจึงนำเสนอหัวข้อวิจัยเพื่อค้นหาถึงวิธีวิเคราะห์การติดกันของตัวอักษรตัวพิมพ์ภาษาไทย พร้อมกับวิธีการตัดแยกตัวอักษรในระดับบนและระดับล่าง

1.2 วัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาถึงวิธีการที่เหมาะสม ให้สามารถแยกตัวอักษรตัวพิมพ์ภาษาไทย ออกจากภาพของประโยคได้อย่างถูกต้อง

2. เพื่อศึกษาถึงลักษณะบางประการของตัวพยัญชนะ ที่จะสามารถจัดกลุ่ม เพื่อนำคุณสมบัตินี้มาใช้ในการแก้ปัญหา การติดกันของภาพตัวอักษรตัวพิมพ์ภาษาไทย ในระดับที่บนและระดับล่าง ยกเว้นระดับกลางได้

3. เพื่อศึกษาถึงวิธีการวิเคราะห์ และตัดแยกภาพของตัวอักษรตัวพิมพ์ภาษาไทยที่ติดกันให้ถูกต้อง โดยไม่จำเป็นต้องส่งไปทำการเปรียบเทียบโดยการรู้จำ(Recognition) หรือเพื่อให้มีการเปรียบเทียบโดยการรู้จำให้น้อยที่สุด

1.3 สมมติฐานของการศึกษา

ภาพของประโยคภาษาไทยที่ได้จากการสแกน เพื่อที่จะนำไปเข้าขบวนการแยกภาพตัวอักษรให้ได้ตัวอักษรเดี่ยวๆ ออกมานั้น ในบางครั้งภาพของกลุ่มตัวอักษรที่ได้มีลักษณะที่ยากต่อการแยกตัวอักษร ซึ่งมีลักษณะต่างๆ ดังนี้

1. **ตัวอักษรเหลื่อมล้ำกัน (Overlap Character)** คือการที่บางส่วนของตัวอักษรสองตัวที่อยู่ติดกันมีลักษณะการวางที่เหลื่อมกัน กรณีนี้มักเกิดกับบางชนิดของตัวอักษร หรือขั้นตอนการเรียงพิมพ์ ตัวอย่างของการเหลื่อมล้ำกันของตัวอักษร ที่ได้จากการแสกน แสดงดังรูปที่ 1.1

รูปที่ 1.1 แสดงการเหลื่อมล้ำกันของตัวอักษรที่ได้จากการแสกน

2. **ตัวอักษรสัมผัสกัน (Touching Character)** คือการที่บางส่วนของตัวอักษรมีการติดกันทั้งในระดับเดียวกันและต่างระดับกัน กรณีนี้สาเหตุส่วนใหญ่จะเกิดจากคุณภาพของการสแกน และบางส่วนเกิดจากรูปแบบของตัวอักษรเอง ตัวอย่างของการติดกันหรือสัมผัสกันของตัวอักษรแสดงดังรูปที่ 1.2

รูปที่ 1.2 แสดงการติดกันหรือสัมผัสกันของตัวอักษร

3. ตัวอักษรซ้อนทับกัน (Crossing Character) คือการที่มีการซ้อนทับกันของตัวอักษร ซึ่งในกรณีนี้ส่วนใหญ่ที่พบเนื่องมาจากชนิดของตัวอักษร ตัวอย่างของตัวอักษรซ้อนทับกัน แสดงดังรูปที่ 1.3

รูปที่ 1.3 แสดงตัวอักษรซ้อนทับกัน

จากปัญหาทั้งสามแบบที่อาจจะเกิดขึ้นในขั้นตอนการแยกภาพตัวอักษรออกจากภาพประโยค (Segmentation) งานวิจัยนี้จึงนำเสนอสมมติฐานที่จะนำมาใช้ ในการแยกภาพตัวอักษรออกจากภาพของประโยค และแก้ปัญหาในแต่ละแบบ โดยนำเสนอการทำงานเป็นสามขั้นตอนคือ ขั้นตอนการแยกภาพตัวอักษรออกจากภาพประโยคให้ได้ภาพตัวอักษรเดี่ยว ขั้นตอนการวิเคราะห์การติดกันของตัวอักษรและเสนอแนวทางการตัดแยก และขั้นตอนการตัดแยกตัวอักษรที่ติดกัน ในขั้นตอนการแยกภาพตัวอักษรออกจากภาพประโยค ให้ได้ภาพตัวอักษรเดี่ยวนั้นเสนอสมมติฐานที่ว่าวิธีการติดตามขอบของตัวอักษรสามารถทำให้ทราบขอบของตัวอักษรตัวพิจารณาอยู่ เมื่อนำมาประยุกต์ใช้ด้วยวิธีการเปลี่ยนรหัสขอบของตัวอักษรให้ทราบถึงขอบด้านซ้าย ด้านขวา และมุมของตัวอักษร การคัดลอกตัวอักษรโดยอาศัยรหัสขอบของเฉพาะตัวอักษรตัวที่พิจารณา ทำให้ทราบว่าพื้นที่ใดเป็นตัวอักษรที่ต้องการคัดลอกหรือพื้นที่ใดเป็นตัวอักษรที่เหลื่อมล้ำกันอยู่จึงสามารถแก้ปัญหาการเหลื่อมล้ำของตัวอักษรได้ ภายใต้เงื่อนไขที่ว่าภาพตัวอักษรที่แยกออกมาได้อาจเป็นตัวอักษรที่มีการติดกัน ดังนั้นในขั้นตอนการวิเคราะห์การติดกันของตัวอักษรเพื่อแก้ปัญหาการติดกันของตัวอักษรทั้งแบบสัมผัสหรือซ้อนทับกันของตัวอักษรจากสมมติฐานหลายข้อดังนี้

1. ตัวอักษรภาษาไทยมีหลายระดับ เราสามารถจำแนกตามระดับความสูงของตัวอักษรที่อาจจะเกิดจากตัวอักษรเดี่ยวหรือตัวอักษรที่ติดกัน ทำให้เราสามารถแยกพิจารณาโอกาสและรูปแบบการติดกันในแต่ละประเภทได้ ภายใต้เงื่อนไขที่ภาพประโยคจะต้องเป็นตัวอักษรตัวพิมพ์ภาษาไทยที่มีขนาดเดียวกันทั้งหมด และต้องการกำหนดตำแหน่งของเส้นแบ่งระดับของตัวอักษรมาให้

2. ตัวอักษรที่อยู่ในระดับกลางสามารถจำแนกเป็นกลุ่มได้ โดยอาศัยลักษณะทางกายภาพที่เด่นชัดคือ แนวเส้นตรงของตัวอักษรทั้งในแนวตั้งและแนวนอน ซึ่งสามารถแสดงคุณสมบัตินี้ให้ปรากฏได้ด้วยวิธีการหาค่าฮีสโตแกรมแบบต่างๆในแนวตั้งและแนวนอนของตัวอักษร ภายใต้เงื่อนไขที่ตัวอักษรจะต้องเป็นตัวอักษรที่ไม่มีลักษณะเป็นตัวเอียง (Italic font)

3. การเหลื่อมล้ำของตัวอักษรไม่ควรมีความกว้างของส่วนที่เลื่อมล้ำเกินกว่าหนึ่งในสามของขนาดความกว้างของตัวอักษร ถ้าเกินให้ถือว่ามีการติดกันของตัวอักษร

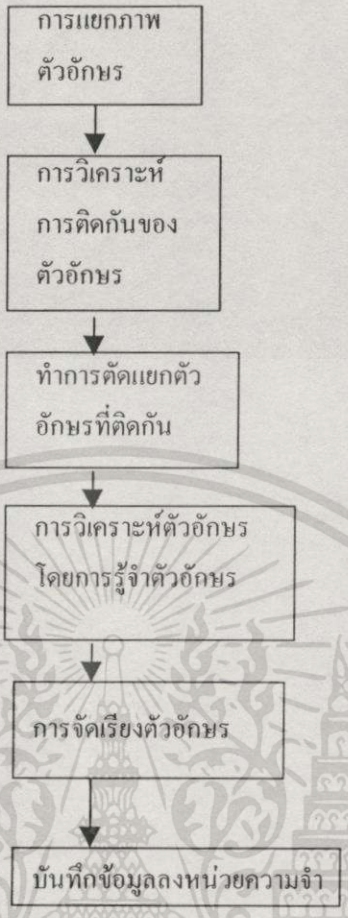
4. จากกลุ่มของตัวอักษรที่แบ่งด้วยวิธีการหาค่าฮีสโตแกรม สามารถจำแนกคุณสมบัติของแต่ละกลุ่ม เพื่อเป็นข้อสังเกตุในการวิเคราะห์การติดกันของตัวอักษรในแต่ละประเภทความสูงของตัวอักษรและกำหนดแนวทางการตัดแยกได้

5. ในบางตัวอักษรที่ไม่สามารถใช้ระดับความสูงและการหาค่าฮีสโตแกรมมาทำการวิเคราะห์ได้ ก็จำเป็นต้องนำขั้นตอนการรู้จำเข้ามาช่วยในการวิเคราะห์ภาพตัวอักษร

6. การวิเคราะห์และกำหนดแนวทางการตัดแยกของตัวอักษรที่ซ้อนทับกัน ให้ใช้วิธีการเดียวกับที่ใช้กับตัวอักษรที่สัมผัสกัน ทำให้การตัดแยกตัวอักษรจะได้ตัวอักษรที่มีบางส่วนขาดไป ดังนั้นขั้นตอนการรู้จำที่จะนำการตัดแยกวิธีการนี้ไปใช้ควรอยู่ภาพใต้เงื่อนไขการรับรู้ตัวอักษรในลักษณะเช่นนี้ด้วย .

1.4 แนวคิดที่ใช้ในการวิจัย

ในขั้นตอนการแยกตัวอักษรออกจากประโยค สามารถทำการแบ่งเป็นขั้นตอนย่อยๆ ได้ 3 ขั้นตอน ได้แก่ ขั้นตอนที่ 1 การแยกภาพตัวอักษรโดยวิธีเปลี่ยนรหัสขอบ ในขั้นตอนนี้ผลของการแยกอาจได้ภาพที่ตัวอักษรติดกัน ขั้นตอนที่ 2 เป็นขั้นตอนที่ทำการวิเคราะห์ภาพตัวอักษรที่ได้ว่ามีการติดกันหรือไม่ ในกรณีที่มีการติดกันก็จะเข้าสู่ขั้นตอนที่ 3 ซึ่งเป็นขั้นตอนที่ใช้ในการตัดแยกตัวอักษรออกจากกัน หลังจากผ่านขั้นตอนการตัดแยกตัวอักษร จนได้ตัวอักษรเดี่ยวแล้ว ก็ถือว่าสิ้นสุดขบวนการจัดการล่วงหน้า (Preprocessing Process) ในระบบ OCR ขั้นตอนต่อมาก็คือ การวิเคราะห์ตัวอักษรด้วยทฤษฎีการรู้จำ เพื่อสามารถจะระบุตัวอักษรได้โดยรหัสเอสกี (ASCII) เพื่อทำการบันทึกข้อมูลลงหน่วยความจำ ก่อนการบันทึกข้อมูลจะต้องมีการจัดเรียงตัวอักษร สำหรับตัวอักษรภาษาไทย จะต้องนำเอาตัวอักษรทั้ง 4 ระดับมาจัดเรียงให้อยู่ในระดับเดียวกันให้ถูกต้องตามลำดับ ขั้นตอนทั้งหมดที่กล่าวมาสามารถสรุปเป็นแผนผังขั้นตอนรูปที่ 1.4



รูปที่ 1.4 แสดงผังขั้นตอนการทำงานเริ่มตั้งแต่แยกตัวอักษร

ดังนั้นเพื่อให้สอดคล้องกับขั้นตอนที่ได้กล่าวมา งานวิจัยนี้จึงนำเสนอแนวคิดในการแยกตัวอักษร ออกเป็นกลุ่มตามระดับของตัวอักษร โดยในขั้นตอนการแยกตัวอักษร เสนอแนวคิดด้วยวิธีการเปลี่ยนรหัสขอบที่สามารถแก้ปัญหาการเหลื่อมล้ำตัวอักษรได้ และการใช้วิธีตัดแยกตามระดับตัวอักษรทำให้ได้แถวของตัวอักษรที่ตัดแยกแล้วแต่ละระดับ เรียงลำดับก่อนหลังอย่างถูกต้อง สำหรับในขั้นตอนการวิเคราะห์การติดกันของตัวอักษร การที่สามารถรู้ระดับของตัวอักษรประกอบกับการเปรียบเทียบขนาดของตัวอักษรกับเส้นแบ่งระดับตัวอักษร และคุณลักษณะทางกายภาพของตัวอักษรไทยที่สามารถแสดงให้เห็นด้วนการแสดงค่าฮีสโตแกรม ผนวกกับการใช้วิธีการรู้จำบางตัวอักษร ทำให้สามารถวิเคราะห์ การติดกันของตัวอักษร และกำหนดแนวทางของการตัดแยก เพื่อที่จะใช้ในขั้นตอนการตัดแยกได้ สุดท้ายงาน วิจัยนี้เสนอแนวคิดของการจัดเรียงตัวอักษรที่อยู่คนละระดับให้อยู่ในระดับเดียวกันได้อย่างสะดวก เพราะแต่ละระดับมีลำดับของการจัดเรียงที่ถูกต้องอยู่แล้ว ไม่จำเป็นต้องมีการจัดเรียงใหม่หมดทั้งบรรทัด

1.5 ขอบเขตของการดำเนินการวิจัย

ขอบเขตของงานวิจัยนี้ จะขึ้นอยู่กับตัวอักษรที่นำมาใช้ในการวิเคราะห์ และการตัดแยก เนื่องจากรูปแบบตัวอักษรตัวพิมพ์ไทยในปัจจุบันมีมากมายหลายรูปแบบ และ หลายขนาด การวิจัยครั้งนี้จึงกำหนดรูปแบบของตัวอักษรที่ใช้ในงานวิจัยมีลักษณะดังต่อไปนี้

1. ตัวอักษรที่ใช้จะต้องเป็นตัวอักษรตัวพิมพ์ภาษาไทย(Thai Printed Character)

2. ขนาดของตัวอักษรใน 1 บรรทัดที่นำมาวิเคราะห์ตัดแยกจะต้องมีขนาดเดียวกัน ในงานวิจัยนี้ไม่ครอบคลุมไปถึงเอกสารที่ประกอบด้วยขนาดของตัวอักษรหลายขนาด(Multi-Size Characters)

3. รูปแบบของตัวอักษร ไม่ครอบคลุมตัวอักษรแบบเอน(Italic Fonts)

4. การวิเคราะห์การตัดแยกไม่ครอบคลุมไปถึงตัวเลขไทย

5. งานวิจัยนี้เป็นการวิเคราะห์ตัดแยกการติดกันของตัวอักษรภาษาไทยเฉพาะในระดับบน และระดับล่างเท่านั้น ซึ่งในระดับกลางได้มีการทำวิจัยเป็นผลงานวิทยานิพนธ์ในหัวข้อ "การตัดแยกตัวอักษรภาษาไทยในระดับกลาง" ซึ่งเรียบเรียงโดย คุณจรรยา เกียรติศิริอนันต์ และ ดร.บุญธีร์ เครือตราชู

1.6 วิธีที่ใช้ในการดำเนินการวิจัย

1.6.1 อุปกรณ์ที่ใช้ในการดำเนินการวิจัย

1. เครื่องแสกนเนอร์ (Scanner) เครื่องแสกนเนอร์ที่ใช้ในงานวิจัยนี้เป็นเครื่องแสกนเนอร์ชนิดตั้งโต๊ะยี่ห้อ Hewlett Packard ผลิตโดยบริษัท Hewlett Packard รุ่น ScanJet 5p สามารถตรวจกวาดเอกสารที่มีขนาดใหญ่ได้ถึงขนาดกระดาษ A4 และสามารถเลือกระดับความละเอียดของภาพ (Scanning Resolution) ได้ตั้งแต่ 75 - 1200 จุดต่อนิ้ว (Dot per Inch) และสามารถปรับค่าความละเอียด ของภาพได้โดยอาศัยซอฟต์แวร์ รูปภาพหรือข้อความที่ตรวจกวาดได้จะถูกจัดเก็บลงในแฟ้มข้อมูลในรูปแบบกราฟิกที่เป็น Windows Bitmap(BMP format)

2. เครื่องคอมพิวเตอร์ (Computer) เครื่องคอมพิวเตอร์ที่ใช้ในงานวิจัยนี้เป็นเครื่องคอมพิวเตอร์ส่วนบุคคล (Personal Computer) ยี่ห้อ Hewlett Packard รุ่น Vectra VL ประกอบด้วย

CPU Pentium ความเร็ว 133 MHz

Memory 32 MB.

Operating System - Windows 95

3. โปรแกรมสำเร็จรูป สำหรับเก็บเอกสารจากเครื่องแสกนเนอร์ โปรแกรมสำเร็จรูป ที่ใช้สำหรับเก็บหน้าเอกสารมีชื่อว่า Adobe Photoshop Version 4.0 ซึ่งประกอบด้วยดีไวซ์ไดรว์เวอร์ (Device Driver) สามารถควบคุมการทำงานของเครื่องแสกนเนอร์ยี่ห้อ Hewlett Packard รุ่น

ScanJet 5p ที่เป็นเครื่องตรวจกวาดที่ใช้ในงานวิจัยนี้ได้เป็นอย่างดี และรูปแบบของการเก็บภาพที่ได้ทำการแสกนสามารถเก็บได้ในหลายรูปแบบได้แก่ TIF, BMP ,PCX และ GIF แต่สำหรับงานวิจัยนี้ได้ทำการเก็บในรูปแบบ BMP

1.6.2 ขั้นตอนการวิจัย

1.6.2.1 ขั้นตอนการเตรียมข้อมูลเพื่อการวิจัย

ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลภาพประโยคของตัวอักษรตัวพิมพ์ภาษาไทยที่มีตัวอักษรขนาดเดียวกันทั้งประโยคและไม่มีรูปแบบตัวอักษรแบบตัวเอียง (Italic) ภาพประโยคที่ได้ก็จะต้องไม่มีลักษณะเอียงเช่นเดียวกัน พร้อมกับมีการระบุตำแหน่งของเส้นแบ่งระดับของตัวอักษรมาให้ด้วย โดยข้อมูลที่ได้ต้องมีการแปลงให้อยู่ในรูปของไฟร์ชนิด TEXT โดยข้อมูลที่มีรหัส "1" แทนเนื้อของภาพ และ "0"แทนพื้นของภาพประโยค ข้อมูลนี้ได้มาจากการแสกนเอกสารจากหน้ากระดาษด้วยความละเอียด 300 dpi และจัดเก็บเอกสารด้วยโปรแกรมสำเร็จรูป Photoshop ซึ่งทำการบันทึกอยู่ในรูปแบบ "BMP" จากนั้นทำการเปลี่ยนไฟร์ข้อมูลที่อยู่ในรูปแบบ "BMP" ให้อยู่ในรูปแบบของไฟร์ชนิด TEXT โดยข้อมูลที่มีรหัส "1" แทนเนื้อของภาพ และ "0"แทนพื้นของภาพประโยค จากนั้นนำข้อมูลที่ได้มาทำการหาตำแหน่งของเส้นแบ่งระดับของตัวอักษร

1.6.2.2 ขั้นตอนการดำเนินงานวิจัย

ขั้นตอนในการทำวิจัยการแยกตัวอักษรนี้ จะเริ่มจากการเก็บข้อมูลการติดกันประเภทต่างๆของตัวอักษร ดังภาคผนวก ก. จากรูปแบบการติดกันแบบต่างๆ ที่พบสามารถนำมาจัดประเภท และทำการหาวิธี หรือขั้นตอนการแยกแต่ละประเภทออกจากกัน งานวิจัยชิ้นนี้จึงได้นำเสนอแนวคิดการใช้เส้นแบ่งระดับ และคุณสมบัติทางกายภาพของฮีสโตแกรมของตัวอักษรระดับกลางมาเป็นตัวแบ่งประเภท และหาแนวคิดที่ใช้ในการตัดแยกตัวอักษร โดยนำการหาค่า Break Cost เพื่อกำหนดจุดตัด จากนั้นจึงทำการพัฒนาโปรแกรม เพื่อทำการตรวจสอบความถูกต้องของแนวคิด และสรุปผลการทดลองเป็นขั้นตอนสุดท้าย

1.7 โครงสร้างวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้แบ่งออกเป็นบทต่างๆ ได้ 6 บท ประกอบด้วย

บทที่ 1 เป็นบทนำที่กล่าวถึง

ความเป็นมาและความสำคัญของปัญหา

วัตถุประสงค์ของการศึกษา

สมมติฐานของการศึกษา

ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

ขอบเขตของการดำเนินการวิจัย

วิธีที่ใช้ในการดำเนินการวิจัย

โครงสร้างวิทยานิพนธ์

บทที่ 2 กล่าวถึงทฤษฎีต่างๆ ที่จะนำไปใช้ในการวิเคราะห์การติดกัน หรือการหาจุดตัดแยกตัวอักษร ได้แก่

ลักษณะตัวอักษรภาษาไทย

ลักษณะตัวพิมพ์ภาษาไทย

การหาค่า Histogram ของภาพตัวอักษร

- Pixel Projection
- Profile Projection
- Modify Pixel Projection

การหาค่า Break Cost เพื่อการกำหนดจุดตัด

บทที่ 3 กล่าวถึงวิธีการแยกภาพตัวอักษรออกจากภาพประโยค โดยวิธีการเปลี่ยนรหัสขอบ โดยมีขั้นตอนต่างๆ ดังนี้

การกำหนดแนวทางและลำดับการค้นหาภาพตัวอักษรในแต่ละระดับ

การแยกภาพตัวอักษร

- การกำหนดรหัสจุดขอบ
- การหาค่าตำแหน่งพิกัด
- การคัดลอกตัวอักษร

บทที่ 4 กล่าวถึงการวิเคราะห์ และตัดแยกตัวอักษรที่ติดกัน มีขั้นตอนต่างๆ ดังนี้

การกำหนดประเภทการติดกันของตัวอักษร

การกำหนดตัวอักษรที่อยู่ระดับกลาง

ข้อสังเกตของตัวอักษรในระดับกลาง

การวิเคราะห์การติดกันของภาพตัวอักษร

การวิเคราะห์การเหลื่อมล้ำของตัวอักษร

การตรวจสอบและแนวทางการตัดแยกตัวอักษร

การตัดแยกตัวอักษรที่ติดกัน

การเรียงรูปประโยคหลังการรู้จำ

บทที่ 5 เป็นการแสดงผลที่ได้ทดสอบ ซึ่งจะแสดงผลการทดสอบของวิธีการ

ผลการทดสอบความถูกต้องของการวิเคราะห์การติดกันของตัวอักษร

ผลการทดสอบความถูกต้องของการตัดแยกตัวอักษร

บทที่ 6 เป็นการกล่าวถึงข้อสรุปของงานวิจัย และแนวทางในการพัฒนาในอนาคต

กล่าวถึงข้อสรุปของงานวิจัยนี้

เสนอแนวทางในการพัฒนางานวิจัยต่อไป

บทที่ 2

หลักการทำงานของทฤษฎี

2.1 ลักษณะตัวอักษรภาษาไทย

ตัวอักษรภาษาไทย จะประกอบด้วยพยัญชนะ 44 ตัว สระ 19 ตัว และวรรณยุกต์ 4 รูป และตัวอักษรพิเศษอีก 3 ตัว ลักษณะประโยคของภาษาไทยดังรูป 4 จะประกอบด้วย สระ และวรรณยุกต์เรียงอยู่ในระดับที่แตกต่างกัน สามารถแบ่งระดับของตัวอักษร ออกเป็น 4 ระดับคือ

ระดับที่ 1 ระดับเหนือบน (ABOVE UPPER)

ประกอบด้วย วรรณยุกต์ และตัวการันต์

ระดับที่ 2 ระดับบน (UPPER)

ประกอบด้วย สระระดับบน และวรรณยุกต์

ระดับที่ 3 ระดับกลาง (MIDDLE)

ประกอบด้วย พยัญชนะ และสระระดับกลาง

ระดับที่ 4 ระดับล่าง (LOWER)

ประกอบด้วย สระล่าง และบางส่วนของพยัญชนะ เช่น ญ

รูปที่ 2.1 แสดงภาพประโยคของภาษาไทย

เมื่อพิจารณาพยัญชนะ และสระบางตัว มีลักษณะที่สูงกว่า 1 ระดับ และบางพยัญชนะมีลักษณะแยกกันอยู่คนระดับ เช่น ฐ, ญ ตามตารางที่ 2.1

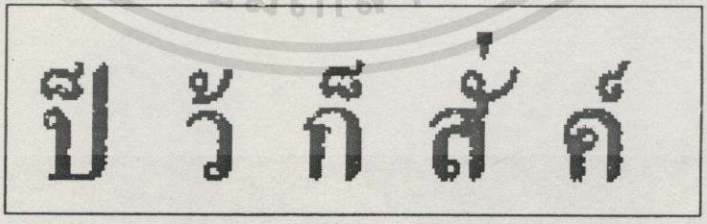
ตารางที่ 2.1 แสดงประเภทของตัวอักษรเดี่ยว

ประเภท	ระดับ	สมาชิก
1	Upper	◌ ◌ ◌ ◌ ◌ ◌ ◌ ◌ ◌ ◌
2	Upper + Middle	ป ฝ ฟ ฬ ไ ใ โ
3	Middle	ฯ ะ เ า ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ผ พ ภ ม ย ร ล ว ศ ษ ส ห อ ฮ
4	Lower + Middle	ฦ ฤ ๆ ฎ ฏ
5	Lower	◌ ◌ ◌ ◌
6	Above upper	◌ ◌ ◌ ◌ ◌ ◌

2.2 ลักษณะของตัวพิมพ์ภาษาไทย

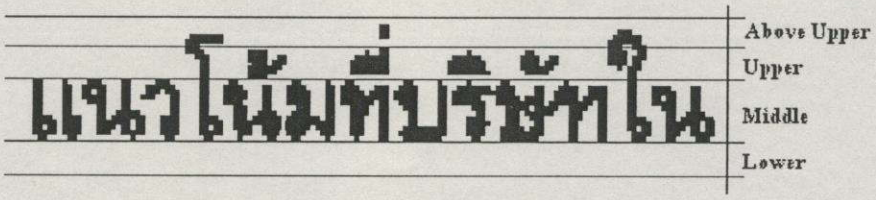
การศึกษาผลของตัวอักษรตัวพิมพ์ภาษาไทย ที่ได้จาการแสกนเอกสาร 3 ประเภท ได้แก่ หนังสือพิมพ์, นิตยสาร, และวรสาร จากผลการศึกษาพบว่า มีคุณลักษณะบางประการที่แตกต่างกัน สามารถสรุปเป็นข้อๆ ดังนี้

1. การจัดเรียงของสระในระดับบนของตัวอักษร มีการจัดเรียงเมื่อเทียบกับตำแหน่งของพยัญชนะในระดับกลางหลายรูปแบบ เช่นอาจอยู่ตรงกับพยัญชนะหรืออยู่ระหว่างพยัญชนะ ตามตัวอย่างรูปที่ 2.2



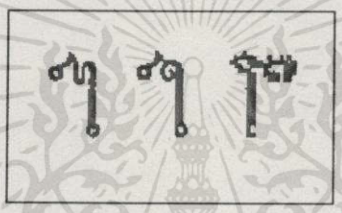
รูปที่ 2.2 แสดงการจัดเรียงของสระในระดับบนของตัวอักษร

2. บางประเภทของเอกสาร การกำหนดความสูงของสระ(ไ,ใ,โ) แตกต่างกัน ในบางรูปแบบตัวอักษร(FONT) มีการกำหนดความสูงถึงระดับเหนือบน ตามตัวอย่างรูปที่ 2.3



รูปที่ 2.3 แสดงตัวอักษรที่มีความสูงถึงระดับเหนือบน

3. บางลักษณะการติดกันของตัวอักษร ไม่ได้ทำให้ความกว้างของตัวอักษรเพิ่มมากขึ้นเกินกว่า 1 ตัวอักษร ตามตัวอย่างรูปที่ 2.4



รูปที่ 2.4 แสดงตัวอักษรติดกันที่ความกว้างไม่เกิน 1 ตัวอักษร

2.3 การหาค่าฮีสโตแกรม ของภาพตัวอักษร

งานวิจัยนี้ได้นำเสนอการหาค่าฮีสโตแกรมแบบต่างๆ ที่เหมาะกับตัวอักษรแต่ละแบบ เพื่อเป็นการแสดงลักษณะของกราฟให้เด่นชัดขึ้น โดยมีวิธีการต่างๆ ดังนี้

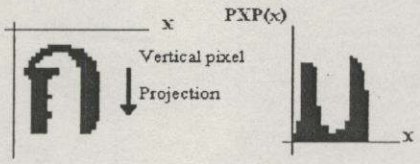
2.3.1 The Pixel Projection [6]

วิธีการนี้เป็นการแสดงค่าจำนวนจุดที่เป็นเนื้อของตัวอักษรในแนวตั้ง(Vertical Pixel Projection) และแนวนอน (Horizontal Pixel Projection) โดยทำการคำนวณจากสมการ

$$\text{Vertical} \quad PXP(x) = \sum_{y} P(x,y)$$

$$\text{Horizontal} \quad PXP(y) = \sum_{x} P(x,y)$$

เมื่อP(x,y) แสดงค่าของจุด ณ.ตำแหน่ง x และ y ผลที่ได้จะแสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 2.5



รูปที่ 2.5 แสดงกราฟ Vertical Pixel Projection

2.3.2 The Profile Projection [6]

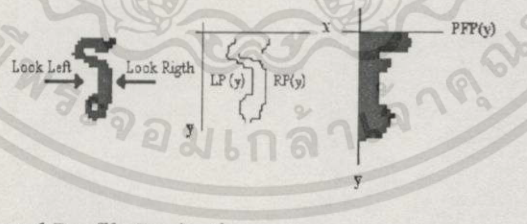
วิธีการนี้เป็นการคำนวณหา ระยะทางจากขอบด้านหนึ่งไปยังขอบอีกด้านหนึ่งของตัวอักษร ในแนวตั้ง (Vertical Profile Projection) และแนวนอน (Horizontal Profile Projection) ในงานวิจัยชิ้นนี้จะขอกล่าวถึงเฉพาะวิธีการ Horizontal Profile Projection เท่านั้น ซึ่งสามารถคำนวณได้จากสมการ

$$\text{Horizontal PFP}(y) = RP(y) - LP(y)$$

เมื่อ $RP(y) = \max_{x} X \in \{x|P(x,y)\}$

$LP(y) = \min_{x} X \in \{x|P(x,y)\}$

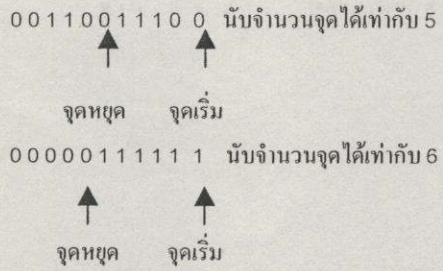
ผลที่ได้จะแสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 2.6



รูปที่ 2.6 แสดงกราฟ Horizontal Profile Projection

2.3.3 The Modify Pixel Projection

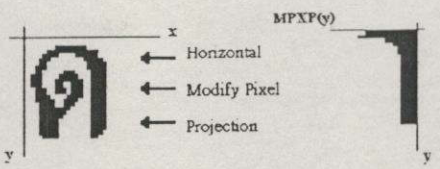
วิธีการนี้ผู้ทำวิจัยได้ทำการปรับปรุงจากการคำนวณแบบ Pixel Projection เพื่อแสดงลักษณะของตัวอักษรด้างข้างในรูปกราฟของฮีสโตแกรมตามแนวนอน ให้เด่นชัดมากขึ้น โดยวิธีการเริ่มจากนับจำนวนจุด จากขอบด้านขวาของกรอบตัวอักษร และนับจำนวนจุดไปทางซ้าย เมื่อพบจุดที่เป็นเนื้อของตัวอักษร ให้นับต่อไปจนถึงพื้นของตัวอักษรจึงหยุด ดังตัวอย่างด้านล่างนี้



แสดงเป็นอัลกอริทึมได้ดังนี้

```
I,J   INTEGER
NUM_ZERO INTEGER
CHAR_FLAG  BOOLEAN
PICT          คือ อาเรย์ 2 มิติขนาด ROW x COL เก็บข้อมูลภาพตัวอักษร
ROW_CNT  คือ อาเรย์ 2 มิติขนาด ROW   เก็บระยะทางจากขอบ
FOR  I = 1 TO ROW DO
      CHAR_FLAG = FALSE
      NUM_ZERO   = 0
      FOR  J = 1 TO COL DO
        IF  PICT[I][J] = '1' THEN
            ROW_CNT[I]  = ROW_CNT[I] + 1
            CHAR_FLAG = TRUE
        ELSE
            IF CHAR_FLAG = TRUE THEN
                  BREAK
            ELSE
                  NUM_ZERO = NUM_ZERO + 1
            END_iF
        END_IF
      END_LOOP
      ROW_CNT[I] =ROW_CNT[I] + NUM_ZERO
END_LOOP
```

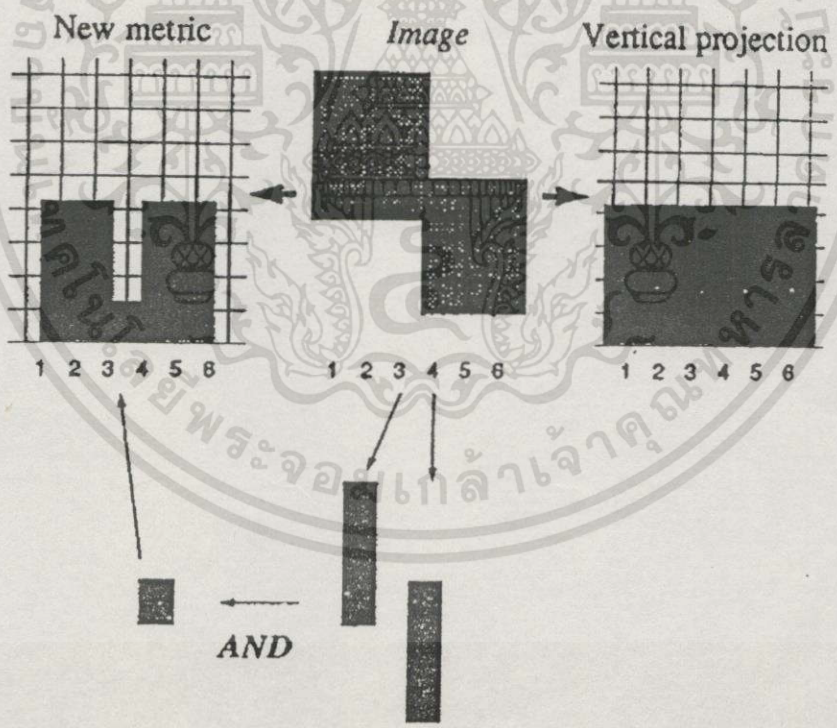
ผลที่ได้แสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 2.7



รูปที่ 2.7 แสดงกราฟ Horizontal Modify Pixel Projection

2.4 การหาค่า Break Cost เพื่อกำหนดจุดตัด [8]

โดยทั่วไปการหาจุดตัดของตัวอักษรที่ติดกันอย่างง่ายๆ ด้วยวิธีการหา Pixel Projection เพื่อกำหนดตำแหน่งของจุดตัดจากตำแหน่งที่มีค่าน้อยที่สุด แต่การใช้วิธีการ Break Cost ค่าที่คำนวณได้จะบอกถึงนัยสำคัญของการสัมผัสกัน (Degree of contact) ของแต่ละคอลัมน์ที่ติดกัน วิธีการนี้คำนวณโดยการนับจำนวนจุดในแนวตั้งที่ได้ จากการทำการ AND กันของเนื้อของภาพในคอลัมน์ที่ติดกัน ดังตัวอย่างรูปที่ 2.8



รูปที่ 2.8 แสดงนัยสำคัญของการสัมผัสกัน

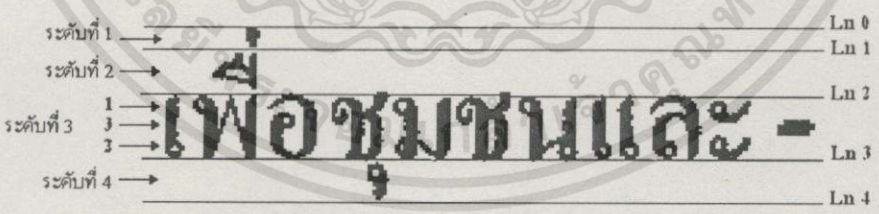
จากรูปที่ 2.8 จะเห็นว่าวิธีการ Break Cost เมื่อแสดงค่า ที่ได้จากการ AND กันจะแสดงจุดสัมผัสได้อย่างชัดเจน ในขณะที่วิธีการ Vertical Projection ไม่สามารถแสดงได้ ดังนั้นวิธีการนี้เราสามารถทราบถึงบริเวณที่มีการสัมผัสกันน้อยที่สุด เพื่อกำหนดเป็นตำแหน่งของจุดตัดของตัวอักษรที่ติดกันได้

บทที่ 3

การแยกภาพตัวอักษรออกจากภาพประโยค ด้วยวิธีการเปลี่ยนรหัสขอบ

การแยกภาพตัวอักษรออกจากภาพประโยคด้วยวิธีการเปลี่ยนรหัสขอบ เป็นการนำเสนอขั้นตอนการแยกภาพตัวอักษรออกจากภาพของประโยค ตามแนวระดับตัวอักษรภาษาไทย 4 ระดับ เริ่มจากการกำหนดแนวทางและลำดับการค้นหาภาพตัวอักษรในแต่ละระดับตามรายละเอียดในหัวข้อ 3.1 จากนั้นในการค้นหาตัวอักษรแต่ละระดับ เมื่อพบตัวอักษรก็ให้ดำเนินวิธีการแยกภาพตัวอักษร ด้วยวิธีการที่พิจารณาขอบภาพตัวอักษรเป็นหลัก พร้อมกับทำการบันทึกพิกัดตำแหน่งของตัวอักษรที่แยกออกมาตามตำแหน่งเดิมที่เรียงอยู่ในภาพประโยค ตามรายละเอียดในหัวข้อ 3.2 ผลที่ได้รับจากวิธีการนี้ คือจะได้แถวของภาพตัวอักษรที่เรียงลำดับหน้าหลังตามลำดับเดิมที่อยู่ในภาพประโยคเป็นจำนวน 4 แถว ในแต่ละแถวหมายถึงตัวอักษรทั้งหมดในแต่ละระดับ และเนื่องจากการแยกภาพตัวอักษรใช้การพิจารณาขอบภาพเป็นหลัก ไม่ได้มีการคำนึงถึงภาพของตัวอักษรที่ทำการแยกออกมาได้ ว่าจะเป็นภาพตัวอักษรเดี่ยว หรือเป็นภาพตัวอักษรที่ติดกัน ดังนั้นขั้นตอนการวิเคราะห์การติดกันของตัวอักษร และขั้นตอนการตัดแยกตัวอักษรที่ติดกัน จึงถูกนำมาใช้ซึ่งจะกล่าวถึงในบทที่ 4

3.1 การกำหนดแนวทางและลำดับการค้นหาภาพตัวอักษรในแต่ละระดับ



รูปที่ 3.1 แสดงภาพระดับของตัวอักษรและเส้นแบ่งระดับ

การแยกภาพตัวอักษรภาษาไทย โดยใช้เส้นแบ่งระดับของตัวอักษร เพื่อกำหนดระดับตัวอักษร จากนั้นให้ทำการแยกตัวอักษรทีละระดับ โดยวิธีการตรวจกวาดจากซ้ายไปขวาตามแนวทางที่กำหนด ดังนี้

ระดับที่ 1 = Ln0 + 1/2(Ln1-Ln0)

ระดับที่ 2 = Ln1 + 1/2(Ln2-Ln1)

ระดับที่ 3 ลำดับที่ 1 = Ln2 + 1/4(Ln3-Ln2)

ระดับที่ 3 ลำดับที่ 2 = Ln2 + 3/4(Ln3-Ln2)

ระดับที่ 3 ลำดับที่ 3 = Ln2 + 1/2(Ln3-Ln2)

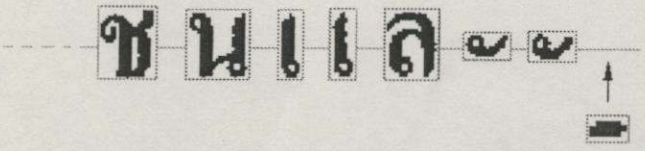
ระดับที่ 4 = Ln3 + 1/2(Ln4-Ln3)

เมื่อ Ln0,Ln1,Ln2,Ln3,Ln4 คือเส้นแบ่งของแต่ละระดับ

เมื่อตรวจกวาดมาจนพบจุดภาพที่เป็นจุดใดจุดหนึ่งของลายเส้นตัวอักษร หรือจุดภาพที่มีค่าเป็น 1 ก็จะใช้วิธีการแยกภาพออกมาด้วยวิธีการเปลี่ยนรหัสขอบ คัดลอกออกมาเรียงต่อกัน จากนั้นจะลบจุดภาพตัวอักษร และค้นหาตัวอักษรตัวถัดไปจนสิ้นสุดประโยค สำหรับลำดับการแยกตัวอักษรทั้ง 4 ระดับนั้น จะเริ่มจากระดับที่ 3 ต่อด้วยระดับที่ 2,1 และ 4 ตามลำดับ ในระดับที่ 3 มีลักษณะพิเศษคือ จะต้องทำการตรวจกวาดถึง 3 ครั้ง เนื่องจากตัวอักษรไทยมีบางตัวที่มีลักษณะทางกายภาพ แยกกันอยู่เช่น "ะ" และบางตัวอักษรพิเศษมีขนาดเล็ก และมีตำแหน่งอยู่กลางระดับ เช่น "-" และการตรวจกวาดในระดับนี้จะเริ่มจากแนวบนเป็นลำดับที่ 1 ต่อด้วยการตรวจกวาดในแนวล่างเป็นลำดับที่ 2 และแนวกลางลำดับสุดท้ายเป็นลำดับที่ 3 โดยแสดงแนวระดับการตรวจกวาดดังรูปที่ 3.1

ตัวอักษรที่ได้จากการตรวจกวาด ครั้งที่ 2 และ 3 ของระดับที่ 3 จะต้องนำมาแทรกในแถวของตัวอักษรที่แยกออกมาแล้วในระดับที่ 3 โดยตำแหน่งที่ทำการคือหน้าตัวอักษรที่มีค่า Xmin มากกว่าค่า Xcen ของตัวอักษรที่นำมาแทรก ถ้าไม่มีให้ใส่เป็นตัวสุดท้าย ตัวอย่างการแทรกภาพที่ได้จากการตรวจหาครั้งที่ 2 และ 3 ของระดับที่ 3 แสดงดังรูป 3.2 และ 3.3

รูปที่ 3.2 แสดงการแทรกภาพที่ได้จากการตรวจหาครั้งที่ 2



รูปที่ 3.3 แสดงการแทรกภาพที่ได้จากการตรวจหาครั้งที่ 3

ดังนั้นด้วยวิธีการนี้เราจะได้ ตัวอักษรที่เรียงลำดับอย่างถูกต้อง แยกเป็นแต่ละระดับดังรูปที่ 3.4 พร้อมที่จะเข้าสู่ขั้นตอนการวิเคราะห์การติดกับของตัวอักษร และขั้นตอนการรู้จำต่อไป

รูปที่ 3.4 แสดงตัวอักษรที่เรียงลำดับอย่างถูกต้องในแต่ละระดับ

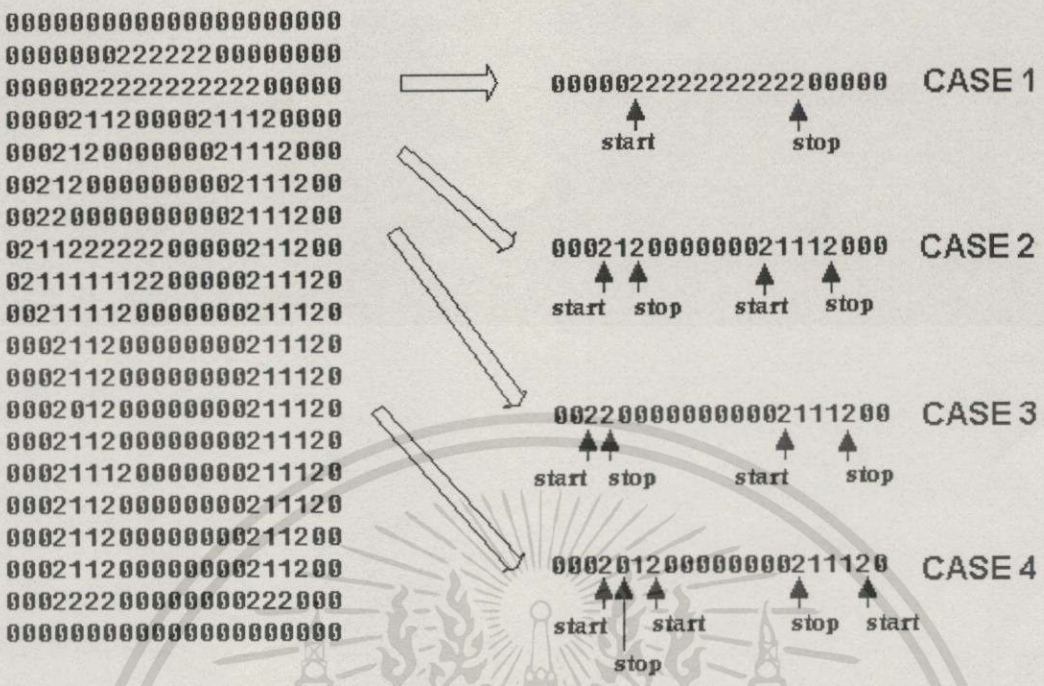
3.2 การแยกภาพตัวอักษร

การแยกตัวอักษรภาษาไทยออกจากประโยค โดยใช้เทคนิคการติดตามขอบของวัตถุเพื่อหาความกว้างและความสูงของตัวอักษร ทำให้สามารถกำหนด กรอบของตัวอักษรเพื่อการคัดลอกได้ ดังรูปที่ 3.5.ก แต่การแยกตัวอักษรโดยอาศัยกรอบเพียงอย่างเดียว ไม่สามารถแยกตัวอักษรที่ลักษณะเหลื่อมล้ำได้ ดังรูปที่ 3.5.ข [1]

ก. ข.

รูปที่ 3.5 แสดงกรอบของตัวอักษร

จึงควรนำขอบของตัวอักษรมาพิจารณาด้วย[2] แต่ถ้าพิจารณาเฉพาะขอบเมื่อต้องการคัดลอกตัวอักษรที่ละจุดจากกรอบด้านซ้าย ไปยังกรอบด้านขวาของตัวอักษร โดยกำหนดจุดเริ่มต้นและจุดสิ้นสุดของการคัดลอกอยู่ที่ขอบของตัวอักษรอาจทำให้เกิดข้อผิดพลาดได้ ดังเช่นตัวอย่างตัวอักษร “ก” รูปที่ 3.6



รูปที่ 3.6 แสดงการคัดลอกโดยใช้รหัสขอบเพียงตัวเดียว

ถ้าเราทำการติดตามขอบแล้วทำการเปลี่ยนรหัสขอบโดยไม่มีการระบุตำแหน่งว่าเป็นขอบซ้ายหรือขอบขวาแต่ให้มีเลขรหัสขอบเป็น 2 เหมือนกันหมด จากตัวอย่างตัวอักษร "ก" ทำการพิจารณาในแต่ละกรณี ในกรณีที่ 1 จะเห็นว่ามีแต่ขอบตัวอักษรดังนั้นในขั้นตอนการคัดลอกถ้าจุดที่ต่อจากรหัสขอบเจอ "0" ก็ให้หยุดการคัดลอก ก็จะได้การคัดลอกที่ถูกต้อง ในกรณีที่ 2 และ 3 เราสามารถใช้ขอบในการกำหนดจุดเริ่มต้นและจุดสิ้นสุดในการคัดลอกได้อย่างถูกต้อง แต่ในกรณีที่ 4 จะเห็นได้ว่ามีสิ่งรบกวน (noise) ที่เนื้อของตัวอักษรทำให้ลำดับการคัดลอกและการหยุดคัดลอกผิดพลาดไปหมดทั้งแถว ดังนั้นจึงสามารถสรุปได้ว่าเราไม่สามารถใช้เพียงรหัสขอบเพียงรหัสเดียวกำหนดการคัดลอกตัวอักษรได้

งานวิจัยนี้จึงเสนอ การแยกตัวอักษรภาษาไทย โดยอาศัยเทคนิคการติดตามขอบวัตถุด้วยการใช้หน้าต่าง[3] ในทิศทางทวนเข็มนาฬิกา และทำการกำหนดตำแหน่งจุดขอบด้านซ้าย จุดขอบด้านขวา และจุดมุมของตัวอักษร ตามรายละเอียดในหัวข้อ 3.2.1 การหาค่าตำแหน่งพิกัดของตัวอักษรตามรายละเอียดในหัวข้อ 3.2.2 และวิธีการคัดลอกตัวอักษรตามรายละเอียดในหัวข้อ 3.2.3

3.2.1 การกำหนดรหัสจุดขอบ

ในการแยกตัวอักษรจะใช้การคัดลอกตัวอักษร โดยเริ่มจากขอบด้านหนึ่งไปยังอีกด้านหนึ่ง (จากซ้ายไปขวา) ดังนั้นจะต้องกำหนดจุดขอบซ้ายและขอบขวาให้ได้เสียก่อน แต่อาจมีบางจุดที่เป็นทั้งขอบซ้ายและขอบขวา เช่น ขอบตัวอักษรที่มีความกว้างเท่ากับ 1 จุด จุดนี้จะกำหนดให้เป็นจุด

พิเศษ โดยกำหนดให้ข้อมูลที่จะทำการวิเคราะห์นั้นประกอบด้วยรหัส "0" และ"1" ให้รหัส "1" เป็นเนื้อของตัวอักษร และรหัส "0" เป็นพื้นหลังของตัวอักษร ในขั้นตอนนี้เราจะทำการแทนที่รหัส "1" ของจุดขอบบางจุดในตัวอักษร ด้วยรหัส "L","R" หรือ "X" โดยกำหนดให้

รหัส "1" แทนจุดขอบที่อยู่ระหว่างขอบด้านซ้ายและขวาของตัวอักษร

รหัส "L" แทนจุดขอบด้านซ้ายของตัวอักษร

รหัส "R" แทนจุดขอบด้านขวาของตัวอักษร

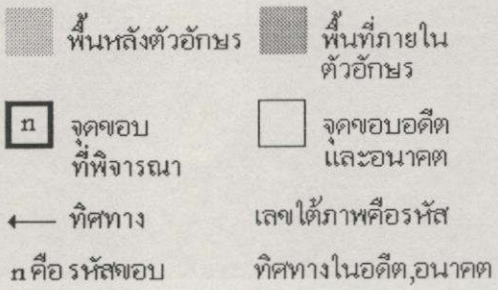
รหัส "X" แทนจุดพิเศษอันได้แก่ จุดปลายของมุมด้านบนและด้านล่าง หรือเนื้อของตัวอักษรที่มีความกว้างเท่ากับ 1 จุด

ใช้รหัสทิศทางของฟรีแมนดังรูปที่ 3.7 ในการกำหนดเงื่อนไขการเปลี่ยนรหัสจุดขอบว่าสมควรเปลี่ยนหรือไม่ ถ้าสมควรเปลี่ยนควรเปลี่ยนเป็นรหัส "L","R" หรือ "X" ดูจากรหัสทิศทางในอดีตและอนาคตของจุดขอบนั้นๆ รหัสทิศทางในอดีตของจุดขอบหมายถึงตำแหน่งจากจุดขอบก่อนหน้าเปรียบเทียบกับจุดขอบที่พิจารณาอยู่ตามทิศทางของฟรีแมน และรหัสทิศทางในอนาคตหมายถึงตำแหน่งจากจุดขอบที่พิจารณาอยู่เปรียบเทียบกับจุดขอบจุดต่อไป

รูปที่ 3.7 แสดงรหัสทิศทางของฟรีแมน

ในขั้นตอนการคัดลอกนั้นกระทำทีละแถวของจุดภาพจากซ้ายไปขวา เมื่อพิจารณาเส้นขอบภาพจะพบว่าจุดปลายซึ่งอาจเป็นมุม หรือปลายของเส้นของภาพนั้น เป็นทั้งจุดขอบซ้ายและขวาของภาพ ดังนั้นเราจึงกำหนดให้เป็นรหัส "X" เพื่อที่จะสามารถบอกในขั้นตอนการคัดลอกได้ว่าจุดถัดไปเป็นจุดที่อยู่นอกขอบภาพ

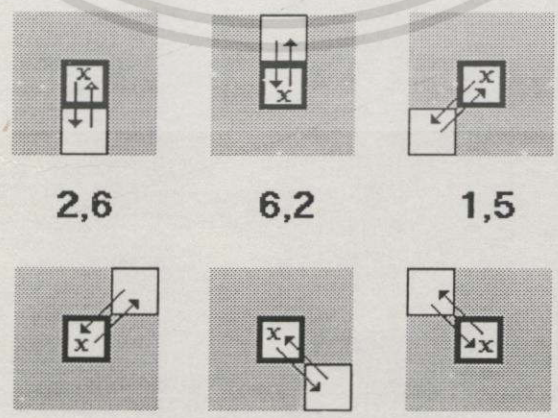
3.2.1.1 กรณีจุดปลายของมุมด้านบน และด้านล่างของตัวอักษร ในกรณีนี้จะทำการเปลี่ยนรหัสขอบเป็น "X" ประกอบด้วยทิศทางในอดีตและอนาคต ดังรูปที่ 3.8 และรูปที่ 3.9



รูปที่ 3.8 แสดงมุมปลายด้านบน

รูปที่ 3.9 แสดงมุมปลายด้านล่าง

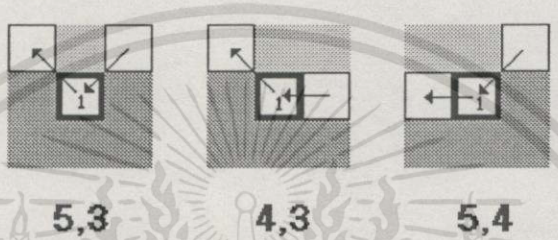
3.2.1.2 กรณีจุดปลายของเนื้อตัวอักษร ในกรณีนี้จะทำการเปลี่ยนรหัสขอบเป็น "X" ประกอบด้วยทิศทางในอดีตและอนาคต ดังรูปที่ 3.10



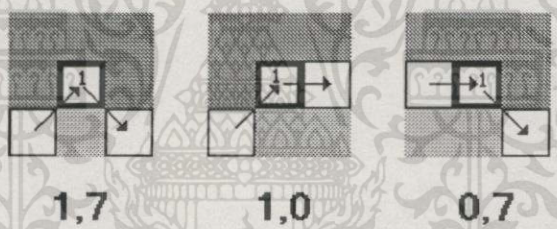
รูปที่ 3.10 แสดงจุดปลายที่มีการเปลี่ยนรหัส

ส่วนที่เป็นมุมเว้าทั้งด้านบน,ด้านล่างของภาพและขอบภาพในแนวนอนถึงแม้เป็นจุดขอบแต่จะเป็นจุดขอบที่อยู่ระหว่างจุดขอบด้านซ้ายและขวาอยู่แล้ว จึงไม่จำเป็นต้องมีการเปลี่ยนรหัสขอบที่จุดนี้ (ยังคงรหัส "1" เหมือนเดิม)

3.2.1.3 กรณีจุดมุมส่วนเว้าด้านบนและด้านล่างของตัวอักษร ในกรณีนี้คงรหัส "1" ไว้เหมือนเดิมประกอบด้วยทิศทางในอดีตและอนาคต ดังรูปที่ 3.11 และ รูปที่ 3.12

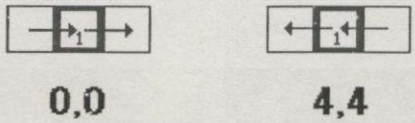


รูปที่ 3.11 แสดงมุมเว้าด้านบน



รูปที่ 3.12 แสดงมุมเว้าด้านล่าง

3.2.1.4 กรณีจุดขอบในแนวนอนของตัวอักษร ในกรณีนี้จะไม่มีการเปลี่ยนแปลงรหัสขอบ ประกอบด้วยทิศทางในอดีตและอนาคต ดังรูปที่ 3.13



รูปที่ 3.13 แสดงขอบในแนวนอน

จุดขอบอื่นที่นอกเหนือจากจุดยกเว้นและจุดพิเศษที่กล่าวมาแล้ว ทำการเปลี่ยนรหัสให้เป็นขอบด้านซ้ายและขวาโดยดูเฉพาะทิศทางในอนาคตของจุดนั้น แต่ในกรณีที่มีซ้อนทับกันของรหัสขอบ

"L" และ "R"แสดงว่าเป็นเนื้อของภาพที่มีความกว้างเท่ากับ1 จุด ดังนั้นจึงทำการเปลี่ยนรหัสขอบเป็นรหัส "X" แทน

3.2.1.5 กรณีจุดขอบด้านซ้าย(รหัส "L") ของตัวอักษร ประกอบด้วยทิศทางในอนาคตได้แก่ 5, 6, 7, 0 และ ไม่ตกอยู่ในกรณีที่ 2.1.1 ถึง 2.1.4 ในกรณีนี้

ถ้ารหัสขอบเดิมเป็น "1" จะเปลี่ยนรหัสเป็น "L"

แต่ถ้ารหัสขอบเดิมเป็น "R" จะเปลี่ยนรหัสเป็น "X"

3.2.1.6 กรณีจุดขอบด้านขวา(รหัส "R") ของตัวอักษร ประกอบทิศทางในอนาคตได้แก่ 1, 2, 3, 4 และ ไม่ตกอยู่ในกรณีที่ 2.1.1 ถึง 2.1.4 ในกรณีนี้

ถ้ารหัสขอบเดิมเป็น "1" จะเปลี่ยนรหัสเป็น "R"

แต่ถ้ารหัสขอบเดิมเป็น "L" จะเปลี่ยนรหัสเป็น "X"

3.2.2 การหาค่าตำแหน่งพิกัด

เมื่อสิ้นสุดขั้นตอนการเปลี่ยนรหัสขอบจะได้ภาพตัวอักษรที่รหัสขอบถูกเปลี่ยนไป แสดงไว้ดังตัวอย่างดังรูปที่ 3.14 และในขณะที่ทำการติดตามขอบและเปลี่ยนรหัสอยู่นั้นก็สามารถเปรียบเทียบพิกัด เพื่อหาจุดที่มีค่าต่ำสุดและสูงสุดของตัวอักษร ได้ดังตัวอย่างรูปที่ 3.15 โดยกำหนดให้

Xmin = ค่าCoordinate ของ x ที่มีค่าต่ำสุด

Xmax = ค่าCoordinate ของ x ที่มีค่าสูงสุด

Ymin = ค่าCoordinate ของ y ที่มีค่าต่ำสุด

Ymax = ค่าCoordinate ของ y ที่มีค่าสูงสุด

Xcen = 1/2(Xmax+Xmin)

```
00000000000    00000000000
00111000110    00L1R000LR0
01111000110    0L11R000LR0
01101100110    0L101R00LR0
01111000110    0L11R000LR0
01100000110    0LR00000LR0
01100100110    0LR00X00LR0
01101110110 ➜  0LR0L1R0LR0
01111111110    0L1111111R0
01111011110    0L11R0L11R0
01110001110    0L1R000L1R0
01100000110    0LR00000LR0
00000000010    000000000X0
00000000000    00000000000
```

รูปที่ 3.14 แสดงรหัสขอบของตัวอักษร

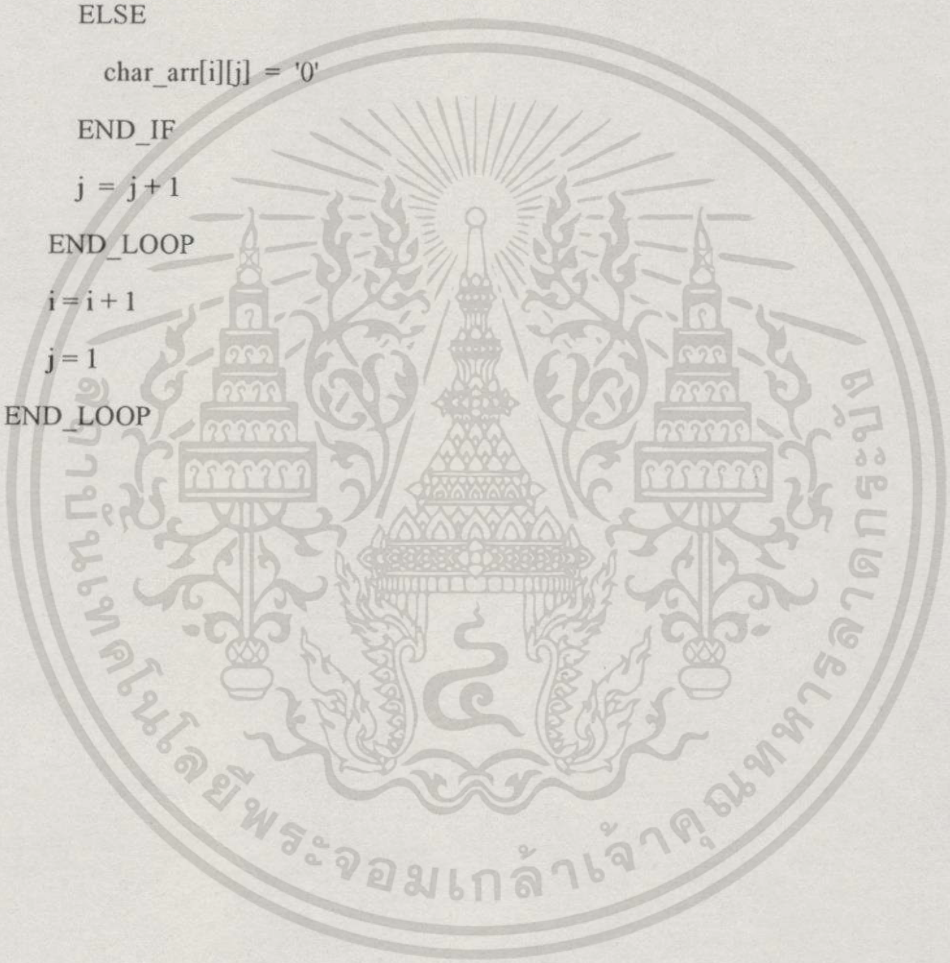
รูปที่ 3.15 แสดงขอบเขตของตัวอักษร

3.2.3 การคัดลอกตัวอักษร

การคัดลอกตัวอักษรออกจากรูปประโยค ในขั้นตอนนี้กระทำการกวาดจุดในภาพจากซ้ายไปขวาและบนลงล่าง โดยเริ่มจากจุด(Xmin,Ymin) ไปสิ้นสุดที่จุด(Xmax,Ymax) ภายในกรอบของตัวอักษร ในแต่ละแถวให้ทำการคัดลอกข้อมูลทั้งหมดตั้งแต่ข้อมูลที่เริ่มต้นด้วยรหัส "L" จนถึงรหัส "R" ในกรณีที่พบรหัส "X" ก็ให้ถือว่าเป็นเนื้อตัวอักษรที่มีความกว้างเท่ากับ 1 บิต จึงสามารถทำการคัดลอกได้ทันที ข้อมูลที่ได้มีการคัดลอกไปแล้วให้ทำการลบโดยเปลี่ยนรหัสเป็น "0" แสดงเป็นอัลกอริทึมได้ดังนี้

```
กำหนดให้
word_arr        เป็นอาเรย์ 2 มิติเก็บข้อมูลภาพประโยค
char_arr        เป็นอาเรย์ 2 มิติเก็บข้อมูลภาพตัวอักษรที่ได้จากการแยกจากประโยค
i,j             INTERGER
sw              CHARACTER
 i=1
 j=1
 FOR row = Ymin TO Ymax DO
 BEGIN
   FOR col = Xmin TO Xmax DO
   BEGIN
     IF word_arr[row][col] = 'L' THEN
       char_arr[i][j] = '1'
       word_arr[row][ col] = '0'
       sw = 'L'
     ELSE IF word_arr[row][ col] = 'R' THEN
       char_arr[i][j] = '1'
       word_arr[row][ col] = '0'
```

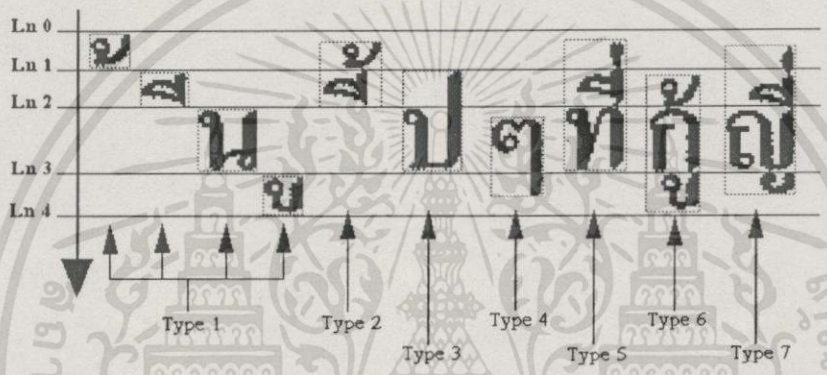
```
        sw = 'R'
      ELSE IF word_arr[row][ col] = 'X' THEN
        char_arr[i][j]  =  '1'
        word_arr[row][ col] = '0'
      ELSE IF sw = 'L' THEN
        char_arr[i][j]  = word_arr[row][ col]
        word_arr[row][ col] = '0'
      ELSE
        char_arr[i][j]  =  '0'
      END_IF
      j  =  j + 1
    END_LOOP
    i = i + 1
    j = 1
END_LOOP
```



บทที่ 4
การวิเคราะห์และการตัดแยกตัวอักษรไทยที่ติดกัน

4.1 การกำหนดประเภทการติดกันของตัวอักษร

รูปแบบการติดกันของตัวอักษรตัวพิมพ์ในแนวระดับบนและล่าง พิจารณาตามความสูงของตัวอักษรเปรียบเทียบกับระดับของตัวอักษรแล้ว[7] สามารถแบ่งประเภทตามความสูงได้ 7 ประเภท แสดงดังรูปที่ 4.1



รูปที่ 4.1 แสดงการแบ่งประเภทตัวอักษรตามความสูง

ประเภทที่ 1 (Type1) หมายถึงตัวอักษรที่มีความสูงอยู่ภายในเส้นแบ่งระดับของตัวอักษรในแต่ละระดับ

ประเภทที่ 2 (Type2) หมายถึงตัวอักษรที่อยู่ในระดับเหนือบนและระดับบน

ประเภทที่ 3 (Type3) หมายถึงตัวอักษรที่อยู่ในระดับบนและระดับกลาง ซึ่งสามารถแบ่งย่อยลงไปเป็นประเภทต่างๆ ได้อีก 6 ประเภท

ประเภทที่ 3.1 หมายถึงพยัญชนะเดี่ยวที่มีความสูงถึงระดับบน

ประเภทที่ 3.2 หมายถึงสระเดี่ยวที่มีความสูงถึงระดับบน

ประเภทที่ 3.3 พยัญชนะเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร

ประเภทที่ 3.4 สระเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร

ประเภทที่ 3.5 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างเกินหนึ่งตัวอักษร

ประเภทที่ 3.6 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร

ดังที่ได้สรุปไว้ในตารางที่ 4.2 พร้อมทั้งยกตัวอย่าง และแนวทางตัดการตัดแยก

ประเภทที่ 4 (Type4) หมายถึงตัวอักษรที่อยู่ในระดับกลางและระดับล่าง ซึ่งสามารถแบ่งย่อยลงไปเป็นประเภทต่างๆ ได้อีก 5 ประเภท

ประเภทที่ 4.1 สระเดี่ยวที่ยาวลงมาจนถึงระดับล่าง

ประเภทที่ 4.2 อักษรเดี่ยวที่ยาวลงมาจนถึงระดับล่าง

ประเภทที่ 4.3 พยัญชนะเดี่ยวที่ยาวลงมาจนถึงระดับล่าง

ประเภทที่ 4.4 พยัญชนะเดี่ยวที่ติดกับสระอู หรือ 'ฐ', 'ญ'

ประเภทที่ 4.5 พยัญชนะเดี่ยวที่ติดกับสระอุ

ดังที่ได้สรุปไว้ในตารางที่ 4.2 พร้อมทั้งยกตัวอย่าง และแนวทางตัดการตัดแยก

ประเภทที่ 5 (Type5) หมายถึงตัวอักษรที่อยู่ในระดับเหนือบน, ระดับบน และระดับกลาง ซึ่งสามารถแบ่งย่อยลงไปเป็นประเภทต่างๆ ได้อีก 5 ประเภท

ประเภทที่ 5.1 สระเดี่ยวที่มีความสูงถึงระดับเหนือบน

ประเภทที่ 5.2 พยัญชนะเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร

ประเภทที่ 5.3 สระเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร

ประเภทที่ 5.4 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างเกินหนึ่งตัวอักษร

ประเภทที่ 5.5 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความไม่กว้างเกินหนึ่งตัวอักษร

ดังที่ได้สรุปไว้ในตารางที่ 4.2 พร้อมทั้งยกตัวอย่าง และแนวทางตัดการตัดแยก

ประเภทที่ 6 (Type6) หมายถึงตัวอักษรที่อยู่ในระดับบน, ระดับกลาง และระดับล่าง

ประเภทที่ 7 (Type7) หมายถึงตัวอักษรอยู่ในทุกระดับ

สามารถหาประเภทของตัวอักษรได้จากตารางที่ 4.1

โดยกำหนดให้

$(x_{min}, y_{min}), (x_{max}, y_{max})$ = พิกัดของภาพตัวอักษรที่ได้จากการแยกภาพตัวอักษร

Ln0 = เส้นบนของระดับเหนือระดับบน

Ln1 = เส้นแบ่งของระดับเหนือบน และระดับบน

Ln2 = เส้นแบ่งของระดับบน และระดับกลาง

Ln3 = เส้นแบ่งของระดับกลาง และระดับล่าง

Ln4 = เส้นล่างของระดับล่าง

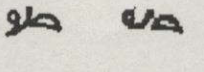
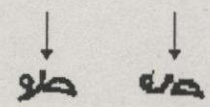
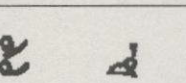
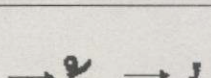
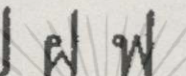
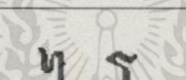
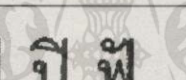
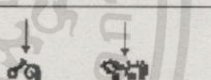
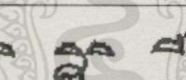
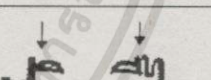
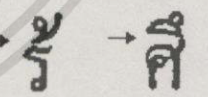
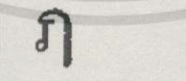
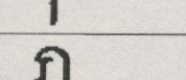
$\in$ = ค่าที่ได้จากการสังเกตประมาณเท่ากับ 1/12 |Ln4-Ln0|

ตารางที่ 4.1 แสดงการคำนวณประเภทของตัวอักษร

ประเภท	เงื่อนไข
1	$\lvert Y_{min} - Ln0 \rvert < \in$ and $\lvert Ln1 - Y_{max} \rvert < \in$ หรือ $\lvert Y_{min} - Ln1 \rvert < \in$ and $\lvert Ln2 - Y_{max} \rvert < \in$ หรือ $\lvert Y_{min} - Ln2 \rvert < \in$ and $\lvert Ln3 - Y_{max} \rvert < \in$ หรือ $\lvert Y_{min} - Ln3 \rvert < \in$ and $\lvert Ln4 - Y_{max} \rvert < \in$
2	$\lvert Y_{min} - Ln0 \rvert < \in$ and $\lvert Ln2 - Y_{max} \rvert < \in$
3	$\lvert Y_{min} - Ln1 \rvert < \in$ and $\lvert Ln3 - Y_{max} \rvert < \in$
4	$\lvert Y_{min} - Ln2 \rvert < \in$ and $\lvert Ln4 - Y_{max} \rvert < \in$
5	$\lvert Y_{min} - Ln0 \rvert < \in$ and $\lvert Ln3 - Y_{max} \rvert < \in$
6	$\lvert Y_{min} - Ln1 \rvert < \in$ and $\lvert Ln4 - Y_{max} \rvert < \in$
7	$\lvert Y_{min} - Ln0 \rvert < \in$ and $\lvert Ln4 - Y_{max} \rvert < \in$

ได้สรุปรายละเอียดของทุกประเภทที่ได้กล่าวมาทั้งหมดไว้ในตารางที่ 4.2 พร้อมทั้งตัวอย่างการติดกันของตัวอักษรและแนวทางการตัดแยกของแต่ละประเภท

ตารางที่ 4.2 แสดงประเภทการติดกันของตัวอักษรและแนวทางการตัดแยก

ประเภทการติดกันของตัวอักษร	ตัวอย่าง	ตัวอย่างแนวทางการตัดแยก
ประเภทที่ 1 หมายถึงตัวอักษรที่มีความสูงอยู่ภายในเส้นแบ่งระดับของตัวอักษรในแต่ละระดับ		
ประเภทที่ 2 หมายถึงตัวอักษรที่อยู่ในระดับเหนือบนและระดับบน		
ประเภทที่ 3.1 หมายถึงพยัญชนะเดี่ยวที่มีความสูงถึงระดับบน		
ประเภทที่ 3.2 หมายถึงสระเดี่ยวที่มีความสูงถึงระดับบน		
ประเภทที่ 3.3 พยัญชนะเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร		
ประเภทที่ 3.4 สระเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร		
ประเภทที่ 3.5 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างเกินหนึ่งตัวอักษร		
ประเภทที่ 3.6 พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร		
ประเภทที่ 4.1 สระเดี่ยวที่ยาวลงมาจนถึงระดับล่าง		
ประเภทที่ 4.2 อักษรเดี่ยวที่ยาวลงมาจนถึงระดับล่าง		
ประเภทที่ 4.3 พยัญชนะเดี่ยวที่ยาวลงมาจนถึงระดับล่าง		
ประเภทที่ 4.4 พยัญชนะเดี่ยวที่ติดกับสระอุ หรือ 'ฐ', 'ญ'		

ตารางที่ 4.2 แสดงประเภทการติดกันของตัวอักษรและแนวทางการตัดแยก(ต่อ)

ประเภทการติดกันของตัวอักษร	ตัวอย่าง	ตัวอย่างแนวทางการตัดแยก
<u>ประเภทที่ 4.5</u> พยัญชนะเดี่ยวที่ติดกับสระอุ	กุ นุ รุ บุ	→รุ →บุ
<u>ประเภทที่5.1</u> สระเดี่ยวที่มีความสูงถึงระดับเหนือบน	ใ ไ โ	
<u>ประเภทที่ 5.2</u> พยัญชนะเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร	ปี่ ปี้	→ปี่ →ปี้
<u>ประเภทที่5.3</u> สระเดี่ยวที่มีความสูงถึงระดับบน ที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร	ใ ไ	↓ใ ↓ไ
<u>ประเภทที่ 5.4</u> พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างเกินหนึ่งตัวอักษร	๕ใ ๕ไ ๑ใ ๑ไ	↓๕ใ ↓๕ไ
<u>ประเภทที่ 5.5</u> พยัญชนะ หรือ สระเดี่ยวที่ติดกับตัวอักษรระดับเหนือบน ในกรณีที่ความกว้างไม่เกินหนึ่งตัวอักษร	ซี่ สี่ ที่	→ซี่ →สี่
<u>ประเภทที่ 6</u> หมายถึงตัวอักษรที่มีความสูงตั้งแต่ระดับบนลงมา จนถึงระดับล่าง	ฝิ ฝู	→ฝิ →ฝู
<u>ประเภทที่ 7</u> หมายถึงตัวอักษรที่มีความสูงจากระดับเหนือบนยาวลงมาถึงระดับล่าง	ญี่	→ญี่

4.2 การกำหนดตัวอักษรที่อยู่ระดับกลาง

เมื่อพิจารณาตัวอักษรที่อยู่ในระดับกลาง มีลักษณะเด่นทางกายภาพที่สังเกตได้ คือมีแนวเส้นตรงของตัวอักษรทั้งในแนวตั้งและแนวนอนในจำนวนและตำแหน่งที่แตกต่างกัน โดยคุณสมบัติเหล่านี้สามารถปรากฏได้เมื่อเราใช้การหาค่าฮีสโตแกรมของตัวอักษร ดังนั้นงานวิจัยนี้ได้นำเสนอวิธีการจัดกลุ่มโดยการนำวิธีการหาค่าฮีสโตแกรมแบบต่างๆมาใช้เพื่อหาจำนวนและตำแหน่งของแนวเส้นตรงในตัวอักษร เพื่อใช้ในการพิจารณาการติดกันของตัวอักษร และแนวทางการตัด

เริ่มต้นจากการแบ่งช่วงระดับกลางออกเป็นสองส่วนดังรูปที่ 4.2 กำหนดให้ h เป็นความสูงของระดับกลาง ช่วงบน(High Level) มีขนาดความสูงเป็นหนึ่งในสี่ของความสูงของระดับกลาง และช่วงล่าง(Low Level) มีขนาดความสูงเป็นสามในสี่ของความสูงของระดับกลาง

HL = 1/4 h เมื่อ HL คือระดับความสูง High Level

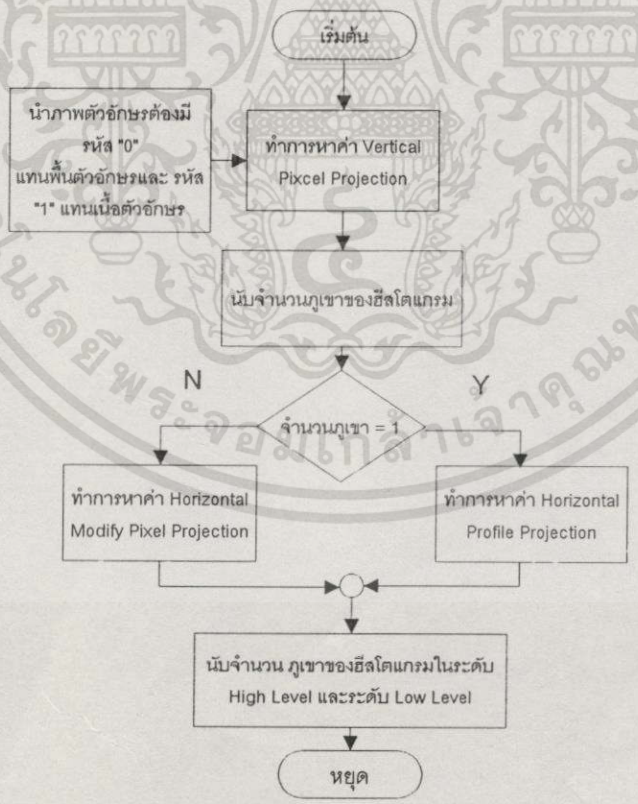
LL = 3/4 h เมื่อ LL คือระดับความสูง Low Level

รูปที่ 4.2 แสดงการแบ่งระดับและกราฟฮีสโตแกรม

จากนั้นทำการหาจำนวนภูเขาของฮีสโตแกรม (Mountain of Histogram) [6] ที่แสดงถึงแนวเส้นตรงของตัวอักษรทั้งในแนวตั้งและแนวนอนในระดับ High Level และ Low Level ของระดับกลางตามวิธีที่แสดงในผังการทำงานรูปที่ 4.3 เพื่อให้ใช้วิธีที่เหมาะสมกับตัวอักษรตามที่งานวิจัยชิ้นนี้ได้นำเสนอดังนี้คือถ้าค่าฮีสโตแกรมในแนวตั้งที่ได้มีจำนวนภูเขาเป็น 1 การหาค่าฮีสโตแกรมด้านแนวนอนควรจะใช้วิธี Profile Projection แต่ถ้ามากกว่า 1 ควรใช้แบบ Modify Pixel Projection จากจำนวนภูเขาในแนวและตำแหน่งต่างๆ สามารถแยกตัวอักษรในระดับกลางออกเป็นกลุ่มต่างๆ ได้ดังตารางที่ 4.3

ตารางที่ 4.3 แสดงกลุ่มพยัญชนะและสระ ในระดับกลาง

กลุ่ม	ภูเขา แนวตั้ง	ภูเขา แนว HL	ภูเขา แนว LL	ตัวอักษร
V1T1	1	>0	=0	า ๅ ๆ
V1T2	1	=0	>0	ไ ใ โ
V1T3	1	>0	>0	ง ร ว
V2T1	2	>0	=0	ก ค ด ต ถ ภ ศ ฤ ฦ ฏ ฐ ท ฑ ห
V2T2	2	=0	>0	ข ฃ ย ษ บ ป
V2T3	2	>0	>0	ช ซ ธ อ ฮ ล ส จ ฐ
V2T4	2	=0	=0	พ ผ ฟ ฝ ม ฆ น ฬ
V3	3	-	-	ฌ ญ ฒ ณ



รูปที่ 4.3 แสดงการหาจำนวนภูเขาของฮีสโตแกรม

4.3 ข้อสังเกตของตัวอักษรในระดับกลาง

การวิเคราะห์การติดกัน และการกำหนดแนวทางของการตัดตัวอักษร จะใช้ข้อสังเกตจากตารางที่ 2 ดังต่อไปนี้ในการตัดสินใจ

1. กลุ่มที่ฮีสโตแกรมมีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 3 แนว ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความสูงหรือต่ำเกินกว่าระดับกลาง กลุ่มนี้ได้แก่ V3 ตัวอย่างเช่น ณ,ฒ เป็นต้น

2. กลุ่มที่ฮีสโตแกรมมีจำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL มากกว่าหรือเท่ากับ 1 แนว ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความสูงเกินกว่าระดับกลาง กลุ่มนี้ได้แก่ V1T1, V1T3, V2T1 และ V2T3 ตัวอย่างเช่น ร,ว,ก,ภ,ธ,อ เป็นต้น

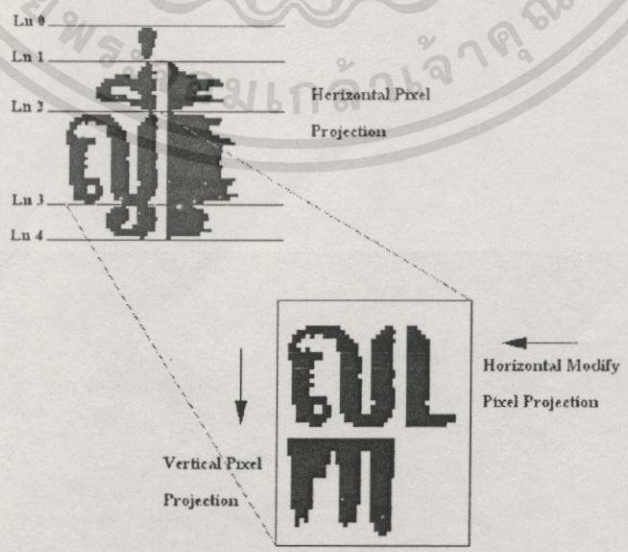
3. กลุ่มที่ฮีสโตแกรมมีจำนวนแนวเส้นตรงในแนวนอน ที่ระดับ LL มากกว่าหรือเท่ากับ 1 แนว ไม่มีตัวอักษร ตัวใดในกลุ่มนี้ที่มีความยาวต่ำกว่าระดับกลาง กลุ่มนี้ได้แก่ V1T3, V2T2 และ V2T3 ตัวอย่างเช่น ร,ว,บ,ย,ป,ธ,อ เป็นต้น

4. กลุ่มที่ฮีสโตแกรมมีจำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL และ LL ของระดับกลางเท่ากับศูนย์ ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความยาวต่ำกว่าระดับกลาง กลุ่มนี้ได้แก่ V2T4 ตัวอย่างเช่น พ,ผ,ฟ เป็นต้น

5. มีเฉพาะสระ "ไ","ใ","โ" เท่านั้นที่มีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 1 และจำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL เท่ากับศูนย์คือกลุ่ม V1T2

6. มีเฉพาะตัวอักษร "า","ๅ","ๆ" เท่านั้นที่มีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 1 จำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL เท่ากับ 1 และระดับ LL เท่ากับศูนย์ คือกลุ่ม V1T1

จะเห็นว่าบางประเภทของการติดกันจะต้องอาศัยคุณสมบัติเหล่านี้วิเคราะห์จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรที่ต้องการวิเคราะห์ตามผังงานรูปที่ 4.3 ดังตัวอย่างรูปที่ 4.4



รูปที่ 4.4 แสดงการหาแนวเส้นตัวอักษรระดับกลาง

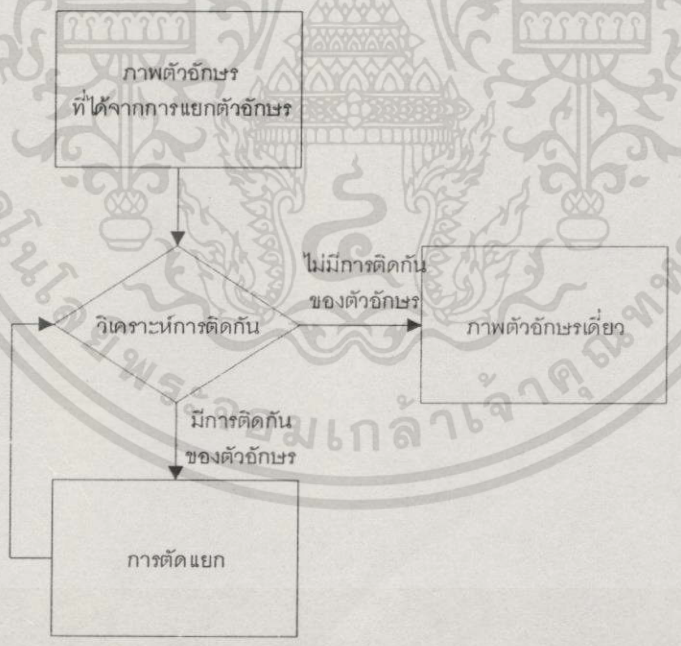
4.4 การวิเคราะห์การติดกันของภาพตัวอักษร

หลังจากที่ได้ภาพตัวอักษรจากการแยกออกจากภาพประโยคแล้วก็เข้าสู่ขั้นตอนการ วิเคราะห์การติดกันของตัวอักษร ประกอบด้วย

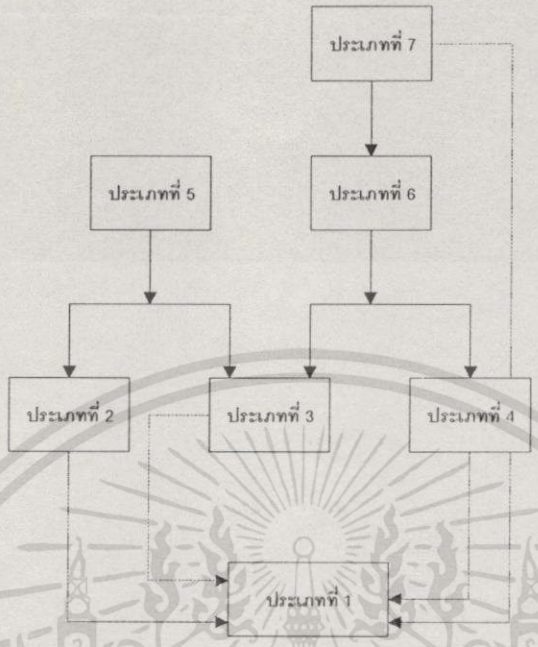
1. ระบุประเภทโอกาสที่จะเกิดการติดกัน ของตัวอักษรว่าจัดอยู่ในประเภทใดในระหว่างประเภทที่1 ถึงประเภทที่ 7 โดยการเปรียบเทียบตำแหน่งพิกัดของภาพตัวอักษรกับเส้นแบ่งระดับของตัวอักษร

2. ตรวจสอบการติดกันของตัวอักษรในแต่ละประเภท ว่ามีการติดกันจริงหรือไม่ และกำหนดแนวทางการตัดแยกตามหัวข้อ 4.6 เพื่อส่งต่อในขั้นตอนการตัดแยกต่อไป

และหลังจากทำการตัดแยกตัวอักษร จนได้ภาพตัวอักษรที่ถูกตัดแยกแล้ว ภาพตัวอักษรเหล่านั้นก็จะถูกส่งกลับเข้าสู่ ขั้นตอนการวิเคราะห์การติดกันของตัวอักษรอีกครั้ง จนกระทั่งทุกภาพตัวอักษรที่ถูกตัดแยกแล้วถูกระบุว่าเป็นตัวอักษรที่ไม่มีการติดกัน ดังรูปภาพที่ 4.5 และสรุปเป็นเส้นทางของตัวอักษรประเภทต่างๆที่ถูกวิเคราะห์และตัดแยกดังแผนภาพรูปที่ 4.6 หมายถึง หลังจากที่ได้ทำการตัดแยกของแต่ละประเภทแล้ว ส่วนที่ตัดแยกได้มีโอกาสที่จะเป็นประเภทใดบ้าง เช่น ประเภทที่ 6 ผลการตัดแยกที่ได้อาจเป็น ประเภทที่ 3 หรือประเภทที่ 4 เป็นต้น

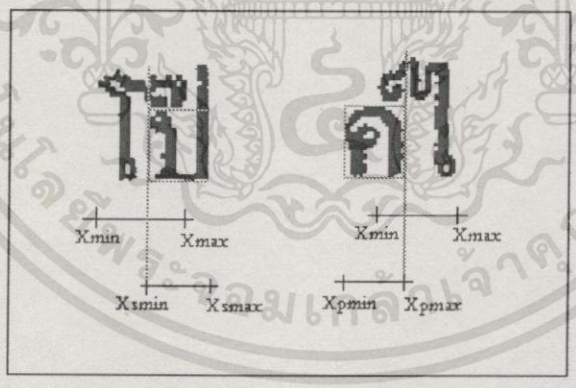


รูปที่ 4.5 แสดงขั้นตอนการวิเคราะห์การติดกันของตัวอักษร



รูปที่ 4.6 สรุปเส้นทางการวิเคราะห์และตัดแยก

4.5 การวิเคราะห์การเหลื่อมล้ำของตัวอักษร



รูปที่ 4.7 แสดงการเหลื่อมล้ำด้านหลังและด้านหน้า

กำหนดให้

Xmin = ค่าCoordinate ของ x ที่มีค่าต่ำสุดของตัวอักษรที่พิจารณา

Xmax = ค่าCoordinate ของ x ที่มีค่าสูงสุดของตัวอักษรที่พิจารณา

Xsmin = ค่าCoordinate ของ x ที่มีค่าต่ำสุดของตัวอักษรลำดับหลัง

Xsmax = ค่าCoordinate ของ x ที่มีค่าสูงสุดของตัวอักษรลำดับหลัง

Xpmin = ค่าCoordinate ของ x ที่มีค่าต่ำสุดของตัวอักษรลำดับก่อน

Xpmax = ค่าCoordinate ของ x ที่มีค่าสูงสุดของตัวอักษรลำดับก่อน

W = ค่าความกว้างของตัวอักษรโดยเฉลี่ย

กรณีที่มีการเหลื่อมล้ำด้านหลัง หมายถึง ตำแหน่งในแนวแกน X ของตัวอักษรที่พิจารณามีการซ้อนอยู่กับกับตำแหน่งของตัวอักษรลำดับหลัง หรือ เมื่อสมการต่อไปนี้เป็นจริง

$$Xmax - Xsmin > 1/3W$$

กรณีที่มีการเหลื่อมล้ำด้านหน้า หมายถึง ตำแหน่งในแนวแกน X ของตัวอักษรที่พิจารณามีการซ้อนอยู่กับกับตำแหน่งของตัวอักษรลำดับหน้า หรือ เมื่อสมการต่อไปนี้เป็นจริง

$$Xpmax - Xmin > 1/3W$$

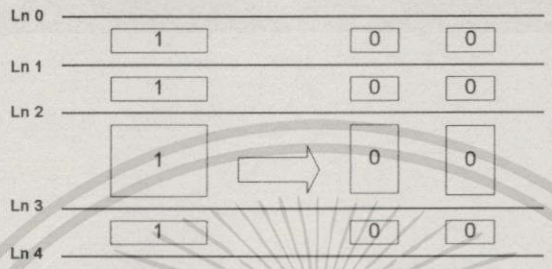
4.6 การตรวจสอบและแนวทางการตัดแยกตัวอักษร

ตารางที่ 4.4 แสดงตัวอักษรเดี่ยวในแต่ละประเภทของตัวอักษร

ประเภทที่ 1	ระดับเหนือบน	่ ้ ๊ ๋ ็
	ระดับบน	ิ ี ึ ื ั ์ ๆ ่ ้ ๊ ๋ ็
	ระดับกลาง	ะ เ า ก ข ค ฆ ง จ ฉ ช ซ ฌ ญ ฒ ณ ด ต ถ ท ธ น บ ผ พ ภ ม ย ร ล ว ศ ษ ส ห อ ฮ ฌ ฐ
	ระดับล่าง	ุ ู ฺ ฯ
ประเภทที่ 2	ระดับเหนือบน และระดับบน	-
ประเภทที่ 3	ระดับกลาง และระดับบน	ป ฝ ฟ ฬ ไ ๅ โ
ประเภทที่ 4	ระดับกลาง และระดับล่าง	ฤ ฦ ฎ ฏ ๅ ๆ ฯ
ประเภทที่ 5	ระดับเหนือบน ,ระดับบน และ ระดับกลาง	ไ ๅ โ
ประเภทที่ 6	ระดับบน ,ระดับกลาง และระดับล่าง	-
ประเภทที่ 7	ระดับเหนือบน, ระดับบน, ระดับกลาง และระดับล่าง	-

จากตารางที่ 4.4 จะเห็นว่ามีตัวอักษรเดี่ยวที่ยาวเกินกว่าหนึ่งระดับ ดังนั้นขั้นตอนนี้จึงเป็นการวิเคราะห์ตรวจสอบตัวอักษรที่ได้ว่า เป็นตัวอักษรเดี่ยวหรือเป็นตัวอักษรที่ติดกัน และถ้าติดกันจะมีแนวทางการตัดแยกอย่างไร แบ่งตามประเภทของตัวอักษรดังนี้

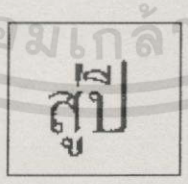
ประเภทที่ 1



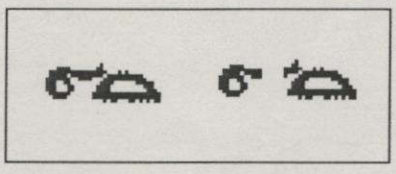
รูปที่ 4.8 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 1

จะเห็นภาพตัวอักษร มีความสูงไม่เกินกว่าระดับที่อยู่ แสดงว่าไม่มีการติดกันในแนวตั้ง จึงพิจารณาเฉพาะความกว้างของตัวอักษร ถ้าภาพตัวอักษรใดมีความกว้างเกินกว่าขนาดความกว้างหนึ่งตัวอักษร สรุปได้ว่ามีการติดกันของตัวอักษรในแนวนอนแต่ถ้าไม่มีแสดงว่าเป็นภาพของตัวอักษรเดี่ยว แนวทางการตัดแยกใช้ขอบของภาพของตัวอักษรระดับกลางเป็นแนวแบ่ง

ในกรณีนี้อาจมีการวิเคราะห์ที่ผิดพลาดได้ เป็นผลมาจาก ความใกล้เคียงกันของตัวอักษร ทำให้การวิเคราะห์ความกว้างของตัวอักษรผิดพลาด เช่น วรรณยุกต์เอก ติดกับสระ เป็นต้น แสดงตัวอย่างดังรูป 4.9 และบางกรณีลักษณะการติดกันของตัวอักษรจากแนวทางการตัดแยกที่กำหนด เมื่อทำการตัดแยกแล้วอาจทำให้ภาพตัวอักษรที่ได้อาจถูกตัดขาดหรือมีส่วนเกินออกมาได้ ดังตัวอย่างที่แสดงดังรูป 4.10

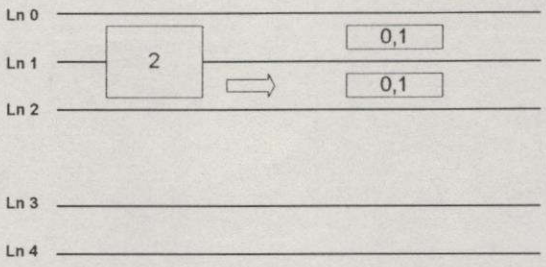


รูปที่ 4.9 แสดงลักษณะการวิเคราะห์ความกว้างผิดพลาด



รูปที่ 4.10 แสดงตัวอย่างการตัดแยกผิดพลาด

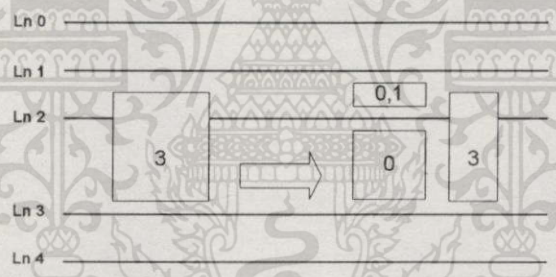
ประเภทที่ 2



รูปที่ 4.11 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 2

ในภาพตัวอักษรที่ได้ มีความสูงอยู่ในระดับบนและเหนือบน จากลักษณะของสระและวรรณยุกต์ในภาษาไทย ไม่มีสระหรือวรรณยุกต์บนตัวใดที่มีความสูงเกินกว่าหนึ่งระดับ ดังนั้นจึงสรุปได้ว่า มีการติดกันของตัวอักษรในแนวตั้ง แนวทางการตัดแยกอยู่ที่เส้นแบ่งระดับบนและเหนือบน (Ln1)

ประเภทที่ 3



รูปที่ 4.12 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 3

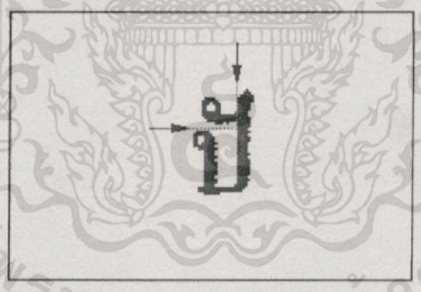
การตรวจสอบการติดกันในประเภทที่ 3 นี้จะเริ่มจากการหาแนวเส้นฮีสโตแกรมของตัวอักษรระดับกลางเพื่อทำการหากลุ่มของตัวอักษร จากนั้นจึงอาศัยข้อสังเกต ทำการวิเคราะห์หาประเภทของการติดกันและกำหนดแนวทางสำหรับการตัดแยก ในการวิเคราะห์การติดกันของตัวอักษรประเภทที่ 3 เพื่อกำหนดประเภทย่อย ที่สามารถบอกได้ว่าเป็นตัวอักษรเดี่ยว หรือเป็นตัวอักษรที่มีการติดกัน โดยเริ่มจากการพิจารณาตามความกว้างของตัวอักษร ถ้ามีความกว้างมากกว่า 1 ตัวอักษรก็จะถูกพิจารณาว่าเป็นประเภทที่ 3.5 แต่ถ้าไม่ใช่ก็จะใช้แนวเส้นตรงทั้งในแนวตั้งและแนวนอนจากข้อสังเกตที่ 1 และ 2 สามารถจำแนกได้ว่าเป็นตัวอักษรที่ติดกันประเภทที่ 3.6 ซึ่งวิเคราะห์ได้จากจำนวนเส้นตรงในแนวตั้งมากกว่า 2 หรือในแนวนอนในระดับ HL มากกว่าศูนย์ แต่ถ้าไม่อยู่ในประเภทที่ 3.6 และมีจำนวนเส้นตรงในแนวตั้งเท่ากับ 1 จากข้อสังเกตที่ 5 ในกรณีนี้อาจจำแนกได้เป็นกลุ่มตัวอักษรเดี่ยวประเภทที่ 3.2 เช่น "ใ" หรือถ้ามีการเหลื่อมล้ำกับตัวอักษรด้านหน้าหรือหลัง

ก็จะกำหนดให้เป็นตัวอักษรที่ติดกันประเภทที่ 3.4 ถ้าไม่อยู่ในประเภทที่ 3.6 และมีจำนวนเส้นตรงในแนวตั้งเท่ากับ 2 กลุ่มนี้ต้องผ่านขบวนการรู้จำ และมีการตรวจสอบค่าฮีสโตแกรมในระดับบนและตรวจสอบการเหลื่อมล้ำของตัวอักษรเพื่อแยกระหว่างตัวอักษรเดี่ยวในประเภทที่ 3.1 กับกลุ่มตัวอักษรที่ติดกันในประเภทที่ 3.3 และ 3.6 สามารถเขียนขั้นตอนการวิเคราะห์การติดกันของประเภทที่ 3 ดังรูปที่ 4.20 ด้วยวิธีการนี้เราสามารถระบุ การติดกันในประเภทที่ 3 ดังนี้

ประเภทที่ 3.1 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T2, V2T4 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("พ","บ","ผ") และตรวจสอบค่าฮีสโตแกรมในระดับบนด้วยวิธี Pixel Projection แล้วว่าไม่มีสระหรือวรรณยุกต์ติดอยู่ให้สรุปว่าเป็นตัวอักษรเดี่ยว

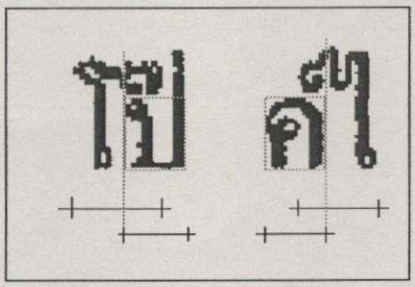
ประเภทที่ 3.2 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V1T2 และไม่มีการเหลื่อมล้ำของตัวอักษรในลำดับหน้าหรือหลัง อาศัยข้อสังเกตในหัวข้อที่ 4.3 ข้อที่ 5 แสดงได้ว่าเป็นตัวอักษรเดี่ยว

ประเภทที่ 3.3 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T2, V2T4 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("พ","บ","ผ") และตรวจสอบค่าฮีสโตแกรมในระดับบนด้วยวิธี Pixel Projection แล้วว่ามีสระหรือวรรณยุกต์ติดอยู่ และตรวจสอบการเหลื่อมล้ำกับตัวอักษรลำดับหลังแล้วพบว่าไม่มี แสดงว่ามีการติดกันของตัวอักษร ให้ใช้ภูเขาที่สองและแนวเส้น Ln2 เป็นการกำหนดแนวตัดดังรูปที่ 4.13



รูปที่ 4.13 การแสดงแนวการตัดในกลุ่ม("บ","พ","ผ")

ประเภทที่ 3.4 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V1T2 และมีการเหลื่อมล้ำของตัวอักษรในลำดับหน้าหรือลำดับหลัง อาศัยข้อสังเกตในหัวข้อที่ 4.3 ข้อที่ 5 แสดงได้ว่ามีตัวอักษรที่ติดกันในระดับบน ตัวอย่างดังรูปที่ 4.14 ให้กำหนดแนวทางการตัดแยกตามขอบของตัวอักษรลำดับหน้าหรือหลังที่มีการเหลื่อมล้ำ

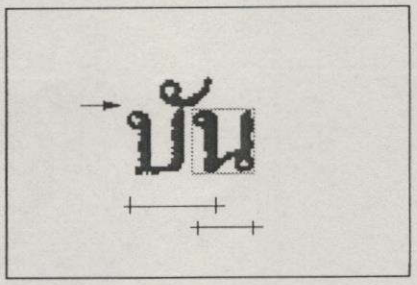


รูปที่ 4.14 แสดงการติดโดยความกว้างไม่เกิน 1 ตัวอักษร

ประเภทที่ 3.5 สามารถระบุได้จากความกว้างของตัวอักษรถ้ามากกว่า 1 ตัวอักษรให้ตรวจสอบว่ามีการเหลื่อมล้ำกับตัวอักษรลำดับหน้าหรือลำดับหลังหรือไม่ ถ้ามีแสดงได้ว่าอยู่ในกลุ่มตัวอักษรประเภทที่ 3.5 ให้กำหนดแนวทางการตัดแยก อยู่ที่กึ่งกลางของความกว้างของตัวอักษรดังรูปที่ 4.15

รูปที่ 4.15 แสดงการติดโดยความกว้างเกิน 1 ตัวอักษร

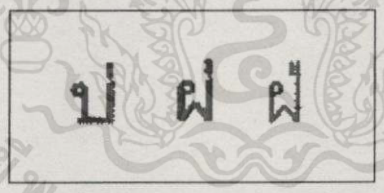
ประเภทที่ 3.6 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V3, V1T3, V2T1, V2T3 อาศัยข้อสังเกตในหัวข้อที่ 4.3 ข้อที่ 1, 2 หรือจากกลุ่ม V2T2, V2T4 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("ข","ฃ","ย","ษ","ม","ฆ","น") แสดงได้ว่าแสดงได้ว่ามีตัวอักษรที่ติดกันในระดับบน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง(Ln2) หรือ สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T2, V2T4 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("พ","บ","ผ") และตรวจสอบค่าฮีสโตแกรมในระดับบนด้วยวิธี Pixel Projection แล้วว่ามีสระหรือวรรณยุกต์ติดอยู่ และตรวจสอบการเหลื่อมล้ำกับตัวอักษรลำดับหลังแล้วพบว่ามีการเหลื่อมล้ำ กันดังตัวอย่างรูปที่ 4.16 แสดงได้ว่ามีตัวอักษรที่ติดกันในระดับบน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบน และระดับกลาง(Ln2) เช่นเดียวกัน



รูปที่ 4.16 แสดงตัวอักษรกลุ่ม ("พ","บ","ผ") ที่มีการเหลื่อมล้ำ

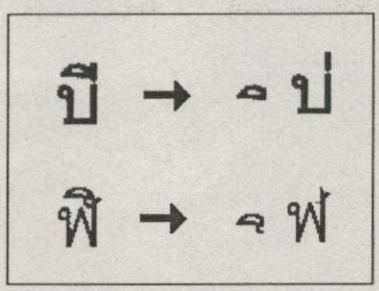
ในกรณีนี้ถ้าพิจารณาตามผังงานขั้นตอนการวิเคราะห์การติดกันของตัวอักษรในประเภทที่ 3 จะเห็นว่ามีการวิเคราะห์ผิดพลาดได้ในบางกรณีดังนี้

กรณีที่หนึ่ง คือกรณีที่ตัวอักษรที่อยู่ในกลุ่ม ("พ" ,"บ", "ผ", "ฟ", "ป" ,"ฝ") ติดกับไม้เอก แสดงได้ดังตัวอย่างรูปที่ 4.17 เมื่อวิเคราะห์ตามแผนงานขั้นตอนการวิเคราะห์แล้ว จะเห็นว่าตัวอักษรที่ติดกันประเภทนี้จะเป็นกลุ่มที่มีแนวเส้นตรงในแนวตั้งเท่ากับ 2 และมีจำนวนเส้นตรงในแนวนอนระดับ HL เท่ากับศูนย์ จากผังการทำงานจึงถูกส่งเข้าสู่ขบวนการรู้จำก็ถูกวิเคราะห์ตัวอักษรระดับกลางว่าอยู่ในกลุ่ม "พ","บ","ผ" เมื่อตรวจสอบฮีสโตแกรมระดับบนจึงไม่พบว่ามีแนวเส้นตรงเพราะตัวอักษรที่มาติดมีขนาดเล็ก(ตัวไม้เอก) จึงถูกจัดเข้ากลุ่มของประเภทพยัญชนะเดี่ยว คือประเภทที่ 3.1 ทั้งที่เป็นประเภทตัวอักษรที่ติดกัน



รูปที่ 4.17 แสดงการวิเคราะห์ผิดพลาดให้ผลเป็นอักษรเดี่ยว

กรณีที่สอง คือกรณีที่มีการติดกันของตัวอักษรที่ภาพตัวอักษรระดับกลางอยู่ในกลุ่ม ("พ" ,"บ", "ผ") ติดกับสระที่มีตำแหน่งการวางที่ไม่เหลื่อมไปข้างหลัง เช่น สระอิ , สระอี แสดงได้ดังตัวอย่างรูปที่ 4.18 เมื่อวิเคราะห์ตามแผนงานขั้นตอนการวิเคราะห์แล้ว เช่นเดียวกับในกรณีที่หนึ่งคือตามผังการทำงานจะต้องถูกส่งเข้าสู่ขบวนการรู้จำและวิเคราะห์ตัวอักษรระดับกลางว่าอยู่ในกลุ่ม "พ"," บ","ผ" เมื่อตรวจสอบฮีสโตแกรมระดับบนจึงพบว่ามีแนวเส้นตรง แต่จะไม่พบการเหลื่อมล้ำของตัวอักษรเนื่องจากสระที่มาติดเป็นสระที่ไม่มีตำแหน่งที่เยื้องไปด้านหลังจึงถูกจัดเข้ากลุ่มของประเภทที่ 3.3 ทำให้เสนอแนวทางการตัดแยกที่ผิดพลาด และแสดงผลการตัดแยกดังรูปที่ 4.18 ทั้งที่ที่ถูกต้องควรจะถูกจัดเข้าประเภทที่ 3.6

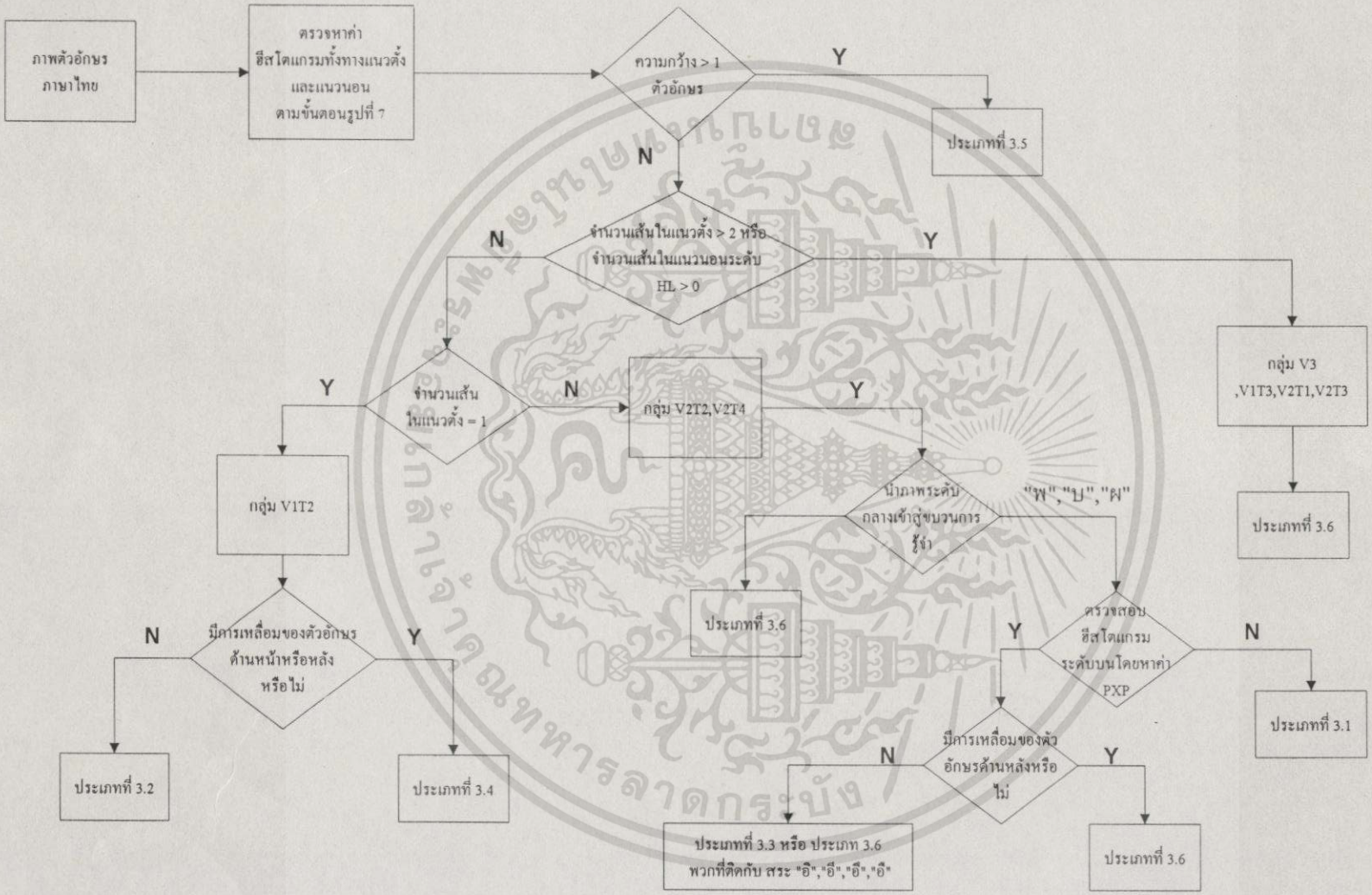


รูปที่ 4.18 การวิเคราะห์ที่ผิดทำให้การตัดผิดพลาด

กรณีที่สาม คือกรณีที่ภาพขนาดของตัวอักษรบางตัว เช่น วรรณยุกต์เอกเมื่อติดกับตัวอักษรในกลุ่ม V1T2 (“ไ”,“ ใ”, “โ”) แล้วบางกรณี การวิเคราะห์การเหลื่อมล้ำกันของตัวอักษรอาจผิดพลาดได้ เช่น ภาพตัวอย่างดังรูปที่ 4.19 เมื่อวิเคราะห์ตามผังการทำงานพวกนี้จะมีแนวเส้นตรงในแนวตั้งเท่ากับ 1 จึงต้องทำการวิเคราะห์การเหลื่อมล้ำกับตัวอักษรด้านหน้าหรือหลัง แต่เนื่องจากเมื่อตรวจสอบแล้วไม่พบการเลื่อมล้ำเพราะขนาดของตัวอักษรที่มาติด(ไม้เอก)มีขนาดเล็ก จึงถูกวิเคราะห์เข้ากลุ่มประเภทที่ 3.2 ซึ่งเป็นกลุ่มอักษรเดี่ยว ซึ่งที่ถูกต้องควรจะเป็นประเภทที่ 3.4

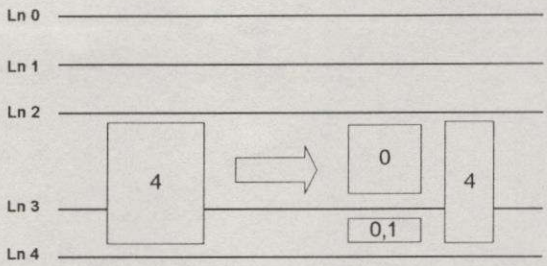
รูปที่ 4.19 แสดงการวิเคราะห์ที่ผิดเนื่องจากขนาดของตัวอักษร

ในทั้งสามกรณีจากผลการทดลองจะพบว่าในกรณีที่หนึ่งและสองจะพบมากในอัตราส่วนที่ใกล้เคียงกันแต่ในกรณีที่ สามจะพบน้อยที่สุด



รูปที่ 4.20 แสดงขั้นตอนในการตรวจสอบในประเภทที่ 3

ประเภทที่ 4



รูปที่ 4.21 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 4

การวิเคราะห์ประเภทที่ 4 ก็ใช้แนวทางเดียวกับประเภทที่ 3 คือจะเริ่มจากการหาแนวเส้นฮีสโตแกรมของตัวอักษรระดับกลางเพื่อใช้ในการวิเคราะห์ดังต่อไปนี้ ในการวิเคราะห์ประเภทที่ 4 สามารถแบ่งกว้างๆเป็นสองกลุ่มคือกลุ่มที่มีแนวเส้นตรงในระดับล่าง และกลุ่มที่ไม่มี ในแต่ละกลุ่มก็มีทั้งตัวอักษรเดี่ยวและตัวอักษรที่ติดกัน ทั้งสองกลุ่มขั้นแรกอาศัยข้อสังเกตที่ 1,3 และ 4 สามารถจำแนกได้ว่าเป็นตัวอักษรที่ติดกันประเภทที่ 4.4 หรือ 4.5 ซึ่งวิเคราะห์ได้จากจำนวนเส้นตรงในแนวตั้งมากกว่า 2 หรือ จำนวนเส้นในแนวนอนระดับ LL มากกว่าศูนย์ หรือระดับ HLและLLเท่ากับศูนย์ ถ้าไม่อยู่ในประเภทเหล่านี้ ในกลุ่มที่ไม่มีแนวเส้นตรงในระดับล่าง ก็จะแยกตัวอักษรเดี่ยวในกลุ่มประเภทที่ 4.2 (“ๆ”,”ๆ”) ได้โดยใช้ข้อสังเกตที่ 6 คือมีจำนวนแนวเส้นตรงในแนวตั้งเท่ากับ 1 แต่ถ้าไม่อยู่ในกลุ่มที่ 4.2 ก็จะต้องเข้าสู่การรู้จำเพื่อแยกพวกอักษรเดี่ยวประเภทที่ 4.1 (“ฤ”,”ฦ”) กับตัวอักษรที่ติดกันในประเภทที่ 4.5 ส่วนในกลุ่มที่มีแนวเส้นตรงในระดับล่างก็จะต้องเข้าสู่การรู้จำเช่นกันเพื่อแยกพวกอักษรเดี่ยวประเภทที่ 4.3 (“ฎ”,”ฏ”) กับตัวอักษรที่ติดกันในประเภทที่ 4.4 สามารถเขียนขั้นตอนการวิเคราะห์การติดกันของประเภทที่ 4 ดังรูปที่ 4.24 ด้วยวิธีการนี้เราสามารถระบุ การติดกันในประเภทที่ 4 ดังนี้

ประเภทที่ 4.1 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T1 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("ถ","ภ") และตรวจสอบค่าฮีสโตแกรมในระดับล่างด้วยวิธี Pixel Projection แล้วว่าไม่มีสระอยู่ให้สรุปว่าเป็นตัวอักษรเดี่ยว

ประเภทที่ 4.2 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V1T1 และอาศัยข้อสังเกตในหัวข้อ 4.3 ข้อที่ 6 ให้สรุปว่าเป็นตัวอักษรเดี่ยว

ประเภทที่ 4.3 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T1 ที่ผ่านขบวนการรู้จำของตัวอักษรระดับกลางแล้วได้ตัวอักษรที่อยู่ในกลุ่ม ("ถ","ภ") และตรวจสอบค่าฮีสโตแกรมในระดับล่างแล้วว่าน่าจะมีสระอยู่ให้สรุปว่าเป็นตัวอักษรเดี่ยว

ประเภทที่ 4.4 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V3,V1T3, V2T2,V2T3,V2T4 และอาศัยข้อสังเกตในหัวข้อ 4.3 ข้อที่ 1, 3 และ 4 หรือ ตัวอักษรที่อยู่ในกลุ่ม V2T1 ที่นำเข้าสู่

ขบวนการรู้จำแล้วอยู่ในกลุ่มที่อยู่ในกลุ่มที่ไม่ใช่ ("ถ","ภ") และตรวจสอบ ค่าฮีสโตแกรมในระดับล่างด้วยวิธี Pixel Projection แล้วว่าน่าจะมีสระอยู่ให้สรุปว่าเป็นตัวอักษรที่ติดกัน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับกลางและระดับล่าง(Ln3)

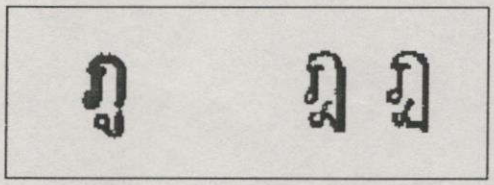
ประเภทที่ 4.5 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V3,V1T3, V2T2,V2T3,V2T4 และอาศัยข้อสังเกตในหัวข้อ 4.3 ข้อที่ 1, 3 และ 4 หรือ ตัวอักษรที่อยู่ในกลุ่ม V2T1 ที่นำเข้าสู่ขบวนการรู้จำแล้วอยู่ในกลุ่มที่อยู่ในกลุ่มที่ไม่ใช่ ("ถ","ภ") และตรวจสอบ ค่าฮีสโตแกรมในระดับล่างด้วยวิธี Pixel Projection แล้วว่าไม่มีสระอยู่ให้สรุปว่าเป็นตัวอักษรที่ติดกัน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับกลางและระดับล่าง(Ln3)

เช่นเดียวกับในประเภทที่ 3 เมื่อพิจารณาตามผังงานการวิเคราะห์ อาจพบข้อผิดพลาดในการวิเคราะห์ได้ในบางกรณีดังนี้

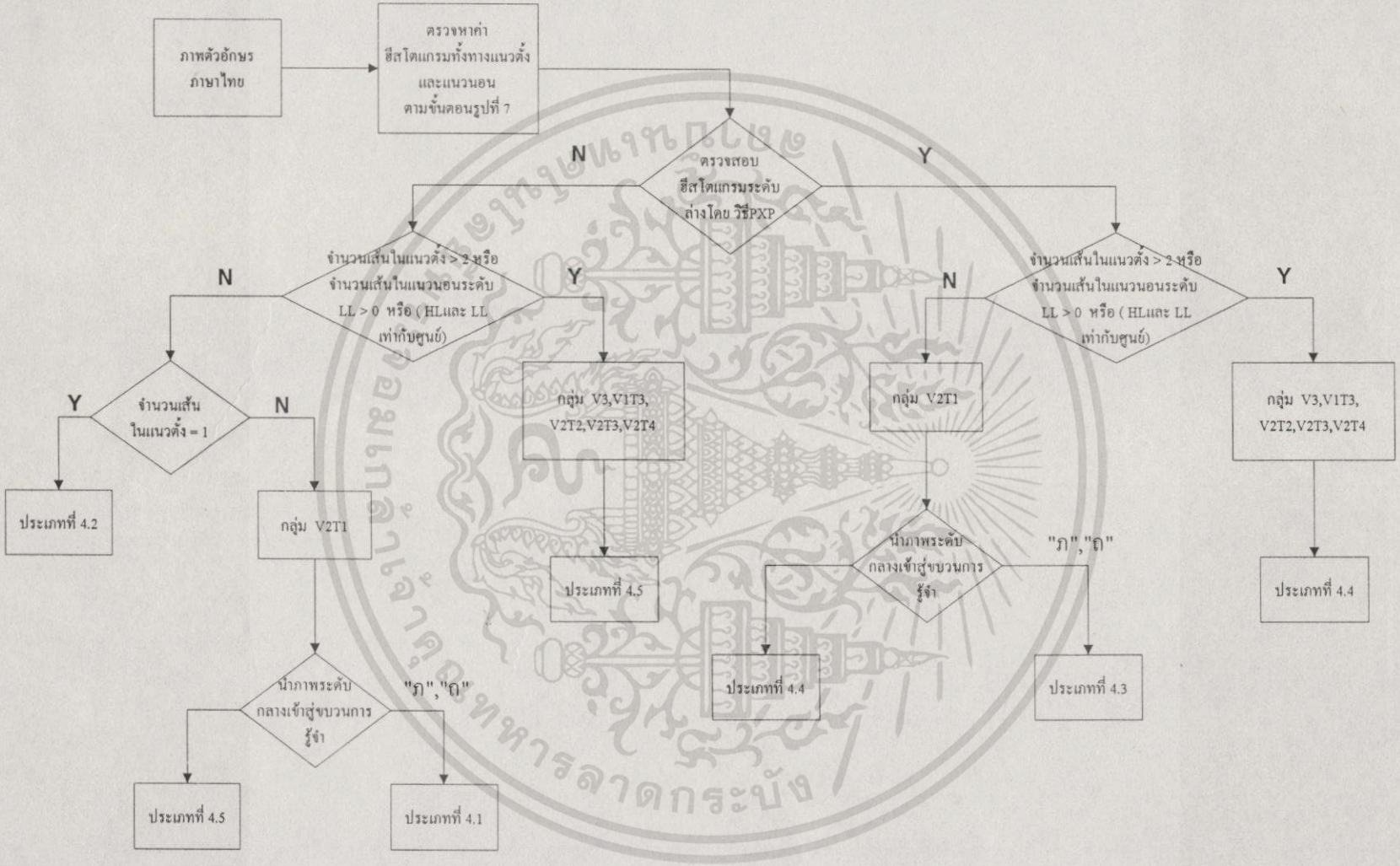
กรณีที่หนึ่ง คือกรณีที่ภาพตัวอักษร “ภ” ติดกับสระ”อุ” แสดงดังรูปที่ 4.22 เมื่อวิเคราะห์ตามผังงาน จะเห็นว่าพวกนี้เป็นกลุ่มที่ไม่มีมีแนวเส้นตรงในระดับล่าง และมีแนวเส้นตรงในแนวตั้งเท่ากับสอง มีจำนวนแนวเส้นตรงระดับ HL เท่ากับ 1 ระดับ LL เท่ากับศูนย์ จึงต้องถูกส่งตัวอักษรระดับกลางเข้าสู่ขบวนการรู้จำ และถูกวิเคราห์ว่าอยู่ในกลุ่ม “ภ”,”ถ” จะถูกจัดเข้ากลุ่มประเภทตัวอักษรเดี่ยวคือประเภท 4.1 คือไม่สามารถแยกระหว่างตัวอักษร “ฦ” กับ “ภ” ที่ติดกับสระ ”อุ” ได้ ทั้งที่ถูกต้องควรถูกจัดเข้ากลุ่ม 4.5

รูปที่ 4.22 แสดงภาพเปรียบเทียบตัวอักษรเดี่ยวและ “ถ” ที่ติดกับสระ ”อุ”

กรณีที่สอง คือกรณีที่ภาพตัวอักษร “ภ” ที่ติดกับสระ”อู “แสดงดังรูปที่ 4.23 เมื่อวิเคราะห์ตามผังงานก็จะให้ผลการวิเคราะห์ที่ผิดพลาด เช่นเดียวกันกับในกรณีที่หนึ่ง แต่ในกรณีนี้จะถูกจัดเข้ากลุ่มอักษรเดี่ยวประเภทที่ 4.3 แทนที่จะเป็นประเภทตัวอักษรติดกันในประเภทที่ 4.4 คือไม่สามารถแยกระหว่างตัวอักษร “ฎ”, “ฏ” กับ “ภ” ที่ติดกับสระ “อู” ได้

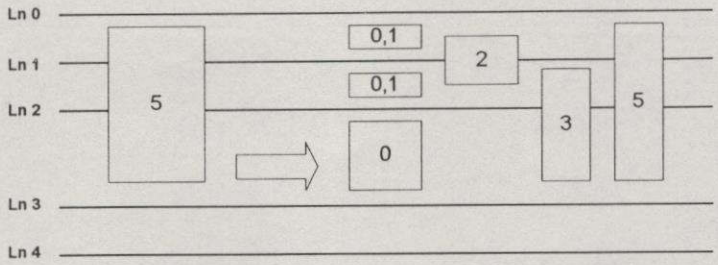


รูปที่ 4.23 แสดงภาพเปรียบเทียบตัวอักษรเดี่ยวและ “ภ” ที่ติดกับสระ “อู”



รูปที่ 4.24 แสดงขั้นตอนในการตรวจสอบในประเภทที่ 4

ประเภทที่ 5



รูปที่ 4.25 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 5

การวิเคราะห์ประเภทที่ 5 ก็ใช้แนวทางเดียวกับประเภทที่ 3 คือจะเริ่มจากการหาแนวเส้นฮีสโตแกรมของตัวอักษรระดับกลาง และนำค่าที่ได้มาทำการวิเคราะห์ดังต่อไปนี้ การวิเคราะห์การติดกันในประเภทที่ 5 จะต่างจากประเภทที่ 3 คือมีตัวอักษรเดียวที่มีโอกาสเกิดได้เพียงกลุ่มเดียวคือประเภทที่ 5.1 การวิเคราะห์จะเริ่มจากการหาค่าความกว้าง ถ้ามากกว่า 1 ตัวอักษรจะถูกกำหนดให้เป็นประเภทที่ 5.4 แต่ถ้าไม่ใช่ก็จะหาตัวอักษรที่ติดกันได้จากข้อสังเกตที่ 1 และ 2 สามารถจำแนกได้ว่าเป็นตัวอักษรที่ติดกันประเภทที่ 5.5 ซึ่งวิเคราะห์ได้จากจำนวนเส้นตรงในแนวตั้งมากกว่า 2 หรือ จำนวนเส้นตรงในแนวนอนระดับ HL มากกว่าศูนย์ และจากข้อสังเกตที่ 3 และ 4 คือเป็นตัวอักษรที่ไม่ได้อยู่ในกลุ่มประเภทที่ 5.5 และมีแนวเส้นตรงในแนวตั้งมากกว่า 1 ให้เป็นตัวอักษรที่ติดกันในประเภทที่ 5.2 ส่วนในกลุ่มที่มีแนวเส้นตรงในแนวตั้งเท่ากับ 1 จะต้องทำการตรวจสอบว่ามีการติดกันของตัวอักษรหรือไม่ด้วยความเหลื่อมล้ำกันหรือไม่ ถ้ามีก็จะเป็นตัวอักษรติดกันประเภทที่ 5.3 แต่ถ้าไม่มีก็จะเป็นตัวอักษรเดี่ยวประเภทที่ 5.1 (“ใ”,”ไ”,”โ”) สามารถเขียนขั้นตอนการวิเคราะห์การติดกันของประเภทที่ 5 ดังรูปที่ 4.26 ด้วยวิธีการนี้เราสามารถระบุ การติดกันในประเภทที่ 5 ดังนี้

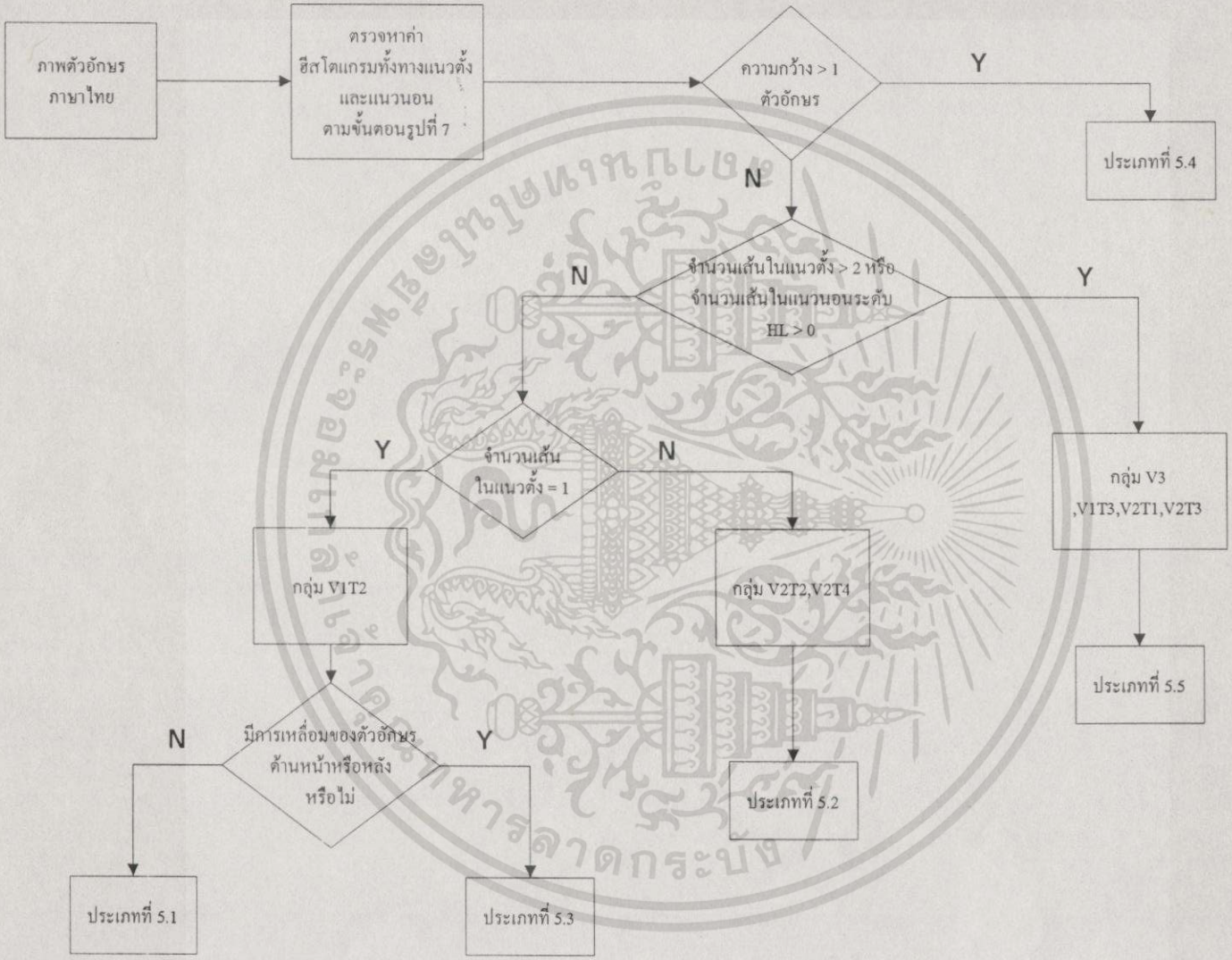
ประเภทที่ 5.1 การวิเคราะห์และกำหนดจุดตัดใช้วิธีการเดียวกับประเภทที่ 3.2

ประเภทที่ 5.2 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม V2T2, V2T4 เนื่องจากกลุ่มนี้ไม่ตัวใดที่สูงถึงระดับเหนือบน แสดงได้ว่ามีตัวอักษรที่ติดกันในระดับเหนือบน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับเหนือบนและระดับบน(Ln1)

ประเภทที่ 5.3 การวิเคราะห์และกำหนดจุดตัดใช้วิธีการเดียวกับประเภทที่ 3.4

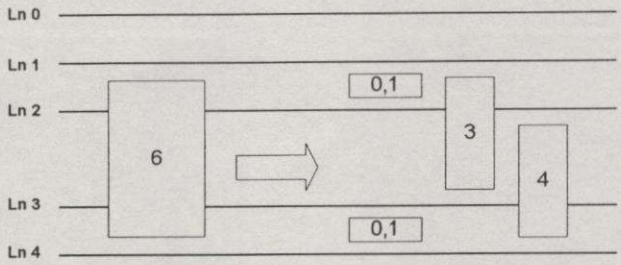
ประเภทที่ 5.4 การวิเคราะห์และกำหนดจุดตัดใช้วิธีการเดียวกับประเภทที่ 3.5

ประเภทที่ 5.5 สามารถระบุได้จากกลุ่มของตัวอักษรที่อยู่ในกลุ่ม อยู่ในกลุ่ม V3 , V1T3, V2T1 และ V2T3 และอาศัยข้อสังเกตในหัวข้อที่ 4.3 ข้อที่ 1,2 แสดงได้ว่ามีตัวอักษรที่ติดกันในระดับบน ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง(Ln2)



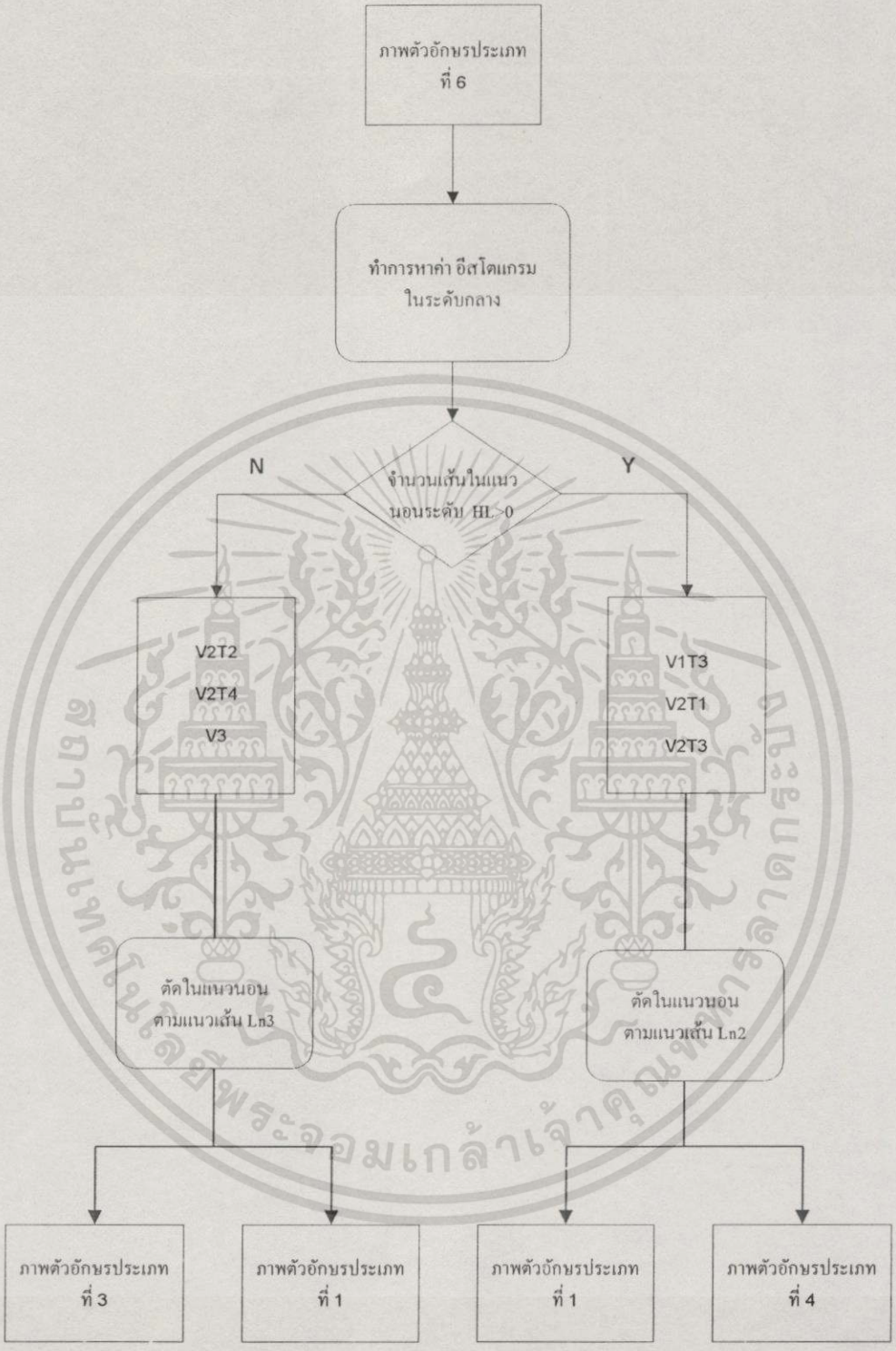
รูปที่ 4.26 แสดงขั้นตอนในการตรวจสอบในประเภทที่ 5

ประเภทที่ 6



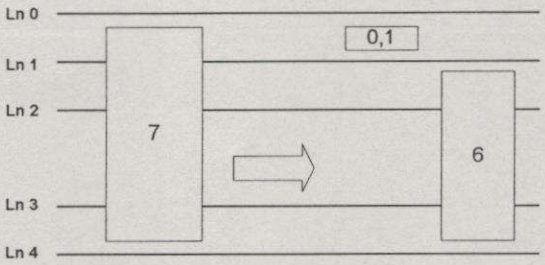
รูปที่ 4.27 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 6

ในประเภทนี้ เมื่อพิจารณาตามระดับของตัวอักษรแล้ว ไม่พบว่าตัวอักษรไทยตัวใดที่มีความสูงที่มีความสูงจากระดับบนจนถึงระดับล่าง จึงสรุปได้ว่า มีการติดกันของตัวอักษรในระดับบนหรือในระดับล่างอย่างแน่นอน แต่เนื่องจากพยัญชนะไทย มีความสูงเกินกว่าระดับกลางได้ทั้งสองระดับ เช่น "ป" หรือ "ฤ" ทำให้ไม่สามารถระบุแนวทางการตัดแยกได้ จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด จากนั้นใช้ข้อสังเกตของตัวอักษรในระดับกลางในหัวข้อที่ 4.3 ข้อที่ 2 คือถ้าตกอยู่ในกลุ่ม V1T3,V2T1 และ V2T3 ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง(Ln2) แต่ถ้าไม่มีตกอยู่ในกลุ่มนี้ ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับกลางและระดับล่าง(Ln3) สามารถเขียนขั้นตอนการวิเคราะห์การติดกันของประเภทที่ 6 ดังรูปที่ 4.28



รูปที่ 4.28 แสดงขั้นตอนการตรวจสอบในประเภทที่ 6

ประเภทที่ 7



รูปที่ 4.29 ผลลัพธ์ที่เป็นไปได้จากการตัดแยกของประเภทที่ 7

จากการที่ไม่มีตัวอักษรไทยตัวใดสูงถึง 4 ระดับ และไม่มีโอกาส ที่ตัวอักษรที่สูงถึง 3 ระดับ จะติดกับสระระดับล่าง เช่น "ใ" ไม่มีโอกาสติดกับสระ "ุ" แสดงว่ามีการติดของสระหรือวรรณยุกต์ในระดับเหนือระดับบนอย่างแน่นอน จึงสามารถที่จะทำการตัดในระดับเหนือระดับบน(Ln1)ได้ และส่วนล่างก็จะถูกจัดเข้าประเภท 6 เพื่อทำการวิเคราะห์อีกครั้ง

4.7 การตัดแยกตัวอักษรที่ติดกัน

จากขั้นตอนการวิเคราะห์การติดกันของตัวอักษร เมื่อมีการวิเคราะห์ได้ผลว่ามีการติดกันของตัวอักษร ก็จะส่งภาพตัวอักษรที่มีการติดกันนั้น เข้าสู่ขั้นตอนการตัดแยกตัวอักษร โดยมีการกำหนดแนวทางการตัดแยกมาด้วยว่า เป็นการตัดแยกในแนวตั้งหรือแนวนอน และกำหนดแนวเส้นของการตัดแยกมาได้ด้วย ในขั้นตอนการตัดแยกตัวอักษรที่ติดกันนี้ งานวิจัยนี้จำนำวิธีการหาค่า Break Cost มาใช้กำหนดเส้นตัดแยกตัวอักษร วิธีนี้เริ่มจากการกำหนดช่วงของการคำนวณ หาค่า Break Cost ในกรณีที่มีการติดกันในแนวนอน กำหนดช่วงของการคำนวณดังนี้

$$\text{ตำแหน่งเริ่มต้น} = Lny - \in$$
$$\text{ตำแหน่งสิ้นสุด} = Lny + \in$$

เมื่อ Lny คือตำแหน่งเส้นแนวนอนของแนวทางการตัดแยกที่ได้มาจากการวิเคราะห์ และ

$\in = 1/8(Ymax - Ymin)$ ตัวอย่างดังรูปที่ 4.30

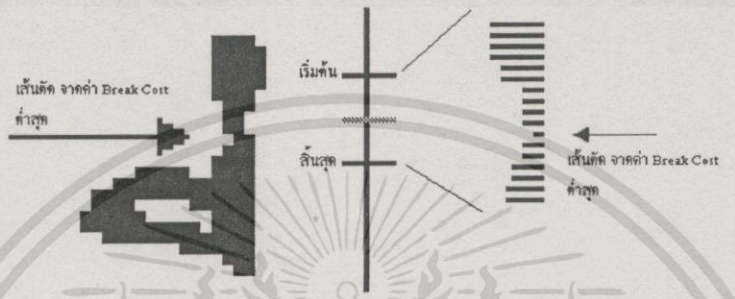
ในกรณีที่มีการติดกันในแนวตั้ง กำหนดช่วงของการคำนวณดังนี้

$$\text{ตำแหน่งเริ่มต้น} = Lnx - \in$$
$$\text{ตำแหน่งสิ้นสุด} = Lnx + \in$$

เมื่อ Lnyx คือตำแหน่งเส้นแนวตั้งของแนวทางการตัดแยกที่ได้มาจากการวิเคราะห์ และ

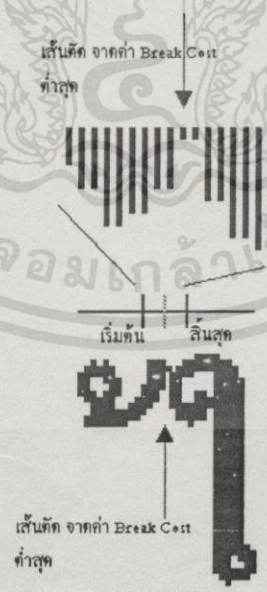
$\in = 1/8(Xmax - Xmin)$ ตัวอย่างดังรูปที่ 4.31

จากนั้นทำการหาค่า Break Cost ในแนวนอน โดยวิธีการนับจำนวนจุดในแนวนอนในช่วงที่มีการซ้อนกันของเนื้อตัวอักษรในแนวดิ่ง(นำ 2 แถวที่ติดกันมาทำการ AND กัน) แสดงผลในรูปของกราฟฮีสโตแกรมดังตัวอย่างรูปที่ 4.30



รูปที่ 4.30 แสดงการหาเส้นตัดในแนวนอน และขอบเขตของการคำนวณ

กำหนดให้เส้นกราฟที่สั้นที่สุดคือ ตำแหน่งที่แบ่งแยกระหว่างแนวนอนสองแนวที่ติดกัน การหาค่าแนวตั้งก็เช่นเดียวกันดังรูปที่ 4.31



รูปที่ 4.31 แสดงการหาเส้นตัดในแนวตั้ง และขอบเขตของการคำนวณ

4.8 การจัดเรียงรูปประโยคหลังการรู้จำ

หลังจากแยกข้อมูลภาพตัวอักษร แต่ละตัวออกมาได้แล้ว และทำการตัดแยกตัวอักษรที่ติดกันออกมาได้หมดแล้ว เราจะได้ ลำดับภาพตัวอักษรในแต่ละระดับ ในลำดับที่ถูกต้อง แสดงดังตัวอย่างรูปที่ 4.32 และในแต่ละตัวอักษรมีการบันทึก

1) เลขระดับ
2) ตำแหน่งพิกัด Xmin, Xmax, Ymin, Ymax
3) ตำแหน่งกึ่งกลางของตัวอักษรในแนวนอน (Xcen)

ก่อนที่จะถูกส่งต่อไปทำการรู้จำตัวอักษร

รูปที่ 4.32 แสดงลำดับของตัวอักษรที่ถูกตัดแยกแล้วในแต่ละระดับ

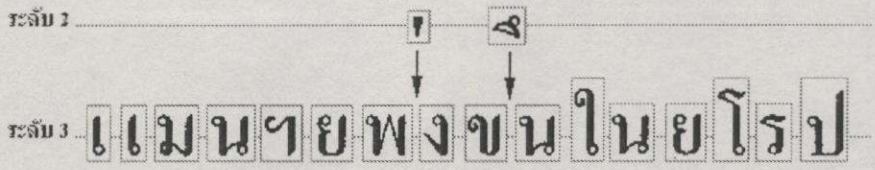
และเมื่อตัวอักษรถูกส่งไปวิเคราะห์ จนทราบว่าเป็นอักษรตัวอะไรแล้ว รหัสแอสกี้ของตัวอักษรของทุกตัวจะถูกนำมาจัดเรียงลำดับให้ในระดับเดียวกัน และกรณีที่อักษรที่อยู่ในคอลัมน์เดียวกัน จะจัดด้วยตำแหน่งดังนี้

ระดับที่ 3 + ระดับที่ 2 + ระดับที่ 1 + ระดับที่ 4

ดังนั้นในงานวิจัยนี้เสนอขั้นตอนการจัดเรียงลำดับตัวอักษร จากรูปแบบของลำดับตัวอักษรที่ถูกจัดเก็บเป็นระดับ ดังนี้

ขั้นตอนที่ 1

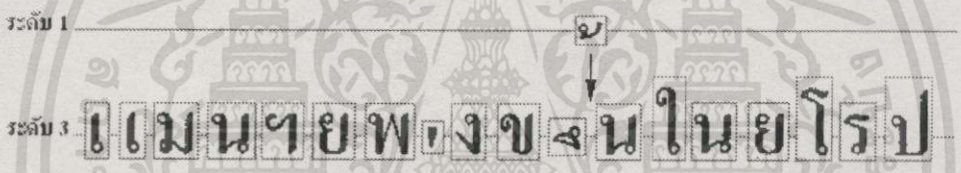
นำตัวอักษรในระดับที่ 2 มาแทรกในลำดับตัวอักษรในระดับที่ 3 โดยตำแหน่งที่แทรกคือ ข้างหน้าของตัวอักษรที่มีเลขระดับเป็น 3 และมีค่า Xmin มากกว่าค่า Xcen ของตัวอักษรระดับ 2 ที่นำมาแทรก ถ้าไม่มีก็ให้ใส่เป็นตัวสุดท้าย



รูปที่ 4.33 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 2

ขั้นตอนที่ 2

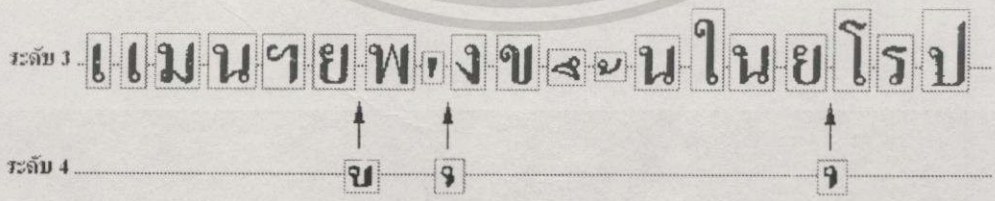
นำตัวอักษรในระดับที่ 1 มาแทรกในลำดับตัวอักษรในระดับที่ 3 โดยตำแหน่งที่แทรกคือ ข้างหน้าของตัวอักษรที่มีเลขระดับเป็น 3 และมีค่า Xmin มากกว่าค่า Xcen ของตัวอักษรระดับ 1 ที่นำมาแทรก ถ้าไม่มีก็ให้ใส่เป็นตัวสุดท้าย



รูปที่ 4.34 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 1

ขั้นตอนที่ 3

นำตัวอักษรในระดับที่ 4 มาแทรกในลำดับตัวอักษรในระดับที่ 3 โดยตำแหน่งที่แทรกคือ ข้างหน้าของตัวอักษรที่มีเลขระดับเป็น 3 และมีค่า Xmin มากกว่าค่า Xcen ของตัวอักษรระดับ 4 ที่นำมาแทรก ถ้าไม่มีก็ให้ใส่เป็นตัวสุดท้าย



รูปที่ 4.35 แสดงลำดับของตัวอักษรระดับที่3 ถูกแทรกโดยตัวอักษรในระดับที่ 4

หลังจากสิ้นสุดขั้นตอนที่ 3 แล้วจะได้ ลำดับของตัวอักษรที่มีการเรียงที่ถูกต้องพร้อมที่จะนำไปบันทึกลงในหน่วยความจำ

ระดับ 3 แเมนฯยขพ,จงข๔ะนในย,โรป

รูปที่ 4.36 แสดงลำดับของตัวอักษรที่เรียงอย่างถูกต้อง

บทที่ 5

ผลการทดสอบ

ผลการทดสอบจากการใช้วิธีวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรตัวพิมพ์ไทยระดับบนและระดับล่าง โดยใช้คุณสมบัติทางแนวตั้งและแนวนอนของฮีสโตแกรม กับรูปภาพประโยคที่ได้จาก หนังสือพิมพ์ไทยรัฐ, นิตยสารมติชนสุดสัปดาห์ และวรสาร INFORACLE ผ่านการแสกนด้วยความละเอียด 300 จุดต่อนิ้ว โดยเลือกเฉพาะบรรทัดที่มีตัวติดกันในระดับบนและระดับล่าง ทดสอบโดยเครื่องคอมพิวเตอร์ส่วนบุคคล(Personal Computer) Pentium 133 MHz, Memory 32 MB โดยได้ทำการทดสอบในหัวข้อต่อไปนี้

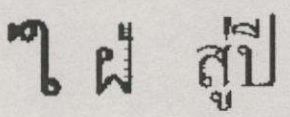
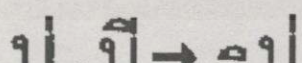
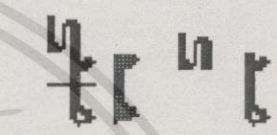
5.1 ผลการทดสอบความถูกต้องของการวิเคราะห์การติดกันของตัวอักษร

ข้อมูลภาพที่ได้จากสิ่งพิมพ์แต่ละประเภท มีการเลือกบทความที่แตกต่างกันออกไป ในบรรทัดที่มีขนาดเดียวกันทั้งหมด และไม่มีตัวหนังสือเอียง หลังจากผ่านการแยกตัวอักษรด้วยวิธีการเปลี่ยนรหัสขอบแล้ว ภาพตัวอักษรที่ได้จะถูกนำเข้าสู่ขั้นตอนการวิเคราะห์ตัวอักษรที่ติดกัน ผลที่ได้จากการวิเคราะห์แสดงได้ดังตารางที่ 5.1 จากนั้นทำการวิเคราะห์ผลของข้อผิดพลาดในการวิเคราะห์สามารถสรุปได้เป็นกรณีต่างๆ ได้ดังตารางที่ 5.2

ตารางที่ 5.1 แสดงผลการวิเคราะห์การติดกันของสิ่งพิมพ์แต่ละประเภท

ประเภทสิ่งพิมพ์	ไทยรัฐ	วรสาร INFORACLE	มติชนสุดสัปดาห์
จำนวนตัวอักษรที่ทดสอบ	7,523	7,140	6,482
จำนวนตัวอักษรที่ติดกันในระดับบนล่าง	707	321	538
จำนวนที่วิเคราะห์ได้อย่างถูกต้อง	640	303	491
เปอร์เซนต์ของการวิเคราะห์ถูกต้อง	90.5%	94.3%	91.2%

ตารางที่ 5.2 แสดงสาเหตุและ เปอร์เซ็นต์ความผิดพลาดในการวิเคราะห์

สาเหตุของการผิดพลาด	เปอร์เซนต์ของความผิดพลาด	ตัวอย่าง
1. ความใกล้เคียงกันของตัวอักษร	13 %	
2. ข้อผิดพลาดของอัลกอลิทึมในวิธีการนี้	83 %	
3. การเกิด noise ที่ทำให้ฮีสโตแกรมผิดพลาด	4 %	

5.2 ผลการทดสอบความถูกต้องของการตัดแยกตัวอักษร

จากตัวอักษรที่ถูกวิเคราะห์ว่าติดกันทั้งหมด จะถูกส่งเข้าสู่ขั้นตอนการตัดแยกตามแนวที่ทำการวิเคราะห์มาแล้ว ผลที่ได้จากการตัดแยกตามแนวทางที่บอกมาแสดงได้ดังตารางที่ 5.3 จากนั้นทำการวิเคราะห์ผลของข้อมูลผิดพลาดในการตัดแยกอันเนื่องมาจากสาเหตุต่างๆ ซึ่งสรุปได้ดังตารางที่ 5.4

ตารางที่ 5.3 แสดงผลการทดสอบความถูกต้องของการตัดแยกตัวอักษรในสิ่งพิมพ์แต่ละประเภท

ประเภทสิ่งพิมพ์	ไทยรัฐ	วรสาร INFORACLE	มติชนสุดสัปดาห์
จำนวนตัวอักษรที่ทดสอบ	7,523	7,140	6,482
จำนวนตัวอักษรที่ติดกันในระดับบนล่าง	707	321	538
จำนวนที่ตัดได้อย่างถูกต้อง	594	294	459
เปอร์เซนต์ของการตัดถูกต้อง	84%	91.5%	85.3%

ตารางที่ 5.4 แสดงสาเหตุและ เปอร์เซ็นต์ความผิดพลาดในการตัดแยก

สาเหตุของการผิดพลาด	เปอร์เซนต์ของความผิดพลาด	ตัวอย่าง
1. การวิเคราะห์ผิดพลาดทำให้บอกแนวการตัดผิดพลาด	38%	
2. โดยลักษณะของการติดกันของตัวอักษรเอง	21%	
3. การเกิด noise ทำให้กำหนดจุดตัดผิดพลาด	41%	

จากตารางที่ 5.1 และ 5.3 เมื่อหาค่าเฉลี่ยแล้ว สามารถสรุปได้ว่าจากผลการทดลอง จะได้เปอร์เซนต์ความถูกต้องของการวิเคราะห์เท่ากับ 92% และทำการตัดแยกได้ถูกต้องคิดเป็น 87 %

บทที่ 6
สรุปผลงานวิจัย และข้อเสนอแนะ

6.1 สรุปผลงานวิจัย

งานวิจัยนี้ประกอบด้วยการทำงานหลายขั้นตอน ซึ่งสามารถสรุปการทำงานในแต่ละขั้นตอนได้ดังต่อไปนี้

1. ขั้นตอนการแยกภาพตัวอักษรด้วยวิธีการเปลี่ยนรหัสขอบ ในขั้นตอนนี้นำเสนอการแยกภาพตัวอักษรออกจากภาพประโยค โดยอาศัยเทคนิคการติดตามขอบภาพ เพื่อทำการเปลี่ยนรหัสที่แสดงถึงตำแหน่งของขอบด้านซ้าย ขอบด้านขวา หรือมุมของภาพ จากนั้นจึงทำการคัดลอกตัวอักษรโดยใช้รหัสขอบที่กำหนดไว้ ดังนั้นด้วยวิธีการนี้ สามารถคัดลอกตัวอักษรที่เหลื่อมกันได้ แต่ในขั้นตอนนี้ยังไม่สามารถแยกตัวอักษรที่ติดกันได้ อีกทั้งในกรณีที่ตัวอักษรมีลักษณะพิเศษเช่น "อำ" หรือ "ญ" ก็จะถูกแยกเป็นภาพตัวอักษรที่แยกกันอยู่คนละระดับ ดังนั้นในขั้นตอนรู้จำที่จะนำผลลัพธ์วิธีนี้ไปใช้ควรจะมีคุณสมบัติรับรู้ผลลัพธ์ในลักษณะนี้ได้ด้วย

2. ขั้นตอนการวิเคราะห์การติดกันของภาพตัวอักษร ในขั้นตอนนี้นำเสนอการวิเคราะห์การติดกันของตัวอักษร และแนวทางการตัดแยกภาพตัวอักษรภาษาไทยในแนวบนและแนวล่าง โดยอาศัยระดับของตัวอักษรเพื่อทำการแบ่งประเภทของตัวอักษรที่ติดกันได้ถึง 7 ประเภท และเสนอแนวทางการตัดแยกของแต่ละประเภท ขั้นตอนนี้มีการนำฮีสโตแกรม มาช่วยในการวิเคราะห์การติดกันของตัวอักษร ซึ่งเป็นขั้นตอนที่ช่วยลดจำนวนตัวอักษรที่จะนำเข้าสู่ขบวนการรู้จำ เพื่อการตัดแยกตัวอักษร และการคำนวณหาค่าฮีสโตแกรม ก็เป็นการคำนวณที่ไม่ซับซ้อน และได้ผลเป็นที่น่าพอใจ แต่ก็ยังมีข้อผิดพลาดที่เป็นผลมาจากการเหลื่อมหรือซ้อนทับกัน ทำให้รูปตัวอักษรที่ตัดแยกได้บางส่วนหาย หรือบางส่วนมี noise เกินมา ดังนั้นเมื่อเข้าสู่ขบวนการรู้จำควรมีการ train ให้รู้จำลักษณะเช่นนี้ด้วย และเนื่องจากวิธีการนี้ใช้เส้นแบ่งระดับในการระบุการติดกันของตัวอักษร ทำให้ไม่สามารถวิเคราะห์ ภาพตัวอักษรที่มีลักษณะหลายขนาดได้ (Multi-Size Characters)

3. ขั้นตอนสุดท้ายเป็นขั้นตอนการตัดแยกตัวอักษร โดยอาศัยแนวทางการแบ่งแยกตัวอักษรที่ได้รับจากขั้นตอนการวิเคราะห์การติดกันของตัวอักษร โดยวิธีการนี้จะทำการหาจุดที่มีค่า Break Cost ต่ำสุดเป็นจุดตัด แต่ผลการตัดอาจจะมีการซ้อนทับกันของตัวอักษรที่ต้องการตัด และจากผลการทดลองที่ได้จากหนังสือพิมพ์ไทยรัฐ, วรสาร INFORACLE และนิตยสารมติชนสุดสัปดาห์ ได้ค่าเฉลี่ยการวิเคราะห์ที่ถูกต้อง 92% และตัดแยกได้ถูกต้อง 87%

6.2 แนวทางในการพัฒนาในอนาคต

งานวิจัยนี้เสนอการตัดแยกเฉพาะระดับบน และระดับล่างเท่านั้น จึงควรมีงานวิจัยต่อที่จะทำการวิเคราะห์และตัดแยกตัวอักษรภาษาไทยในระดับกลางด้วย และวิธีการที่นำเสนอนี้ใช้ได้เฉพาะเอกสารภาษาไทยที่มีขนาดเดียวกันเท่านั้น แต่เอกสารโดยทั่วไปแล้วมักจะประกอบด้วยตัวหนังสือภาษาอังกฤษปนอยู่ด้วย ดังนั้นหากได้รับการปรับปรุงให้ทำการวิเคราะห์ตัดแยกทั้งภาษาไทย และภาษาอังกฤษ ไปพร้อมๆ กันในขั้นตอนเดียวกัน ก็จะเป็นการช่วยให้การแยกตัวอักษรได้อย่างสมบูรณ์ และนำไปประยุกต์ใช้ในงานต่างๆ ได้มากยิ่งขึ้น

เอกสารอ้างอิง

[1] อนุชิต จารุวนาวัฒน์, สุรสิทธิ์ ราตรี และรศ.ชม กิ้มปาน. "การแยกภาพตัวอักษรภาษาไทยออกจากภาพของประโยค." การประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 10, 2530. หน้า 379-385

[2] ประกาศิต ชาติบุรุษ, ธนา หงษ์สุวรรณ์, วัชระ ฉัตวิธิยะ และ ดร.บุญวัฒน์ อัตชู. "การแยกตัวอักษรตัวพิมพ์ภาษาไทยทีละตัวออกจากคำ." การประชุมวิชาการการประยุกต์ใช้คอมพิวเตอร์ในทางวิศวกรรม มหาวิทยาลัยสงขลานครินทร์, กุมภาพันธ์ 2531.

[3] ชาญชัย พิสิทธิวิทยานนท์, ฟูศักดิ์ ชีวสุวิทย์ และกิตติ ตีรเศรษฐ. "การติดตามขอบวัตถุโดยการใช้ตารางหน้าต่าง." การประชุมวิชาการวิศวกรรมไฟฟ้า ครั้งที่ 10, 2530. หน้า 148-157

[4] Hiromichi F.,Yasuaki N. and Kiyomichi K. "Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis." Proceeding of the IEEE, Vol.80, No.7, July 1992. pp. 1079-1092

[5] Panich, W., Jitapunkul S. and Choruengwiwat P. "Segmentation of Connected Characters using Distinctive features of Thai Characters in Thai Character Recognition System." 20th Electrical Engineering Conference, Bangkok, Thailand,1997. pp. 338-342.

[6] Su Liang, Shridhar M. and Ahmadi M. "Efficient Algorithms For Segmentation and Recognition Of Printed Characters In Document Processing." Proceedings of the 2nd ICDAR 1993. pp. 240-243.

[7] Abdelwahab Z. and Rolf I. "Optical Font Recognition Using Typographical Features." IEEE Transaction on pattern analysis and machine intelligence, vol. 20, No.8, August 1998. pp. 877-882

[8] LU Y. "Machine Printed Character Segmentation-An Overview." Pattern Recognition , Vol.28, No.1, 1995. pp. 67-80

ภาคผนวก ก.

ข้อสรุปของประเภทการติดกันของตัวอักษรในรูปแบบต่างๆ โดยทำการสุ่มตรวจนับจำนวนที่ติดกันจาก หนังสือพิมพ์, นิตยสาร และวารสาร และแสดงผลที่ได้เป็นอัตราส่วนของการติดกันในแต่ละแบบดังนี้

ประเภท	มีอัตราที่พบ

1. การติดกันของสระหรือวรรณยุกต์ ที่ระดับบน

หนังสือพิมพ์	0.62 %
นิตยสาร	1.15 %
วารสาร	0.86 %

2. การติดของพยัญชนะกับสระหรือวรรณยุกต์ ที่ระดับบน

หนังสือพิมพ์	4.33 %
นิตยสาร	20.00 %
วารสาร	16.90 %

3. การติดของสระนำกับสระหรือวรรณยุกต์ ที่ระดับบน

หนังสือพิมพ์	6.20 %
นิตยสาร	4.23 %
วารสาร	2.29 %

4. การติดของพยัญชนะกับสระหรือวรรณยุกต์ ที่ระดับบนทางขวา

หนังสือพิมพ์	0.62 %
นิตยสาร	0.38 %
วารสาร	0.29 %

5. การติดกันของสระหรือวรรณยุกต์ ที่ระดับบนกับระดับเหนือบน

หนังสือพิมพ์	9.90 %
นิตยสาร	26.92 %
วารสาร	32.38 %

6. การติดของพยัญชนะกับสระหรือวรรณยุกต์ ที่ระดับกลางกับระดับบน

หนังสือพิมพ์	21.67 %
นิตยสาร	33.46 %
วารสาร	12.61 %

7. การติดของสระนำกับพยัญชนะ ที่ระดับกลาง

หนังสือพิมพ์	0.62 %
นิตยสาร	0.77 %
วารสาร	0.29 %

8. การติดของพยัญชนะกับสระ ที่ระดับกลางกับระดับล่าง

หนังสือพิมพ์	11.15 %
นิตยสาร	4.23 %
วารสาร	9.74 %

9. การติดกันของพยัญชนะ ที่ระดับกลาง

หนังสือพิมพ์	44.89 %
นิตยสาร	8.86 %
วารสาร	24.64 %

ภาคผนวก ข.

การวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรพิมพ์ไทย โดยใช้คุณลักษณะทางแนวตั้งและแนวนอนของฮีสโตแกรม

Connected Analysis and Segmentation of Thai Characters With Horizontal and Vertical Histogram of Character Feature

นายศุภกร รัตนปราการ* ดร.บุญธีร์ เครือตราชู**

บทคัดย่อ

บทความนี้นำเสนอการวิเคราะห์การติดกันของตัวอักษร และแนวทางการตัดแยกภาพตัวอักษรภาษาไทยในระดับที่นอกเหนือจากระดับกลาง โดยอาศัยระดับของตัวอักษรเพื่อทำการแบ่งประเภทของตัวอักษรที่ติดกัน ทำให้สามารถแบ่งประเภทของตัวอักษรได้ 7 ประเภท จากนั้นใช้คุณสมบัติของฮีสโตแกรมมาวิเคราะห์การติดกัน และการตัดแยกตัวอักษรของแต่ละประเภท

ผลการทดสอบกับภาพของประโยคตัวอักษรพิมพ์ไทยจากนิตยสาร จำนวน 4,250 ตัวอักษร สามารถวิเคราะห์ ตัวอักษรที่ติดกันได้ถูกต้อง 93 % และตัดแยกได้ถูกต้อง 86 %

Abstract

This paper presents the connected analysis and segmentation of Thai character in the position aside from the middle level. By using level of Thai characters for grouping the connected Thai characters. So, we can group the Thai characters into 7 groups. Later, we use characteristics of histogram to analyse and segment the Thai characters for each groups.

From the experimentation result, analyzing 4,250 characters from magazine, the accuracy rate of connected analyzing is 93% and segmentation is 86% .

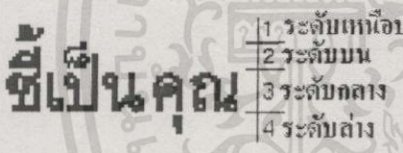
1.บทนำ

ขั้นตอนการแยกภาพตัวอักษรออกจากภาพของประโยค (Character Segmentation) เป็นขั้นตอนการดึงภาพเฉพาะ 1 ตัวอักษร ออกมาจากภาพประโยคของเอกสารในแต่ละบรรทัด สำหรับตัวอักษรไทยการแยกตัวอักษร แบบอัตโนมัติยิ่งเพิ่มความซับซ้อน เนื่องจากตัวอักษรไทยประกอบด้วย พยัญชนะ, สระ และวรรณยุกต์ ที่มีระดับที่แตกต่างกัน ถึง 4 ระดับ ดังรูปที่ 1. ในขั้นตอนนี้อาจจะพบตัวอักษรที่สัมผัสหรือซ้อนทับกัน ทั้งในระดับเดียวกัน (ตามแนวนอน) และในระดับต่างกัน (ตามแนวตั้ง)[1] สาเหตุอาจจะเกิดจากกลไกในการพิมพ์, ขนาด (size), สิ่งรบกวน

* นักศึกษาปริญญาโท คณะเทคโนโลยีสารสนเทศ สจล.

** ผู้ช่วยศาสตราจารย์ ภาควิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สจล.

(noise) หรือแม้แต่รูปแบบของตัวอักษรเอง จึงจำเป็นต้องเพิ่มกระบวนการเพื่อทำการวิเคราะห์ว่าตัวอักษรที่ได้มีการติดกันหรือไม่ ถ้าพบว่ามีการติดกันจะตัดแยกตัวอักษรที่ติดกันได้อย่างไร และจากการเก็บรวบรวมตรวจนับตัวอักษรจากสิ่งพิมพ์บางประเภท เช่น นิตยสาร Byte Thailand พบว่ามีอัตราการติดกันของตัวอักษร ทั้งในระดับบนและล่างอาจมีมากถึง 90.37 % ดังนั้นงานวิจัยนี้จึงนำเสนอการวิเคราะห์การติดกันของตัวอักษรตัวพิมพ์ภาษาไทย พร้อมกับแนวทางการตัดแยกตัวอักษรในระดับที่นอกเหนือจากระดับกลาง (ซึ่งในระดับกลางได้มีผลงานเป็นวิทยานิพนธ์ในหัวข้อ "การตัดแยกตัวอักษรภาษาไทยในระดับกลาง" เขียนโดยคุณจรรยา เกียรติศิริอนันต์ และดร. บุญธีร์ เครือตราชู)



รูปที่ 1. แสดงระดับของตัวอักษรไทย

2. การวิเคราะห์และการตัดแยกตัวอักษรไทยที่ติดกัน

งานวิจัยนี้ได้นำเสนอการหาค่าฮีสโตแกรมแบบต่างๆ ที่เหมาะกับตัวอักษรแต่ละแบบ เพื่อเป็นการแสดงลักษณะของกราฟให้เด่นชัดขึ้น โดยมีวิธีการต่างๆ ดังนี้

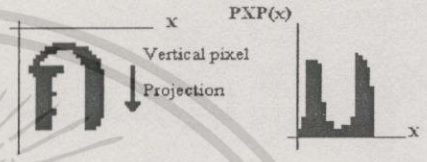
2.1 The Pixel Projection [2]

วิธีการนี้เป็นการแสดงค่าจำนวนจุดที่เป็นเนื้อของตัวอักษรในแนวตั้ง(Vertical Pixel Projection) และแนวนอน (Horizontal Pixel Projection) โดยทำการคำนวณจากสมการ

$$\text{Vertical} \quad PXP(x) = \sum_{y} P(x,y)$$

$$\text{Horizontal} \quad PXP(y) = \sum_{x} P(x,y)$$

เมื่อP(x,y) แสดงค่าของจุด ณ.ตำแหน่ง x และ y ผลที่ได้จะแสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 2



รูปที่ 2 แสดงกราฟ Vertical Pixel Projection

2.2 The Profile Projection [2]

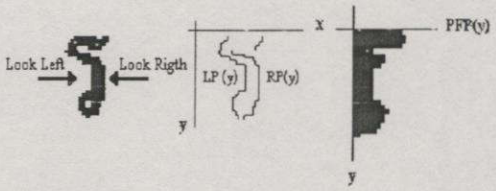
วิธีการนี้เป็นการคำนวณหา ระยะทางจากขอบด้านหนึ่งไปยังขอบอีกด้านหนึ่งของตัวอักษร ในแนวตั้ง (Vertical Profile Projection) และแนวนอน (Horizontal Profile Projection) ในงานวิจัยชิ้นนี้จะขอกล่าวถึงเฉพาะวิธีการ Horizontal Profile Projection เท่านั้น ซึ่งสามารถคำนวณได้จากสมการ

$$\text{Horizontal PFP}(y) = RP(y) - LP(y)$$

เมื่อ $RP(y) = \max X \in \{x|P(x,y)\}$

$LP(y) = \min X \in \{x|P(x,y)\}$

ผลที่ได้จะแสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 3



รูปที่ 3 แสดงกราฟ Horizontal Profile Projection

2.3 The Modify Pixel Projection

วิธีการนี้ผู้ทำวิจัยได้ทำการปรับปรุงจากการคำนวณแบบ Pixel Projection เพื่อแสดงลักษณะของตัวอักษรด้างข้างในรูปกราฟของฮีสโตแกรมตามแนวนอน ให้เด่นชัดมากขึ้น โดยวิธีการเริ่มจากนับจำนวนจุด จากขอบด้านขวาของกรอบตัวอักษร และนับจำนวนจุดไปทางซ้าย เมื่อพบจุดที่เป็นเนื้อของตัวอักษร ให้นับต่อไปจนถึงพื้นของตัวอักษรจึงหยุด ดังตัวอย่าง

0 0 1 1 0 0 1 1 1 0 0 นับจำนวนจุดได้เท่ากับ 5

จุดหยุด จุดเริ่ม

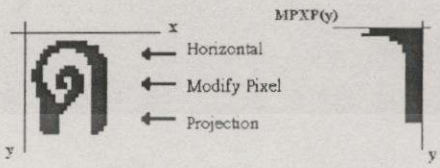
0 0 0 0 0 1 1 1 1 1 1 นับจำนวนจุดได้เท่ากับ 6

จุดหยุด จุดเริ่ม

แสดงเป็นอัลกอริทึมได้ดังนี้

```
กำหนดให้
I,J  INTEGER
NUM_ZERO INTEGER
CHAR_FLAG  BOOLEAN
PICT        คือ อาเรย์ 2 มิติขนาด ROW x COL เก็บข้อมูลภาพตัวอักษร
ROW_CNT  คือ อาเรย์ 2 มิติขนาด ROW  เก็บระยะทางจากขอบ
FOR  I = 1 TO ROW DO
     CHAR_FLAG = FALSE
     NUM_ZERO  = 0
     FOR  J = 1 TO COL DO
       IF  PICT[I][J] = '1' THEN
             ROW_CNT[I]  = ROW_CNT[I] + 1
             CHAR_FLAG = TRUE
       ELSE
             IF CHAR_FLAG = TRUE THEN
                   BREAK
             ELSE
                   NUM_ZERO = NUM_ZERO + 1
          END_IF
       END_IF
     END_LOOP
     ROW_CNT[I] =ROW_CNT[I] + NUM_ZERO
END_LOOP
```

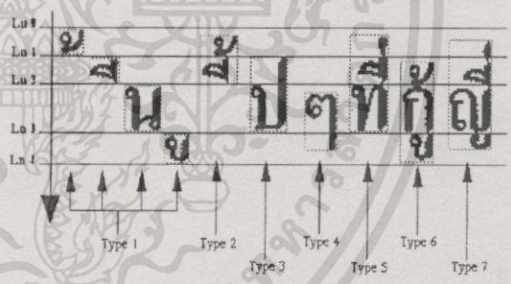
ผลที่ได้แสดงอยู่ในรูปกราฟดังตัวอย่างรูปที่ 4



รูปที่ 4 แสดงกราฟ Horizontal Modify Pixel Projection

การกำหนดประเภทการติดกันของตัวอักษร

รูปแบบการติดกันของตัวอักษรตัวพิมพ์ในแนวระดับบนล่าง พิจารณาตามความสูงของตัวอักษรเปรียบเทียบกับระดับของตัวอักษรแล้ว สามารถแบ่งประเภทตามความสูงได้ 7 ประเภท แสดงดังรูปที่ 5



รูปที่ 5 แสดงการแบ่งประเภทตัวอักษรตามความสูง

ประเภทที่ 1 (Type1) หมายถึงตัวอักษรที่มีความสูงอยู่ภายในเส้นแบ่งระดับของตัวอักษรในแต่ละระดับ ดังนั้นประเภทนี้จึงไม่มีโอกาสที่จะเกิดการติดกันในแนวตั้ง แต่มีโอกาสที่จะเกิดได้ในแนวนอนได้ถ้ามีความกว้างมากกว่า 1 ตัวอักษร

ประเภทที่ 2 (Type2) หมายถึงตัวอักษรที่อยู่ในระดับเหนือบนและระดับบน ดังนั้นประเภทนี้จึงเป็นตัว

อักษรที่ติดกันแนวดิ่ง เพราะไม่มีสระหรือวรรณยุกต์บนที่มีความยาวเกิน 1 ระดับ

ประเภทที่ 3 (Type3) หมายถึงตัวอักษรที่มีความสูงจากระดับกลางสูงขึ้นไปจนถึงระดับบน ประเภทนี้อาจเป็นได้ทั้งตัวอักษรเดี่ยวเช่น "ป", "ฟ" หรือเป็นพยัญชนะที่ติดกับวรรณยุกต์หรือสระได้

ประเภทที่ 4 (Type4) หมายถึงตัวอักษรที่มีความสูงจากระดับกลางยาวลงมาจนถึงระดับล่าง ประเภทนี้อาจเป็นได้ทั้งตัวอักษรเดี่ยว เช่น "ฦ", "ฦ" หรือเป็นพยัญชนะที่ติดกับสระล่างได้

ประเภทที่ 5 (Type5) หมายถึงตัวอักษรที่มีความสูงจากระดับกลางสูงขึ้นไปจนถึงระดับเหนือบน ประเภทนี้อาจเป็นตัวอักษรเดี่ยว เช่น "ไ", "โ", "ใ" ได้ในบางรูปแบบ(FONT) หรืออาจเป็นพยัญชนะติดกับสระหรือวรรณยุกต์ระดับบนได้

ประเภทที่ 6 (Type6) หมายถึงตัวอักษรที่มีความสูงตั้งแต่ระดับบนลงมา จนถึงระดับล่าง ประเภทนี้จะต้องมีการติดกันของตัวอักษร เพราะไม่มีตัวอักษรไทยที่สูงในระดับนี้

ประเภทที่ 7 (Type7) หมายถึงตัวอักษรที่มีความสูงจากระดับเหนือบนยาวลงมาถึงระดับล่าง ประเภทนี้มีการติดกันในระดับเหนือบน เพราะไม่มีตัวอักษรตัวใดที่มีความสูงในระดับเหนือบนและติดกับระดับล่าง

และสามารถหาประเภทของตัวอักษรได้ โดยกำหนดให้

$(x_{min}, y_{min}), (x_{max}, y_{max})$ = พิกัดของภาพตัวอักษรที่ได้จากการแยกภาพตัวอักษร

Ln0 = เส้นบนของระดับเหนือบน

Ln1 = เส้นแบ่งของระดับเหนือบน และระดับบน

Ln2 = เส้นแบ่งของระดับบน และระดับกลาง

Ln3 = เส้นแบ่งของระดับกลาง และระดับล่าง

Ln4 = เส้นล่างของระดับล่าง

การกำหนดตัวอักษรที่อยู่ระดับกลาง

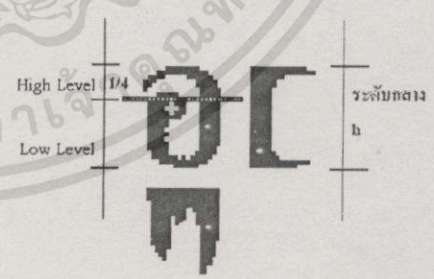
เมื่อพิจารณาตัวอักษรที่อยู่ในระดับกลาง มีลักษณะเด่นทางกายภาพที่สังเกตได้ คือมีแนวเส้นตรงของตัวอักษรทั้งในแนวตั้งและแนวนอนในจำนวนและตำแหน่งที่แตกต่างกัน โดยคุณสมบัติเหล่านี้สามารถปรากฏได้เมื่อเราใช้การหาค่าฮีสโตแกรมของตัวอักษร ดังนั้นงานวิจัยนี้ได้นำเสนอวิธีการจัดกลุ่มโดยการนำวิธีการหาค่าฮีสโตแกรมแบบต่างๆมาใช้เพื่อหาจำนวนและตำแหน่งของแนวเส้นตรงในตัวอักษร เพื่อใช้ในการพิจารณาการติดกันของตัวอักษรและแนวทางการตัด

เริ่มต้นจากการแบ่งช่วงระดับกลางออกเป็นสองส่วนดังรูปที่ 6 กำหนดให้ h เป็นความสูงของระดับกลาง ช่วงบน(High Level) มีขนาดความสูงเป็นหนึ่งในสี่ของความสูงของระดับกลาง และช่วงล่าง (Low Level) มีขนาดความสูงเป็นสามในสี่ของความสูงของระดับกลาง

HL = 1/4 h เมื่อ HL คือระดับความสูง High Level

LL = 3/4 h เมื่อ LL คือระดับความสูง Low Level

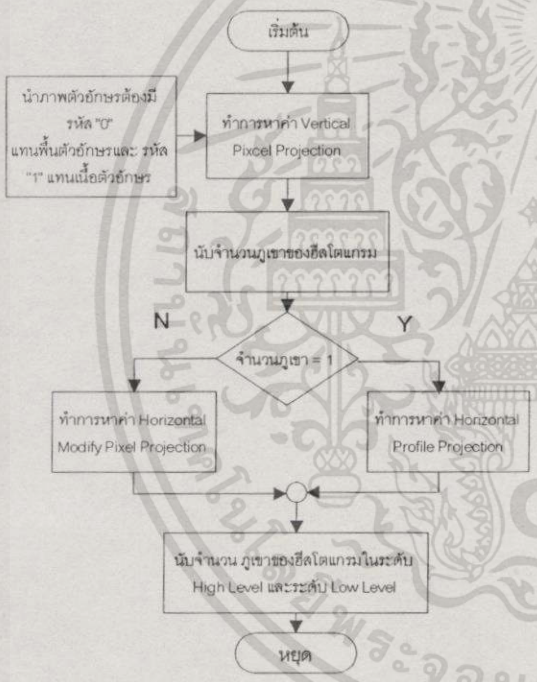
แสดงดังรูปที่ 6



รูปที่ 6 แสดงการแบ่งระดับและกราฟฮีสโตแกรม

จากนั้นทำการหาจำนวนภูเขาของฮีสโตแกรม (Mountain of Histogram) [2] ที่แสดงถึงแนวเส้นตรง

ของตัวอักษรทั้งในแนวตั้งและแนวนอนในระดับ High Level และ Low Level ของระดับกลาง ตามวิธีที่แสดงในผังการทำงานรูปที่ 7 เพื่อให้ใช้วิธีที่เหมาะสมกับตัวอักษรตามที่งานวิจัยชิ้นนี้ได้นำเสนอดังนี้คือ ถ้าค่าฮีสโตแกรมในแนวตั้งที่ได้มีจำนวนภูเขาเป็น 1 การหาค่าฮีสโตแกรมด้านแนวนอนควรจะใช้วิธี Profile Projection แต่ถ้ามากกว่า 1 ควรใช้แบบ Modify Pixel Projection



รูปที่ 7 แสดงการหาจำนวนภูเขาของฮีสโตแกรม

จากจำนวนภูเขาในแนวตั้งและแนวนอนและตำแหน่งต่างๆ สามารถแยกตัวอักษรในระดับกลางออกเป็นกลุ่มต่างๆ ได้ดังตารางที่ 1

กลุ่ม	ภูเขาแนวตั้ง	ภูเขาแนว HL	ภูเขาแนว LL	ตัวอักษร
V1T1	1	>0	=0	เ ๅ ๆ
V1T2	1	=0	>0	ไ ใ โ
V1T3	1	>0	>0	ง ร ว
V2T1	2	>0	=0	ก ค ฅ ด ต ถ ภ ศ ฤ ฦ ฎ ฏ ฉ ท ฑ ห
V2T2	2	=0	>0	ข ฃ ย บ ป
V2T3	2	>0	>0	ช ซ ธ อ ฮ ล ส จ ฐ
V2T4	2	=0	=0	พ ผ ฟ ฝ ม ฆ น ฬ
V3	3	-	-	ฌ ญ ฒ ณ

ตารางที่ 1 แสดงกลุ่มตัวอักษรระดับกลาง

การวิเคราะห์การติดกันของภาพตัวอักษร

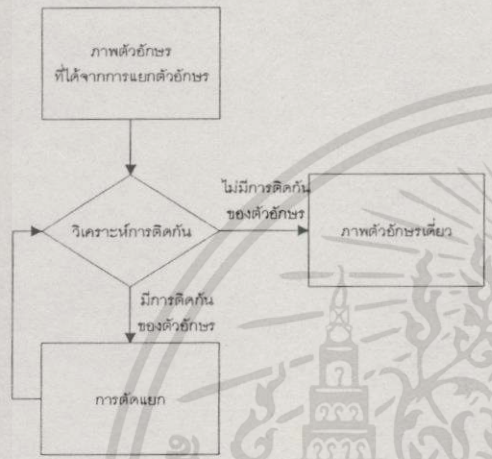
หลังจากที่ได้ภาพตัวอักษรจากการแยกออกจากภาพประโยคแล้ว ก็เข้าสู่ขั้นตอนการวิเคราะห์การติดกันของตัวอักษร ประกอบด้วย

1.ระบุประเภทโอกาสที่จะเกิดการติดกันของตัวอักษรว่าจัดอยู่ในประเภทใดในระหว่างประเภทที่1 ถึงประเภทที่ 7 โดยการเปรียบเทียบตำแหน่งพิกัดของภาพตัวอักษรกับเส้นแบ่งระดับของตัวอักษร

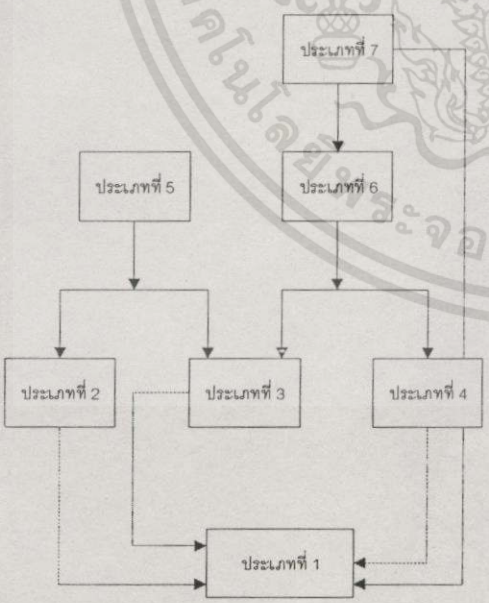
2.ตรวจสอบการติดกันของตัวอักษรในแต่ละประเภท ว่ามีการติดกันจริงหรือไม่ และกำหนดแนวทางการตัดแยกตามหัวข้อ 3 เพื่อส่งต่อในขั้นตอนการตัดแยกต่อไป

และหลังจากทำการตัดแยกตัวอักษร จนได้ภาพตัวอักษรที่ถูกตัดแยกแล้ว ภาพตัวอักษรเหล่านั้นก็จะถูกส่งกลับเข้าสู่ ขั้นตอนการวิเคราะห์การติดกันของตัวอักษรอีกครั้ง จนกระทั่งทุกภาพตัวอักษรที่ถูกตัดแยกแล้วถูกระบุว่าเป็นตัวอักษรที่ไม่มีการติดกัน ดังรูปที่ 8 และสรุปเป็นเส้นทางของตัวอักษรประเภทต่างๆที่ถูกวิเคราะห์และตัดแยกดังแผนภาพรูปที่ 9 หมายถึง หลังจากที่ได้ทำการตัดแยกของแต่ละ

ประเภทแล้ว ส่วนที่ตัดแยกได้มีโอกาสที่จะเป็นประเภทใดบ้าง เช่น ประเภทที่ 6 ผลการตัดแยกที่ได้อาจเป็น ประเภทที่ 3 หรือประเภทที่ 4 เป็นต้น



รูปที่ 8 แสดงขั้นตอนการวิเคราะห์การติดกันของตัวอักษร



รูปที่ 9 สรุปเส้นทางการวิเคราะห์และตัดแยก

การวิเคราะห์การติดกัน และการกำหนดแนวทางของการตัดแยกตัวอักษร จะใช้ข้อสังเกตดังต่อไปนี้ในการตัดสินใจ

ข้อสังเกตของตัวอักษรในระดับกลาง

จากตารางที่ 1 มีข้อสังเกตดังนี้

1. กลุ่มที่มีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 3 แนว ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความสูงหรือต่ำเกินกว่าระดับกลาง กลุ่มนี้ได้แก่ V3 ตัวอย่างเช่น ฉ,ฒ เป็นต้น

2. กลุ่มที่มีจำนวนแนวเส้นตรงในแนวนอนที่ระดับ HL มากกว่าหรือเท่ากับ 1 แนว ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความสูงเกินกว่าระดับกลาง กลุ่มนี้ได้แก่ V1T1, V1T3, V2T1 และ V2T3 ตัวอย่างเช่น ร,ว,ก,ภ,ธ,อ เป็นต้น

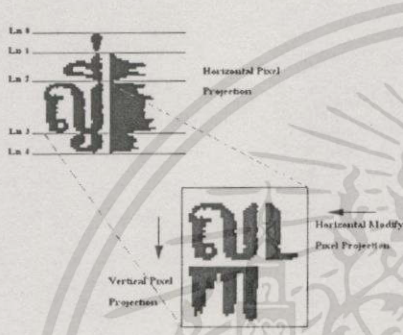
3. กลุ่มที่มีจำนวนแนวเส้นตรงในแนวนอนที่ระดับ LL มากกว่าหรือเท่ากับ 1 แนว ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความยาวต่ำกว่าระดับกลาง กลุ่มนี้ได้แก่ V1T3, V2T2 และ V2T3 ตัวอย่างเช่น ร,ว,บ,ย,ป,ธ,อ เป็นต้น

4. กลุ่มที่มีจำนวนแนวเส้นตรงในแนวนอนที่ระดับ HL และ LL ของระดับกลาง เท่ากับศูนย์ ไม่มีตัวอักษรตัวใดในกลุ่มนี้ที่มีความยาวต่ำกว่าระดับกลาง กลุ่มนี้ได้แก่ V2T4 ตัวอย่างเช่น พ,ผ,ฟ เป็นต้น

5. มีเฉพาะสระ "ไ","ใ","โ" เท่านั้นที่มีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 1 และจำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL เท่ากับศูนย์คือ กลุ่ม V1T2

6. มีเฉพาะตัวอักษร "า","ๅ","ๆ" เท่านั้นที่มีจำนวนแนวเส้นตรงตามแนวตั้งเท่ากับ 1 จำนวนแนวเส้นตรงในแนวนอน ที่ระดับ HL เท่ากับ 1 และระดับ LL เท่ากับศูนย์ คือกลุ่ม V1T1

จะเห็นว่าบางประเภทของการติดกันจะต้องอาศัยคุณสมบัติเหล่านี้วิเคราะห์จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรที่ต้องการวิเคราะห์ตามผังงานรูปที่ 7 ดังตัวอย่างรูปที่ 10



รูปที่ 10 แสดงการหาแนวเส้นตัวอักษรระดับกลาง

3. การตรวจสอบและแนวทางการตัดแยกตัวอักษร

แบ่งได้ตามประเภทของตัวอักษรได้ดังนี้

ประเภทที่ 1

จะเห็นภาพตัวอักษร มีความสูงไม่เกินกว่าระดับที่อยู่ แสดงว่าไม่มีการติดกันในแนวตั้ง จึงพิจารณาเฉพาะความกว้างของตัวอักษร ถ้าภาพตัวอักษรใดมีความกว้างเกินกว่าขนาดความกว้างหนึ่งตัวอักษร สรุปได้ว่ามีการติดกันของตัวอักษรในแนวนอน แนวทางการตัดแยกใช้ขอบของภาพของตัวอักษรระดับกลางเป็นแนวแบ่ง แต่ถ้าไม่มีแสดงว่าเป็นภาพของตัวอักษรเดี่ยว

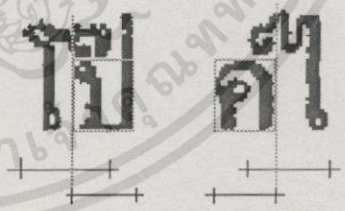
ประเภทที่ 2

ในภาพตัวอักษรที่ได้ มีความสูงอยู่ในระดับบนและเหนือบน จากลักษณะของสระและวรรณยุกต์ในภาษาไทย ไม่มีสระหรือวรรณยุกต์บนตัวใดที่มีความสูงเกินกว่าหนึ่งระดับ ดังนั้นจึงสรุปได้ว่า มีการติดกันของตัวอักษรในแนวตั้ง และแนวทางการตัดแยกอยู่ที่เส้นแบ่งระดับบนและเหนือบน (Ln1)

ประเภทที่ 3

ตัวอักษรประเภทนี้อาจเป็นตัวอักษรเดี่ยวที่มีความสูงจากระดับกลางถึงระดับบนได้ เช่น "ป", "ฟ" จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด จากนั้นใช้ข้อสังเกตข้อที่ 1,2 คือถ้าอยู่ในกลุ่ม V3,V1T3,V2T1,V2T3 ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง (Ln2)

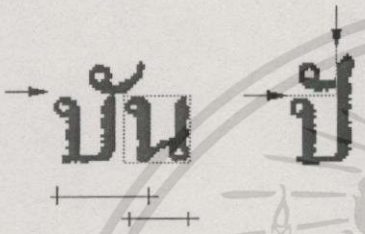
ถ้าตกในกลุ่ม V1T2 ("ไ","ใ","โ") จากข้อสังเกต 5 ในกรณีนี้อาจมีการติดกับตัวอักษรระดับบนบางตัวที่มีขนาดเล็กทำให้ความกว้างไม่เกินหนึ่งตัวอักษรได้เช่น ตัวการันต์ ในตัวอย่างดังรูป 11 ในกรณีนี้ให้ตรวจสอบว่ามีการเหลื่อมล้ำกับตัวอักษรลำดับหน้าหรือหลังหรือไม่ ถ้ามีให้กำหนดแนวทางการตัดแยกตามขอบของตัวอักษรลำดับหน้าหรือหลังที่มีการเหลื่อมล้ำ



รูปที่ 11 แสดงการติดโดยความกว้างไม่เกิน 1 ตัวอักษร

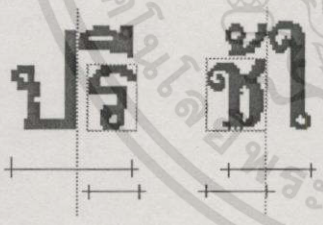
ถ้าอยู่ในกลุ่ม V2T2,V2T4 จะต้องนำเข้าสู่ขบวนการรู้จำ ผลที่ได้ถ้าอยู่ในกลุ่ม ("ข","ฃ","ย","ษ","ม","ฒ","น") ให้กำหนดการตัดตามแนวเส้น Ln2 แต่ถ้าอยู่ในกลุ่ม ("บ","พ","ผ") ให้ตรวจสอบว่าสระ

อยู่ระดับบนหรือไม่โดยดูจากค่าฮีสโตแกรมถ้าไม่มีถือว่าเป็นอักษรเดี่ยวแต่มีให้ตรวจสอบว่ามีการเหลื่อมล้ำกับตัวอักษรลำดับหลังหรือไม่ ถ้ามีให้กำหนดแนวทางการตัดตามแนวเส้น Ln2 แต่ถ้าไม่มีให้ใช้ภูเขาที่สองและแนวเส้น Ln2 เป็นการกำหนดแนวตัด ดังรูปที่ 12



รูปที่ 12 การแสดงแนวการตัดในกลุ่ม("บ","พ","ผ")

แต่ถ้าความกว้างของตัวอักษรมากกว่า 1 ตัวอักษรให้ตรวจสอบว่ามีการเหลื่อมล้ำกับตัวอักษรลำดับหน้าหรือหลังหรือไม่ ถ้ามีให้กำหนดแนวทางการตัดแยก อยู่ที่กึ่งกลางของความกว้างของตัวอักษร ดังรูปที่ 13



รูปที่ 13 แสดงการติดกันโดยความกว้างเกิน 1 ตัวอักษร

แต่การวิเคราะห์นี้ไม่สามารถแยกกรณีที่มีการติดกันของตัวอักษรในกลุ่ม ("พ","บ","ผ","ฟ","ป","ฝ") กับไม้เอกได้ เช่น การที่ "บ" ,"ผ" และ"ฝ" ติดกับไม้เอก ดังตัวอย่างรูปที่ 14 และกรณีที่มีการติดกันของตัวอักษรในกลุ่ม ("พ","บ","ผ") กับสระที่มีตำแหน่งการวางที่ไม่เหลื่อมไปข้างหลังได้ เช่น สระ "อิ","อี" จะให้การวิเคราะห์และตัดแยกไม่ถูกต้อง ดังตัวอย่างรูปที่ 16

ประเภทที่ 4

ประเภทนี้มีตัวอักษรบางตัวที่มีความสูงตั้งแต่ระดับกลางยาวลงมาจนถึงระดับล่าง เช่น "ฤ","ฦ" จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด จากนั้นใช้ข้อสังเกตข้อที่ 1,3 และ 4 คือ กลุ่ม V3,V1T3, V2T2,V2T3,V2T4 ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับกลางและระดับล่าง(Ln3) ถ้าอยู่ในกลุ่ม V1T1 จากข้อสังเกตข้อ 6 สรุปว่าเป็นอักษรเดี่ยว แต่ถ้าอยู่ในกลุ่ม V2T1 จะต้องนำเข้าสู่ขบวนการรู้จำถ้าอยู่ในกลุ่ม ("ก","ค","ศ","ด","ต","ถ","ท","ฑ","ห") ให้กำหนดการตัดตามแนวเส้น Ln3 เช่นกัน กรณีนอกเหนือจากนี้ให้ถือว่าเป็นตัวอักษรเดี่ยว แต่การวิเคราะห์นี้ไม่สามารถแยกระหว่างตัวอักษร "ฦ","ฦ" กับ "ภ"ที่ติดกับสระ "อู" ได้

ประเภทที่ 5

ในกลุ่มนี้ บางรูปแบบ(Font) ของตัวอักษร “ไ”,”ไ”,”โ” มีความสูงถึงระดับเหนือระดับบน ทำให้ไม่สามารถแยกระดับเหนือระดับบนออกไปได้ จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด จากนั้นใช้ข้อสังเกตของตัวอักษรในระดับกลางข้อที่ 1 อยู่ในกลุ่ม V3 หรือข้อที่ 2 อยู่ในกลุ่ม V1T3, V2T1 และ V2T3 ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง(Ln2) ถ้าตกในกลุ่ม V1T2 (“ไ”,”ไ”,”โ”) จากข้อสังเกต 5 ให้หาตัวอักษรที่อาจมีการติดเช่นเดียวกับประเภทที่ 3 แต่ถ้าไม่มีตกอยู่ในกลุ่มที่กล่าวมา ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับเหนือบนและระดับบน (Ln1)

ประเภทที่ 6

ในประเภทนี้ เมื่อพิจารณาตามระดับของตัว

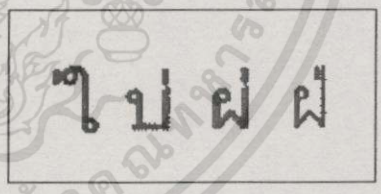
อักษรแล้ว ไม่พบว่าตัวอักษรไทยตัวใดที่มีความสูงจากระดับบนจนถึงระดับล่าง จึงสรุปได้ว่ามีการติดกันของตัวอักษรในระดับบนหรือในระดับล่างอย่างแน่นอน แต่เนื่องจากพยัญชนะไทย มีความสูงเกินกว่าระดับกลางได้ทั้งสองระดับ เช่น “ป” หรือ “ฤ” ทำให้ไม่สามารถระบุแนวทางการตัดแยกได้ จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด จากนั้นใช้ข้อสังเกตของตัวอักษรในระดับกลางข้อที่ 2 คือถ้าตกอยู่ในกลุ่ม V1T3,V2T1 และ V2T3 ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับบนและระดับกลาง(Ln2) แต่ถ้าไม่มีตกอยู่ในกลุ่มนี้ ให้กำหนดแนวทางการตัดในแนวนอน ตามแนวเส้นแบ่งระดับกลางและระดับล่าง(Ln3)

ประเภทที่ 7

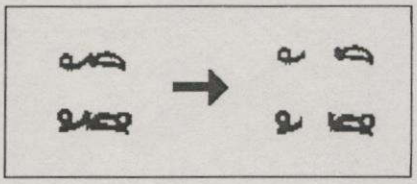
จากการที่ไม่มีตัวอักษรไทยตัวใดสูงถึง 4 ระดับ และไม่มีโอกาส ที่ตัวอักษรที่สูงถึง 3 ระดับ จะติดกับสระระดับล่าง เช่น “ใ” ไม่มีโอกาสติดกับสระ “อุ” แสดงว่ามีการติดกันของสระหรือวรรณยุกต์ ในระดับเหนือระดับบนอย่างแน่นอน จึงสามารถที่จะทำการตัดในระดับเหนือบน(Ln1)ได้ และส่วนล่างก็จะถูกจัดเข้าประเภท 6 เมื่อทำการวิเคราะห์อีกครั้ง

3. ผลการทดลอง

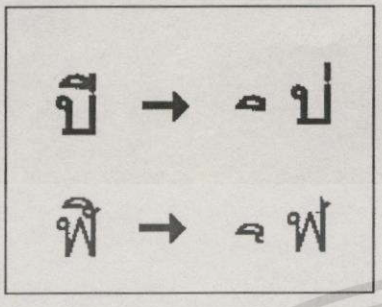
ผลการทดสอบจากการใช้วิธีการวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรพิมพ์ไทย โดยใช้คุณลักษณะทางแนวตั้ง และแนวนอนของฮีสโตแกรม กับรูปภาพประโยคที่ได้จากนิตยสาร BYTE Thailand ผ่านการสแกนด้วยความละเอียด 300 จุดต่อนิ้ว โดยทำการเลือกเฉพาะบรรทัดที่มีตัวติดกันในระดับบนและล่าง นับจำนวนตัวอักษรได้ 4,250 ตัวอักษร มีจำนวนตัวอักษรที่ติดกันในระดับบนและล่าง 340 ตัวอักษร สามารถแสดงได้ถูกต้อง 316 ตัวอักษร คิดเป็น 93% และทำการตัดแยกได้ถูกต้อง 292 ตัวอักษร คิดเป็น 86% ทดสอบโดยเครื่อง PC Pentium 133 MHz , Memory 32 MB แต่อย่างไรก็ตามผลที่ได้ก็ยังมีข้อผิดผลาดอันเนื่องมาจากความใกล้เคียงกันของภาษาไทย เช่นจากรูปที่ 14 มี ไม้เอก อยู่ติดกับ "บ" จะถูกวิเคราะห์ได้ว่าเป็น "ป" นอกจากนี้ขนาดที่เล็กมากของอักษรบางตัวเช่น ไม้เอก ซึ่งอยู่ติดกับ "ใ" จะถูกวิเคราะห์เป็น "ไ" เป็นต้น และจากรูปที่ 15 แสดงถึงการตัดในแนวเส้นตรงที่ผิดพลาด โดยบางส่วนของตัวอักษรที่เหลื่อมล้ำกันจะถูกตัดขาด และอาจไปติดกับส่วนอื่น และจากรูปที่ 16 แสดงถึงการวิเคราะห์ที่ผิดจึงทำให้การตัดผิดพลาดได้ เช่น "บี" จะถูกวิเคราะห์เป็น "ป" และ สระอิ , "พิ" จะถูกวิเคราะห์เป็น "ฟ" และ สระอิ ซึ่งทำให้การตัดตามการวิเคราะห์ที่ได้ผิดพลาด



รูปที่ 14 แสดงการวิเคราะห์ผิดพลาดให้ผลเป็นอักษรเดี่ยว



รูปที่ 15 การตัดที่ผิดพลาด



รูป 16 การวิเคราะห์ที่ผิดทำให้การตัดผิดพลาด

4.สรุป

วิธีที่นำเสนอนี้มีลักษณะที่น่าสนใจคือเป็นวิธีที่มีการนำฮิสโตแกรมมาช่วยในการวิเคราะห์การติดกันของตัวอักษร ซึ่งเป็นขั้นตอนที่ช่วยลดจำนวนตัวอักษร ที่จะนำเข้าสู่ขบวนการรู้จำเพื่อการตัดแยกตัวอักษร และถึงแม้ว่าผลการทดลองการวิเคราะห์และตัดแยกจะให้ผลที่น่าพอใจ แต่ก็ยังมีข้อผิดพลาดอยู่โดยเฉพาะการตัดที่ผิดพลาดที่เป็นผลมาจากการเหลื่อมหรือซ้อนทับกันทำให้รูปตัวอักษรที่ได้บางส่วนหายไป หรือมี noise เพิ่มขึ้นมา ดังนั้นเมื่อเข้าสู่ขบวนการรู้จำจึงควรที่จะสอน(train)ให้รู้จำลักษณะเช่นนี้ด้วย และบางครั้งการเอียงของหน้ากระดาษก็อาจทำให้ระดับตัวอักษรผิดพลาดซึ่งเป็นผลให้วิธีการนี้มีความคลาดเคลื่อนได้ ซึ่งโดยทั่วไปในระบบของ OCR จะมีขั้นตอนของการปรับความเอียงของบรรทัดอยู่ในขั้นตอนของการเตรียมข้อมูลอยู่แล้ว นอกจากนี้วิธีการที่นำเสนอจะต้องอาศัยระดับของตัวอักษรที่มีขนาดเดียวกัน ดังนั้นจึงไม่สามารถนำวิธีการนี้ไปใช้กับตัวอักษรที่มีหลายขนาด(Multi-size character) ในหนึ่งบรรทัดได้

5.เอกสารอ้างอิง

[1] W. Panich, S. Jitapunkul, P. Choruengwiwat, "Segmentation of Connected Characters using Distinctive features of Thai Characters in Thai Character Recognition System" 20th Electrical Engineering Conference, Bangkok, Thailand,1997 , pp.338-342

[2] Su Liang, M. Shridhar, M Ahmadi, "Efficient Algorithms For Segmentation and Recognition Of Printed Characters In Document Processing", Proceedings of the 2nd ICDAR 1993,pp.240-243

[3] A. Zramdini, R. Ingold, "Optical Font Recognition Using Typographical Features", IEEE Transaction on pattern analysis and machine intelligence, vol. 20, No.8, August 1998

ประวัติผู้เขียน

นายศุภกร รัตนปราการ เกิดวันที่ 17 มกราคม 2511 ณ อำเภอหาดใหญ่ จังหวัดสงขลา สำเร็จการศึกษาวิทยาศาตร์ สาขาคณิตศาสตร์ จากมหาวิทยาลัยสงขลานครินทร์ จังหวัดสงขลา ปีการศึกษา 2532