การตรวจจับและตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน

DETECTION AND SEGMENTATION OF TOUCHING AND CROSSING THAI PRINTED CHARACTERS

ภิญโญ พงษ์โพธิ์
PINYO PONGPO

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2549

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

การตรวจจับและตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน

DETECTION AND SEGMENTATION OF TOUCHING AND CROSSING THAI PRINTED CHARACTERS

ภิญโญ พงษ์โพธิ์

PINYO PONGPO

เลขหมู่.................................
เลขทะเบียน 69085
วัน,เดือน,ปี - 7 ก.พ. 2550

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2549

DETECTION AND SEGMENTATION OF TOUCHING AND CROSSING THAI PRINTED CHARACTERS

PINYO PONGPO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2006

COPYRIGHT 2006

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

หัวข้อวิทยานิพนธ์	การวิเคราะห์และตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน
นักศึกษา	นายภิญโญ พงษ์โพธิ์
รหัสประจำตัว	44067459
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2549
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ.ดร. ภัทรชัย ลลิตโรจน์วงศ์

บทคัดย่อ

การเพิ่มความถูกต้องของระบบรู้จำอักษรภาษาไทย จำเป็นต้องมีระบบประมวลผลก่อนการรู้จำ ได้แก่กระบวนการปรับเอียงเอกสารรูปภาพ กระบวนการปรับความคมชัด กระบวนการลดสัญญาณรบกวน กระบวนการตัดแบ่งบรรทัดและตัวอักษร และกระบวนการวิเคราะห์ประเภทและตัดรูปภาพตัวอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน สำหรับการรู้จำอักษรภาษาไทย ภาพอักษรที่ติดกันมีผลกระทบต่อความถูกต้องของระบบรู้จำอักษรภาษาไทย ดังนั้น โครงงานวิจัยนี้จึงมุ่งเน้นในการออกแบบและพัฒนาโปรแกรมสำหรับตัดภาพอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน กระบวนการตัดภาพอักษรที่ติดกันประกอบด้วย 3 ขั้นตอน ได้แก่ การสกัดภาพและจำแนกภาพอักษร การวิเคราะห์การติดกันของตัวอักษร และการหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ ในการดึงภาพอักษรจะประกอบด้วยการใช้ค่าสถิติของจุดดำในแนวนอนและการติดตามจุดดำเพื่อแยกตัวอักษรออกจากเอกสาร และใช้ระดับอักษรเพื่อจำแนกภาพอักษร ในการวิเคราะห์การติดกันของตัวอักษรจะใช้ข้อมูลภาพอักษร ได้แก่ ลักษณะเส้นอักษร อัตราส่วนความกว้างและความสูงของรูปภาพ สำหรับการหาตำแหน่งจุดตัดจะคำนวณหาจุดตัดวิกฤตหรือจุดตัดที่มีจุดร่วมน้อยสุดเพื่อตัดภาพอักษรติด โดยขึ้นกับประเภทอักษรติด ในกรณีที่การตัดภาพอักษรติด ทำให้ภาพอักษรไม่สมบูรณ์ จำเป็นต้องมีการปรับแต่งภาพอักษร ซึ่งงานวิจัยนี้ใช้ 2 เทคนิคคือ การคำนวณขอบเขตและชดเชยส่วนที่หายไปหลังการตัดจุดตัดวิกฤต หรือการเพิ่มพื้นที่อักษรเพื่อชดเชยอักษรหลังการตัด เพื่อเพิ่มความสมบูรณ์ให้กับตัวอักษร

Thesis Title	Analysis and segmentation of touching and crossing Thai Printed Characters
Student	Mr. Pinyo Pongpo
Student ID.	44067459
Degree	Master of Science
Program	Information Technology
Year	2006
Thesis Advisor	Asst.Prof.Dr. Pattarachai Lalitrojwong

ABSTRACT

The accuracy of Thai OCR needs binarization, noise reduction, angle detection, document analysis, main line detection and segmentation of a cross-touch connected character image. In Thai language, a connected character image has a significant effect on the accuracy of Thai OCR. Problems of previous studies derive from the unrecognition of connected character in the image. Thus, this research aims at designing and developing the segmentation of cross-touch connected character image. The segmentation of a cross-touch connected character image consists of three steps, detecting and grouping character images, analyzing character connection patterns, and locating the connected character cutting lines and compensating character. Detecting and grouping character images requires horizontal projection and black point connected, to extract each character from the original document, and leveling each character to identify its type. The analysis of connected characters uses the characteristics of Thai characters and their Width-Height ratio. To locate the cutting point of each character, either peak to valley or break cost technique is used. If cutting causes a character defect due to overlapping of connected characters, the compensation technique will be used to restore and to refine them. Two compensation techniques used in this step are the identification of a connected character overlapping area and a compensation area, and area enlargement in order to increase the compensation area by using the peak to valley technique.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยความเมตตา ให้คำแนะนำและคำปรึกษาการจัดทำวิทยานิพนธ์มาโดยตลอดจากอาจารย์ที่ปรึกษา ผศ.ดร.ภัทรชัย ลลิตโรจน์วงศ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่านและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณคณะกรรมการสอบหัวข้อและโครงร่างวิทยานิพนธ์ที่ได้กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ จนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี

ขอขอบพระคุณเจ้าหน้าที่ของมหาวิทยาลัยขอนแก่นที่คอยให้คำแนะนำ และให้คำปรึกษาต่างๆเกี่ยวกับงานประชุมวิชาการ NCSEC2006 ด้วยดีมาตลอด

สุดท้ายนี้ขอขอบพระคุณเจ้าหน้าที่ทุกคนในฝ่ายงานบริการนักศึกษา คณะเทคโนโลยีสารสนเทศ และบัณฑิตวิทยาลัย ที่ได้อำนวยความสะดวกเสมอมา

ภิญโญ พงษ์โพธิ์

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	I
บทคัดย่อภาษาอังกฤษ	II
กิตติกรรมประกาศ	III
สารบัญ	IV
สารบัญตาราง	VI
สารบัญรูป	VIII
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการศึกษา	2
1.3 สมมติฐานของการศึกษา	2
1.4 ขอบเขตของการดำเนินการวิจัย	4
1.5 ขั้นตอนของการศึกษา	4
บทที่ 2 งานวิจัยที่เกี่ยวข้อง	5
บทที่ 3 ทฤษฎีและแนวทางที่ใช้	12
3.1 การแยกภาพอักษรออกจากเอกสาร	12
3.2 ส่วนตรวจสอบการติดกันของตัวอักษร	13
3.2.1 การหาค่าความกว้างเฉลี่ยของตัวอักษรในระดับกลาง	13
3.2.2 การหาค่าความสูงเฉลี่ยของตัวอักษรในระดับบน	13
3.3 การหาตำแหน่งที่ติดกันหรือทับซ้อนกัน	14
3.3.1 การหาจุดตัดวิกฤต	14
3.3.2 การหาจุดตัดที่มีจุดร่วมน้อยที่สุด	15
3.4 การปรับปรุงภาพให้สมบูรณ์	15
3.4.1 การชดเชยขอบเขตส่วนที่หายไป	15
3.4.1 การเพิ่มจุดตัดหลังการหาจุดตัดวิกฤต	16
บทที่ 4 การวิเคราะห์การสัมผัสกันและตัดแยกตัวอักษร	17
4.1 ลักษณะของประโยคในภาษาไทย	17

สารบัญ (ต่อ)

หน้า

4.2 การวิเคราะห์และจัดกลุ่มตัวอักษรที่สัมผัสกัน………………………………. 19

4.3 การวิเคราะห์การสัมผัสกัน และการกำหนดจุดตัดแยก.. 20

4.3.1 กลุ่มที่ 1 อักษรระดับบน ติดกับ อักษรระดับบนในแนวนอน....................... 22

4.3.2 กลุ่มที่ 2 อักษรระดับบน ติดกับ อักษรระดับบนในแนวตั้ง…………………. 23

4.3.3 กลุ่มที่ 3 อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์เดียวกัน…….... 23

4.3.4 กลุ่มที่ 4 อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน……….... 25

4.3.5 กลุ่มที่ 5 อักษรระดับกลาง ติดกับ อักษรระดับบน…………………….... 27

4.3.6 กลุ่มที่ 6 อักษรระดับกลาง ติดกับ อักษรระดับกลาง……………………..... 29

4.3.7 กลุ่มที่ 7 อักษรระดับกลาง ติดกับ อักษรระดับล่าง………………………. 30

บทที่ 5 การทดลอง………………………………………………………………. 32

5.1 โปรแกรมตรวจจับและตัดแยกตัวอักษรที่สัมผัสและทับซ้อนกัน………........... 32

5.2 รายละเอียดการทดลอง………………………………………………...... 35

5.3 การวิเคราะห์และสรุปผลการทดลอง……………………………………... 58

บทที่ 6 สรุปผลงานวิจัยและข้อเสนอแนะ……………………………………….... 60

6.1 สรุปผลการวิจัย……………………………………………….......................... 60

6.2 แนวทางการพัฒนาต่อไปในอนาคต……………………………………...... 60

เอกสารอ้างอิง……………………………………………………………………. 62

ภาคผนวก……………………………………………………………………………. 63

ผลงานตีพิมพ์……………………………………………………………………….. 64

ประวัติผู้เขียน... 74

สารบัญตาราง

ตารางที่ หน้า

2.1 รายละเอียดวิธีการแบ่งกลุ่ม และการตัดตัวอักษรของวิชา [1].. 5

2.2 รายละเอียดวิธีการแบ่งกลุ่ม และการตัดตัวอักษรของจักริน [4]…………………………. 8

2.3 สรุปวิธีการที่ใช้ในการตัดแยก และเปอร์เซ็นต์ความถูกต้องของงานวิจัย……………….. 10

2.4 สรุปปัญหาต่างๆที่เกิดขึ้นในแต่ละงานวิจัย.. 10

4.1 ประเภทของอักษรเดี่ยวที่แบ่งกลุ่มตามระดับตัวอักษร…………………………….. 18

4.2 รูปแบบการติดกันของตัวอักษรในแบบต่าง ๆ 20

5.1 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว... 36

5.2 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 36
AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว..

5.3 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 37
AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว..

5.4 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 37
BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว...

5. 5 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 38
BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว...

5.6 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 38
CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว……………………………...

5.7 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 39
CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว..

5.8 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 39
DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว…………………………….

5.9 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 40
DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว…………………………….

5.10 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 40
EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว..

5.11 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร 41
EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว..

สารบัญตาราง (ต่อ)

ตารางที่	หน้า
5.12 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 200 จุด..........	41
5.13 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 300 จุด….......	42
5.14 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 400 จุด..........	42
5.15 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด......	43
5.16 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด......	43
5.17 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด......	44
5.18 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด......	44
5.19 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด......	45
5.20 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด......	45
5.21 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด.....	46
5.22 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด.....	46
5.23 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด.....	47
5.24 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด.....	47
5.25 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด.....	48
5.26 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด.....	48

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

5.27 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด...... 49

5.28 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด........ 49

5.29 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด........ 50

5.30 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด........ 50

5.31 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด........ 51

5.32 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด........ 51

5.33 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด...... 52

5.34 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด..... 52

5.35 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด..... 53

5.36 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด...... 53

5.37 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด...... 54

5.38 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด...... 54

5.39 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด.... 55

5.40 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด..... 55

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

5.41 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด..... 56

5.42 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด..... 56

5.43 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด..... 57

5.44 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด..... 57

5.45 สรุปผลการทดลองทั้งหมด 33 ตาราง จำนวน 669,855 ตัวอักษร.................................... 58

สารบัญรูป

รูปที่	หน้า
1.1 ตัวอย่างของตัวอักษรที่มีการเหลื่อมล้ำกัน..........	3
1.2 ตัวอย่างของตัวอักษรที่มีการสัมผัสกัน..........	3
1.3 ตัวอย่างของตัวอักษรที่มีการทับซ้อนกัน..........	3
3.1 ตัวอย่างการใช้การฉายแนวนอนเพื่อแยกบรรทัด..........	12
3.2 การหาจุดแบ่งโดยใช้จุดตัดวิกฤต..........	14
3.3 นัยสำคัญของการสัมผัสกัน..........	15
3.4 วิธีการหาจุดตัดเพื่อชดเชย และการชดเชย..........	16
4.1 การแบ่งระดับของประโยคในภาษาไทย..........	17
4.2 กระบวนการแยกภาพตัวอักษรออกเป็นตัวอักษรเดี่ยว..........	18
4.3 ตัวอย่างภาพตัวอักษรที่แยกออกเป็นตัวอักษรเดี่ยว..........	18
4.4 ภาพรวมในการวิเคราะห์และตัดแยกตัวอักษร..........	21
4.5 ตัวอย่างการตัดในกลุ่มที่ 1..........	22
4.6 วิธีการตัดแยกอักษรระดับบน ติดกับ อักษรระดับบนในแนวนอน..........	22
4.7 ตัวอย่างการตัดในกลุ่มที่ 2..........	23
4.8 วิธีการตัดแยกอักษรระดับบน กับ อักษรระดับบนในแนวตั้ง..........	23
4.9 ตัวอย่างการตัดในกลุ่มที่ 3..........	24
4.10 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์เดียวกัน..........	25
4.11 ตัวอย่างการตัดในกลุ่มที่ 4 แบบที่ 1..........	25
4.12 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน แบบที่ 1..........	26
4.13 ตัวอย่างการตัดในกลุ่มที่ 4 แบบที่ 2..........	26
4.14 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน แบบที่ 2..........	27
4.15 ตัวอย่างการตัดในกลุ่มที่ 5..........	27
4.16 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับบน..........	28
4.17 ตัวอย่างการตัดในกลุ่มที่ 6..........	29
4.18 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับกลาง..........	30
4.19 ตัวอย่างการตัดในกลุ่มที่ 7..........	30
4.20 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับล่าง..........	31
5.1 หน้าจอหลัก..........	32
5.2 ผลที่ได้จากการเลือกไฟล์รูปภาพ..........	33

สารบัญรูป(ต่อ)

รูปที่ หน้า

5.3 ผลที่ได้จากการเลือกคลิกปุ่ม Segmentation………………………………………………. 34
5.4 ผลที่ได้จากการเลือกคลิกปุ่ม Square (Before Segmentation)………………………….. 34
5.5 ผลที่ได้จากการเลือกคลิกปุ่ม Square (After Segmentation)... 35

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เอกสารเป็นแหล่งข้อมูลและความรู้ที่สำคัญเพื่อช่วยประกอบการค้นคว้าและตัดสินใจในด้านต่างๆ การจัดเก็บเอกสารสามารถจัดเก็บได้ทั้งแบบรูปภาพเอกสารหรือเอกสารข้อความ การจัดเก็บแบบรูปภาพเอกสารสามารถทำได้รวดเร็ว แต่เปลืองเนื้อที่ในการจัดเก็บมาก อีกทั้งไม่สามารถแก้ไขรูปแบบและเนื้อหาได้ง่าย สามารถนำไปประยุกต์ในการค้นหาเอกสารได้ แต่สิ้นเปลืองเวลาและกำลังคนในการนำเข้าข้อมูลมาก ดังนั้น จึงมีการนำโปรแกรมประยุกต์รู้จำอักษรมาช่วยเพิ่มประสิทธิภาพในการจัดเก็บเอกสารมากขึ้น ในประเทศไทยนั้นยังไม่มีการนำระบบรู้จำอักษรภาษาไทยมาใช้อย่างแพร่หลาย เนื่องจากมีข้อจำกัดในเรื่องรูปแบบอักษร สัญญาณรบกวนและตัวอักษรติดกัน การพัฒนาระบบรู้จำอักษรภาษาไทยที่ผ่านมามีการวิจัยด้านการจัดการภาพเอกสารก่อนการรู้จำ (OCR Preprocessing) เพราะมักเกิดจากรูปภาพเอกสารก่อนนำเข้าแก่นรู้จำ (Character Recognition Kernel) มีลักษณะที่ไม่สมบูรณ์หรือไม่เหมาะสม กระบวนการจัดการภาพเอกสารก่อนการรู้จำจึงเป็นสิ่งจำเป็นต่อการรู้จำ

การจัดการภาพเอกสารก่อนการรู้จำ เป็นกระบวนการในการปรับปรุงรูปภาพเอกสารเพื่อนำไปใช้ในส่วนแก่นรู้จำอักษรภาษาไทย ปัญหาที่พบบ่อยจำแนกได้ 3 ประเภท ได้แก่ ปัญหาจากเอกสารต้นฉบับ ปัญหาจากการสแกนเอกสาร และปัญหาจากรูปอักขระของภาษาไทย

ปัญหาจากเอกสารต้นฉบับ เป็นปัญหาที่เกี่ยวกับเอกสารไม่คมชัดหรือมีสัญญาณรบกวนต่างๆในเอกสาร ส่วนใหญ่ปัญหานี้จะเกิดในต้นฉบับที่ได้จากการถ่ายสำเนา สามารถปรับปรุงโดยใช้กระบวนการปรับความชัดของเอกสาร (Binarization) และกระบวนการลดสัญญาณรบกวน (Noise Reduction)

ปัญหาจากการสแกนเอกสาร เป็นปัญหาที่เกี่ยวกับรูปเอกสารที่มีความเอียง และความละเอียดของรูปที่ได้ไม่มากพอ สาเหตุมักเกิดจากผู้ใช้ไม่ระมัดระวัง สามารถแก้ไขโดยใช้กระบวนการปรับความเอียงของเอกสาร (Angle Detection) หรือสแกนรูปเอกสารอีกครั้งด้วยความระมัดระวังมากขึ้น

สำหรับปัญหาจากรูปอักขระของภาษาไทย เป็นปัญหาที่เกี่ยวกับการติดกันของตัวอักษร ซึ่งการติดกันของอักษรภาษาไทยสามารถเกิดได้หลายลักษณะ ทั้งในแนวนอนและแนวตั้ง หรือทั้งสองแนวพร้อมกัน หากมองจากโครงสร้างของภาษาไทยจะเห็นได้ว่ามีระดับอักษรถึง 4 ระดับ ซึ่งทำให้เกิดการทับซ้อน การสัมผัส และการเหลื่อมกันของอักษร ซึ่งเป็นปัญหาที่เกิดจากลักษณะเฉพาะของตัวอักษรภาษาไทยและรูปแบบของตัวอักษร ดังนั้น แนวคิดในการแก้ปัญหาการติดกันที่ใช้

ได้ผลในต่างประเทศ จึงไม่สามารถนำมาใช้ได้โดยตรง จำเป็นต้องมีการปรับหรือใช้เทคนิคผสมกับการแก้ปัญหาการติดของตัวอักษรภาษาไทย

งานวิจัยนี้จึงได้ศึกษาและวิเคราะห์ลักษณะการติดกันของอักษรไทย ทั้งแบบทับซ้อนและสัมผัสกัน โดยศึกษาถึงวิธีการแยกภาพที่ติดกัน เพื่อหาแนวทางการแยกตัวอักษรให้สมบูรณ์มากขึ้น โดยแบ่งออกเป็น 3 ขั้นตอน ขั้นตอนแรกคือ การแบ่งประเภทรูปภาพอักษร โดยใช้การฉายแนวนอน (Horizontal Projection) เพื่อแยกระดับของตัวอักษรทั้ง 3 ระดับ และใช้วิธีติดตามจุดดำในการแยกตัวอักษรในแนวตั้ง ขั้นตอนที่สองคือ การวิเคราะห์ลักษณะและประเภทการติดของตัวอักษร โดยใช้ลักษณะสำคัญของภาษาไทย เช่น ลักษณะปลายตัวอักษรและหัวอักษร เป็นต้น ค่าความกว้างและความสูงของตัวอักษรในระดับต่างๆ และลักษณะการติดของตัวอักษรจากขั้นตอนแรก และขั้นตอนที่สามคือ การหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ ซึ่งประกอบด้วยการหาจุดตัดวิกฤต (Peak to Valley) การหาจุดตัดที่มีจุดร่วมน้อยที่สุด (Break Cost) ตำแหน่งที่ลึกที่สุด (Deepest Hole) หากพบว่าภาพอักษรบางส่วนขาดหายไปจากการตัด จำเป็นต้องการการชดเชยอักษรภาพให้สมบูรณ์ โดยใช้การคำนวณขอบเขตและชดเชยส่วนที่หายไปหลังการตัดหรือการเพิ่มพื้นที่อักษรเพื่อชดเชยอักษรหลังการตัดตัวอักษร เพื่อเพิ่มความสมบูรณ์ให้กับตัวอักษร

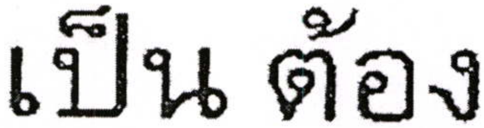
1.2 วัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาปัญหาต่างๆที่เกิดขึ้นในกระบวนการจัดการภาพเอกสารก่อนการรู้จำ ที่ส่งผลกระทบต่อความถูกต้องของการรู้จำอักษร เช่น ตัวอักษรในเอกสารทับซ้อนกัน อักษรสัมผัสกัน เป็นต้น
2. เพื่อศึกษาลักษณะการทับซ้อนและสัมผัสกันของอักษรตัวพิมพ์ภาษาไทย ในระดับบน ระดับล่าง และระดับกลาง
3. เพื่อศึกษาปัญหาที่เกิดขึ้นพร้อมทั้งหาแนวทางในการแก้ไขเพื่อนำไปใช้งานได้อย่างมีประสิทธิภาพ โดยไม่จำเป็นต้องส่งไปทำการเปรียบเทียบโดยการรู้จำหรือเพื่อให้มีการเปรียบเทียบโดยการรู้จำให้น้อยที่สุด
4. เพื่อเป็นแนวทางในการพัฒนาระบบการแยกตัวอักษรที่ติดกันต่อไปในอนาคต

1.3 สมมติฐานของการศึกษา

จากภาพของเอกสารที่ได้จากการสแกนเพื่อที่จะนำเข้าสู่กระบวนการแยกภาพตัวอักษรให้ได้ตัวอักษรเดี่ยวๆออกมานั้น ภาพตัวอักษรมีลักษณะต่างๆกันดังนี้

1. ตัวอักษรที่มีการเหลื่อมล้ำกัน คือ การที่บางส่วนของขอบเขตตัวอักษรสองตัวที่อยู่ต่อเนื่องกันมีลักษณะเหลื่อมล้ำกัน ดังรูปที่ 1.1 เป็นการเหลื่อมล้ำกันระหว่าง ป กับ ี หรือ ต กับ ้



รูปที่ 1.1 ตัวอย่างของตัวอักษรที่มีการเหลื่อมล้ำกัน

2. ตัวอักษรที่มีการสัมผัสกัน คือ การที่บางส่วนของตัวอักษรมีการสัมผัสกัน สามารถเกิดได้ทั้งในระดับเดียวกันและต่างระดับกัน สาเหตุเนื่องมาจากคุณภาพของการสแกน และรูปแบบของตัวอักษร ดังรูปที่ 1.2 เป็นการสัมผัสกันระหว่าง ี กับ ้ หรือ ี กับ ่ หรือ ข กับ ง

รูปที่ 1.2 ตัวอย่างของตัวอักษรที่มีการสัมผัสกัน

3. ตัวอักษรที่มีการทับซ้อนกัน คือ การที่มีการทับซ้อนกันของตัวอักษร ซึ่งในกรณีนี้จะเกิดจากคุณภาพของการสแกน และรูปแบบของตัวอักษร ดังรูปที่ 1.3 เป็นการทับซ้อนกันระหว่าง ป กับ ั

รูปที่ 1.3 ตัวอย่างของตัวอักษรที่มีการทับซ้อนกัน

สำหรับงานวิจัยนี้ จะพิจารณาเฉพาะการสัมผัสกันของตัวอักษรเท่านั้น เมื่อทำการศึกษาในเบื้องต้นพบว่า ความกว้างและความสูงของตัวอักษร สามารถนำมาใช้พิจารณาระบุถึงการสัมผัสกันของตัวอักษรในรูปแบบต่างๆ ประกอบกับคุณสมบัติทางกายภาพของตัวอักษร ซึ่งจะเป็นแนวทางในการแก้ปัญหาการสัมผัสกันของตัวอักษรได้

1.4 ขอบเขตของการดำเนินการวิจัย

1. กลุ่มของตัวอักษรที่ใช้ประกอบด้วยตัวอักษรตัวพิมพ์ภาษาไทย (Thai Printed Character) ที่นิยมใช้กันอยู่ทั่วไป ซึ่งมีแบบอักษร AngsanaUPC, CordiaUPC, BrowalliaUPC, DilleniaUPC และ EucrosiaUPC โดยมีขนาดของตัวอักษร 14 และ 16 จุด

2. งานวิจัยนี้ทำการทดลองกับข้อมูลที่เป็นบทความภาษาไทย เช่น หนังสือพิมพ์ วารสาร จำนวน 58,575 ตัว และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์อิงค์เจ็ท และเครื่องพิมพ์เลเซอร์ จำนวน 611,280 ตัว เป็นต้น

3. ขนาดของตัวอักษรใน 1 บรรทัดที่นำมาวิเคราะห์จะต้องมีขนาดเดียวกัน

4. รูปแบบของตัวอักษร ไม่ครอบคลุมตัวอักษรแบบเอน (Italic Fonts) ขีดเส้นใต้ และตัวเลขไทย ไม่มีภาพ และตาราง

5. ข้อมูลเอกสารไม่ครอบคลุมถึงเอกสารเอียง

6. กรณีตัวอักษร 1 ตัว แต่มีหลายส่วนประกอบ เช่น แ ฐ ญ ะ งานวิจัยนี้ จะถือว่าเป็นคนละตัวอักษร

1.5 ขั้นตอนของการศึกษา

1. ศึกษาบทความและทฤษฎีต่างๆที่มีความเกี่ยวข้องกับงานวิจัยนี้

2. เก็บข้อมูลตัวอย่างจากหนังสือพิมพ์ วารสาร และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์อิงค์เจ็ทและเครื่องพิมพ์เลเซอร์ พร้อมทั้งนำไปทำการสแกนด้วยเครื่องสแกนเนอร์

3. ศึกษาลักษณะการสัมผัสและทับซ้อนกันของตัวอักษรในแบบต่างๆ พร้อมทั้งแบ่งแยกเป็นกลุ่มๆ

4. เขียนโปรแกรมเพื่อหาจุดที่ใช้ในการแบ่งแยกตัวอักษรออกจากกัน ทั้งในระดับบน ระดับกลาง และระดับล่าง กับข้อมูลตัวอย่าง

5. รวบรวมผลการทดลองที่ได้จากโปรแกรม พร้อมทั้งวิเคราะห์ข้อผิดพลาดจากการทำงานของอัลกอริธึม และแก้ไขข้อผิดพลาดให้การทำงานมีประสิทธิภาพมากที่สุด

6. สรุปผลที่ได้ทั้งหมด พร้อมทั้งทำเอกสารนำเสนอเป็นงานวิจัย

บทที่ 2
งานวิจัยที่เกี่ยวข้อง

เอกสารภาษาไทยที่จะนำไปประมวลผลการรู้จำตัวอักษร มักจะพบกับปัญหาการติดกันของตัวอักษร และเนื่องจากลักษณะของตัวอักษรภาษาไทยมีหลายระดับ ทำให้มักจะเกิดปัญหาการทับซ้อนหรือติดกันของตัวอักษรทั้งในแนวนอนและในแนวตั้งด้วย ในช่วงทศวรรษที่ผ่านมา มีหลายงานวิจัยพยายามแก้ไขปัญหานี้ โดยการวิเคราะห์หาจุดตัดของตัวอักษรที่ติดกัน จากนั้นจึงแยกตัวอักษรออกจาก

วิชา [1] ใช้ลักษณะบ่งความต่างของอักษรไทย ในการแบ่งลักษณะการติดกันของตัวอักษรพิมพ์ โดยเริ่มจากการแบ่งกลุ่มของอักษรเดี่ยวออกเป็น 6 กลุ่มตามระดับเส้นบรรทัดของการเขียนอักษรได้แก่ อักษรระดับบน อักษรระดับบนกลาง อักษรระดับกลาง อักษรระดับกลางล่าง อักษรระดับล่าง และอักษรระดับเหนือระดับบน แบ่งลักษณะการติดกันของอักษรเดี่ยวออกเป็น 10 กลุ่ม โดยใช้ลักษณะบ่งความต่างในการวิเคราะห์ว่าอักษรนั้นเป็นอักษรเดี่ยวหรืออักษรติดกันและติดกันในกลุ่มใด ลักษณะบ่งความต่างที่ใช้ได้แก่ ความกว้าง ความสูง เส้นตรง และพื้นที่ว่าง เป็นต้น เมื่อวิเคราะห์กลุ่มการติดกันได้แล้วในแต่ละกลุ่ม จะใช้วิธีต่างๆในการตัดที่แตกต่างกันออกไป เช่น ตำแหน่งที่มีค่าการฉายแนวตั้งน้อยที่สุด (Minimum Vertical Projection) ตำแหน่งที่มีค่าฉายแนวนอนน้อยสุด (Minimum Horizontal Projection) ตำแหน่งที่ลึกที่สุด หรือขอบของตัวอักษรในการแบ่ง มีความถูกต้องในการตัด 95.6% จากตัวอักษรตัวอย่าง 18,000 ตัว และมีตัวอักษรที่ติดกัน 5% วิธีการเหล่านี้ ใช้ได้กับการติดกันของอักษรในบางกรณี เช่น ตัวอักษรที่ไม่มีสัญญาณรบกวนมากนัก และลักษณะที่ติดกันจะต้องเป็นลักษณะเดียวกับที่คาดไว้ เช่น จุดที่เกิดอักษรติดจะมีจุดภาพไม่มาก และไม่เหมาะกับอักษรเอียง เนื่องจากวิธีฉาย ตำแหน่งที่ลึกที่สุด หรืออื่นๆ จะใช้ได้ดีเมื่อตำแหน่งค่าที่น้อยที่สุดที่ได้มีความแตกต่างกับค่าที่ใช้เปรียบเทียบสูง เมื่อพบอักษรที่ติดกันมากจะไม่ได้ผลตามที่ต้องการ

ตารางที่ 2.1 รายละเอียดวิธีการแบ่งกลุ่ม และการตัดตัวอักษรของวิชา [1]

กลุ่มที่	กลุ่มอักษร	วิธีการวิเคราะห์	วิธีการตัดแยก	ตัวอย่าง
1	ระดับบน+ ระดับบน	ความกว้างมากกว่าปกติ	ตำแหน่งกึ่งกลางแนวนอน	บัติ มีดี
2	ระดับบน+ ระดับบนกลาง	ความกว้าง 1 ตัวอักษร ไม่มีหางทางซ้าย	ใช้ช่องว่างทางซ้ายที่มากที่สุดบริเวณเส้น	ปี ปี

ตารางที่ 2.1 (ต่อ)

กลุ่มที่	กลุ่มอักษร	วิธีการวิเคราะห์	วิธีการตัดแยก	ตัวอย่าง
		ส่วนล่างไม่เหมือน"เ" ส่วนบนมีสระ ทางขวาเรียบ	ระดับบน	
3	ระดับบน+ ระดับบนกลางต่างคอลัมน์	ความกว้าง 1 ตัวอักษร ไม่มีหางทางซ้าย ส่วนล่างเป็น "เ"	พิจารณาเฉพาะส่วนบนช่วงกึ่งกลางภาพ ตัดที่ช่องว่างจากข้างล่างที่มีค่ามากที่สุด	รัไ ด้ไ
4	ระดับบนกลาง+ ระดับกลางต่างคอลัมน์	ความกว้างมากกว่า 1 ตัวอักษร มีช่องว่างส่วนขวาล่างของภาพ	ช่องว่างทางซ้ายแรกจากล่างที่ยาวถึง w/2 หางของพยัญชนะ	ปรี ฟรี
5	ระดับบน+ ระดับเหนือบน	ความสูงมากกว่าปกติ	ส่วนที่แคบที่สุดบริเวณที่เป็นความสูงปกติของสระระดับบน	พี้ ที่
6	ระดับบน+ ระดับกลาง	กว้าง 1 ตัวอักษร ไม่มีหางทางซ้าย ส่วนล่างไม่เหมือน "เ" ส่วนบนมีสระ ทางขวาไม่เรียบ	จุดที่มีความกว้างน้อยที่สุดบริเวณเส้นระดับบน	ศี ด้
7	ระดับบนกลาง+ ระดับกลาง	ความกว้างมากกว่า 1 ตัวอักษร ไม่มีช่องว่างส่วนขวาล่างของภาพ	ตัดตำแหน่งที่มีความหนาในแนวตั้งน้อยที่สุดในส่วนหน้าของภาพ	ไม
8	ระดับกลาง+ ระดับล่าง	มีส่วนแคบ (เส้นตรงในแนวดิ่งของอักษร) น้อยกว่าส่วนกว้าง	ตัดที่ช่องว่างทางซ้ายที่ลึกที่สุดในบริเวณเส้นบรรทัดล่าง	นุ ดู
9	ระดับกลาง+ ระดับกลาง	กว้างมากกว่าอักษรที่กว้างที่สุด	Breakpoint Criterion Function ของค่าเฉลี่ยการฉายจุดภาพแนวตั้ง	ยา รท
10	นอกเหนือจากนี้	การติดที่มากกว่าสองตัวอักษร	ใช้วิธีตัดจากกลุ่มต่างๆ ผสมกัน	พี้

ศุภกร [2] นำเสนอในส่วนของการจัดเตรียมข้อมูล ในขั้นตอนการตัดแยกตัวอักษรภาษาไทยออกจากภาพของประโยคในระดับที่นอกเหนือจากระดับกลาง โดยแบ่งเป็น 3 ขั้นตอน ได้แก่ ขั้นตอนการแยกภาพตัวอักษรโดยวิธีเปลี่ยนรหัสขอบ ในขั้นตอนนี้ใช้การติดตามขอบ และแทนที่ขอบด้วยเครื่องหมายที่แสดงตำแหน่งของขอบด้านซ้าย ด้านขวา และมุม เพื่อสามารถที่จะคัดลอกตัวอักษรที่เหลื่อมล้ำกันได้ โดยคำนึงลำดับภาพตัวอักษรที่เรียงลำดับหน้าหลังตามลำดับเดิมที่อยู่ในภาพประโยค ในขั้นตอนนี้ ผลของการแยกอาจได้ภาพตัวอักษรติดกัน หลังจากนั้นจึงผ่านเข้าสู่ขั้นตอนการวิเคราะห์การติดกันของตัวอักษรและเสนอแนวทางการตัดแยก โดยอาศัยระดับของตัวอักษรเพื่อทำการแบ่งประเภทของตัวอักษรที่ติดกัน ทำให้สามารถแบ่งแยกประเภทของตัวอักษรทั้งหมดได้ 7 ประเภท จากนั้นใช้การหาค่าฮิสโทแกรมแบบต่างๆ ได้แก่ การฉายจุดภาพ (Pixel Projection) การฉายขอบภาพ (Profile Projection) และการฉายจุดภาพประยุกต์ (Modify Pixel Projection) มาวิเคราะห์ว่าเป็นตัวอักษรที่ติดกันหรือไม่ และอยู่ในกลุ่มย่อยกลุ่มใด โดยใช้วิธีการระบุกลุ่มของอักษรในส่วนบรรทัดหลักออกเป็น 8 กลุ่มตามลักษณะของฮิสโทแกรมที่หาได้ เช่น ถ้าอักษรไม่อยู่ในกลุ่ม ป ฟ ฝ ไ ใ โ และพบว่ามีขอบเขตในระดับกลางบน จะถือว่าเป็นตัวติดกัน ให้แบ่งที่เส้นบรรทัดบน ซึ่งต้องใช้ขั้นตอนการรู้จำร่วมด้วยหลังจากพบว่าเป็นตัวติดแบบใด วิธีการตัดตัวอักษร ถ้าเป็นการติดกันในแนวตั้ง จะใช้เส้นแบ่งระดับเป็นเกณฑ์ในการตัด ถ้าเป็นการติดกันในแนวนอน จะใช้ขอบของตัวอักษรเป็นเกณฑ์ในการแบ่ง หรือบางครั้งใช้ค่าความแตกต่างระหว่างจุดข้างเคียงที่ต่ำที่สุดเพื่อกำหนดจุดตัด การทดลองใช้ตัวอย่างตัวอักษรจากหนังสือพิมพ์นิตยสารต่างๆ 21,145 ตัวอักษร มีตัวอักษรที่ติดกัน 1,566 ตัวอักษร สามารถวิเคราะห์ตัวอักษรที่ติดกันได้ถูกต้อง 91.57% และตัดแยกได้ถูกต้อง 86.01% ข้อดีของวิธีนี้คือ การระบุกลุ่มอักษรยิ่งระบุได้ละเอียดเท่าใด จะทำให้การตัดแยกผิดพลาดน้อยลง แต่ในบางครั้ง การแบ่งกลุ่มยิ่งละเอียด ถ้าแบ่งแล้วไม่ตรงกลุ่มจะเกิดการตัดที่ผิดพลาดได้ และยังมีปัญหากับตัวอักษรที่เหลื่อม หรือซ้อนทับกัน ทำให้รูปตัวอักษรที่ได้บางส่วนหายไป และไม่สามารถใช้กับอักษรที่มีหลายขนาดได้

จรรยา [3] เน้นระบบการแยกตัวอักษรภาษาไทยที่มีลักษณะการติดกันในแนวนอน ซึ่งมีจำนวนการติดกันของตัวอักษรตั้งแต่ 2 ตัวอักษรขึ้นไป โดยใช้วิธีการหาเส้นทางเดินที่สั้นที่สุด (Shortest Path) ซึ่งได้มีการทดลองมาแล้วในภาษาอังกฤษมาลองใช้ในภาษาไทย พบว่ามีความผิดพลาดเกิดขึ้นมาก จึงได้เสนอวิธีการแก้ไขเป็นแบบเฉียงแต่ยังได้ผลไม่ดี จึงเปลี่ยนมาใช้วิธีฮิสโทแกรม การแสดงผลจะเป็นการสรุปความเป็นไปได้ในการแยกตัวอักษรแต่ละวิธี ตัวอักษรที่ใช้เป็นตัวอย่างอักษรที่ติดกันแนวนอน 2 ตัวอักษรทั้งหมด 381 ตัวอย่าง วิธีการหาทางเดินที่สั้นที่สุดตัดถูกต้อง 57.7% วิธีฮิสโทแกรมตัดได้ถูกต้อง 91.08% และเฉพาะวิธีฮิสโทแกรม ตัวอย่างที่เป็นอักษร 1 ตัวมีจำนวน 200 ตัวอย่าง ตัดถูกต้อง 92% และตัวอย่างที่มีติดกันมากกว่า 2 ตัวอักษรขึ้นไปมีอยู่ 109 ตัวอย่าง ตัดถูกต้อง 82.57%

จักริน [4] ได้เสนอวิธีการตัดแยกตัวอักษรที่สัมผัสกันในแนวตั้งและแนวนอน เริ่มจากใช้ลักษณะของตัวอักษรภาษาไทยซึ่งมีหลายระดับ แตกต่างจากภาษาอื่นมาใช้วิเคราะห์ตัวอักษร แบ่งออกเป็น 3 กลุ่มคือ อักษรระดับบน อักษรระดับกลาง และอักษรระดับล่าง หลังจากนั้น นำค่าทางสถิติมาใช้หาฐานนิยมของความกว้างตัวอักษรแต่ละตัว เพื่อนำมาใช้เป็นความกว้างมาตรฐานค่าอัตราส่วนความกว้าง และความสูงเฉลี่ยใช้ในการตรวจสอบว่าตัวอักษรนั้นเป็นอักษรเดี่ยว หรือตัวอักษรที่สัมผัสกัน โดยแบ่งรูปแบบการสัมผัสกันของตัวอักษรแบบต่างๆออกเป็น 8 กลุ่ม และคำนวณหาจุดที่ใช้แบ่งตัวอักษรออกจากกันเป็นตัวอักษรเดี่ยวๆ ด้วยวิธีเฉพาะของแต่ละกลุ่ม เช่น ค่าความแตกต่างระหว่างจุดข้างเคียง (Peak to Valley) โดยใช้สมการที่พัฒนาโดย Kahan, Pavlidis และ Baird [7] มาประยุกต์ใช้กับตัวอักษรไทย และพิจารณาร่วมกับค่าการฉาย ค่าความกว้างมาตรฐาน ค่า Break Cost และการกำหนดค่าช่วงที่เหมาะสมในการตัดแยก จำนวนตัวอักษรที่ใช้ทดสอบ 20,000 ตัวอักษร พบอักษรที่ติดกัน 883 ตัวอักษร คิดเป็น 4.42% ตัดได้ถูกต้อง 98.41% การแบ่งชนิดของอักษรใช้เพียงอัตราส่วนความกว้างและความสูง ไม่ได้วิเคราะห์ให้ละเอียด ทำให้อักษรเดี่ยวบางตัว เช่น ฤ ฦ ๆ ถูกตัดแยกไปด้วย

ตารางที่ 2.2 รายละเอียดวิธีการแบ่งกลุ่ม และการตัดตัวอักษรของจักริน [4]

กลุ่มที่	กลุ่มอักษร	วิธีการวิเคราะห์	วิธีการตัดแยก	ตัวอย่าง
1	ระดับบน+ ระดับบน	ความกว้างมากกว่าปกติ	หาจุดตัดวิกฤตบริเวณกึ่งกลาง	บัติ มีดี
2	ระดับบน+ ระดับบนกลาง	ความกว้าง 1 ตัวอักษร ส่วนบนกว้างมากกว่าปกติ ขาหลังทั้งส่วนบนและส่วนล่างต้องเป็นแท่งตรง ด้านบนต้องเปิด	ตัดตามแนวแกน y บริเวณขาหลังด้านบน	ปี ปี
3	ระดับบน+ ระดับบนกลางต่างคอลัมน์	ส่วนบนกว้างมากกว่าปกติ ส่วนล่างเป็น “ไ”	พิจารณาเฉพาะส่วนบนช่วงกึ่งกลางภาพ ตัดที่ช่องว่างจากข้างล่างที่มีค่ามากที่สุด	รไ ด้ไ
4	ระดับบน+ ระดับเหนือบน	ความสูงมากกว่าปกติ	จุดตัดวิกฤต + New Matric	พี้ ที่

ตารางที่ 2.2 (ต่อ)

กลุ่มที่	กลุ่มอักษร	วิธีการวิเคราะห์	วิธีการตัดแยก	ตัวอย่าง
5	ระดับบน+ ระดับกลาง	ความกว้าง 1 ตัวอักษร ส่วนบนกว้างมากกว่าปกติ ด้านบนปิด	หาจุดตัดวิกฤตบริเวณด้านบน + ตรวจสอบการตัดผ่านจุดดำ	ศี ด้
6	ระดับกลาง+ ระดับล่าง	ระดับล่าง ความกว้างน้อยกว่าค่าที่กำหนด	ตัดตามเส้นแบ่งอักษรระดับล่าง	นุ ดู
7	ระดับกลาง+ ระดับกลาง	กว้างมากกว่าอักษรที่กว้างที่สุด	จุดตัดวิกฤต+ขนาดที่ตัด > 0.6W	ยา รท
8	นอกเหนือจากนี้	การติดที่มากกว่าสองตัวอักษร	ใช้วิธีตัดจากกลุ่มต่างๆ ผสมกัน	ฟี้

ดวงเพ็ญ [6] ได้เสนอการวิเคราะห์และตัดแยกอักษรที่ติดกันในลายมือภาษาไทย เพื่อเพิ่มประสิทธิภาพของระบบรู้จำลายมือ การวิเคราะห์ลักษณะการติดกัน แบ่งเป็น 2 ขั้นตอน ขั้นตอนแรก หาแนวระดับอักษร ซึ่งมีทั้งหมด 5 กลุ่ม เช่น อักษรระดับบน อักษรระดับกลาง เป็นต้น ขั้นตอนที่สอง วิเคราะห์หาลักษณะการติดกันในแนวนอนและแนวตั้งในแต่ละระดับ ซึ่งมีทั้งหมด 14 กลุ่ม เช่น อักษรระดับบนติดกับอักษรระดับบนในแนวนอน อักษรระดับบนติดกับอักษรระดับกลาง เป็นต้น สำหรับขั้นตอนแรก ยังพบปัญหาแนวระดับอักษรลายมือที่ไม่เป็นเส้นตรง งานวิจัยนี้จึงได้พัฒนาเทคนิคการฉายตำแหน่งเส้นขอบภาพ เพื่อทำให้การวิเคราะห์หาระดับอักษรมีความถูกต้องมากยิ่งขึ้น สำหรับการวิเคราะห์หาลักษณะการติดกัน ได้ใช้ข้อมูลสำคัญของตัวอักษร ได้แก่ ความกว้าง ความสูง เส้นตรง พื้นที่ว่าง เป็นต้น ส่วนการแยกตัวอักษรหมายถึง การหาตำแหน่งตัดแยกอักษรที่ติดกันให้ถูกต้อง โดยนำเทคนิคการฉายจุดภาพมาประยุกต์ใช้ในการตัดแยกตัวอักษรระดับบนที่ติดกันในแนวนอน และเทคนิคการฉายขอบภาพ ประยุกต์ใช้ในการตัดแยกตัวอักษรระดับบนที่ติดกันในแนวตั้ง รวมทั้งการใช้ค่าอนุพันธ์อันดับหนึ่งสำหรับตัดแยกตัวอักษรที่ติดกันระหว่างแนวระดับ นอกจากนี้ ยังได้พัฒนาวิธีฉายและอัลกอริธึมเพื่อให้เหมาะสมสำหรับการตัดแยกลายมือในกลุ่มอักษรระดับกลางติดกับอักษรระดับกลาง ได้แก่ การฉายจุดภาพอย่างน้อยที่สุด และฟังก์ชันการหาจุดตัดวิกฤตแบบประยุกต์ ผลการทดสอบกับลายมือ 5,360 ตัว พบตัวอักษรที่ติดกัน 806 ตัว วิเคราะห์ตัวอักษรติดกันได้ 75.93 % ตัดแยกได้ 63.52 %

จากงานวิจัยต่างๆที่ได้กล่าวมาข้างต้นนั้น สามารถสรุปวิธีการที่ใช้ในการตัดแยกตัวอักษรและเปอร์เซ็นต์ความถูกต้องในการตัดแยก ดังตารางที่ 2.3

ตารางที่ 2.3 สรุปวิธีการที่ใช้ในการตัดแยก และเปอร์เซ็นต์ความถูกต้องของงานวิจัย

ปี พ.ศ.	ผู้วิจัย	วิธีการที่ใช้	% ความถูกต้อง
2539	วิชา	- การฉายแนวนอนและการฉายแนวตั้ง - ช่องว่างที่ลึกที่สุด - ขอบของตัวอักษรที่อยู่ต่างระดับ	95.60
2542	ศุภกร	- เส้นระดับบรรทัด - ความกว้างของตัวอักษร - ค่าสถิติของจุดดำ - ระยะห่างจากขอบ	87.00
2543	จรรยา	- เส้นทางที่สั้นที่สุด - ค่าสถิติของจุดดำ	57.70 91.08
2543	จักริน	- ความกว้างของตัวอักษร - จุดตัดวิกฤต - New Metric	98.41
2545	ดวงเพ็ญ	- ฉายขอบภาพ - ฉายจุดภาพ - ฉายขอบภาพอย่างน้อยที่สุด - การหาจุดตัดวิกฤตแบบประยุกต์	63.52

งานวิจัยต่างๆที่ได้กล่าวมาแล้วนั้น ย่อมมีข้อดี ข้อเสีย และปัญหาต่างๆที่เกิดขึ้น เราสามารถสรุปปัญหาต่างๆที่เกิดขึ้นในแต่ละงานวิจัย ดังตารางที่ 2.4

ตารางที่ 2.4 สรุปปัญหาต่างๆที่เกิดขึ้นในแต่ละงานวิจัย

ปี พ.ศ.	ผู้วิจัย	ปัญหา
2539	วิชา	1. เกิดความผิดพลาดเนื่องจากความคล้ายคลึงกันจนแยกไม่ออก เช่น ฏ กับ ฎ , ฤ กับ ฤ เป็นต้น 2. เกิดความผิดพลาดจากการหาจุดตัดผิดเมื่อมีสัญญาณรบกวน หรือตัวอักษรที่เหลื่อมล้ำกัน (Overlapped Characters) ซึ่งจะทำให้บริเวณที่เป็นจุดตัดนั้นไม่ถูกต้อง คือจะมีบางส่วนขาด บางส่วนเกิน 3. แก้ปัญหา ฝ่ ไม่ได้ 4. แก้ปัญหา ปี หรือ บี , ป หรือ บ่ ไม่ได้

ตารางที่ 2.4 (ต่อ)

ปี พ.ศ.	ผู้วิจัย	ปัญหา
		5. ถ้าตัวอักษรมีการติดกันมาก จะทำให้ตัดผิด
2542	ศุภกร	1. ไม่สามารถตัดอักษรระดับกลางได้ 2. มีการแบ่งกลุ่มละเอียด ซึ่งถ้าแบ่งผิดพลาด จะทำให้การตัดแยกผิดพลาดไปด้วย 3. ใช้กับอักษรที่มีหลายขนาดไม่ได้
2543	จรรยา	1. ไม่สามารถตัดอักษรระดับบนและระดับล่างได้ 2. ถ้าจำนวนจุดยอดที่ได้จากการหาค่าฮิสโทแกรมผิดจากจำนวนที่กำหนดไว้ จะทำให้การตัดตัวอักษรผิดพลาด 3. ถ้าตัวอักษรมีการติดกันมาก จะทำให้ตัดผิด
2543	จักริน	1. ฤ ฦ ๆ ถูกตัดไปด้วย 2. แก้ปัญหา ฝ่ ไม่ได้ 3. แก้ปัญหา ปี หรือ บี , ป่ หรือ บ่ ไม่ได้ 4. ถ้าตัวอักษรมีการติดกันมาก จะทำให้ตัดผิด
2545	ดวงเพ็ญ	1. ใช้กับลายมือ ทำให้เส้นบรรทัดไม่เป็นแนวเส้นตรง ทำให้เกิดความผิดพลาดในการหาระดับบรรทัด 2. ความกว้างของตัวอักษร ขนาดโดยรวมแตกต่างกันมาก และปัญหาตัวอักษรเอียง ส่งผลให้ค่ากลางความกว้างของตัวอักษรที่คำนวณได้ ไม่สามารถใช้ในการวิเคราะห์การติดกันของอักษรทั้งหมดได้

บทที่ 3

ทฤษฎีและแนวทางที่ใช้

การวิเคราะห์และตัดแยกตัวอักษรที่ติดกัน ประกอบด้วย 3 ส่วน ส่วนแรกคือการแยกภาพอักษรออกจากเอกสาร โดยใช้การฉายแนวนอน เพื่อแยกระดับอักษร และใช้การติดตามจุดดำแยกตัวอักษรออกมา ส่วนต่อมาคือส่วนตรวจสอบการติดกันของตัวอักษร โดยใช้ความกว้าง ความสูง และลักษณะสำคัญของประเภทระดับอักษร เช่น ลักษณะหัวหรือหาง เป็นต้น ส่วนต่อมาคือส่วนการหาตำแหน่งที่ติดกัน โดยใช้การหาค่าจุดตัดวิกฤต หรือการหาจุดตัดที่มีจุดร่วมน้อยที่สุด แล้วชดเชยตัวอักษร เมื่อบางส่วนของภาพอักษรขาดหายไป ในการตัดตัวอักษรแบบทับซ้อน โดยหาจุดที่ลึกที่สุดจากเส้นขอบตัวอักษร กระบวนการที่กล่าวมาทั้งหมดมีใจความสำคัญต่อไปนี้

3.1 การแยกภาพอักษรออกจากเอกสาร

การแยกภาพอักษรออกจากเอกสารจะใช้อัลกอริธึมการฉายแนวนอน [2] เพื่อแยกตัวอักษรจากเอกสารให้เป็นบรรทัด และใช้การติดตามจุดดำแยกย่อยบรรทัดเป็นตัวอักษรเดี่ยว หลังจากที่ได้ตัวอักษร เราจะนำไปใช้ในกระบวนการต่อไป

การฉายแนวนอน เป็นการเก็บค่าสถิติจุดดำโดยใช้ค่าสถิติมาช่วยในการตัดสินใจ แนวคิดคือ จุดที่แบ่งบรรทัด หรือตัวอักษรจะมีค่าสถิติน้อย หรือค่าสถิติเท่ากับศูนย์ในแนวระดับแต่ละแถวตามลำดับ เราสามารถคำนวณจากสูตรดังนี้

$$\text{Horizontal: P(y)} = \sum_{X} P(x,y) \qquad (3.1)$$

ซึ่ง P(x,y) เป็นค่าเม็ดสีของจุด ณ ตำแหน่ง (x,y)

P(x) เป็นค่าผลรวมของจำนวนเม็ดสีในแนวแกน X

P(y) เป็นค่าผลรวมของจำนวนเม็ดสีในแนวแกน Y

รูปที่ 3.1 ตัวอย่างการใช้การฉายแนวนอนเพื่อแยกบรรทัด

ส่วนการติดตามจุดดำ คือการดูว่าจุดที่เราพิจารณานั้น จุดที่ติดกับจุดที่เราพิจารณา 8 จุด เป็นจุดขาวหรือจุดดำ ถ้าเป็นจุดดำ ก็จะพิจารณาซ้ำเหมือนกับเริ่มต้นไปเรื่อยๆจนได้จุดดำทั้งหมดที่ประกอบเป็นหนึ่งตัวอักษร วิธีนี้จะแก้ปัญหาการเหลื่อมล้ำกันของตัวอักษรได้ เช่น ปี ถ้าใช้วิธีการฉายแนวตั้งและแนวนอน จะได้ชุดของตัวอักษรชุดเดียว เนื่องจากตัวอักษรมีการเหลื่อมล้ำกันอยู่ต้องนำไปตัดแยกอีกทีในภายหลัง แต่ถ้าใช้วิธีการติดตามจุดดำ จะได้ตัวอักษรสองชุดแยกกันเลย ไม่ต้องนำไปตัดแยกอีกทีในภายหลัง

3.2 ส่วนตรวจสอบการติดกันของตัวอักษร

การวิเคราะห์การติดกันของตัวอักษรในแต่ละแบบนั้น ต้องใช้วิธีการที่แตกต่างกันไป ซึ่งจะใช้ค่าความกว้างเฉลี่ยของอักษรระดับกลาง (W) และความสูงเฉลี่ยของอักษรระดับบน (H) มาใช้ในการพิจารณา ค่าที่ได้นั้นได้มาจากกลุ่มของตัวอักษรที่นิยมใช้กันอยู่ทั่วไป ซึ่งมีแบบอักษร AngsanaUPC, CordiaUPC, BrowalliaUPC, DilleniaUPC และ EucrosiaUPC โดยมีขนาดของตัวอักษร 14 และ 16 จุด และได้จากหนังสือพิมพ์ สื่อสิ่งพิมพ์ต่างๆ สำหรับการหาค่า W และ H มีรายละเอียดดังนี้

3.2.1 การหาค่าความกว้างเฉลี่ยของตัวอักษรในระดับกลาง

เนื่องจากความแตกต่างของความกว้างของตัวอักษร ไทย เช่น “เ” , “ก” และ “ฌ” เป็นต้น ทำให้ไม่สามารถใช้ความกว้างของตัวอักษรตัวใดตัวหนึ่งมาเป็นมาตรฐานในการเปรียบเทียบได้ ดังนั้น งานวิจัยนี้จึงได้ทดลองหาความกว้างเฉลี่ยของตัวอักษรไทยในรูปแบบต่างๆกัน ซึ่งพบว่าอักษรไทยโดยส่วนใหญ่จะมีความกว้างเป็น 0.8 เท่าของความสูงของตัวอักษรในระดับกลาง จึงใช้ค่านี้เป็นค่าความกว้างเฉลี่ย (W) เพื่อใช้ในการวิเคราะห์การติดกันและตัดแยกตัวอักษรต่อไป

3.2.2 การหาค่าความสูงเฉลี่ยของตัวอักษรในระดับบน

ความสูงเฉลี่ยของตัวอักษรในระดับบน คือความสูงของสระและวรรณยุกต์ในระดับบน ซึ่งจากการทดลองพบว่า สระและวรรณยุกต์ในระดับบนจะมีความสูงประมาณ 0.5 เท่าของของความสูงของตัวอักษรในระดับกลาง ในงานวิจัยนี้จึงใช้ค่านี้เป็นค่าความสูงเฉลี่ยของตัวอักษรในระดับบน (H) เพื่อใช้ในการวิเคราะห์การติดกันและตัดแยกตัวอักษรต่อไป

3.3 การหาตำแหน่งที่ติดกันหรือทับซ้อนกัน

3.3.1 การหาจุดตัดวิกฤต

จุดตัดวิกฤต คือจุดเชื่อมของสองอักษรซึ่งจะมีค่าของการฉายแนวตั้ง (V(x)) เปลี่ยนแปลงแบบ Sharp Minimum ดังนั้น Kahan, Palvidis และ Baird [7] จึงเสนอให้ใช้อัตราส่วนระหว่างอนุพันธ์อันดับสอง คือ V(x-1)-2V(x)+V(x+1) กับค่าของการฉายเป็นสมการเงื่อนไขในการหาจุดตัด

$$PV(x) = \frac{V(x-1) - 2V(x) + V(x+1)}{V(x)} \tag{3.2}$$

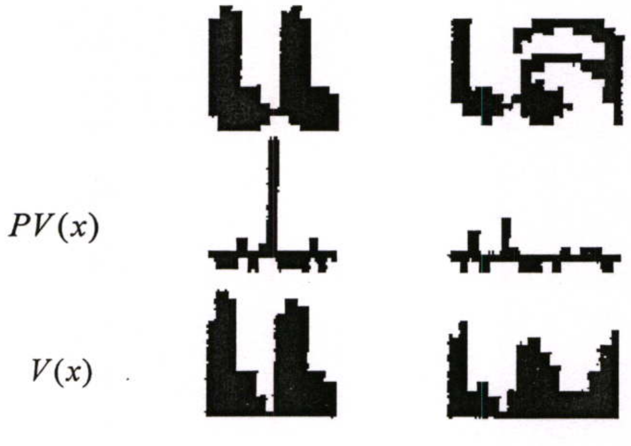
เมื่อ $PV(x)$ เป็นค่าจุดตัดวิกฤตของตำแหน่ง x โดยเปรียบเทียบกับตำแหน่ง x+1 และ x-1

$V(x-1)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x-1

$V(x)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x

$V(x+1)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x+1

จากสมการค่า $PV(x)$ จะมีค่ามากเมื่อค่าผลรวมจุดดำตามแนวตั้ง $V(x)$ มีค่าน้อยกว่าผลรวมของจุดดำตำแหน่ง x+1 และ x-1 แสดงว่าตำแหน่งไหนที่ได้ค่า $PV(x)$ มาก ก็จะตัด ณ ตำแหน่งนั้น แต่ถ้าตำแหน่งที่ติดกันมีจุดภาพมาก อาจจะทำให้ตำแหน่งที่ตัดคลาดเคลื่อนได้

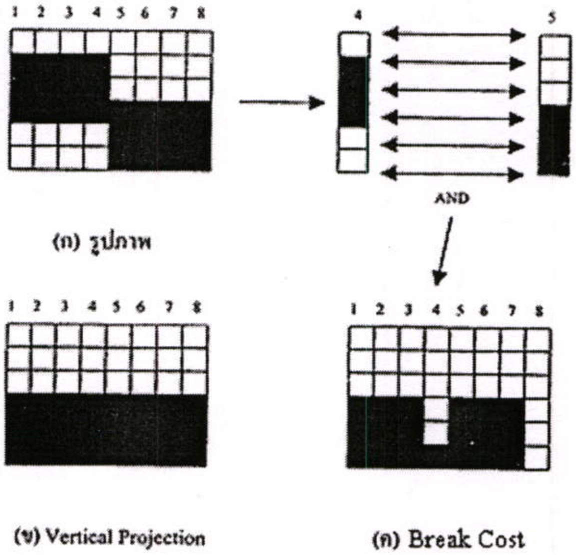


รูปที่ 3.2 การหาจุดแบ่งโดยใช้จุดตัดวิกฤต

จากรูปที่ 3.2 จะเห็นได้ว่า "แ" จะได้ค่า $PV(x)$ สูงสุดจะอยู่ระหว่าง "เ" และ "เ" แสดงว่าตำแหน่งนั้นเป็นตำแหน่งที่สัมผัสกันของตัวอักษร และจะตัด ณ ตำแหน่งนั้น

3.3.2 การหาจุดตัดที่มีจุดร่วมน้อยที่สุด

โดยทั่วไป การหาจุดตัดของตัวอักษรที่สัมผัสกันอย่างง่ายๆ จะใช้วิธีการหาค่าการฉายจุดภาพ เพื่อกำหนดตำแหน่งของจุดตัด จากตำแหน่งที่มีค่าการฉายที่น้อยที่สุด แต่การใช้การหาจุดตัดที่มีจุดร่วมน้อยที่สุด [8] ค่าที่ได้จะบอกถึงนัยสำคัญของการสัมผัสกัน (Degree of Contact) ของแต่ละคอลัมน์ที่ติดกัน ดังรูปที่ 3.3



รูปที่ 3.3 นัยสำคัญของการสัมผัสกัน

จากรูปที่ 3.3 รูป (ก) คือรูปภาพของเนื้อตัวอักษรที่ติดกัน โดยจะแบ่งทั้งหมดเป็น 8 เซลตามแนวตั้ง ถ้าเราใช้วิธีตามรูป (ข) คือวิธีฉายภาพแนวตั้ง เราจะได้ค่าต่างๆในแต่ละเซลเท่ากับ 3 เหมือนกันหมด ทำให้ไม่สามารถหาตำแหน่งของจุดตัดแยกได้ ดังนั้นเราจึงใช้วิธีตามรูป (ค) คือ จะนำเซลที่อยู่ติดกันมา AND กัน จะเห็นว่า ในเซลที่ 4 ค่าที่ได้เท่ากับ 1 นอกนั้นเท่ากับ 3 ดังนั้นวิธีการนี้เราสามารถทราบถึงบริเวณที่มีการสัมผัสกันน้อยที่สุด เพื่อกำหนดเป็นตำแหน่งของจุดตัดตัวอักษรที่ติดกันได้ จากรูป (ค) จุดที่ใช้ในการตัดแยกคือเซลที่ 4

3.4 การปรับปรุงภาพให้สมบูรณ์

3.4.1 การชดเชยขอบเขตส่วนที่หายไป

การชดเชยขอบเขตส่วนที่หายไป (Area Compensation) [5] ใช้สำหรับปรับปรุงภาพหลังการตัดอักษรติดกันระดับบนกับระดับกลางบน เนื่องจากเป็นประเภทที่เกิดการติดกันแบบทับซ้อนหลังการตัด บางส่วนของตัวอักษรจะหายไป เราจำเป็นต้องชดเชยรูปภาพหลังตัด โดยหาจุดตัด

อักษร 4 จุดก่อนการตัด หลังจากตัดแยกตัวอักษรตัวแรกแล้ว เราจะหาเส้นขอบของจุดทั้ง 4 จุด ซึ่งเป็นจุดเชื่อมของตัวอักษรที่ 2 โดยอาศัยสมการเส้นตรง แล้วถมสี (Floodfill) ในบริเวณรอบเส้นขอบและตัวอักษรที่เหลือ เพื่อชดเชยส่วนที่ทับซ้อนกัน เช่น ปี ฝี เป็นต้น ดังรูปที่ 3.4

รูปที่ 3.4 วิธีการหาจุดตัดเพื่อชดเชย และการชดเชย

3.4.2 การเพิ่มจุดตัดหลังการหาจุดตัดวิกฤต

การเพิ่มจุดตัดหลังการหาจุดตัดวิกฤต (Compensation of Peak to Valley) [5] เหมาะสำหรับการติดกันของตัวอักษรระดับบนกับระดับบนเหลื่อมกัน ภาพส่วนหน้าจะเกิดการขาดหายไปของเนื้ออักษร และภาพส่วนหลังเกิดการเกินของเนื้ออักษร ดังนั้น การต่อเติมเนื้อที่ขาดหายไปจะทำเฉพาะภาพส่วนหน้าเท่านั้น โดยพิจารณาหาความชันของหางที่ขาด หาความหนาของเนื้อหาง และคำนวณหาความยาวของหางที่ขาดหายไป เช่น ˜̂ เป็นต้น

สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

บทที่ 4

การวิเคราะห์การสัมผัสกันและตัดแยกตัวอักษร

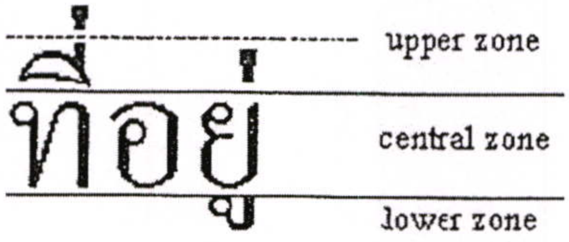
การสัมผัสของตัวอักษรสามารถเกิดขึ้นได้หลายรูปแบบ ทั้งในระดับเดียวกัน และต่างระดับกัน เมื่อเกิดการสัมผัสกันของตัวอักษร วิธีการที่จะหาจุดตัดแยกตัวอักษรนั้น จะต้องใช้วิธีการที่แตกต่างกันไป ตามลักษณะการสัมผัสกันของตัวอักษรนั้นๆ ซึ่งได้จากการสำรวจข้อมูลจากแหล่งข้อมูลต่างๆ เช่น หนังสือพิมพ์ และวารสารต่างๆ เป็นต้น

4.1 ลักษณะของประโยคในภาษาไทย

เมื่อศึกษาลักษณะของประโยคในภาษาไทยพบว่า ประโยคภาษาไทยประกอบด้วยการเรียงกันของพยัญชนะ สระ และวรรณยุกต์ในระดับต่างๆกัน ซึ่งสามารถแบ่งออกได้เป็น 3 ระดับใหญ่ๆ ตามระดับของเส้นบรรทัดดังต่อไปนี้

1. ระดับบน (Upper Zone) ประกอบด้วยสระระดับบน และวรรณยุกต์
2. ระดับกลาง (Central Zone) ประกอบด้วย พยัญชนะ และสระในระดับกลาง แบ่งย่อยได้เป็น กลางบน กลาง และ กลางล่าง
3. ระดับล่าง (Lower Zone) ประกอบด้วยสระระดับล่าง

ตัวอย่างแสดงในรูปที่ 4.1



รูปที่ 4.1 การแบ่งระดับของประโยคในภาษาไทย

ในแต่ละระดับของตัวอักษรนั้น สามารถแสดงประเภทของอักษรเดี่ยวที่แบ่งกลุ่มตามระดับตัวอักษรได้ดังตารางที่ 4.1

69085

ตารางที่ 4.1 ประเภทของอักษรเดี่ยวที่แบ่งกลุ่มตามระดับตัวอักษร

กลุ่ม	ระดับอักษร	อักษร
1	บน	ิ ี ึ ื ์ ั ํ ็ ่ ้ ๊ ๋
2	กลางบน	ป ฝ ฟ พ โ ไ ใ
3	กลาง	ฯ ะ า เ แ ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ 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	กลางล่าง	ญ ฎ ฏ ฐ ฤ ฦ ๆ
5	ล่าง	ุ ู

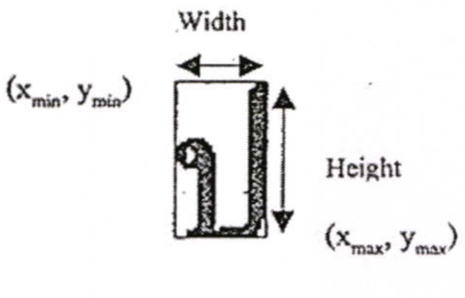
เมื่อแบ่งระดับของตัวอักษรแล้ว ก็ทำการแยกภาพของตัวอักษรออกเป็นอักษรเดี่ยว โดยใช้วิธีการติดตามจุดดำ โดยที่อักษรที่สัมผัสกันยังไม่สามารถแยกกันได้ ดังรูปที่ 4.2

รูปที่ 4.2 กระบวนการแยกภาพตัวอักษรออกเป็นตัวอักษรเดี่ยว

เมื่อทำการแยกตัวอักษรออกเป็นอักษรเดี่ยวแล้ว เราก็จะเก็บค่าต่างๆที่จำเป็น เพื่อใช้ในการวิเคราะห์ ดังต่อไปนี้

- Level เป็นค่าที่บอกระดับของตัวอักษรว่าอยู่ระดับอะไร
- Xmin , Xmax เป็นค่าพิกัดตามแนวแกน X ของภาพตัวอักษร
- Ymin , Ymax เป็นค่าพิกัดตามแนวแกน Y ของภาพตัวอักษร
- Height เป็นค่าความสูงของตัวอักษร
- Width เป็นค่าความกว้างของตัวอักษร

ดังแสดงตัวอย่างในรูปที่ 4.3



รูปที่ 4.3 ตัวอย่างภาพตัวอักษรที่แยกออกเป็นตัวอักษรเดี่ยว

4.2 การวิเคราะห์และจัดกลุ่มตัวอักษรที่สัมผัสกัน

จากตารางที่ 4.1 ระดับที่แตกต่างกันของตัวพิมพ์อักษรภาษาไทย เมื่อเกิดการสัมผัสและทับซ้อนกัน จะเกิดได้ในหลายรูปแบบ เราสามารถวิเคราะห์รูปแบบการติดกันแยกได้เป็น 14 กลุ่มดังนี้

A1. อักษรระดับบน ติดกับ อักษรระดับบน เช่น วิธี ที่

A2. อักษรระดับบน ติดกับ อักษรระดับกลางบน เช่น ปั ปริ

A3. อักษรระดับบน ติดกับ อักษรระดับกลาง เช่น ศึ

A4. อักษรระดับบน ติดกับ อักษรระดับกลางล่าง เช่น ฐี

A5. อักษรระดับกลางบน ติดกับ อักษรระดับกลางบน เช่น ปไ

A6. อักษรระดับกลางบน ติดกับ อักษรระดับกลาง เช่น ปา

A7. อักษรระดับกลางบน ติดกับ อักษรระดับกลางล่าง เช่น ปฎ

A8. อักษรระดับกลางบน ติดกับ อักษรระดับล่าง เช่น ปู

A9. อักษรระดับกลาง ติดกับ อักษรระดับกลาง เช่น กาง

A10. อักษรระดับกลาง ติดกับ อักษรระดับกลางล่าง เช่น ษฐ

A11. อักษรระดับกลาง ติดกับ อักษรระดับล่าง เช่น ฤ

A12. อักษรระดับกลางล่าง ติดกับ อักษรระดับกลางล่าง ไม่เกิดขึ้น

A13. อักษรระดับกลางล่าง ติดกับ อักษรระดับล่าง ไม่เกิดขึ้น

A14. อักษรระดับล่าง ติดกับ อักษรระดับล่าง ไม่เกิดขึ้น

เมื่อทำการรวบรวมข้อมูลจากเอกสารประเภทหนังสือพิมพ์ต่างๆ และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์หรืออิงค์เจทในรูปแบบตัวอักษรต่างๆ เราสามารถตัดแบ่งและรวมกลุ่มรูปแบบการติดกัน โดยใช้วิธีการในการตัดแยกเป็นเกณฑ์ ได้ดังนี้

- กลุ่ม A1 สามารถแยกได้เป็นติดกันในแนวนอน และติดกันในแนวตั้ง จัดใหม่เป็นกลุ่มที่ 1 และ 2
- กลุ่ม A2 สามารถแยกได้เป็นคอลัมน์เดียวกัน และคอลัมน์ติดกัน จัดใหม่เป็นกลุ่มที่ 3 และ 4
- กลุ่ม A3, A4 สามารถยุบรวมกันได้ จัดใหม่เป็นกลุ่มที่ 5
- กลุ่ม A5, A6, A7, A9, A10 สามารถยุบรวมกันได้ จัดใหม่เป็นกลุ่มที่ 6
- กลุ่ม A8, A11 สามารถยุบรวมกันได้ จัดใหม่เป็นกลุ่มที่ 7
- กลุ่ม A12, A13, A14 ไม่เกิดขึ้น เนื่องจากมีอักษรอยู่ติดกันน้อยมาก หรืออาจจะไม่มีเลย สามารถตัดทิ้งได้

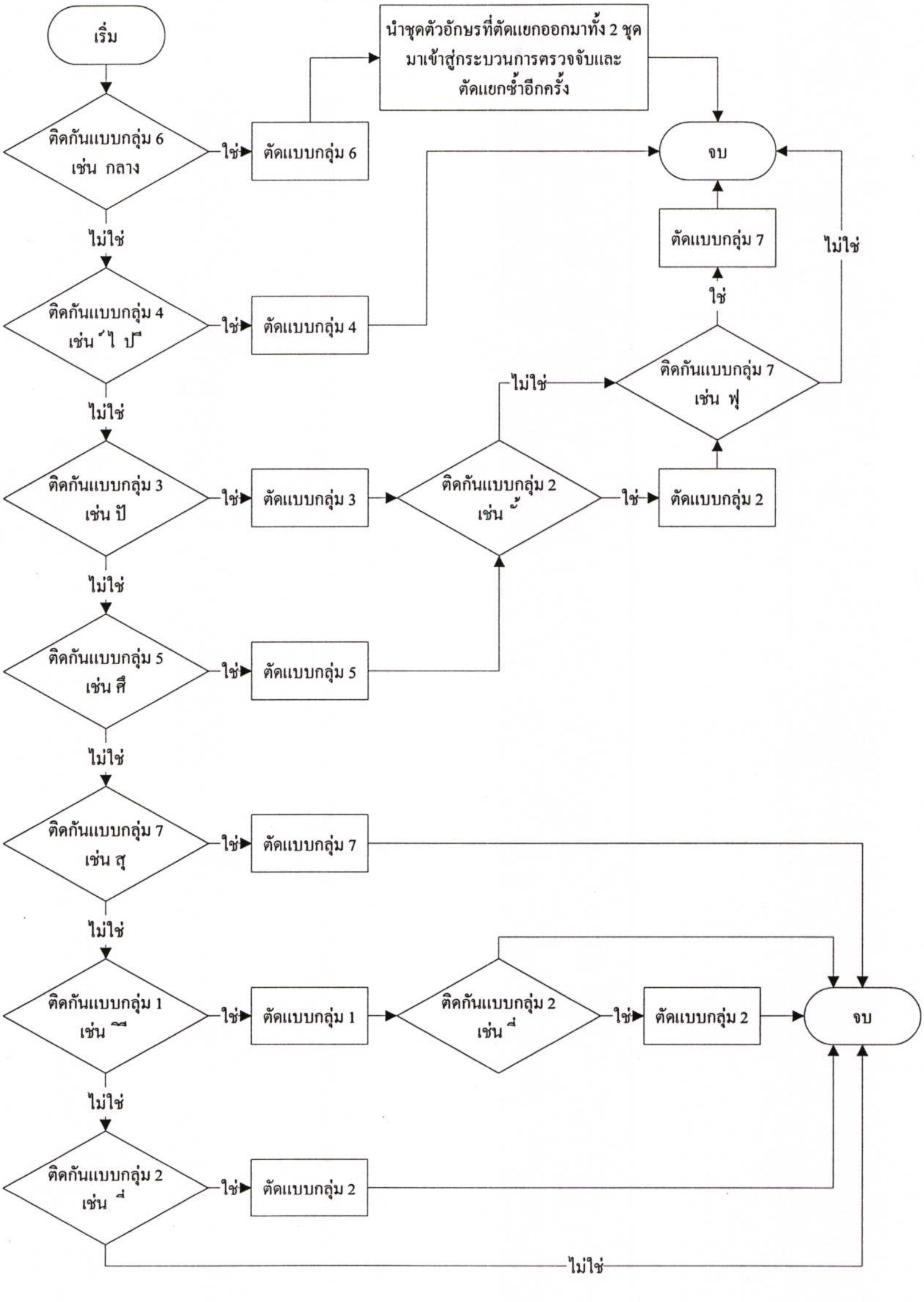
ดังนั้น เราสามารถแบ่งกลุ่มได้ทั้งหมด 7 กลุ่ม ดังตารางที่ 4.2

ตารางที่ 4.2 รูปแบบการติดกันของตัวอักษรในแบบต่างๆ

กลุ่มที่	ลักษณะการสัมผัสและทับซ้อน	ตัวอย่าง
1	อักษรระดับบน ติดกับ อักษรระดับบนในแนวนอน	ีี ัี
2	อักษรระดับบน ติดกับ อักษรระดับบนในแนวตั้ง	ั้ ี่
3	อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์เดียวกัน	ปั ปี ฝ่ ปิ
4	อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน	ั ไ ปี
5	อักษรระดับกลาง ติดกับ อักษรระดับบน	ศึ ด้
6	อักษรระดับกลาง ติดกับ อักษรระดับกลาง	ดา กง
7	อักษรระดับกลาง ติดกับ อักษรระดับล่าง	สุ รู

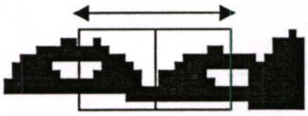
4.3 การวิเคราะห์การสัมผัสกัน และการกำหนดจุดตัดแยก

การวิเคราะห์และตัดแยกตัวอักษรติดกัน จำเป็นจะต้องมีวิธีเฉพาะในแต่ละกลุ่ม ซึ่งในงานวิจัยนี้ใช้ค่าความกว้างและความสูงของตัวอักษรในระดับต่างๆมาใช้ในการพิจารณา สามารถอธิบายภาพรวมในการวิเคราะห์และตัดแยกตัวอักษรได้ดังรูปที่ 4.4



รูปที่ 4.4 ภาพรวมในการวิเคราะห์และตัดแยกตัวอักษร

4.3.1 กลุ่มที่ 1 อักษรระดับบน ติดกับ อักษรระดับบนในแนวนอน ตัวอย่างเช่น ิี ัิ



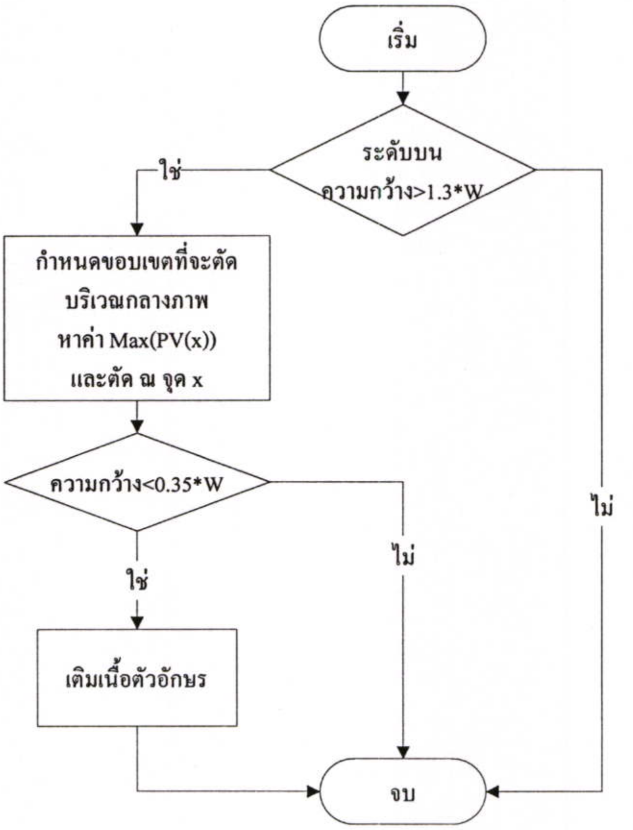
รูปที่ 4.5 ตัวอย่างการตัดในกลุ่มที่ 1

ลักษณะที่ตรวจสอบ ถ้าอยู่ระดับบนและมีความกว้างมาก กว่า 1.3 เท่าของความกว้างเฉลี่ย แสดงว่าติดกัน ถ้าน้อยกว่าจะวิเคราะห์เป็นตัวอักษรเดี่ยว

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

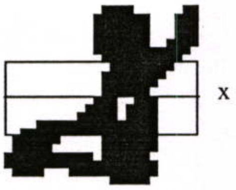
1. กำหนดขอบเขตที่จะตัดบริเวณกลางภาพ
2. หาค่าจุดตัดวิกฤตสูงสุด และตัด ณ ตำแหน่งนั้น
3. ถ้าส่วนที่ตัดได้มีค่าความกว้างน้อยกว่า 0.35 เท่าของความกว้างเฉลี่ย แสดงว่ามีบางส่วนขาดหายไป เนื่องจากการทับซ้อน ให้เติมเนื้อตัวอักษร เช่น ัิ

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.6



รูปที่ 4.6 วิธีการตัดแยกอักษรระดับบน ติดกับ อักษรระดับบนในแนวนอน

4.3.2 กลุ่มที่ 2 อักษรระดับบน ติดกับ อักษรระดับบนในแนวตั้ง ตัวอย่างเช่น ั้ ี่



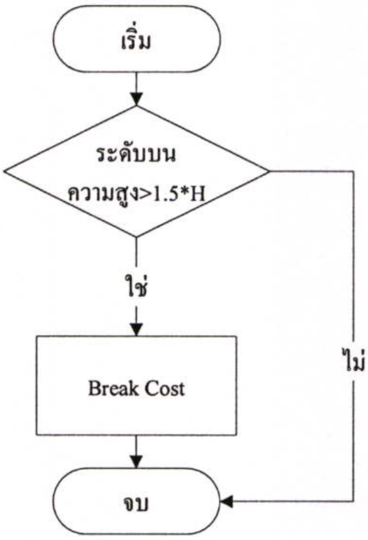
รูปที่ 4.7 ตัวอย่างการตัดในกลุ่มที่ 2

ลักษณะที่ตรวจสอบ ถ้าอยู่ระดับบนและมีความสูงมากกว่า 1.5 เท่าของความสูงตัวอักษรระดับบน แสดงว่าติดกัน

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

1. หาจุดตัดที่มีจุดร่วมน้อยที่สุดในแนวนอน บริเวณกลางภาพ
2. ตัด ณ ตำแหน่งที่มีจุดร่วมน้อยที่สุด

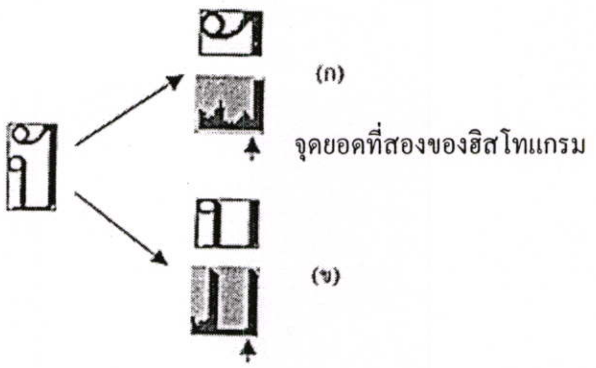
สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.8



รูปที่ 4.8 วิธีการตัดแยกอักษรระดับบน กับ อักษรระดับบนในแนวตั้ง

4.3.3 กลุ่มที่ 3 อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์เดียวกัน ตัวอย่างเช่น ปั ปี ฝ่ ปิ

ลักษณะที่ตรวจสอบ พิจารณาระดับบน (รูป 4.9 ก) และระดับกลาง (รูป 4.9 ข) ถ้าจุดยอดของแท่งฮิสโทแกรมที่สองของทั้งสองระดับตรงกัน และตัวอักษรมีลักษณะเปิดด้านบน เช่น ป ฟ ฝ ถ้าไม่ใช่ก็จะเป็นการติดกันแบบกลุ่มที่ 5

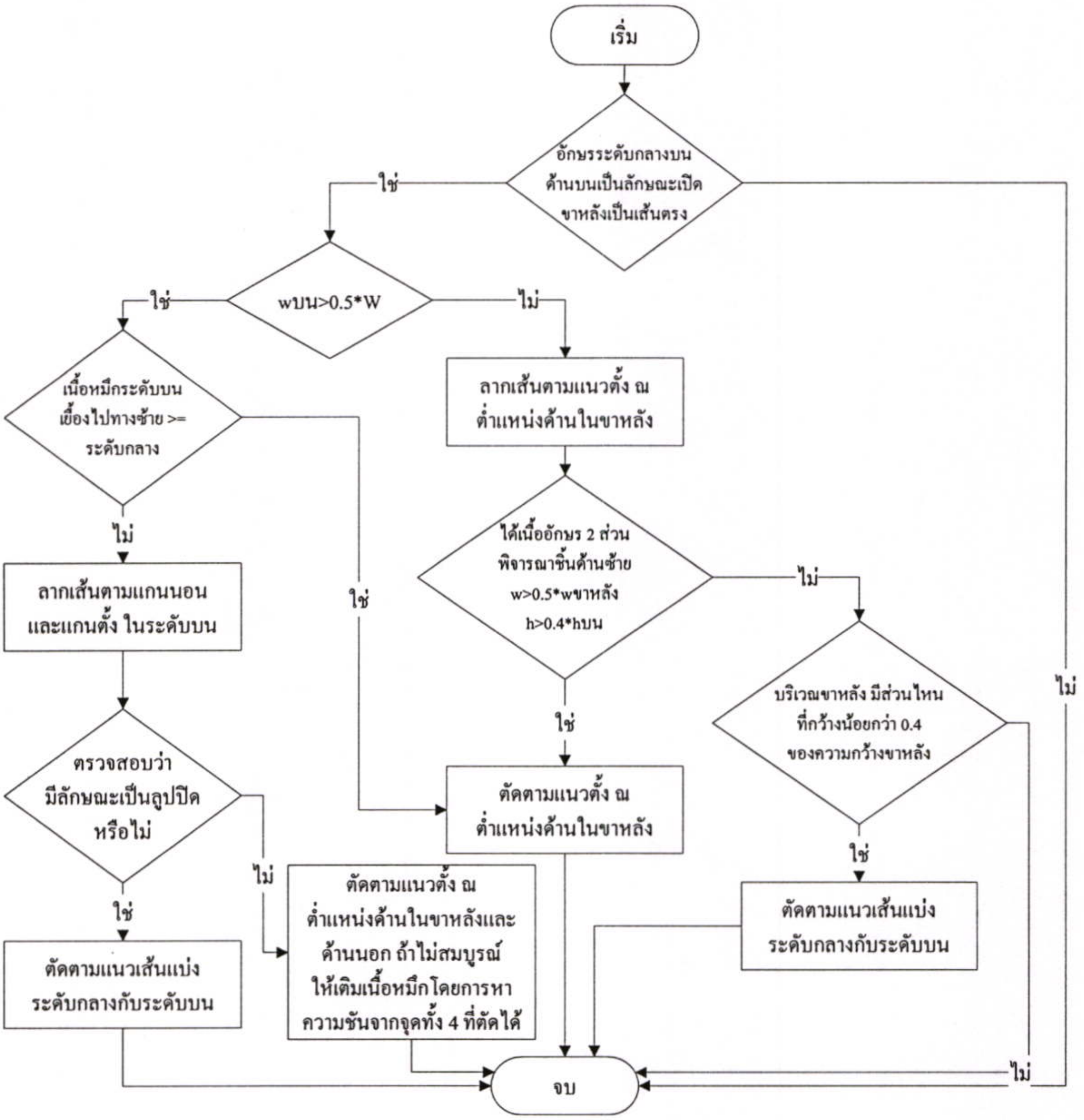


รูปที่ 4.9 ตัวอย่างการตัดในกลุ่มที่ 3

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

1. ถ้าความกว้างส่วนบนน้อยกว่า 0.5*W แสดงว่าตัวอักษรที่ติดกันคือ ไม้เอก หรือสัญญาณรบกวน ให้พิจารณาดังนี้
   1.1 ตัดตามแนวแกนตั้งระดับบน ด้านในขาหลังของตัวอักษร ว่าได้เนื้ออักษรเป็นสองชุดหรือไม่
   1.2 ถ้าได้สองชุด ตรวจสอบว่าเนื้ออักษรที่ได้จากการตัดทางซ้ายมีความกว้างเกิน 0.5 เท่าของความกว้างขาหลังหรือไม่ และความสูงเกิน 0.4 ของความสูงของส่วนบนหรือไม่
   1.3 ถ้าใช่ แสดงว่าเป็นการติดกันในลักษณะ ฝ่ ป่ แต่ถ้าไม่ใช่ อาจจะเกิดจากสัญญาณรบกวน
   1.4 ถ้าเนื้อตัวอักษรขาดหาย ก็เติมเนื้อจุดดำให้เท่ากับความกว้างของขาหลัง
   1.5 ถ้าตัดตามข้อ 1.1 แล้วได้เนื้ออักษรชุดเดียว ให้ตรวจสอบว่าขาหลังของตัวอักษรบริเวณกึ่งกลางมีจุดไหนที่เนื้อตัวอักษรน้อยกว่า 0.4 ของความกว้างของขาหลังหรือไม่
   1.6 ถ้ามี ก็ตัดตามแนวเส้นแบ่งระดับกลางกับระดับบน ซึ่งจะแก้ปัญหา บ่ ในกลุ่มที่ 6 ได้ ถ้าไม่มีก็หยุด
2. ถ้าความกว้างส่วนบนมากกว่า 0.5*W แสดงว่าไม่ใช่ไม้เอกหรือสัญญาณรบกวน ให้พิจารณาดังนี้
   2.1 ถ้าเนื้อหมึกระดับบนเยื้องไปทางซ้ายมากกว่าหรือเท่ากับระดับกลาง เช่น ปี ให้ตัดตามแนวตั้งด้านในขาหลัง
   2.2 ถ้าไม่ใช่ อาจจะเป็น บี ในกลุ่มที่ 6 ให้ลากเส้นตามแนวแกนนอนและแกนตั้งในระดับบนแล้วตรวจสอบว่ามีลักษณะเป็นวงหรือลูปปิดหรือไม่ ถ้ามีตัดตามเส้นแบ่งระดับกลางกับระดับบน
   2.3 นอกเหนือจากนี้ให้ตัดตามแนวตั้งด้านในและด้านนอกขาหลัง ถ้าตัวอักษรไม่สมบูรณ์ เช่น ปี ให้เติมเนื้อหมึก โดยการหาความชันจากจุด 4 จุดที่ได้

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.10



รูปที่ 4.10 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์เดียวกัน

4.3.4 กลุ่มที่ 4 อักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน แบ่งเป็น 2 แบบ คือ
<u>แบบที่ 1</u> ตัวอย่างเช่น ปี

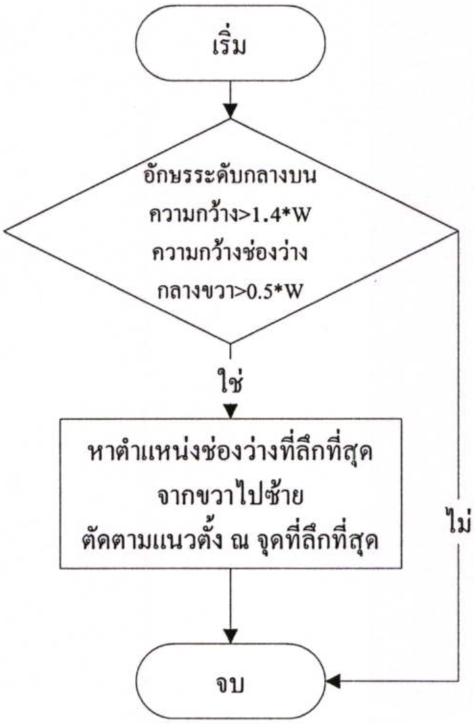
รูปที่ 4.11 ตัวอย่างการตัดในกลุ่มที่ 4 แบบที่ 1

ลักษณะที่ตรวจสอบ เป็นอักษรระดับกลางบน มีความกว้างมากกว่า 1.4*W และความกว้างช่องว่างระดับกลางด้านขวามากกว่า 0.5*W

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

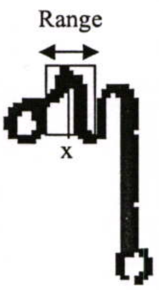
1. หาตำแหน่งช่องว่างที่ลึกที่สุดจากขวาไปซ้าย
2. ตัดตามแนวตั้ง ณ จุดที่ลึกที่สุด

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.12



รูปที่ 4.12 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน แบบที่ 1

<u>แบบที่ 2</u> ตัวอย่างเช่น ˊไ



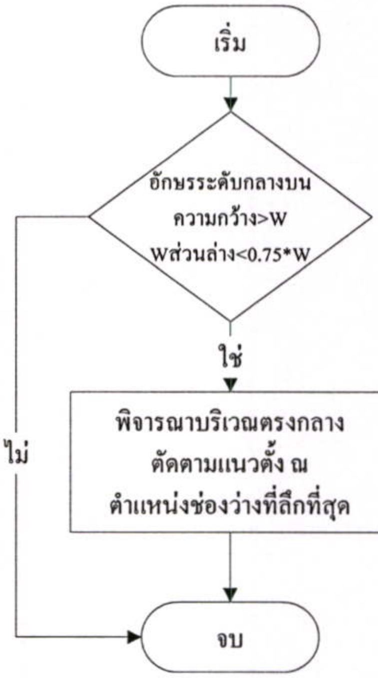
รูปที่ 4.13 ตัวอย่างการตัดในกลุ่มที่ 4 แบบที่ 2

ลักษณะที่ตรวจสอบ เป็นอักษรระดับกลางบน มีความกว้างมากกว่า W และความกว้างระดับล่าง น้อยกว่า 0.75*W

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

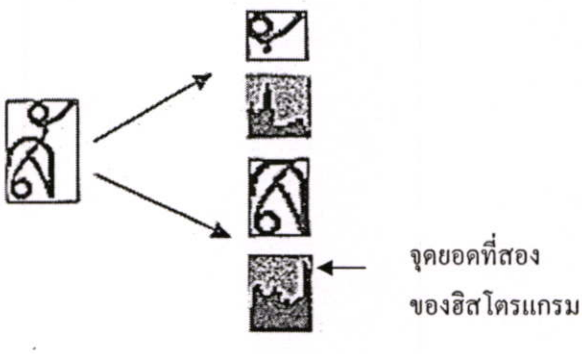
1. พิจารณาบริเวณกลางภาพ
2. ตัด ณ ตำแหน่งช่องว่างที่ลึกที่สุด

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.14



รูปที่ 4.14 วิธีการตัดแยกอักษรระดับกลางบน ติดกับ อักษรระดับบนคอลัมน์ติดกัน แบบที่ 2

4.3.5 กลุ่มที่ 5 อักษรระดับกลาง ติดกับ อักษรระดับบน ตัวอย่างเช่น สั ศึ ด้



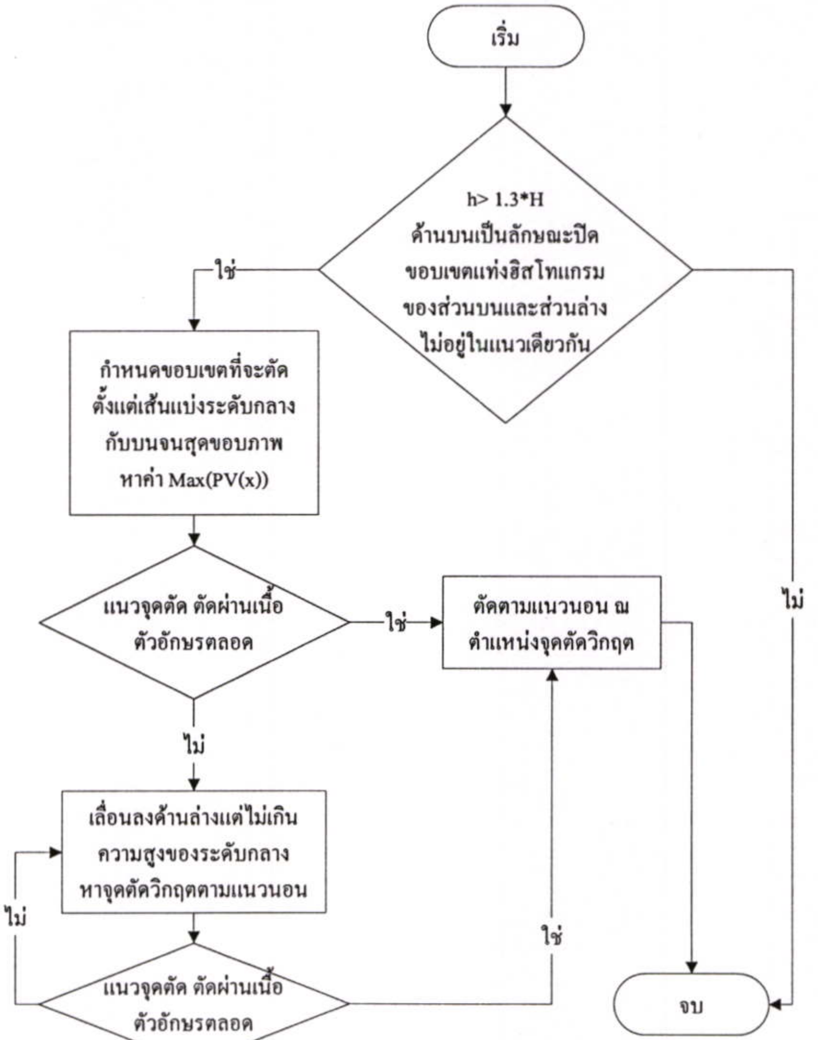
รูปที่ 4.15 ตัวอย่างการตัดในกลุ่มที่ 5

ลักษณะที่ตรวจสอบ พิจารณาระดับบนและระดับกลาง ดังรูปที่ 4.15 ถ้าจุดยอดที่สองของทั้งสองส่วนไม่ตรงกัน และตัวอักษรมีลักษณะปิดด้านบน เช่น ศ ด ถ้าไม่ใช่ ก็จะเป็นการติดกันแบบกลุ่มที่ 3

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

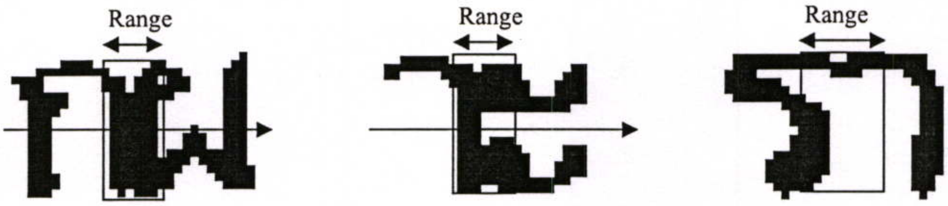
1. หาค่าจุดตัดวิกฤตสูงสุดตามแนวนอนช่วงระดับบน
2. ตรวจสอบว่ามีการตัดผ่านเนื้อตลอดช่วงหรือไม่ เพราะอาจจะตัดโดน ิ ี ได้
3. ถ้าตัดผ่านเนื้อตลอดช่วง ให้ตัดตรงจุดที่มีค่าจุดตัดวิกฤตสูงสุด ถ้าไม่ ให้เลื่อนลงมาเรื่อยๆ จนกว่าจะตัดผ่านเนื้อตลอดช่วง แต่ไม่เกินความสูงของระดับกลาง
4. ถ้าเลื่อนมาจนสุดแล้วยังไม่สามารถตัดผ่านเนื้อตลอดช่วงได้ ให้ตัดที่ระดับความสูงของระดับ กลาง

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.16



รูปที่ 4.16 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับบน

4.3.6 กลุ่มที่ 6 อักษรระดับกลาง ติดกับ อักษรระดับกลาง ตัวอย่างเช่น ดา กง



(ก) (ข) (ค)

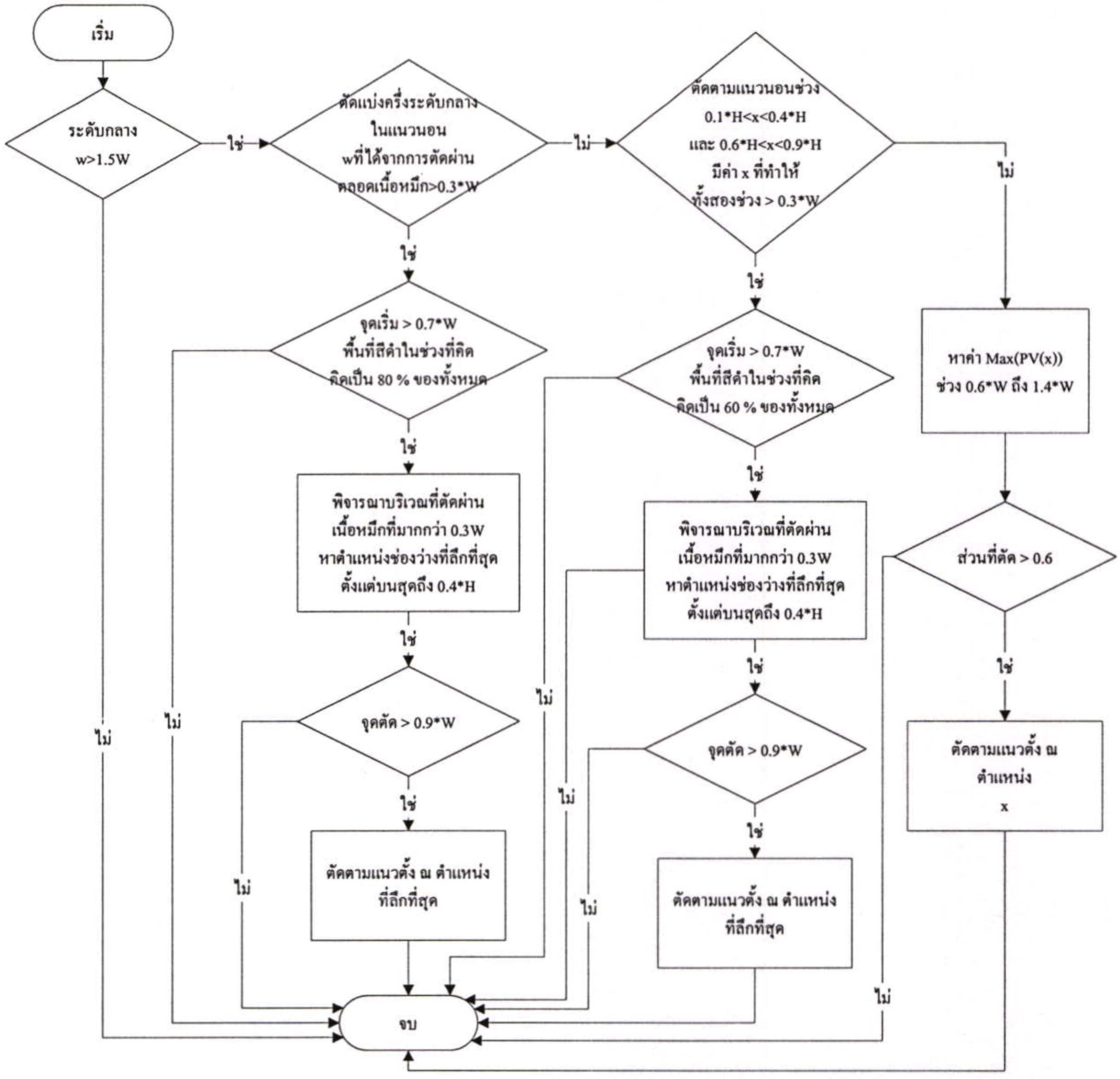
รูปที่ 4.17 ตัวอย่างการตัดในกลุ่มที่ 6

ลักษณะที่ตรวจสอบ มีความกว้างมากกว่า 1.5*W แสดงว่าติดกัน

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

1. ตัดแบ่งครึ่งระดับกลางในแนวนอน ตรวจสอบความหนาของตัวอักษรว่าเกิน 0.3*W หรือไม่
2. ถ้าความหนาเกิน แสดงว่าเกิดการติดกันดังรูป (ก) ให้หาพื้นที่บริเวณรอบๆว่ามีเนื้ออักษรเกิน 80% ของพื้นที่ทั้งหมดที่พิจารณาหรือไม่ ถ้าเกิน ให้หาช่องว่างที่ลึกที่สุดบริเวณที่พิจารณาเป็นจุดตัด ถ้าตำแหน่งที่ตัดมากกว่า 0.9*W ให้ตัดได้ แต่ถ้าไม่เกินก็หยุด
3. ถ้าความหนาไม่เกิน 0.3*W แสดงว่าเกิดการติดกันดังรูป (ข) ให้ตรวจดูช่วงบนและช่วงล่างว่ามีเกินช่วงละ 1 หรือไม่ ถ้ามี ให้หาพื้นที่บริเวณรอบๆว่ามีเนื้ออักษรเกิน 60% หรือไม่ ถ้าเกินให้หาช่องว่างที่ลึกที่สุดบริเวณนั้นเป็นจุดตัด ถ้าระยะที่ตัดมากกว่า 0.9*W ตัด ถ้าไม่เกินหยุด
4. นอกเหนือจากนั้น จะเป็นดังรูป (ค)ให้หาค่าจุดตัดวิกฤตสูงสุด และตัด ณ ตำแหน่งนั้น

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.18



รูปที่ 4.18 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับกลาง

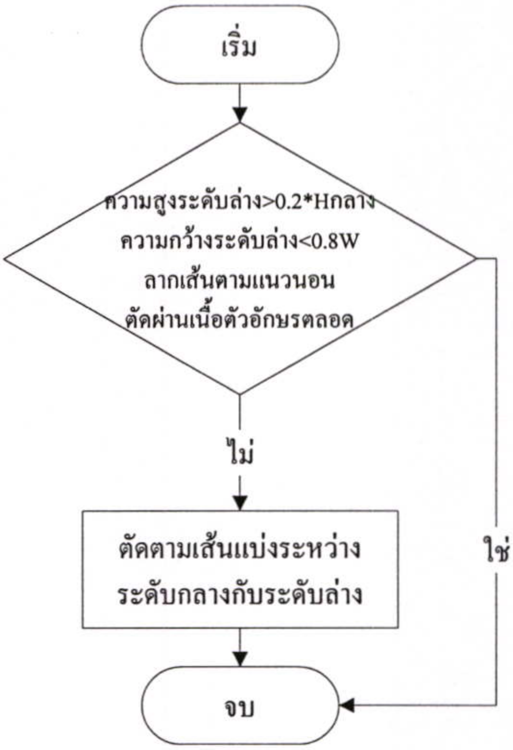
4.3.7 กลุ่มที่ 7 อักษรระดับกลาง ติดกับ อักษรระดับล่าง ตัวอย่างเช่น สุ รู

รูปที่ 4.19 ตัวอย่างการตัดในกลุ่มที่ 7

ขั้นตอนการตัดแยกตัวอักษร มีดังนี้

1. ความสูงระดับล่างมากกว่า 0.2 เท่าของความสูงระดับกลาง ความกว้างระดับล่างน้อยกว่า 0.8 เท่าของความกว้างระดับกลาง และการตัดผ่านในแนวนอนต้องมีส่วนที่ไม่ตัดผ่านเนื้อตัวอักษรตลอด เพราะตัวอักษร ฤ ฦ ๆ จะได้ไม่โดนตัดไปด้วย
2. ตัดตามเส้นแบ่งระหว่างระดับกลางกับระดับล่าง

สามารถเขียนกระบวนการตัดแยกตัวอักษรได้ ดังรูปที่ 4.20



รูปที่ 4.20 วิธีการตัดแยกอักษรระดับกลาง ติดกับ อักษรระดับล่าง

บทที่ 5

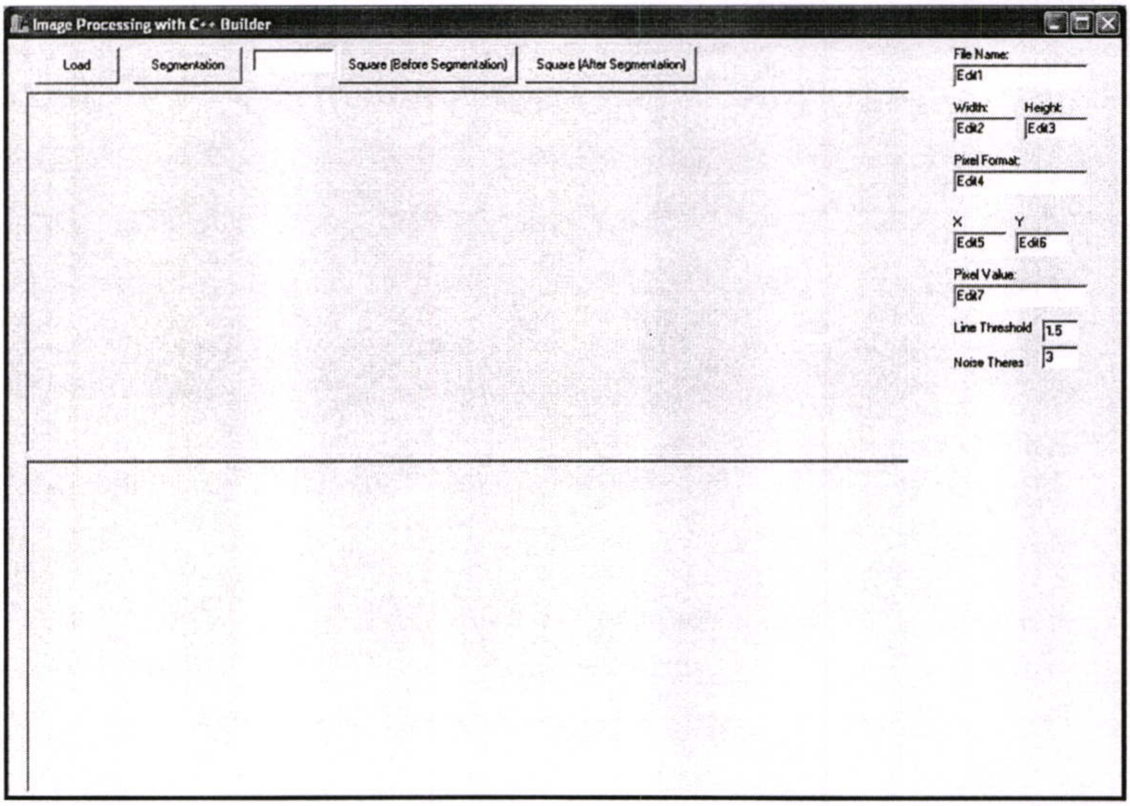
การทดลอง

จากการวิเคราะห์การสัมผัสกัน และกำหนดจุดตัดแยกในรูปแบบต่างๆ ทั้ง 7 กลุ่มนั้น งานวิจัยนี้ ได้ทำการทดลองวิธีดังกล่าว โดยพัฒนาโปรแกรมการตรวจจับและตัดแยกตัวอักษรไทยด้วยภาษา C++ โดยทดลองกับกับข้อมูลจากแหล่งต่างๆ เช่น หนังสือพิมพ์ เอกสารที่ได้จากการพิมพ์ในรูปแบบอักษรที่แตกต่างกันคือ AngsanaUPC, CordiaUPC, BrowalliaUPC, DilleniaUPC และ EucrosiaUPC โดยมีขนาดของตัวอักษร 14 และ 16 จุด เป็นต้น ผ่านการสแกนด้วยความละเอียด 200, 300 และ 400 จุด มีเอกสารที่ใช้ทดสอบทั้งหมด 33 ชุด ซึ่งบทนี้จะแสดงรายละเอียดของผลการทดลองแยกตามประเภทของเอกสาร ได้ผลดังนี้

5.1 โปรแกรมตรวจจับและตัดแยกตัวอักษรที่สัมผัสและทับซ้อนกัน

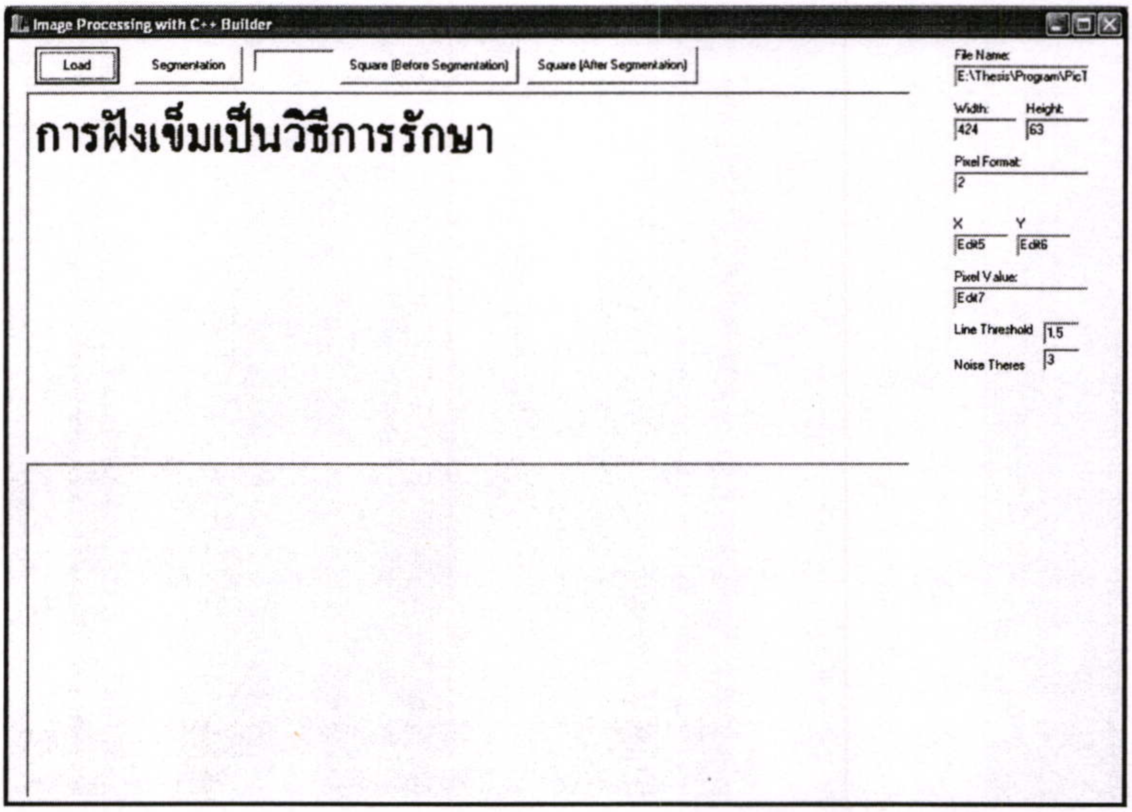
โปรแกรมนี้ เป็นโปรแกรมที่พัฒนาขึ้นมาเพื่อใช้ตรวจจับและตัดแยกตัวอักษรที่สัมผัสและทับซ้อนกัน มีขั้นตอนการทำงานของระบบดังนี้

เมื่อรันโปรแกรมขึ้นมา จะได้หน้าจอดังรูปที่ 5.1



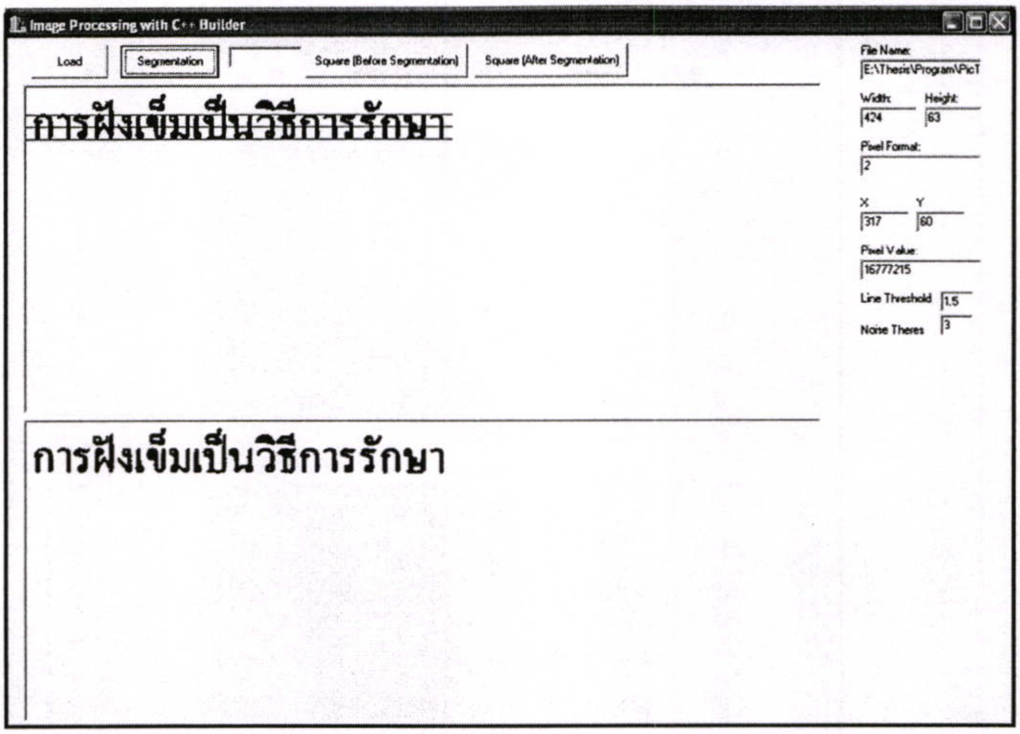
รูปที่ 5.1 หน้าจอหลัก

คลิกปุ่ม Load เพื่อเลือกไฟล์รูปภาพแผนที่บิต (Bitmap) ที่ต้องการตรวจจับและตัดแยก ได้ผลดังรูปที่ 5.2



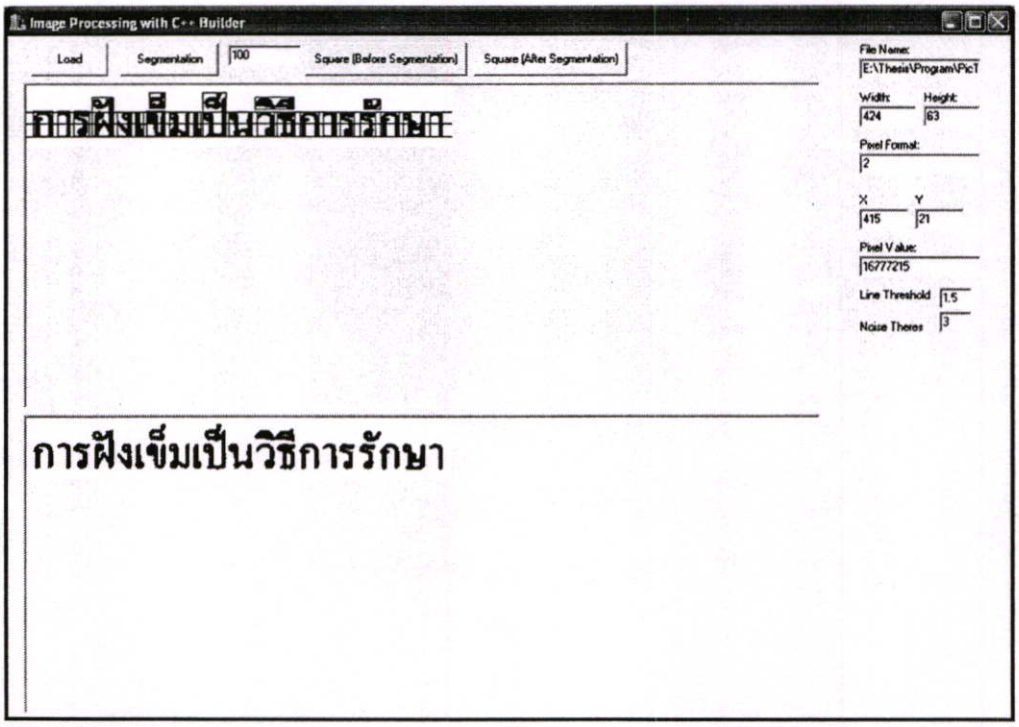
รูปที่ 5.2 ผลที่ได้จากการเลือกไฟล์รูปภาพ

คลิกปุ่ม Segmentation โปรแกรมจะคำนวณหาเส้นแบ่งระดับบน ระดับกลาง และระดับล่าง ในแต่ละบรรทัด แยกอักขระออกเป็นอักษรเดี่ยว ตรวจสอบการติดกันของตัวอักษร และหาตำแหน่งที่ติดกันพร้อมทั้งตัดแยก ได้ผลดังรูปที่ 5.3



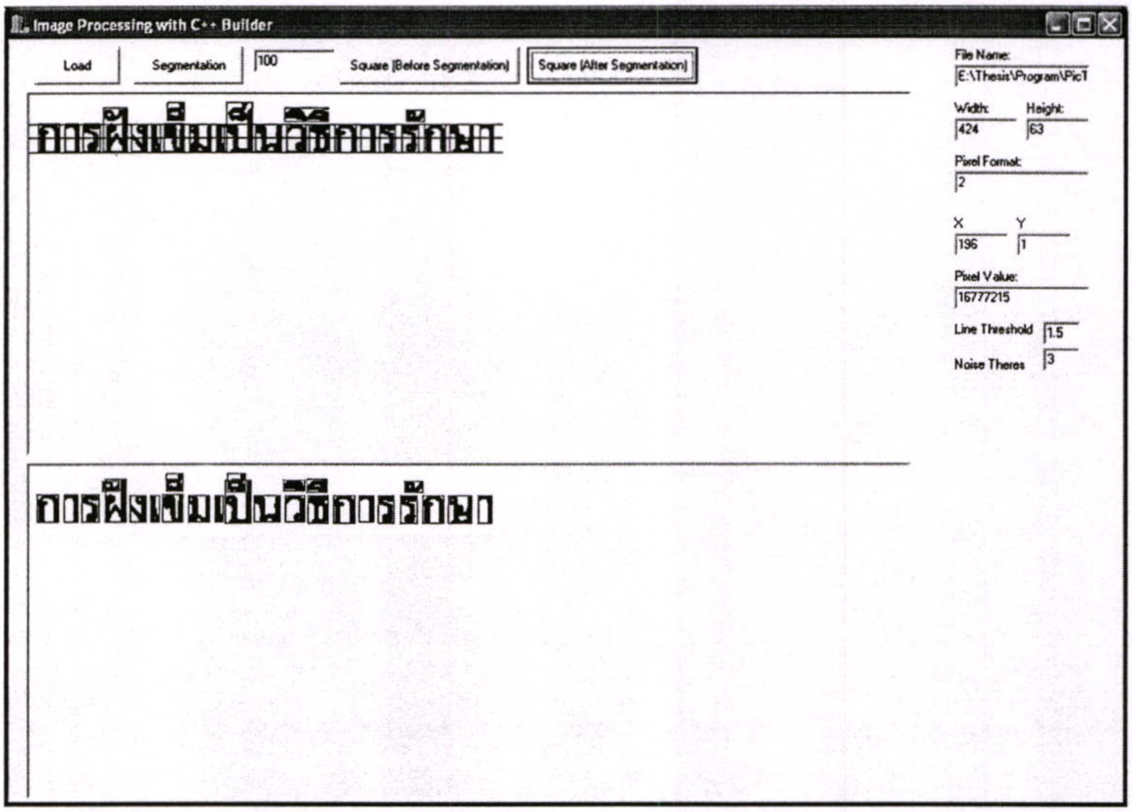
รูปที่ 5.3 ผลที่ได้จากการเลือกคลิกปุ่ม Segmentation

จากรูปที่ 5.3 กรอบด้านบนจะเป็นภาพต้นฉบับ ส่วนกรอบด้านล่างจะเป็นภาพผลที่ได้จากการตัดแยกตัวอักษรที่ติดกัน เมื่อคลิกปุ่ม Square (Before Segmentation) จะได้ผลดังรูปที่ 5.4



รูปที่ 5.4 ผลที่ได้จากการเลือกคลิกปุ่ม Square (Before Segmentation)

จากรูปที่ 5.4 เมื่อคลิกปุ่ม Square (Before Segmentation) โปรแกรมจะใส่กรอบสี่เหลี่ยมสีน้ำเงินล้อมรอบตัวอักษรแต่ละตัวของภาพต้นฉบับ จะสังเกตุได้ว่า ฝ ปี ~~ มีการติดกันอยู่ เพราะอยู่ในกรอบสี่เหลี่ยมเดียวกัน เมื่อคลิกปุ่ม Square (After Segmentation) จะได้ผลดังรูปที่ 5.5



รูปที่ 5.5 ผลที่ได้จากการเลือกคลิกปุ่ม Square (After Segmentation)

จากรูปที่ 5.5 เมื่อคลิกปุ่ม Square (After Segmentation) โปรแกรมจะใส่กรอบสี่เหลี่ยมสีแดงล้อมรอบตัวอักษรแต่ละตัวของภาพที่อยู่กรอบด้านล่าง จะสังเกตุได้ว่า ฝ ปี ~~ มีการตัดแยกออกเป็นสองตัวอักษรเรียบร้อยแล้ว เพราะอยู่คนละกรอบสี่เหลี่ยมกัน ดังนั้น โปรแกรมได้ตรวจจับและตัดแยกตัวอักษรที่สัมผัสและทับซ้อนกันเสร็จสมบูรณ์แล้ว

5.2 รายละเอียดการทดลอง

เอกสารที่ใช้ทดสอบมาจากหนังสือพิมพ์ต่างๆ และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์หรืออิงค์เจท โดยมีความละเอียดจากการสแกน ขนาดตัวอักษร และรูปแบบตัวอักษรต่างๆกันรวม 33 ชุด มีจำนวนตัวอักษรที่ติดกันดังตารางที่ 5.1-5.11

ตารางที่ 5.1 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	56 (0.29%)	48 (0.25%)	37 (0.19%)
2	446 (2.28%)	407 (2.08%)	385 (1.97%)
3	338 (1.73%)	312 (1.60%)	304 (1.56%)
4	63 (0.32%)	55 (0.28%)	52 (0.27%)
5	548 (2.81%)	398 (2.04%)	373 (1.91%)
6	729 (3.73%)	612 (3.13%)	598 (3.06%)
7	354 (1.81%)	334 (1.71%)	323 (1.65%)
รวม	**2,534** (12.98%)	**2,166** (11.09%)	**2,072** (10.61%)

ตารางที่ 5.2 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	47 (0.23%)	36 (0.18%)	32 (0.16%)
2	298 (1.46%)	251 (1.23%)	236 (1.16%)
3	184 (0.90%)	172 (0.84%)	166 (0.81%)
4	42 (0.21%)	34 (0.17%)	31 (0.15%)
5	766 (3.76%)	604 (2.96%)	589 (2.89%)
6	53 (0.26%)	51 (0.25%)	49 (0.24%)
7	396 (1.94%)	377 (1.85%)	362 (1.78%)
รวม	**1,786** (8.77%)	**1,525** (7.48%)	**1,465** (7.19%)

ตารางที่ 5.3 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	40 (0.20%)	32 (0.16%)	30 (0.15%)
2	273 (1.34%)	231 (1.13%)	225 (1.10%)
3	165 (0.81%)	153 (0.75%)	148 (0.73%)
4	39 (0.19%)	30 (0.15%)	27 (0.13%)
5	748 (3.67%)	588 (2.88%)	572 (2.81%)
6	46 (0.23%)	43 (0.21%)	41 (0.20%)
7	367 (1.80%)	358 (1.76%)	351 (1.72%)
รวม	**1,678** (8.24%)	**1,435** (7.04%)	**1,394** (6.84%)

ตารางที่ 5.4 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	75 (0.37%)	52 (0.26%)	44 (0.22%)
2	744 (3.65%)	678 (3.33%)	641 (3.15%)
3	148 (0.73%)	113 (0.55%)	104 (0.51%)
4	67 (0.33%)	49 (0.24%)	42 (0.21%)
5	685 (3.36%)	567 (2.78%)	521 (2.56%)
6	1,062 (5.21%)	675 (3.31%)	487 (2.39%)
7	10 (0.05%)	7 (0.03%)	6 (0.03%)
รวม	**2,791** (13.70%)	**2,141** (10.51%)	**1,845** (9.05%)

ตารางที่ 5.5 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	58 (0.28%)	41 (0.20%)	35 (0.17%)
2	677 (3.32%)	532 (2.61%)	486 (2.39%)
3	132 (0.65%)	92 (0.45%)	74 (0.36%)
4	59 (0.29%)	41 (0.20%)	33 (0.16%)
5	612 (3.00%)	488 (2.39%)	432 (2.12%)
6	756 (3.71%)	422 (2.07%)	367 (1.80%)
7	8 (0.04%)	6 (0.03%)	5 (0.02%)
รวม	**2,302** (11.30%)	**1,622** (7.96%)	**1,432** (7.03%)

ตารางที่ 5.6 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	96 (0.47%)	75 (0.37%)	71 (0.35%)
2	479 (2.35%)	387 (1.90%)	365 (1.79%)
3	301 (1.48%)	274 (1.34%)	259 (1.27%)
4	57 (0.28%)	38 (0.19%)	32 (0.16%)
5	1,582 (7.76%)	1,376 (6.75%)	1,284 (6.30%)
6	744 (3.65%)	635 (3.12%)	562 (2.76%)
7	383 (1.88%)	351 (1.72%)	334 (1.64%)
รวม	**3,642** (17.87%)	**3,136** (15.39%)	**2,907** (14.27%)

ตารางที่ 5.7 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	78 (0.38%)	56 (0.27%)	51 (0.25%)
2	422 (2.07%)	311 (1.53%)	286 (1.40%)
3	267 (1.31%)	232 (1.14%)	201 (0.99%)
4	48 (0.24%)	31 (0.15%)	25 (0.12%)
5	1,462 (7.18%)	1,274 (6.25%)	1,192 (5.85%)
6	675 (3.31%)	521 (2.56%)	472 (2.32%)
7	341 (1.67%)	310 (1.52%)	289 (1.42%)
รวม	**3,293** (16.16%)	**2,735** (13.42%)	**2,516** (12.35%)

ตารางที่ 5.8 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	39 (0.19%)	31 (0.15%)	28 (0.14%)
2	81 (0.40%)	52 (0.26%)	44 (0.22%)
3	476 (2.34%)	408 (2.00%)	387 (1.90%)
4	122 (0.60%)	96 (0.47%)	82 (0.40%)
5	283 (1.39%)	241 (1.18%)	217 (1.06%)
6	4,120 (20.22%)	2,259 (11.09%)	1,475 (7.24%)
7	440 (2.16%)	364 (1.79%)	331 (1.62%)
รวม	**5,561** (27.29%)	**3,451** (16.94%)	**2,564** (12.58%)

ตารางที่ 5.9 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	41 (0.20%)	28 (0.14%)	22 (0.11%)
2	67 (0.33%)	41 (0.20%)	35 (0.17%)
3	402 (1.97%)	355 (1.74%)	331 (1.62%)
4	95 (0.47%)	64 (0.31%)	51 (0.25%)
5	224 (1.10%)	178 (0.87%)	152 (0.75%)
6	1,956 (9.60%)	957 (4.70%)	435 (2.13%)
7	323 (1.59%)	256 (1.26%)	214 (1.05%)
รวม	**3,108** (15.25%)	**1,879** (9.22%)	**1,240** (6.09%)

ตารางที่ 5.10 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	27 (0.13%)	18 (0.09%)	15 (0.07%)
2	528 (2.59%)	453 (2.22%)	421 (2.07%)
3	312 (1.53%)	283 (1.39%)	271 (1.33%)
4	94 (0.46%)	78 (0.38%)	65 (0.32%)
5	456 (2.24%)	376 (1.85%)	341 (1.67%)
6	891 (4.37%)	647 (3.18%)	533 (2.61%)
7	143 (0.70%)	102 (0.50%)	83 (0.41%)
รวม	**2,451** (12.03%)	**1,957** (9.60%)	**1,729** (8.49%)

ตารางที่ 5.11 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน		
	ความละเอียดภาพ 200 จุด	ความละเอียดภาพ 300 จุด	ความละเอียดภาพ 400 จุด
1	20 (0.10%)	12 (0.06%)	10 (0.05%)
2	463 (2.27%)	385 (1.89%)	354 (1.74%)
3	266 (1.31%)	231 (1.13%)	204 (1.00%)
4	71 (0.35%)	59 (0.29%)	53 (0.26%)
5	372 (1.83%)	288 (1.41%)	247 (1.21%)
6	675 (3.31%)	425 (2.09%)	311 (1.53%)
7	106 (0.52%)	76 (0.37%)	55 (0.27%)
รวม	**1,973** (9.68%)	**1,476** (7.24%)	**1,234** (6.06%)

ผลจากการทดลองกับเอกสาร 33 ชุด ได้ผลสรุปดังตารางที่ 5.12-5.45

ตารางที่ 5.12 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	56	56	56	0	0	0
2	446	446	446	0	0	0
3	338	338	338	0	0	0
4	63	57	57	0	0	6
5	548	548	548	0	0	0
6	729	718	718	0	0	11
7	354	354	354	0	0	0
รวม	**2,534**	**2,517**	**2,517**	**0**	**0**	**17**

จากตารางที่ 5.12 สรุปได้ว่า กลุ่มที่ 1, 2, 3, 5 และ 7 สามารถตรวจจับและตัดแยกได้ถูกต้องทั้งหมด กลุ่มที่ 4 และ 6 ตรวจจับไม่ได้รวมทั้งหมด 11 ตัวอักษร เช่น “ ่ใ ” และ “ เภ ” ดังนั้นตัวอักษรที่ตรวจจับได้ถูกต้องสามารถตัดแยกได้ถูกต้องทั้งหมด

ตารางที่ 5.13 – 5.44 ผลที่ได้จะมีลักษณะเช่นเดียวกับตารางที่ 5.12 แต่รูปแบบอักษร CordiaUPC และ DilleniaUPC นั้น ในกลุ่มที่ 3 จะมีตัวอักษรที่ตรวจจับผิด คือ ฬ และตัดแยกผิด เพราะมีความสูงมากกว่าความสูงเฉลี่ย ดังนั้นโปรแกรมจึงเข้าใจว่าเป็นการติดกันระหว่างอักษรระดับกลางติดกับอักษรระดับบน ทำให้ถูกตัดแยกไปด้วย

ตารางที่ 5.13 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	48	48	48	0	0	0
2	407	407	407	0	0	0
3	312	312	312	0	0	0
4	55	50	50	0	0	5
5	398	398	398	0	0	0
6	612	603	603	0	0	9
7	334	334	334	0	0	0
รวม	**2,166**	**2,152**	**2,152**	**0**	**0**	**14**

ตารางที่ 5.14 ข้อมูลจากหนังสือพิมพ์ จากตัวอักษรทั้งหมด 19,525 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	37	37	37	0	0	0
2	385	385	385	0	0	0
3	304	304	304	0	0	0
4	52	48	48	0	0	4
5	373	373	373	0	0	0
6	598	590	590	0	0	8
7	323	323	323	0	0	0
รวม	**2,072**	**2,060**	**2,060**	**0**	**0**	**12**

ตารางที่ 5.15 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	47	47	47	0	0	0
2	298	298	298	0	0	0
3	184	184	184	0	0	0
4	42	38	38	0	0	4
5	766	766	766	0	0	0
6	53	47	47	0	0	6
7	396	396	396	0	0	0
รวม	**1,786**	**1,776**	**1,776**	**0**	**0**	**10**

ตารางที่ 5.16 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	36	36	36	0	0	0
2	251	251	251	0	0	0
3	172	172	172	0	0	0
4	34	31	31	0	0	3
5	604	604	604	0	0	0
6	51	47	47	0	0	4
7	377	377	377	0	0	0
รวม	**1,525**	**1,518**	**1,518**	**0**	**0**	**7**

ตารางที่ 5.17 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	32	32	32	0	0	0
2	236	236	236	0	0	0
3	166	166	166	0	0	0
4	31	28	28	0	0	3
5	589	589	589	0	0	0
6	49	45	45	0	0	4
7	362	362	362	0	0	0
รวม	**1,465**	**1,458**	**1,458**	**0**	**0**	**7**

ตารางที่ 5.18 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	40	40	40	0	0	0
2	273	273	273	0	0	0
3	165	165	165	0	0	0
4	39	36	36	0	0	3
5	748	748	748	0	0	0
6	46	41	41	0	0	5
7	367	367	367	0	0	0
รวม	**1,678**	**1,670**	**1,670**	**0**	**0**	**8**

ตารางที่ 5.19 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	32	32	32	0	0	0
2	231	231	231	0	0	0
3	153	153	153	0	0	0
4	30	27	27	0	0	3
5	588	588	588	0	0	0
6	43	39	39	0	0	4
7	358	358	358	0	0	0
รวม	**1,435**	**1,428**	**1,428**	**0**	**0**	**7**

ตารางที่ 5.20 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร AngsanaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	30	30	30	0	0	0
2	225	225	225	0	0	0
3	148	148	148	0	0	0
4	27	24	24	0	0	3
5	572	572	572	0	0	0
6	41	37	37	0	0	4
7	351	351	351	0	0	0
รวม	**1,394**	**1,387**	**1,387**	**0**	**0**	**7**

ตารางที่ 5.21 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	75	75	75	0	0	0
2	744	744	744	0	0	0
3	148	148	148	0	0	0
4	67	63	63	0	0	4
5	685	685	685	0	0	0
6	1,062	1,051	1,051	0	0	11
7	10	10	10	0	0	0
รวม	**2,791**	**2,776**	**2,776**	**0**	**0**	**15**

ตารางที่ 5.22 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	52	52	52	0	0	0
2	678	678	678	0	0	0
3	113	113	113	0	0	0
4	49	46	46	0	0	3
5	567	567	567	0	0	0
6	675	667	667	0	0	8
7	7	7	7	0	0	0
รวม	**2,141**	**2,130**	**2,130**	**0**	**0**	**11**

ตารางที่ 5.23 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	44	44	44	0	0	0
2	641	641	641	0	0	0
3	104	104	104	0	03	0
4	42	40	40	0	0	2
5	521	521	521	0	0	0
6	487	481	481	0	0	6
7	6	6	6	0	0	0
รวม	**1,845**	**1,837**	**1,837**	**0**	**0**	**8**

ตารางที่ 5.24 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	58	58	58	0	0	0
2	677	677	677	0	0	0
3	132	132	132	0	0	0
4	59	56	56	0	0	3
5	612	612	612	0	0	0
6	756	747	747	0	0	9
7	8	8	8	0	0	0
รวม	**2,302**	**2,290**	**2,290**	**0**	**0**	**12**

ตารางที่ 5.25 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	41	41	41	0	0	0
2	532	532	532	0	0	0
3	92	92	92	0	0	0
4	41	39	39	0	0	2
5	488	488	488	0	0	0
6	422	415	415	0	0	7
7	6	6	6	0	0	0
รวม	**1,622**	**1,613**	**1,613**	**0**	**0**	**9**

ตารางที่ 5.26 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	35	35	35	0	0	0
2	486	486	486	0	0	0
3	74	74	74	0	0	0
4	33	31	31	0	0	2
5	432	432	432	0	0	0
6	367	361	361	0	0	6
7	5	5	5	0	0	0
รวม	**1,432**	**1,424**	**1,424**	**0**	**0**	**8**

ตารางที่ 5.27 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	96	96	96	0	0	0
2	479	479	479	0	0	0
3	301	304	301	0	3	0
4	57	52	52	0	0	5
5	1,582	1,582	1,582	0	0	0
6	744	732	732	0	0	12
7	383	383	383	0	0	0
รวม	**3,642**	**3,628**	**3,625**	**0**	**3**	**17**

ตารางที่ 5.28 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	75	75	75	0	0	0
2	387	387	387	0	0	0
3	274	277	274	0	3	0
4	38	34	34	0	0	4
5	1,376	1,376	1,376	0	0	0
6	635	625	625	0	0	10
7	351	351	351	0	0	0
รวม	**3,136**	**3,125**	**3,122**	**0**	**3**	**14**

ตารางที่ 5.29 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	71	71	71	0	0	0
2	365	365	365	0	0	0
3	259	262	259	0	3	0
4	32	29	29	0	0	3
5	1,284	1,284	1,284	0	0	0
6	562	553	553	0	0	9
7	334	334	334	0	0	0
รวม	**2,907**	**2,898**	**2,895**	**0**	**3**	**12**

ตารางที่ 5.30 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	78	78	78	0	0	0
2	422	422	422	0	0	0
3	267	270	267	0	3	0
4	48	44	44	0	0	4
5	1,462	1,462	1,462	0	0	0
6	675	665	665	0	0	10
7	341	341	341	0	0	0
รวม	**3,293**	**3,282**	**3,279**	**0**	**3**	**14**

ตารางที่ 5.31 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	56	56	56	0	0	0
2	311	311	311	0	0	0
3	232	235	232	0	3	0
4	31	28	28	0	0	3
5	1,274	1,274	1,274	0	0	0
6	521	512	512	0	0	9
7	310	310	310	0	0	0
รวม	**2,735**	**2,726**	**2,723**	**0**	**3**	**12**

ตารางที่ 5.32 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร CordiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	51	51	51	0	0	0
2	286	286	286	0	0	0
3	201	204	201	0	3	0
4	25	22	22	0	0	3
5	1,192	1,192	1,192	0	0	0
6	472	464	464	0	0	8
7	289	289	289	0	0	0
รวม	**2,516**	**2,508**	**2,505**	**0**	**3**	**11**

ตารางที่ 5.33 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	39	39	39	0	0	0
2	81	81	81	0	0	0
3	476	479	476	0	3	0
4	122	115	115	0	0	7
5	283	283	283	0	0	0
6	4,120	4,100	4,100	0	0	20
7	440	440	440	0	0	0
รวม	**5,561**	**5,537**	**5,534**	**0**	**3**	**27**

ตารางที่ 5.34 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	31	31	31	0	0	0
2	52	52	52	0	0	0
3	408	411	408	0	3	0
4	96	91	91	0	0	5
5	241	241	241	0	0	0
6	2,259	2,243	2,243	0	0	16
7	364	364	364	0	0	0
รวม	**3,451**	**3,433**	**3,430**	**0**	**3**	**21**

ตารางที่ 5.35 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	28	28	28	0	0	0
2	44	44	44	0	0	0
3	387	390	387	0	3	0
4	82	78	78	0	0	4
5	217	217	217	0	0	0
6	1,475	1,462	1,462	0	0	13
7	331	331	331	0	0	0
รวม	**2,564**	**2,550**	**2,547**	**0**	**3**	**17**

ตารางที่ 5.36 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	41	41	41	0	0	0
2	67	67	67	0	0	0
3	402	405	402	0	3	0
4	95	89	89	0	0	6
5	224	224	224	0	0	0
6	1,956	1,939	1,939	0	0	17
7	323	323	323	0	0	0
รวม	**3,108**	**3,088**	**3,085**	**0**	**3**	**23**

ตารางที่ 5.37 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	28	28	28	0	0	0
2	41	41	41	0	0	0
3	355	358	355	0	3	0
4	64	59	59	0	0	5
5	178	178	178	0	0	0
6	957	942	942	0	0	15
7	256	256	256	0	0	0
รวม	**1,879**	**1,862**	**1,859**	**0**	**3**	**20**

ตารางที่ 5.38 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร DilleniaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	22	22	22	0	0	0
2	35	35	35	0	0	0
3	331	334	331	0	3	0
4	51	47	47	0	0	4
5	152	152	152	0	0	0
6	435	421	421	0	0	14
7	214	214	214	0	0	0
รวม	**1,240**	**1,225**	**1,222**	**0**	**3**	**18**

ตารางที่ 5.39 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	27	27	27	0	0	0
2	528	528	528	0	0	0
3	312	312	312	0	0	0
4	94	90	90	0	0	4
5	456	456	456	0	0	0
6	891	881	881	0	0	10
7	143	143	143	0	0	0
รวม	**2,451**	**2,437**	**2,437**	**0**	**0**	**15**

ตารางที่ 5.40 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	18	18	18	0	0	0
2	453	453	453	0	0	0
3	283	283	283	0	0	0
4	78	75	75	0	0	3
5	376	376	376	0	0	0
6	647	639	639	0	0	8
7	102	102	102	0	0	0
รวม	**1,957**	**1,946**	**1,946**	**0**	**0**	**11**

ตารางที่ 5.41 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 14 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	15	15	15	0	0	0
2	421	421	421	0	0	0
3	271	271	271	0	0	0
4	65	63	63	0	0	2
5	341	341	341	0	0	0
6	533	527	527	0	0	6
7	83	83	83	0	0	0
รวม	**1,729**	**1,721**	**1,721**	**0**	**0**	**8**

ตารางที่ 5.42 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 200 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	20	20	20	0	0	0
2	463	463	463	0	0	0
3	266	266	266	0	0	0
4	71	68	68	0	0	3
5	372	372	372	0	0	0
6	675	667	667	0	0	8
7	106	106	106	0	0	0
รวม	**1,973**	**1,962**	**1,962**	**0**	**0**	**11**

ตารางที่ 5.43 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 300 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	12	12	12	0	0	0
2	385	385	385	0	0	0
3	231	231	231	0	0	0
4	59	57	57	0	0	2
5	288	288	288	0	0	0
6	425	419	419	0	0	6
7	76	76	76	0	0	0
รวม	**1,476**	**1,468**	**1,468**	**0**	**0**	**8**

ตารางที่ 5.44 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์และอิงค์เจท ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 16 จุด จากตัวอักษรทั้งหมด 20,376 ตัว ที่ความละเอียด 400 จุด

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	10	10	10	0	0	0
2	354	354	354	0	0	0
3	204	204	204	0	0	0
4	53	51	51	0	0	2
5	247	247	247	0	0	0
6	311	306	306	0	0	5
7	55	55	55	0	0	0
รวม	**1,234**	**1,227**	**1,227**	**0**	**0**	**7**

ตารางที่ 5.45 สรุปผลการทดลองทั้งหมด 33 ตาราง จำนวน 669,855 ตัวอักษร

กลุ่มที่	จำนวนตัวอักษรที่ติดกัน	จำนวนตัวอักษรที่ตรวจจับได้				จำนวนตัวอักษรที่ตรวจจับไม่ได้
		ทั้งหมด	ตัดแยกได้ถูกต้อง	ตัดแยกไม่ถูกต้อง		
				ตรวจจับถูกต้อง	ตรวจจับผิด	
1	1,381	1,381	1,381	0	0	0
2	11,684	11,684	11,684	0	0	0
3	8,065	8,101	8,065	0	36	0
4	1,825	1,706	1,706	0	0	119
5	20,036	20,036	20,036	0	0	0
6	24,284	23,986	23,986	0	0	298
7	7,765	7,765	7,765	0	0	0
รวม	**75,040**	**74,659**	**74,623**	**0**	**36**	**417**

5.3 การวิเคราะห์และสรุปผลการทดลอง

จากผลการทดลอง สามารถสรุปค่าความถูกต้องต่างๆได้ดังนี้

% ความถูกต้องในการตรวจจับตัวอักษร = [อักษรทั้งหมด – (อักษรที่ตรวจจับผิด+อักษรตรวจจับไม่ได้)] / อักษรทั้งหมด

= [[669,855-(36+417)]/ 669,855]*100

= 99.93 %

% ความถูกต้องในการตรวจจับและตัดแยก = [(อักษรที่ติดกัน+อักษรตรวจจับผิด) – (อักษรที่ตัดแยกไม่ถูกทั้งตรวจจับถูกและตรวจจับผิด+อักษรตรวจจับไม่ได้)] / (อักษรที่ติดกัน+อักษรตรวจจับผิด)

= [[(75,040+36)-(0+36+417)]/ (75,040+36)]*100

= 99.40%

% ความถูกต้องในการตัดแยกเฉพาะที่ตรวจจับได้ถูกต้อง ไม่รวมตรวจจับผิดและตรวจจับไม่ได้ = อักษรที่ตัดแยกถูก / (อักษรที่ติดกัน-อักษรที่ตรวจจับไม่ได้)

= [74,623/ (75,040-417)]*100

= 100%

การตัดแยกตัวอักษรที่ติดกันในแต่กลุ่ม สามารถตัดแยกได้อย่างถูกต้อง แต่บางกรณีก็ยังมีข้อผิดพลาดอยู่ ดังนี้

1. เกิดจากการติดกันของตัวอักษรที่มีขนาดเล็กกว่าความกว้างหรือความสูงเฉลี่ย เช่น " ่ใ " หรือ " เภ " ซึ่งมีความกว้างน้อยกว่าความกว้างเฉลี่ย จึงถูกพิจารณาเป็น 1 ตัวอักษร ทำให้ไม่ได้ถูกตัดแยกไปด้วย ดังในกลุ่มที่ 4 และ 6
2. ตัวอักษร " พ " จะถูกตัดแยกในรูปแบบอักษร CordiaUPC และ DilleniaUPC เพราะสองรูปแบบนี้จะมีความสูงมากกว่าความสูงเฉลี่ย จึงถูกพิจารณาเป็นอักษร 2 ตัวติดกันคือ อักษรระดับกลางติดกับอักษรระดับบน ทำให้ถูกตัดแยกไปด้วย ดังในกลุ่มที่ 3

นอกจากนี้จากผลการทดลองสรุปได้ว่า ทั้งรูปแบบอักษร ขนาดอักษร และความละเอียดจากการสแกน จะมีผลต่อการติดกันของตัวอักษรแตกต่างกันไป ทำให้การติดกันในแต่ละกลุ่มแตกต่างกันไป ดังนี้

1. รูปแบบตัวอักษรในแต่ละแบบ จะมีระยะห่างในการเรียงกันมากน้อยแตกต่างกันไป เช่น DilleniaUPC จะมีระยะห่างในการเรียงกันน้อยกว่า AngsanaUPC มาก ทำให้โอกาสที่จะติดกันในระดับกลางมีสูงกว่ามากไปด้วย
2. ขนาดของอักษร ถ้ามีขนาดเล็ก ระยะห่างในการเรียงกันก็จะมีน้อยกว่าอักษรที่มีขนาดใหญ่ ดังนั้น ตัวอักษรขนาด 14 จุด จึงติดกันมากกว่าตัวอักษรขนาด 16 จุด
3. ความละเอียดจากการสแกน ถ้าสแกนที่ความละเอียดสูงๆ ก็จะได้ภาพที่มีคุณภาพมากกว่าภาพที่สแกนที่ความละเอียดน้อยๆ ดังนั้น ภาพที่ได้จากการสแกนที่ความละเอียดสูงๆ ตัวอักษรจะติดกันน้อยกว่าภาพที่ได้จากการสแกนที่ความละเอียดต่ำๆ

เพราะฉะนั้น ภาพที่ได้จากการสแกนที่ความละเอียดสูงๆ และตัวอักษรมีขนาดใหญ่ จะช่วยลดปัญหาตัวอักษรติดกันได้มากขึ้น และช่วยลดการทำงานของโปรแกรมที่ใช้ในการตรวจจับและตัดแยกได้มากยิ่งขึ้นไปด้วย

บทที่ 6
สรุปผลงานวิจัยและข้อเสนอแนะ

6.1 สรุปผลงานวิจัย

งานวิจัยนี้จึงได้ศึกษาและวิเคราะห์ลักษณะการติดกันของอักษรไทย ทั้งแบบทับซ้อนและสัมผัสกัน โดยศึกษาถึงวิธีการแยกภาพที่ติดกัน เพื่อหาแนวทางการแยกตัวอักษรให้สมบูรณ์มากขึ้น โดยแบ่งออกเป็น 3 ขั้นตอน ขั้นตอนแรกคือ การแบ่งประเภทรูปภาพอักษร โดยใช้การฉายแนวนอน เพื่อแยกระดับของตัวอักษรทั้ง 3 ระดับ และใช้วิธีติดตามจุดดำ ในการแยกตัวอักษรในแนวตั้ง ขั้นตอนที่สองคือ การวิเคราะห์ลักษณะและประเภทการติดของตัวอักษร โดยใช้ลักษณะสำคัญของภาษาไทย เช่น ลักษณะปลายตัวอักษร และหัวอักษร เป็นต้น ค่าความกว้างและความสูงของตัวอักษรในระดับต่างๆ และลักษณะการติดของตัวอักษรจากขั้นตอนแรก และขั้นตอนที่สามคือ การหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ ซึ่งประกอบด้วย การหาจุดตัดวิกฤต การหาจุดตัดที่มีจุดร่วมน้อยที่สุด และตำแหน่งที่ลึกที่สุด หากพบว่าภาพอักษรบางส่วนขาดหายไปจากการตัด จำเป็นต้องการชดเชยอักษรภาพให้สมบูรณ์โดยใช้การคำนวณขอบเขตและชดเชยส่วนที่หายไปหลังการตัด หรือการเพิ่มพื้นที่อักษรเพื่อชดเชยอักษรหลังการตัดตัวอักษร เพื่อเพิ่มความสมบูรณ์ให้กับตัวอักษร งานวิจัยนี้สามารถแบ่งกลุ่มการติดกันได้ทั้งหมด 7 กลุ่ม และในแต่ละกลุ่มก็ใช้วิธีในการตัดแบ่งแตกต่างกันออกไป ขึ้นอยู่กับรูปแบบการติดกัน ผลที่ได้จากกระบวนการนี้ จะนำตัวอักษรที่ตัดแยกออกมาได้ ส่งให้กับกระบวนการรู้จำตัวอักษรต่อไป ซึ่งวิธีการรู้จำจะมีการคำนวณอย่างไรก็ได้ แต่จะทำให้กระบวนการรู้จำทำงานได้อย่างมีประสิทธิภาพและถูกต้องมากขึ้นเมื่อเทียบกับไม่มีการตรวจจับและตัดแยกตัวอักษรที่ติดกันก่อน

6.2 แนวทางการพัฒนาต่อไปในอนาคต

1. ทำการสำรวจข้อมูลและรวบรวมให้มากขึ้นทั้งในรูปแบบและขนาดของตัวอักษรต่างๆ
2. ปรับปรุงและเพิ่มวิธีการที่สามารถใช้กับตัวหนังสือภาษาอังกฤษได้ด้วย เพื่อเพิ่มประสิทธิภาพมากยิ่งขึ้น
3. นำวิธีการเหล่านี้ไปใช้ร่วมกับกระบวนการรู้จำตัวอักษรไทย โดยเป็นกระบวนการจัดการล่วงหน้า (Preprocessing) ก่อนส่งเข้าสู่กระบวนการรู้จำ ซึ่งจะทำให้ผลของความถูกต้องมีมากยิ่งขึ้น
4. กรณีตัวอักษร 1 ตัว แต่มีหลายส่วนประกอบ เช่น แ ฐ ญ ะ ได้เสนอแนะวิธีการรวมกันเป็นอักษรเดี่ยว ยกตัวอย่าง เช่น ะ แต่ละส่วนประกอบจะมีความสูงน้อยกว่าค่าความสูง

ของตัวอักษรในระดับกลางค่อนข้างมาก ดังนั้นเราสามารถรู้ได้ทันทีว่าเป็น ะ เราก็จะทำการรวมให้เป็นตัวอักษรตัวเดียว

5. ส่วนที่ยังตรวจสอบผิดพลาดหรือตัดแยกผิดพลาด จะนำไปพิจารณาเพื่อหาทฤษฎี หรือวิธีการแก้ปัญหาเพิ่มเติม เพื่อให้การตรวจสอบและตัดแยกถูกต้องมากที่สุด

เอกสารอ้างอิง

[1] วิชา พานิช. 2539. "ระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย" วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย.

[2] ศุภกร รัตนปราการ. 2542. "การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย" วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[3] จรรยา เกียรติศิริอนันต์. 2543. "การแยกตัวอักษรภาษาไทยที่ติดกัน" วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[4] จักริน สุขสวัสดิ์ชน. 2543. "การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่สัมผัสกัน" วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[5] วรวิทย์ เปรมรัตนชัย. 2543. "การแยกตัวพิมพ์อักษรภาษาไทยที่ซ้อนทับกัน" วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[6] ดวงเพ็ญ เจตน์พิพัฒนพงษ์. 2545. "การวิเคราะห์และตัดแยกตัวอักษรลายมือภาษาไทยโดยใช้ข้อมูลสำคัญของตัวอักษร" วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์, ภาควิชาวิศวกรรมคอมพิวเตอร์, มหาวิทยาลัยเกษตรศาสตร์.

[7] S. Kahan, T. Pavlidis and H. S. Baird. 1987. "On the Recognition of Printed Characters of Any Font and Size", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 2, pp. 274-287.

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

ภาคผนวก

บทความที่ตีพิมพ์ในการประชุมวิชาการวิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 10 (NCSEC 2006)

P. Pongpo, P. Lalitrojwong. 2006. "Analysis and segmentation of touching and and crossing Thai printed characters", *Proceedings of the 10th National Computer Science and Engineering Conference (NCSEC 2006)*, Khon Kaen, October 25-27, pp.296-303.

NCSEC 2006

The 10th National Computer Science and Engineering Conference, NCSEC 2006

การประชุมวิชาการ วิทยาการคอมพิวเตอร์และวิศวกรรมคอมพิวเตอร์แห่งชาติ ครั้งที่ 10

The 10th National Computer Science and Engineering Conference

25-27 ตุลาคม 2549

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
และภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยขอนแก่น

จัดโดย

ร่วมกับ

SIRDC

E-SAAN SOFTWARE PARK

สนับสนุนโดย

NECTEC a member of NSTDA

REUTERS

Sun microsystems We make the net work.

IRCP group

TOT

SYSTEM INTEGRATOR C.S.I.

MFEC Public Company Limited

Website: http://ncsec2006.kku.ac.th

» Contents

P77 การพยากรณ์ความต้องการใช้ไฟฟ้าระยะสั้นโดยใช้การแปลงเวฟเลทและการตัดเล็มโครงข่ายประสาทเทียม. 239
(Short -Term Load Forecasting by Using Wavelet Transform and Pruned Neural Networks)
กวีพจน์ บันลือวงศ์ สิรภัทร เชี่ยวชาญวัฒนา และ คำรณ สุนัติ

P78 การตัดคำภาษาไทยโดยใช้กฎและตัวแบบฮิดเดนมาร์คอฟ 247
(Thai Word Segmentation Using Rule Based and Hidden Markov Model)
ฐาปนี เฮงสนั่นกูล และ พุธษดี ศิริแสงตระกูล

P83 Neural Network Based Greedy Job Scheduler 257
Kothalil Gopalakrishnan Anilkumar and Thitipong Tanprasert

P86 ผลกระทบของความล่าช้าในการแสดงผล และการใช้ข้อมูลป้อนกลับต่อประสิทธิภาพของการใช้ฐานข้อมูลผ่านเวปไซด์ 263
(Impacts of Delay in Output Display and Use of Feedback on Efficiency of Web Database Usage)
ชัชพงศ์ ตั้งมณี

P87 ระบบการจัดการความรู้ของกองทัพอากาศโดยใช้เทคนิคเครือข่ายประสาทเทียม 273
(The Royal Thai Air Force Knowledge Management System by using Neural Network)
ธนพันธุ์ หร่ายเจริญ วีรนันท์ หาญสวธา และ อัญชลี อนันตศิริ

P88 การระบุตำแหน่งและสร้างแผนที่ด้วยการทำงานร่วมกันของหุ่นยนต์สองตัวโดยอาศัยการตรวจจับเส้นขอบแนวดิ่งด้วยกล้องออมนิ 280
Cooperative Localization and Mapping using the Vertical Edge Detection from an Omni-Directional Camera
Mahisorn Wongphati and Attawith Sudsang

P90 การตรวจจับแพ็คเก็ตรูปแบบผิดปกติโดยใช้ข่ายงานเบย์ 288
(Packet Header Anomaly Detection Using Bayesian Network)
มงคล ทะกอง และ ศาสตรา วงศ์ธนวสุ

P112 การวิเคราะห์และตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน 296
(Analysis and Segmentation of Touching and Crossing Thai Printed Characters)
ภิญโญ พงษ์โพธิ์ และ ภัทรชัย ลลิตโรจน์วงศ์

P118 การคำนวณรูปแบบการวางตัวของกลุ่มหุ่นยนต์จากข้อมูลเชิงมุม 304
(Computing Positional Pattern of Multiple Robots from Angular Measures)
สรวิชญ์ ธนพานิชกูล ฌฤทธิ์ บุญให้เจริญ และ อรรถวิทย์ สุดแสง

P139 การตรวจสอบกราฟถอดแบบด้วยวิธีทางพันธุกรรมและการเทียบเซ็กเมนท์ 312
(Genetic Algorithm & Segmentation to Discover Graph Isomorphism)
สิริมนัส เอกวศิน เสรี ชิโนดม และ ประวิทย์ บุญมี

การวิเคราะห์และตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน

Analysis and Segmentation of Touching and Crossing Thai Printed Characters

ภิญโญ พงษ์โพธิ์ ภัทรชัย ลลิตโรจน์วงศ์

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ถนนฉลองกรุง เขตลาดกระบัง กรุงเทพฯ 10520

Email: pinyo_pp@hotmail.com, pattarachai@it.kmitl.ac.th

บทคัดย่อ

การเพิ่มความถูกต้องของระบบรู้จำอักษรไทย จำเป็นต้องมีระบบประมวลผลก่อนการรู้จำ ซึ่งภาพอักษรที่ติดกันมีผลกระทบต่อความถูกต้องของระบบรู้จำ ดังนั้น งานวิจัยนี้จึงมุ่งเน้นในการออกแบบและพัฒนาโปรแกรมสำหรับตัดภาพอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน กระบวนการตัดภาพอักษรที่ติดกันประกอบด้วย 3 ขั้นตอน ได้แก่ การสกัดภาพและจำแนกภาพอักษร การวิเคราะห์การติดกันของตัวอักษร และการหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ ในการดึงภาพอักษร จะประกอบด้วยการฉายแนวนอน และการติดตามจุดดำ เพื่อแยกตัวอักษรออกจากเอกสาร และใช้ระดับอักษรเพื่อจำแนกภาพอักษร ในการวิเคราะห์การติดกันของตัวอักษร จะใช้ข้อมูลภาพอักษร ได้แก่ ลักษณะเส้นอักษร อัตราส่วนความกว้างและความสูงของรูปภาพ สำหรับการหาตำแหน่งจุดตัด จะคำนวณหาจุดตัดวิกฤต หรือจุดตัดที่มีจุดร่วมน้อยสุด เพื่อตัดภาพอักษร ในกรณีที่การตัดภาพอักษรติดทำให้ภาพอักษรไม่สมบูรณ์ จำเป็นต้องมีการปรับแต่งภาพอักษรให้สมบูรณ์มากยิ่งขึ้น

คำสำคัญ: การฉายแนวนอน, จุดตัดวิกฤต, จุดร่วมน้อยสุด, การตัดแยกตัวอักษรไทย

Abstract

The accuracy of Thai OCR needs Thai Pre-OCR. In Thai language, an image of connected characters has a significant effect on the accuracy of Thai OCR. Problems of previous studies derive from the inability of recognition of connected characters in the image. Thus, this research aims at developing an approach to segment cross-touch connected character images. The segmentation of cross-touch connected character images consists of three steps: detecting and grouping character images, analyzing connected characters, and locating the connected character cutting and character compensation. Detecting and grouping character images requires 1) horizontal projection and black point connected; 2) extracting each character from the original document; and 3) the leveling of each character to identify its type. The analysis of connected characters uses the characteristics of Thai characters and their Width-Height ratio. To locate the cutting point of each character, either peak to valley or break cost techniques is used. If cutting causes a character defect due to overlapping of connected characters, the compensation technique will be used to restore and to refine them. Two types of compensation technique are the identification of connected character overlapping area and compensation area; and the area enlargement in order to increase the compensation area.

Key Words: horizontal projection, peak to valley, minimum break cost, Thai character segmentation

1. บทนำ

การรู้จำอักษรไทย ภาพอักษรที่ติดกันมีผลกระทบต่อความถูกต้องของระบบรู้จำอักษรภาษาไทย เนื่องจากการรู้จำนั้น ถ้ามีการรู้จำรูปแบบของตัวอักษรที่ผิด ผลที่ได้ก็จะผิดไปด้วย ดังนั้น โครงงานวิจัยนี้จึงมุ่งเน้นในการออกแบบและพัฒนาโปรแกรมสำหรับตัดภาพอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน เพื่อให้ผลที่ได้จากการรู้จำถูกต้องและมีประสิทธิภาพมากยิ่งขึ้น

ปัญหาการติดกันของตัวอักษร มักจะเกิดจากการสแกน และการถ่ายเอกสาร และเนื่องจากลักษณะของตัวอักษรภาษาไทยมีหลายระดับ ทำให้มักจะเกิดปัญหาการทับซ้อนหรือติดกันของตัวอักษรทั้งในแนวนอนและในแนวตั้งด้วย โดยสามารถแก้ไขได้โดยการวิเคราะห์หาจุดตัดของตัวอักษรที่ติดกัน จากนั้นจึงแยกตัวอักษรออกจากกัน สำหรับปัญหานี้ มีนักวิจัยพยายามแก้ปัญหาโดยนำเสนอเทคนิคต่างๆ ซึ่งพอสรุปได้ดังนี้

วิชา [1] ใช้การฉายแนวนอนและแนวตั้ง การหาช่องว่างที่ลึกที่สุด และขอบของตัวอักษรที่อยู่ต่างระดับเพื่อแยกตัวอักษร แต่ยังพบข้อจำกัดคือ ยังไม่สามารถแยกความแตกต่างระหว่าง ฤ กับ ฤ และยังหาจุดตัดผิดเมื่อมีสัญญาณรบกวน

ศุภกร [2] วิเคราะห์หาเส้นระดับบรรทัด และใช้ความกว้างของตัวอักษรและค่าสถิติของจุดดำช่วยในการตัดแยก แต่ยังไม่สามารถตัดอักษรระดับกลางได้ และไม่สามารถแยกอักษรหลายขนาดได้

จรรยา [3] หาเส้นทางที่สั้นที่สุดและค่าสถิติของจุดดำเพื่อใช้แยกตัวอักษร แต่ไม่ครอบคลุมการตัดอักษรระดับบนและล่าง และถ้าตัวอักษรติดกันมาก ก็ยังตัดไม่ถูกต้อง

จักริน [4] ใช้ความกว้างของตัวอักษร จุดตัดวิกฤต และ New Metric ช่วยในการแยกตัวอักษร แต่ข้อจำกัดคือ ตัวอักษร ฤ ฦ ๆ จะถูกตัดไปด้วย และถ้าตัวอักษรติดกันมาก ก็ยังตัดไม่ถูกต้อง

2. วิธีการที่ใช้ในการวิเคราะห์และตัดแยกตัวอักษร

ในงานวิจัยนี้ การวิเคราะห์และตัดแยกตัวอักษรที่ติดกัน ประกอบด้วยกระบวนการ 3 ขั้นตอน คือการแยกภาพอักษรออกจากเอกสาร โดยใช้การฉายแนวนอนเพื่อแยกระดับอักษร และใช้การติดตามจุดดำเพื่อแยกตัวอักษรออกมา กระบวนการต่อมาคือ การตรวจสอบการติดกันของตัวอักษร โดยใช้ความกว้าง ความสูง และลักษณะสำคัญของประเภทระดับอักษร เช่น ลักษณะหัวหรือหางเป็นต้น และกระบวนการสุดท้ายคือ การหาตำแหน่งที่ติดกัน โดยใช้การหาค่าจุดตัดวิกฤตหรือการหาจุดตัดที่มีจุดร่วมน้อยที่สุด แล้วชดเชยตัวอักษรเมื่อบางส่วนของภาพอักษรขาดหายไปในการตัดตัวอักษรแบบทับซ้อน โดยหาจุดที่ลึกที่สุดจากเส้นขอบตัวอักษร กระบวนการที่กล่าวมาทั้งหมดมีใจความสำคัญดังต่อไปนี้

2.1 การแยกภาพอักษรออกจากเอกสาร

การแยกภาพอักษรออกจากเอกสารจะใช้อัลกอริทึมการฉายแนวนอน (Horizontal Projection) [2] เพื่อแยกตัวอักษรจากเอกสารให้เป็นบรรทัด และใช้การติดตามจุดดำแยกย่อยบรรทัดเป็นตัวอักษรเดี่ยว หลังจากที่ได้ตัวอักษรเราจะนำไปใช้ในกระบวนการต่อไป

การฉายแนวนอน เป็นการเก็บค่าสถิติจุดดำโดยใช้ค่าสถิติมาช่วยในการตัดสินใจ แนวคิดคือ จุดที่แบ่งบรรทัดหรือตัวอักษรจะมีค่าสถิติน้อย หรือค่าสถิติเท่ากับศูนย์ในแนวระดับแต่ละแถวตามลำดับ เราสามารถคำนวณจากสูตรดังนี้

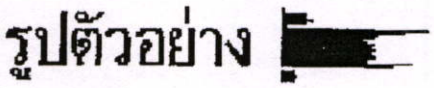
$$P(x,y) = \sum_{X} P(x,y) \quad (1)$$

ซึ่ง P(x,y) เป็นค่าเม็ดสีของจุด ณ ตำแหน่ง (x,y)

P(x) เป็นค่าผลรวมของจำนวนเม็ดสีในแนวแกน X

P(y) เป็นค่าผลรวมของจำนวนเม็ดสีในแนวแกน Y

จากรูปที่ 1 จะเห็นได้ว่า เราสามารถแยกตัวอักษรเป็น 3 ระดับตามแนวนอน โดยดูจากค่าสถิติจากการฉายแนวนอน



รูปที่ 1 ตัวอย่างการใช้การฉายแนวนอน

ส่วนการติดตามจุดดำ คือการดูว่าจุดที่เราพิจารณานั้น จุดที่ติดกับจุดที่เราพิจารณา 8 จุด เป็นจุดขาวหรือจุดดำ ถ้าเป็นจุดดำ ก็จะพิจารณาซ้ำเหมือนกับเริ่มต้นไปเรื่อยๆจนได้จุดดำทั้งหมดที่ประกอบเป็นหนึ่งตัวอักษร วิธีนี้จะแก้ปัญหาการเหลื่อมล้ำกันของตัวอักษรได้

2.2 ส่วนตรวจสอบการติดกันของตัวอักษร

การวิเคราะห์การติดกันของตัวอักษรในแต่ละแบบนั้นต้องใช้วิธีการที่แตกต่างกันไป ซึ่งจะใช้ค่าความกว้างเฉลี่ยของอักษรระดับกลาง (W) และความสูงเฉลี่ยของอักษรระดับบน (H) มาใช้ในการพิจารณา ส่วนรายละเอียดการนำสองค่านี้ไปใช้เพื่อแยกตัวอักษร จะอธิบายในหัวข้อที่ 3 สำหรับการหาค่า W และ H มีรายละเอียดดังนี้

2.2.1 การหาค่าความกว้างเฉลี่ยของตัวอักษรระดับกลาง

เนื่องจากความแตกต่างของความกว้าง ของตัวอักษรไทย เช่น "เ" "ก" และ "ฌ" เป็นต้น ทำให้ไม่สามารถใช้ความกว้างของตัวอักษรตัวใดตัวหนึ่งมาเป็นมาตรฐานในการเปรียบเทียบได้ ดังนั้น งานวิจัยนี้จึงได้ทดลองหาความกว้างเฉลี่ยของตัวอักษรไทยในรูปแบบต่างๆกัน ซึ่งพบว่าอักษร ไทยโดยส่วนใหญ่จะมีความกว้างเป็น 0.8 เท่าของความสูงของตัวอักษรในระดับกลาง จึงใช้ค่านี้เป็นค่าความกว้างเฉลี่ย (W) เพื่อใช้ในการวิเคราะห์การติดกันและตัดแยกตัวอักษรต่อไป

2.2.2 การหาค่าความสูงเฉลี่ยของตัวอักษรระดับบน

ความสูงเฉลี่ยของตัวอักษรในระดับบน คือความสูงของสระและวรรณยุกต์ในระดับบน ซึ่งจากการทดลองพบว่าสระและวรรณยุกต์ในระดับบนจะมีความสูงประมาณ 0.5 เท่าของของความสูงของตัวอักษรในระดับกลาง ในงานวิจัยนี้จึงใช้ค่านี้เป็นค่าความสูงเฉลี่ยของตัวอักษรในระดับบน (H) เพื่อใช้ในการวิเคราะห์การติด กันและตัดแยกตัวอักษรต่อไป

2.3 การหาตำแหน่งที่ติดกันหรือทับซ้อนกัน

2.3.1 การหาจุดตัดวิกฤต

จุดตัดวิกฤต (Peak to Valley) คือจุดเชื่อมของสองอักษรซึ่งจะมีค่าของการฉายแนวตั้ง (Vertical Projection หรือ V(x)) เปลี่ยนแปลงแบบ Sharp Minimum ดังนั้น Kahan, Palvidis และ Baird [6] จึงเสนอให้ใช้อัตราส่วนระหว่างอนุพันธ์อันดับสอง คือ V(x-1)-2V(x)+V(x+1) กับค่าของการฉายเป็นสมการเงื่อนไขในการหาจุดตัด

$$PV(x) = \frac{V(x-1) - 2V(x) + V(x+1)}{V(x)} \quad (2)$$

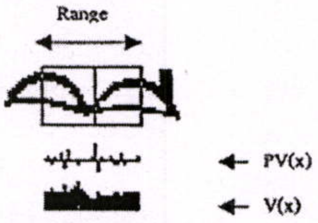
เมื่อ $PV(x)$ เป็นค่าการฉายแนวตั้งของตำแหน่ง x โดยเปรียบเทียบกับตำแหน่ง x+1 และ x-1

$V(x-1)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x-1

$V(x)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x

$V(x+1)$ เป็นค่าการฉายแนวตั้งหรือผลรวมจุดดำของตำแหน่ง x+1

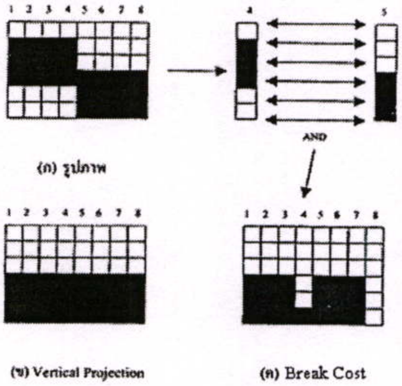
จากสมการค่า $PV(x)$ จะมีค่ามากเมื่อค่าผลรวมจุดดำตามแนวตั้ง $V(x)$ มีค่าน้อยกว่าผลรวมของจุดดำตำแหน่ง x+1 และ x-1 แสดงว่าตำแหน่งไหนที่ได้ค่า $PV(x)$ มาก ก็จะตัด ณ ตำแหน่งนั้น แต่ถ้าตำแหน่งที่ติดกันมีจุดภาพมากอาจจะทำให้ตำแหน่งที่ตัดคลาดเคลื่อนได้



รูปที่ 2 การหาจุดแบ่งโดยใช้จุดตัดวิกฤต

2.3.2 การหาจุดตัดที่มีจุดร่วมน้อยที่สุด

โดยทั่วไป การหาจุดตัดของตัวอักษรที่สัมผัสกันอย่างง่ายๆ จะใช้วิธีการหาค่าการฉายจุดภาพ (Pixel Projection) เพื่อกำหนดตำแหน่งของจุดตัด จากตำแหน่งที่มีค่าโปรเจคชันที่น้อยที่สุด แต่การใช้การหาจุดตัดที่มีจุดร่วมน้อยที่สุด (Minimum Break Cost) [7] ค่าที่คำนวณได้จะบอกถึงนัยสำคัญของการสัมผัสกัน (Degree of Contact) ของแต่ละคอลัมน์ที่ติดกัน ดังตัวอย่างในรูปที่ 3



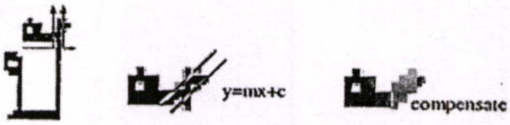
รูปที่ 3 นัยสำคัญของการสัมผัสกัน

จากรูปที่ 3 จะเห็นว่าการหาจุดตัดที่มีจุดร่วมน้อยที่สุด ที่เกิดจากจากการ AND กันของจุดภาพที่อยู่ติดกัน จะแสดงจุดสัมผัสได้อย่างชัดเจน ในขณะที่วิธีการฉายภาพแนวตั้งไม่สามารถแสดงได้ ดังนั้นวิธีการนี้เราสามารถทราบถึงบริเวณที่มีการสัมผัสกันน้อยที่สุด เพื่อกำหนดเป็นตำแหน่งของจุดตัดตัวอักษรที่ติดกันได้

2.4 การปรับปรุงภาพให้สมบูรณ์

2.4.1 การชดเชยขอบเขตส่วนที่หายไป

การชดเชยขอบเขตส่วนที่หายไป (Area Compensation) [5] ใช้สำหรับปรับปรุงภาพหลังการตัดอักษรติดกันระดับบนกับระดับกลางบน เนื่องจากเป็นประเภทที่เกิดการติดกันแบบทับซ้อน หลังการตัด บางส่วนของตัวอักษรจะหายไป เราจำเป็นต้องชดเชยรูปภาพหลังตัด โดยหาจุดตัดอักษร 4 จุดก่อนการตัด หลังจากตัดแยกตัวอักษรตัวแรกแล้ว เราจะหาเส้นขอบของจุดทั้ง 4 จุด ซึ่งเป็นจุดเชื่อมของตัวอักษรที่ 2 โดยอาศัยสมการเส้นตรง แล้วถมสี (Floodfill) ในบริเวณรอบเส้นขอบและตัวอักษรที่เหลือ เพื่อชดเชยส่วนที่ทับซ้อนกัน เช่น ปิ ฝี เป็นต้น ดังรูปที่ 4



รูปที่ 4 วิธีการหาจุดตัดเพื่อชดเชย และการชดเชย

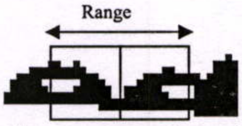
2.4.2 การเพิ่มจุดตัดหลังการหาจุดตัดวิกฤต

การเพิ่มจุดตัดหลังการหาจุดตัดวิกฤต (Compensation of Peak to Valley) [5] เหมาะสำหรับการติดกันของตัวอักษรระดับบนกับระดับบนเหลื่อมกัน ภาพส่วนหน้าจะเกิดการขาดหายไปของเนื้ออักษร และภาพส่วนหลังเกิดการเกินของเนื้ออักษร ดังนั้น การต่อเติมเนื้อที่ขาดหายไปจะทำเฉพาะภาพส่วนหน้าเท่านั้น โดยพิจารณาหาความชันของหางที่ขาด หาความหนาของเนื้อหาง และคำนวณหาความยาวของหางที่ขาดหายไป เช่น เป็นต้น

3. การตัดแยกตัวอักษรติดกันในรูปแบบต่างๆ

จากระดับที่แตกต่างกันของตัวพิมพ์อักษรภาษาไทย เมื่อเกิดการสัมผัสและทับซ้อนกัน จะเกิดได้ในหลายรูปแบบ โดยงานวิจัยนี้เสนอวิธีแก้ปัญหาแบบศึกษาสำนึก (Heuristic Approach) โดยสร้างกฎแบบศึกษาสำนึกสำหรับกรณีต่างๆดังนี้

กลุ่มที่ 1 อักษรระดับบนติดกับอักษรระดับบนในแนวนอน ตัวอย่างเช่น



รูปที่ 5 ตัวอย่างห้ารตัดในกลุ่มที่ 1

ลักษณะที่ตรวจสอบ ถ้าอยู่ระดับบนและมีความกว้างมากกว่า 1.3 เท่าของความกว้างเฉลี่ย แสดงว่าติดกัน

ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. กำหนดขอบเขตที่จะตัดบริเวณกลางภาพ
2. หาค่าจุดตัดวิกฤตสูงสุด และตัด ณ ตำแหน่งนั้น
3. ถ้าส่วนที่ตัดได้มีค่าความกว้างน้อยกว่า 0.35 เท่าของความกว้างเฉลี่ยแสดงว่ามีบางส่วนขาดหายไป

เนื่องจากการทับซ้อน ให้เติมเนื้อตัวอักษร เช่น ~ี

กลุ่มที่ 2 อักษรระดับบนติดกับอักษรระดับบนในแนวตั้ง ตัวอย่างเช่น ้ ี่

รูปที่ 6 ตัวอย่างการตัดในกลุ่มที่ 2

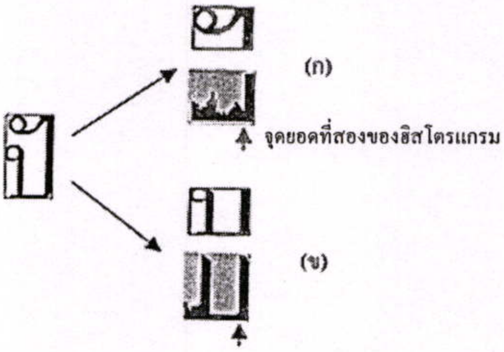
ลักษณะที่ตรวจสอบ ถ้าอยู่ระดับบนและมีความสูงมากกว่า 1.5 เท่าของความสูงตัวอักษรระดับบน แสดงว่าติดกัน ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. หาจุดตัดที่มีจุดร่วมน้อยที่สุดในแนวนอน บริเวณกลางภาพ
2. ตัด ณ ตำแหน่งที่มีจุดร่วมน้อยที่สุด

กลุ่มที่ 3 อักษรระดับกลาง-บนติดกับอักษรระดับบนคอลัมน์เดียวกัน ตัวอย่างเช่น ปั ปี ฝ่ ปิ

ลักษณะที่ตรวจสอบ พิจารณาระดับบน (รูป 7ก) และระดับกลาง (รูป 7ข) ถ้าจุดยอดของแท่งฮิสโตรแกรมที่สองของทั้งสองระดับตรงกัน และตัวอักษรมีลักษณะเปิดด้านบน เช่น ป ฟ ฝ ถ้าไม่ใช่ ก็จะเป็นการติดกันแบบกลุ่มที่ 6 ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. ถ้าความกว้างส่วนบนน้อยกว่า 0.5*W แสดงว่าตัวอักษรที่ติดกันคือ ไม้เอก หรือสัญญาณรบกวน ให้พิจารณาดังนี้
   - 1.1 ตัดตามแนวแกนตั้งระดับบน ด้านในขาหลังของตัวอักษรว่าได้เนื้ออักษรเป็นสองชุดหรือไม่



รูปที่ 7 ตัวอย่างการตัดในกลุ่มที่ 3

   - 1.2 ถ้าได้สองชุด ตรวจสอบว่าเนื้ออักษรที่ได้จากการตัดทางซ้ายมีความกว้างเกิน 0.5 เท่าของความกว้างขาหลังหรือไม่ และความสูงเกิน 0.4 ของความสูงของส่วนบนหรือไม่
   - 1.3 ถ้าใช่ แสดงว่าเป็นการติดกันในลักษณะ ฝ่ ป่ แต่ถ้าไม่ใช่ อาจจะเกิดจากสัญญาณรบกวน
   - 1.4 ถ้าเนื้อตัวอักษรขาดหาย ก็เติมเนื้อจุดดำให้เท่ากับความกว้างของขาหลัง
   - 1.5 ถ้าตัดตามข้อ 1.1 แล้วได้เนื้ออักษรชุดเดียว ให้ตรวจสอบว่าขาหลังของตัวอักษร บริเวณกึ่งกลางมีจุดไหนที่เนื้อตัวอักษรน้อยกว่า 0.4 ของความกว้างของขาหลังหรือไม่
   - 1.6 ถ้ามีก็ตัดตามแนวเส้นแบ่งระดับกลางกับระดับบน ซึ่งจะแก้ปัญหา บ่ ในกลุ่มที่ 6 ได้ ถ้าไม่มีก็หยุด
2. ถ้าความกว้างส่วนบนมากกว่า 0.5*W แสดงว่าไม่ใช่ไม้เอกหรือสัญญาณรบกวน ให้พิจารณาดังนี้
   - 2.1 ถ้าเนื้อหมึกระดับบนเยื้องไปทางซ้ายมากกว่าหรือเท่ากับระดับกลาง เช่น ปิ ให้ตัดตามแนวตั้งด้านในขาหลัง
   - 2.2 ถ้าไม่ใช่ อาจจะเป็น บี ในกลุ่มที่ 6 ให้ลากเส้นตามแนวแกนนอนและแกนตั้งในระดับบน แล้วตรวจสอบว่ามีลักษณะเป็นวงหรือลูปปิดหรือไม่

ถ้ามีตัดตามเส้นแบ่งระดับกลางกับระดับบน

2.3 นอกเหนือจากนี้ให้ตัดตามแนวตั้งด้านในและด้านนอกขาหลัง ถ้าตัวอักษรไม่สมบูรณ์ เช่น ปี ให้เติมเนื้อหมึก โดยการหาความชันจากจุด 4 จุดที่ได้

กลุ่มที่ 4 อักษรระดับกลาง-บนติดกับอักษรระดับบนคอลัมน์ติดกัน

ตัวอย่างเช่น ่ไ ป๊

รูปที่ 8 ตัวอย่างการตัดในกลุ่มที่ 4

<u>แบบที่ 1</u>

ลักษณะที่ตรวจสอบ เป็นอักษรระดับกลางบน มีความกว้างมากกว่า W และความกว้างระดับล่างน้อยกว่า 0.75*W จะติดกันแบบ ่ไ

ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. พิจารณาบริเวณกลางภาพ
2. ตัด ณ ตำแหน่งช่องว่างที่ลึกที่สุด

<u>แบบที่ 2</u>

ลักษณะที่ตรวจสอบ เป็นอักษรระดับกลางบน มีความกว้างมากกว่า W และความกว้างระดับล่างมากกว่า 0.75*W แต่น้อยกว่า 1.2*W จะติดกันแบบ ป๊

ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. หาตำแหน่งช่องว่างที่ลึกที่สุดจากขวาไปซ้าย
2. ตัดตามแนวตั้ง ณ จุดที่ลึกที่สุด

กลุ่มที่ 5 อักษรระดับกลาง-บนติดกับอักษรระดับกลาง-บน

ตัวอย่างเช่น ปไ

รูปที่ 9 ตัวอย่างการตัดในกลุ่มที่ 5

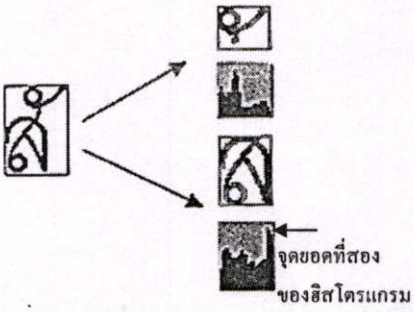
ลักษณะที่ตรวจสอบ เป็นอักษรระดับกลางบน มีความกว้างมากกว่า W และความกว้างระดับล่างมากกว่า 1.2*W

ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. พิจารณาบริเวณกึ่งกลางภาพ
2. ตัด ณ ตำแหน่งช่องว่างที่ลึกที่สุด

กลุ่มที่ 6 อักษรระดับกลางติดกับอักษรระดับบน

ตัวอย่างเช่น สั ศี ด้



รูปที่ 10 ตัวอย่างการตัดในกลุ่มที่ 6

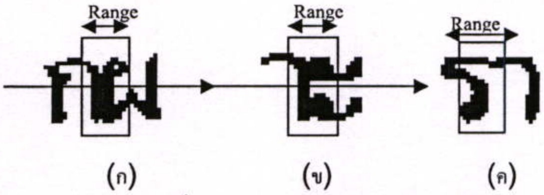
ลักษณะที่ตรวจสอบ พิจารณาระดับบนและระดับกลาง ดังรูปที่ 10 ถ้าจุดยอดที่สองของทั้งสองส่วนไม่ตรงกัน และตัวอักษรมีลักษณะปิดด้านบน เช่น ศ ด ถ้าไม่ใช่ ก็จะเป็นการติดกันแบบกลุ่มที่ 3 ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. หาค่าจุดตัดวิกฤตสูงสุดตามแนวนอนช่วงระดับบน
2. ตรวจสอบว่ามีการตัดผ่านเนื้อตลอดช่วงหรือไม่ เพราะอาจจะตัดโดน ิ ี ได้
3. ถ้าตัดผ่านเนื้อตลอดช่วง ให้ตัดตรงจุดที่มีค่าจุดตัดวิกฤตสูงสุด ถ้าไม่ ให้เลื่อนลงมาเรื่อยๆจนกว่าจะตัดผ่านเนื้อตลอดช่วง แต่ไม่เกินความสูงของระดับกลาง

4. ถ้าเลื่อนมาจนสุดแล้วยังไม่สามารถตัดผ่านเนื้อตลอดช่วงได้ ให้ตัดที่ระดับความสูงของระดับกลาง

กลุ่มที่ 7 อักษรระดับกลางติดกับอักษรระดับกลาง

ตัวอย่างเช่น คา กง



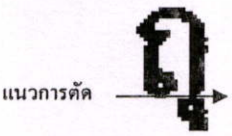
รูปที่ 11 ตัวอย่างการตัดในกลุ่มที่ 7

ลักษณะที่ตรวจสอบ มีความกว้างมากกว่า 1.5*W แสดงว่าติดกัน ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. ตัดแบ่งครึ่งระดับกลางในแนวนอน ตรวจสอบความหนาของตัวอักษรว่าเกิน 0.3*W หรือไม่
2. ถ้าความหนาเกิน แสดงว่าเกิดการติดกันดังรูป (ก) ให้หาพื้นที่บริเวณรอบๆว่ามีเนื้ออักษรเกิน 80% ของพื้นที่ทั้งหมดที่พิจารณาหรือไม่ ถ้าเกิน ให้หาช่องว่างที่ลึกที่สุดบริเวณที่พิจารณาเป็นจุดตัด ถ้าตำแหน่งที่ตัดมากกว่า 0.9*W ให้ตัดได้ แต่ถ้าไม่เกินก็หยุด
3. ถ้าความหนาไม่เกิน 0.3*W แสดงว่าเกิดการติดกันดังรูป (ข) ให้ตรวจดูช่วงบนและช่วงล่างว่ามีเกินช่วงละ 1 หรือเปล่า ถ้ามี ให้หาพื้นที่บริเวณรอบๆว่ามีเนื้ออักษรเกิน 60% หรือไม่ ถ้าเกินให้หาช่องว่างที่ลึกที่สุดบริเวณนั้นเป็นจุดตัด ถ้าระยะที่ตัดมากกว่า 0.9*W ตัด ถ้าไม่เกิน หยุด
4. นอกเหนือจากนั้น จะเป็นดังรูป (ค)ให้หาค่าจุดตัดวิกฤตสูงสุด และตัด ณ ตำแหน่งนั้น

กลุ่มที่ 8 อักษรระดับกลางติดกับอักษรระดับล่าง

ตัวอย่าง เช่น สุ ฐ



รูปที่ 12 ตัวอย่างการตัดในกลุ่มที่ 8

ขั้นตอนการตัดแยกตัวอักษรมีดังนี้

1. ความสูงระดับล่างมากกว่า 0.2 เท่าของความสูงระดับกลาง ความกว้างระดับล่างน้อยกว่า 0.8 เท่าของความกว้างระดับกลาง และการตัดผ่านในแนวนอนต้องมีส่วนที่ไม่ตัดผ่านเนื้อตัวอักษรตลอด เพราะตัวอักษร ฤ ฦ ๆ จะได้ไม่โดนตัดไปด้วย
2. ตัดตามเส้นแบ่งระหว่างระดับกลางกับระดับล่าง

4. ผลการทดลอง

ผู้วิจัยได้พัฒนาระบบด้วยโปรแกรม C++ Builder5 ตามหลักการที่เสนอ ผลการทดลองจากการใช้วิธีวิเคราะห์การตัดกันในแต่ละกลุ่มทั้ง 8 กลุ่ม กับรูปภาพประโยคที่ได้จากวารสาร และสิ่งพิมพ์ต่างๆ ผ่านการสแกนด้วยความละเอียด 300 dpi ได้ผลดังตารางที่ 1

ตารางที่ 1 ข้อมูลที่ได้จากการสแกนวารสาร สิ่งพิมพ์ต่างๆจำนวน 20,000 ตัดอักษร ที่ความละเอียด 300 dpi

กลุ่มที่	จำนวนตัวอักษรที่สัมผัสกัน	จำนวนตัวอักษรที่ตัดแยกได้ถูกต้อง	% ความถูกต้อง
1	12	12	100
2	180	180	100
3	44	42	95.45
4	40	39	97.5
5	8	8	100
6	20	20	100
7	340	332	97.65
8	36	36	100
รวม	680	669	98.38

จากผลการทดลอง ยังมีข้อผิดพลาดเช่นกัน เนื่องจากมีสัญญาณรบกวน เกิดจากตัวอักษรที่ไม่สมบูรณ์ หรือเกิดการติดกันของตัวอักษรที่มีขนาดเล็กกว่าความกว้างเฉลี่ย เช่น ' ไ ทำให้คำนวณค่าผิดพลาดและตัดแยกได้ไม่ถูกต้อง

5. สรุป

งานวิจัยนี้ได้นำเสนอวิธีการวิเคราะห์และตัดแยกอักษรพิมพ์ภาษาไทยแบบสัมผัสและทับซ้อนกัน ประกอบด้วย 3 ขั้นตอน ได้แก่ การสกัดภาพและจำแนกภาพอักษร การวิเคราะห์การติดกันของตัวอักษร และการหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ สามารถแบ่งกลุ่มการติดกันได้ทั้งหมด 8 กลุ่ม ซึ่งวิธีการที่ใช้ในการตัดแยกในแต่ละกลุ่มนั้น สามารถใช้ได้กับตัวอักษรพิมพ์ไทยได้เป็นอย่างดี แต่ก็ยังมีข้อผิดพลาดเกิดขึ้น และจะพัฒนาให้ถูกต้องมากยิ่งขึ้นต่อไปในอนาคต

6. เอกสารอ้างอิง

[1] วิชา พานิช. 2539. “ระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย” วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต, จุฬาลงกรณ์มหาวิทยาลัย.

[2] ศุภกร รัตนปราการ. 2542. “การแยกสระและวรรณยุกต์ระดับบนล่างออกจากสายอักขระตัวพิมพ์ไทย” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[3] จรรยา เกียรติศิริอนันต์. 2543. “การแยกตัวอักษรภาษาไทยที่ติดกัน” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[4] จักริน สุขสวัสดิ์ชน. 2543. “การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่สัมผัสกัน” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[5] วรวิทย์ เปรมรัตนชัย. 2543. “การแยกตัวพิมพ์อักษรภาษาไทยที่ซ้อนทับกัน” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[6] S. Kahan, T. Pavlidis and H. S. Baird. 1987. “On the Recognition of Printed Characters of Any Font and Size”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 9, No 2, pp 274-287.

[7] Y. Lu. 1995. “Machine Printed Character Segmentation: An Overview”, *Pattern Recognition*, Vol 28, No 1, pp 67-80.

ประวัติผู้เขียน

ชื่อ-นามสกุล นายภิญโญ พงษ์โพธิ์

วัน เดือน ปีเกิด 12 มิถุนายน พ.ศ. 2518

ที่อยู่ 92 หมู่ 1 ต.ดอนยายหอม ถ.เศรษฐกิจ2 อ.เมือง จ.นครปฐม 73000

ประวัติการศึกษา 2541 วิศวกรรมศาสตรบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์