การเชื่อมต่อลายเส้นอักษรตัวพิมพ์ไทยที่ขาดหายในแนวตั้ง

CONNECTION OF VERTICAL BROKEN LINE OF THAI PRINTED CHARACTERS

อุบลรัตน์ พาชิยานุกูล

UBOLRAT PACHIYANUKUL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2546

ISBN 974-324-448-4

การเชื่อมต่อลายเส้นอักษรตัวพิมพ์ไทยที่ขาดหายในแนวตั้ง

CONNECTION OF VERTICAL BROKEN LINE OF THAI PRINTED CHARACTERS

อุบลรัตน์ พาชิยานุกูล

UBOLRAT PACHIYANUKUL

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2546

ISBN 974-324-448-4

CONNECTION OF VERTICAL BROKEN LINE OF THAI PRINTED CHARACTERS

UBOLRAT PACHIYANUKUL

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

2003

ISBN 974-324-448-4

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การเชื่อมต่อลายเส้นอักษรตัวพิมพ์ไทยที่ขาดหายในแนวตั้ง

CONNECTION OF VERTICAL BROKEN LINE OF THAI PRINTED CHARACTERS

ชื่อนักศึกษา นางสาวอุบลรัตน์ พาชิยานุกูล

รหัสประจำตัว 42067022

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ผู้ควบคุมวิทยานิพนธ์ รศ.ดร.วิเชียร เปรมชัยสวัสดิ์

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.วิเชียร	เปรมชัยสวัสดิ์	
รศ.ดร.บุญวัฒน์	อัตชู	
ผศ.ดร.ประจวบ	วานิชชัชวาล	
ผศ.ดร.วรพจน์	กรีสุระเดช	
ผศ.ดร.อาริต	ธรรมโน	

วัน/เดือน/ปี ที่สอบ 21 พฤษภาคม 2546 เวลา 10.30 น. เป็นต้นไป

สถานที่สอบ ณ ห้อง M03 (ชั้นลอย) อาคารเรียนรวมและปฏิบัติการคณะเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.บุญวัฒน์ อัตชู)

คณบดีบัณฑิตวิทยาลัย

วันที่.....29.....**เดือน**.....พฤษภาคม.....**พ.ศ**.....2546.....

หัวข้อวิทยานิพนธ์	การเชื่อมต่อลายเส้นอักษรตัวพิมพ์ไทยที่ขาดหายในแนวตั้ง
ชื่อนักศึกษา	นางสาว อุบลรัตน์ พาชิยานุกูล
รหัสประจำตัว	42067022
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2546
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ. ดร. วิเชียร เปรมชัยสวัสดิ์

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอวิธีการใหม่เพื่อทำการซ่อมแซมอักษรตัวพิมพ์ไทยที่ขาดในแนวตั้ง ซึ่งเป็นการเตรียมตัวอักษรที่มีโครงสร้างสมบูรณ์ก่อนที่จะเข้าสู่กระบวนการรู้จำตัวอักษร กระบวนการซ่อมแซมตัวอักษรขาด แบ่งออกเป็น 2 ขั้นตอนหลัก คือ ขั้นตอนการตรวจสอบตัวอักษรขาดและการซ่อมแซมตัวอักษร ขั้นตอนการตรวจสอบตัวอักษรขาดพิจารณาจาก 2 ลักษณะ คือ พิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x และพิจารณาจากความสัมพันธ์ของลักษณะเด่นของโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร ลักษณะเด่นที่นำมาพิจารณาร่วม ได้แก่ ตำแหน่งหัวของตัวอักษร ขาของตัวอักษร และจุดปลายของตัวอักษร วิธีการซ่อมแซมจะขึ้นกับลักษณะการขาดของตัวอักษรทั้งสองแบบ ตำแหน่งของการเชื่อมต่อหาได้จากจุดปลายของภาพตัวอักษร

Thesis Title	Connection of Vertical Broken Line of Thai Printed Characters
Student	Miss Ubolrat Pachiyanukul
Student ID.	42067022
Degree	Master of Science
Programme	Information Technology
Year	2003
Thesis Advisor	Assoc. Prof. Dr. Wichian Premchaiswadi

ABSTRACT

This thesis presents a new scheme for repairing vertically broken Thai font characters. That is to prepare a perfect font structure before passing it to the character recognition process. There are two steps in the proposed scheme: determining the broken characters and repairing them. There are two methods to determine broken characters: to consider from overlapping area of the image block in x-axis and using specific characteristics of the character with character structure code. There characteristic includes: position of the head, the leg and the broken location of the character. The method we use to fix broken characters depends on how they are broken. The location where we join the lines can be found at the end points of the broken images.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยคำแนะนำและคำปรึกษาจาก รศ. ดร.วิเชียร เปรมชัยสวัสดิ์ ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่าน และขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณอาจารย์พงษ์สุรีย์ ลิ้มมณีวิจิตร อาจารย์ประจำคณะวิทยาการสารสนเทศศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร ที่ได้อธิบายและให้คำปรึกษาในงานวิจัยการเชื่อมต่อลายเส้นที่ขาดหายไปในแนวนอนของตัวอักษรภาษาไทย รวมทั้งให้ความรู้ต่าง ๆ ที่เป็นประโยชน์ต่องานวิจัย

ขอขอบคุณอาจารย์สุรการ ดวงผาสุข อาจารย์ประจำคณะวิทยาการสารสนเทศศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร ที่ได้ให้ข้อมูลที่เป็นประโยชน์ต่อการทำงานวิจัย

ขอขอบคุณเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศทุกท่าน ที่ได้ช่วยดำเนินงานในการจัดการสอบและให้ข้อมูลที่เป็นประโยชน์ในการเตรียมตัวในการเสนอวิทยานิพนธ์

ขอขอบคุณเพื่อน ๆ (IS 7) ทุกคนที่ให้กำลังใจและให้คำแนะนำที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

อุบลรัตน์ พาชิยานุกูล

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VI

สารบัญรูป VII

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของงานวิจัย 2

1.3 ทฤษฎีและหลักการที่เกี่ยวข้อง 2

1.3.1 นิยามตัวอักษรขาด 2

1.3.2 กระบวนการซ่อมแซมตัวอักษรขาด 3

1.4 แผนการดำเนินงาน 11

บทที่ 2 การประมวลผลภาพเบื้องต้น 12

2.1 กระบวนการรู้จำตัวอักษร 12

2.2 การประมวลผลภาพ 13

2.2.1การแยกบรรทัด 13

2.2.2 การหาระดับของตัวอักษร 14

2.2.3 การหาตำแหน่งและขนาดของภาพตัวอักษร 16

บทที่ 3 การวิเคราะห์ตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด 18

3.1 นิยามตัวอักษรขาด 18

3.2 การพิจารณาตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x 20

3.3 การพิจารณาตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นโครงสร้าง
ตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร 21

3.3.1 การหาตำแหน่งหัวของตัวอักษร 21

3.3.2 การหารหัสแทนโครงสร้างตัวอักษร 21

สารบัญ (ต่อ)

หน้า

3.3.3 การวิเคราะห์รหัสแทนโครงสร้างตัวอักษร 22

บทที่ 4 การซ่อมแซมตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด 27

4.1 การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x 27

4.1.1 การจับคู่ภาพกรอบตัวอักษร 27

4.1.2. การพิจารณาหาตำแหน่งเชื่อมต่อ 27

4.1.3 การจับคู่จุดปลายและลากเส้นเชื่อมจุด 31

4.2 การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร 33

4.2.1 การจับคู่ภาพกรอบตัวอักษร 33

4.2.2. การพิจารณาหาตำแหน่งเชื่อมต่อ 36

4.2.3 การจับคู่จุดปลายและลากเส้นเชื่อมจุด 36

บทที่ 5 ผลการทดลอง 37

5.1 การทดลองหารหัสแทนโครงสร้างตัวอักษรปกติ 37

5.2 การทดลองหารหัสแทนโครงสร้างตัวอักษรที่ขาด 37

5.3 การทดลองซ่อมตัวอักษรขาด 38

บทที่ 6 สรุปผลการวิจัยและข้อเสนอแนะ 42

บรรณานุกรม 45

ภาคผนวก ก บทความที่ได้รับการตีพิมพ์ในการประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 25 (EECON-25) 47

ประวัติผู้เขียน 53

สารบัญตาราง

ตารางที่ หน้า

1.1 ผลการทดสอบด้วยโปรแกรม OCR ในการรู้จำตัวอักษรขาด 2

3.1 รหัสแทนโครงสร้างตัวอักษร มี 16 รูปแบบ 23

5.1 ตัวอย่างตัวอักษรทั้ง 7 ชนิด 37

5.2 ผลการทดสอบด้วยโปรแกรม ThaiOCR และ ArnThai ในการรู้จำตัวอักษร 39

5.3 ผลการทดสอบด้วยโปรแกรม OCR ในการรู้จำตัวอักษรขาดที่ซ่อมแซมแล้ว 40

5.4 สถิติการพบตัวอักษรขาดในเอกสารประเภทต่าง ๆ 41

สารบัญรูป

รูปที่ หน้า

1.1 ข้อมูลภาพตัวอักษรขาด 1

1.2 ตัวอย่างข้อมูลภาพที่ใช้ทดสอบด้วยซอฟต์แวร์รู้จำตัวอักษร 1

1.3 แสดงกระบวนการซ่อมแซมตัวอักษรขาด 3

1.4 การแยกบรรทัดและตัวอักษรโดยใช้ฮิสโตแกรม และการหาขอบภาพ 4

1.5 การพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x 5

1.6 การพิจารณาการเหลื่อมล้ำของกรอบภาพในแนวแกน x 6

1.7 ตำแหน่งหัวของตัวอักษร 6

1.8 การสแกนจากจุดสแกน 4 ทิศทาง 7

1.9 วิธีหาจุดสแกนของกรอบภาพและหารหัสแทนโครงสร้างตัวอักษร 8

1.10 ขอบเขตของตัวอักษรและส่วนที่เหลื่อมล้ำ 9

1.11 การเชื่อมต่อจุดขาดกรณีเหลื่อมล้ำในแนวแกน x 9

1.12 การรวมกรอบภาพที่เหลื่อมล้ำแต่ไม่มีการซ้อนทับกัน 10

1.13 การเชื่อมต่อจุดขาดกรณีพิจารณาจากความสัมพันธ์ของ
ลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร 10

2.1 กระบวนการรู้จำตัวอักษร 12

2.2 กระบวนการประมวลผลภาพเบื้องต้น 13

2.3 การแบ่งบรรทัดโดยการวิเคราะห์ค่าฮีสโตแกรมในแนวนอน 14

2.4 การแบ่งระดับของตัวอักษรทั้ง 3 ส่วน 15

2.5 ค่าฮีสโตแกรมในแนวนอนเปรียบเทียบ 90 % ของค่าเฉลี่ยฮีสโตแกรมในแต่ละบรรทัด 15

2.6 ฮิสโตแกรมในแนวแกน x 16

2.7 การหา Character Block 17

2.8 Character Frame 17

3.1 ภาพตัวอักษรขาด 18

3.2 Character Frame ที่ได้จากการประมวลผลภาพเบื้องต้น 18

3.3 ขั้นตอนการตรวจสอบตัวอักษรภาษาไทยที่ขาด 19

3.4 กรอบภาพที่ซ้อนทับกัน 20

3.5 กรอบภาพที่ไม่ซ้อนทับกัน 20

3.6 การหาตำแหน่งหัวของตัวอักษร 21

สารบัญรูป(ต่อ)

รูปที่ หน้า

3.7 การหารหัสแทนโครงสร้างตัวอักษร 22

3.8 ตัวอย่างการวิเคราะห์ตัวอักษรขาดด้วยรหัสแทนโครงสร้างตัวอักษร 24

3.9 ตัวอักษร ฒ ขาดจากการวิเคราะห์รหัส “1110” 24

3.10 ตัวอักษรขาดจากการวิเคราะห์รหัส “1011” 24

3.11 ตัวอักษรขาดจากการวิเคราะห์รหัส “1111” 24

3.12 ตัวอักษรขาดจากการวิเคราะห์รหัส “0110” 25

3.13 ตัวอักษรขาดจากการวิเคราะห์รหัส “1001” 25

3.14 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “0111” 25

3.15 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “1101” 25

3.16 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “1101” 25

3.17 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “0011” 26

3.18 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “1000” 26

3.19 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “0010” 26

3.20 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “0100” 26

3.21 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส “0001” 26

4.1 แสดงกรอบภาพที่มีขนาดเล็กของตัวอักษร จ ขาด 27

4.2 การพิจารณาส่วนบนและส่วนล่างของตำแหน่งเชื่อมต่อ 28

4.3 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “1100” ทำให้ได้การเชื่อมต่อบริเวณส่วนล่าง 28

4.4 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “1100” ทำให้ได้การเชื่อมต่อบริเวณส่วนส่วนบน 29

4.5 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “1100” ในกรณีพบหัวในตำแหน่งที่ 1 หรือ 2 29

4.6 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “1101” 29

4.7 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “0110” กรณีพบหัวในตำแหน่งที่ 1 30

4.8 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “0110” กรณีไม่พบหัวในตำแหน่งที่ 1 30

4.9 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “0011” 30

4.10 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส “1011” 30

4.11 ขอบเขตของตัวอักษรและส่วนที่เหลื่อมล้ำ 32

4.12 การเชื่อมต่อจุดขาดบริเวณส่วนบน 32

สารบัญรูป(ต่อ)

รูปที่ หน้า

4.13 การเชื่อมต่อจุดขาดบริเวณส่วนล่าง 32

4.14 การจับคู่จากการวิเคราะห์ด้วยรหัส “1100” 33

4.15 การจับคู่จากการวิเคราะห์ด้วยรหัส “0111” 33

4.16 การจับคู่จากการวิเคราะห์ด้วยรหัส “0110” 33

4.17 การจับคู่จากการวิเคราะห์ด้วยรหัส “1101” 34

4.18 การจับคู่จากการวิเคราะห์ด้วยรหัส “1011” 34

4.19 การจับคู่จากการวิเคราะห์ด้วยรหัส “0011” 34

4.20 การจับคู่จากการวิเคราะห์ด้วยรหัส “1001” 35

4.21 การจับคู่จากการวิเคราะห์ด้วยรหัส “0010” 35

4.22 การจับคู่จากการวิเคราะห์ด้วยรหัส “0010” 35

4.23 การจับคู่จากการวิเคราะห์ด้วยรหัส “1111” 35

4.24 การขาดในกรณีไม่พบจุดปลาย 36

4.25 การลากเส้นเชื่อมต่อกรณีไม่พบจุดปลาย 36

5.1 ตัวอย่างข้อมูลภาพที่ใช้ในการทดสอบ 38

5.2 ตัวอย่างข้อมูลภาพที่ใช้ในการทดสอบหลังการเชื่อมต่อ 39

6.1 ตัวอักษรที่ต้องทำการเชื่อมต่อ 3 จุด 43

6.2 กรณีเชื่อมต่อที่จุดที่ 1 43

6.3 กรณีเชื่อมต่อที่จุดที่ 2 44

บทที่ 1
บทนำ

1.1 ความเป็นมา และความสำคัญของปัญหา

ระบบการรู้จำตัวอักษร ความถูกต้องของการรู้จำจะขึ้นกับข้อมูลภาพตัวอักษร ซึ่งต้องเป็นภาพที่ครบถ้วนและมีโครงสร้างสมบูรณ์ ในงานวิจัยที่ผ่านมากล่าวถึง การแก้ไขตัวอักษรที่ติดกัน [6][7][8] การลดสัญญาณรบกวน [8] การทำขอบตัวอักษรให้เรียบ [9] และการเชื่อมต่อลายเส้นที่ขาดหายไปของอักษรตัวพิมพ์ภาษาไทยในแนวนอน [4] แต่ไม่ได้กล่าวถึงการเชื่อมต่อลายเส้นที่ขาดหายไปของอักษรตัวพิมพ์ภาษาไทยในแนวตั้ง ดังรูปที่ 1.1 ดังนั้นในงานวิจัยนี้จึงมีจุดประสงค์หลักเพื่อทำการปรับปรุงซ่อมแซมข้อมูลภาพตัวอักษรที่ขาดในแนวตั้ง ซึ่งอาจเกิดขึ้นเนื่องจากการสแกน หรือเกิดจากเอกสารคุณภาพต่ำ การพิมพ์ด้วยเครื่องพิมพ์เลเซอร์ เป็นผลให้กระบวนการรู้จำเกิดความผิดพลาดไม่สามารถรู้จำได้ถูกต้อง

รูปที่ 1.1 ข้อมูลภาพตัวอักษรขาด

เมื่อนำตัวอักษรที่ขาดไปทดสอบด้วยซอฟต์แวร์ OCR สำหรับภาษาไทยที่ใช้กันทั่วไปในขณะนี้ คือ ArnThai และ ThaiOCR พบว่าซอฟต์แวร์ดังกล่าวไม่สามารถรู้จำตัวอักษรขาดได้ ตัวอย่างข้อมูลภาพที่ใช้ทดสอบ ดังรูปที่ 1.2 และผลการรู้จำด้วยซอฟต์แวร์ดังกล่าวแสดงดังตารางที่ 1.1

รูปที่ 1.2 ตัวอย่างข้อมูลภาพที่ใช้ทดสอบด้วยซอฟต์แวร์รู้จำตัวอักษร

ตารางที่ 1.1 ผลการทดสอบด้วยโปรแกรม OCR ในการรู้จำตัวอักษรขาด

ข้อความทดสอบ	ซอฟต์แวร์ทดสอบ	
	ThaiOCR	ArnThai
ข้อความทดสอบ	jWr) 1 JY5W ▌ชา ข่	▌jภ El flฑ)า]jTl ไ /LYา e t1
ผู้ที่มีตะกั่วสะสมอยู่สูง	4้ที่3Jตะr ั ะJสะสมPe)▌ูสูะร	▌ที่@Jตะโ▌ศ3สะสมM E)Eiสู\'
คำถวายผ้าป่า	า] ะ3าย 4้▌]	r ▌ }วายฒโ้า\111
ถูกอักขระ	ถู่กe▌f1 ่ระ	[lกE าf)t'ร

งานวิจัยนี้ได้นำเสนอแนวทางเพื่อแก้ปัญหาภาพตัวอักษรขาดอันเป็นสาเหตุหนึ่งที่ทำให้กระบวนการรู้จำตัวอักษรผิดพลาดหรือไม่สามารถรู้จำได้ ซึ่งจะทำให้ระบบการรู้จำตัวอักษรภาษาไทยมีความสมบูรณ์และมีประสิทธิภาพเพิ่มขึ้น

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาวิธีการแยกข้อมูลภาพตัวอักษรออกจากเอกสาร
2. เพื่อศึกษาวิธีการในการระบุภาพตัวอักษรว่าเป็นตัวอักษรขาดหรือไม่
3. เพื่อศึกษาโครงสร้างของตัวอักษรภาษาไทย ในการแบ่งประเภทการขาดของตัวอักษร
4. เพื่อศึกษาแนวทางการซ่อมแซมตัวอักษรขาดในแต่ละประเภท
5. เพื่อเพิ่มประสิทธิภาพการรู้จำในระบบการรู้จำตัวอักษรภาษาไทย

1.3 ทฤษฎีและหลักการที่เกี่ยวข้อง

1.3.1 นิยามตัวอักษรขาด

ตัวอักษรขาด หมายถึง ภาพตัวอักษรที่ถูกแยกออกเป็นชิ้นส่วนภาพย่อย ๆ โดยขอบเขตของงานวิจัยนี้มีดังนี้

- ใช้กับฟอนต์ของตัวอักษรภาษาไทยที่มีหัวไม่ปิดทึบ
- ใช้กับตัวอักษรตัวปรกติ (ตัวตรง)
- เป็นการขาดในแนวตั้ง
- วิเคราะห์เฉพาะตัวอักษรใน Central Zone ซึ่งประกอบด้วยพยัญชนะไทย 44 ตัว และสระ 6 ตัว คือ า, เ, แ, ไ, ใ และ โ
- รูปแบบและตำแหน่งของการขาด คือ ขาดในส่วนที่เป็นเส้นเชื่อมระหว่างขาของตัวอักษร

จากนิยามดังกล่าวทำให้พบลักษณะการขาดของตัวอักษรพิมพ์ภาษาไทยมีลักษณะดังนี้

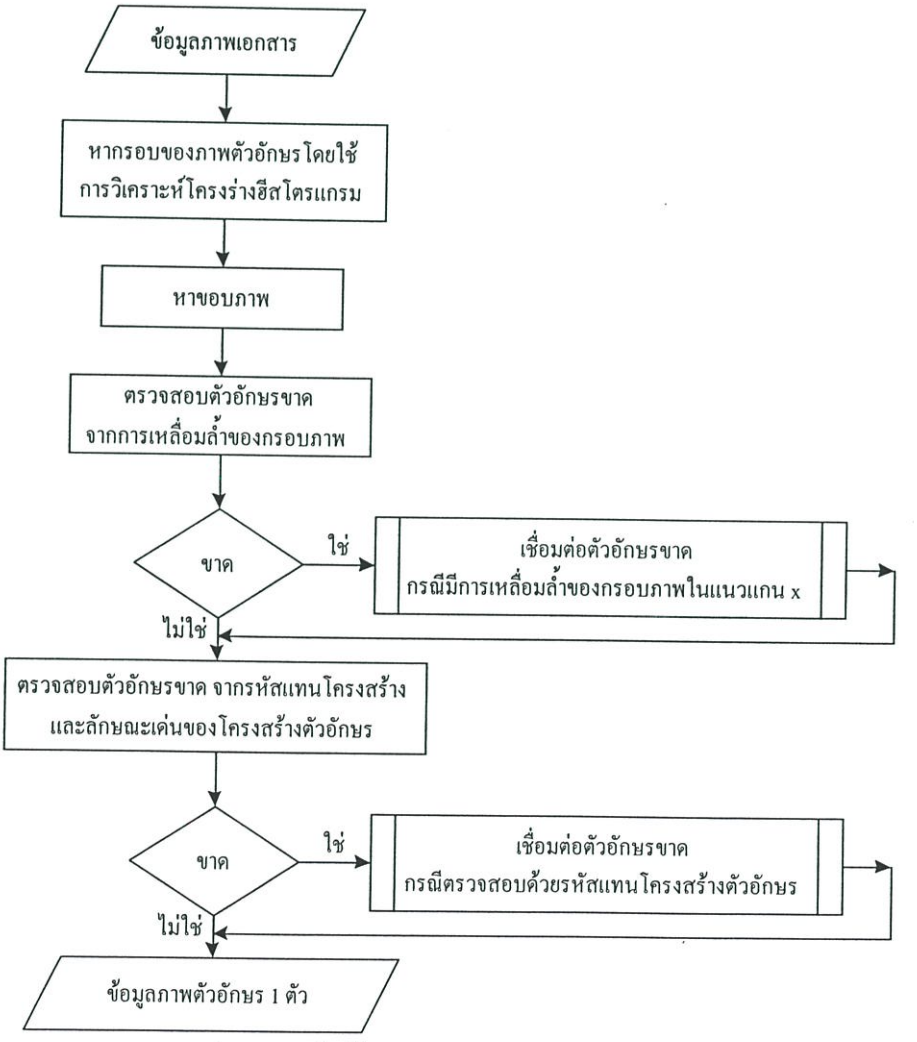
- ตัวอักษรที่มีขาเดียว จะมีลักษณะของกรอบภาพเป็นดังนี้

- ตัวอักษรที่มีสองขา จะมีลักษณะของกรอบภาพเป็นดังนี้

- ตัวอักษรที่มีสามขา จะมีลักษณะกรอบภาพเป็นดังนี้

1.3.2 กระบวนการซ่อมแซมตัวอักษรขาด

กระบวนการซ่อมแซมอักษรตัวอักษรขาดแสดงได้ดังรูปที่ 1.3

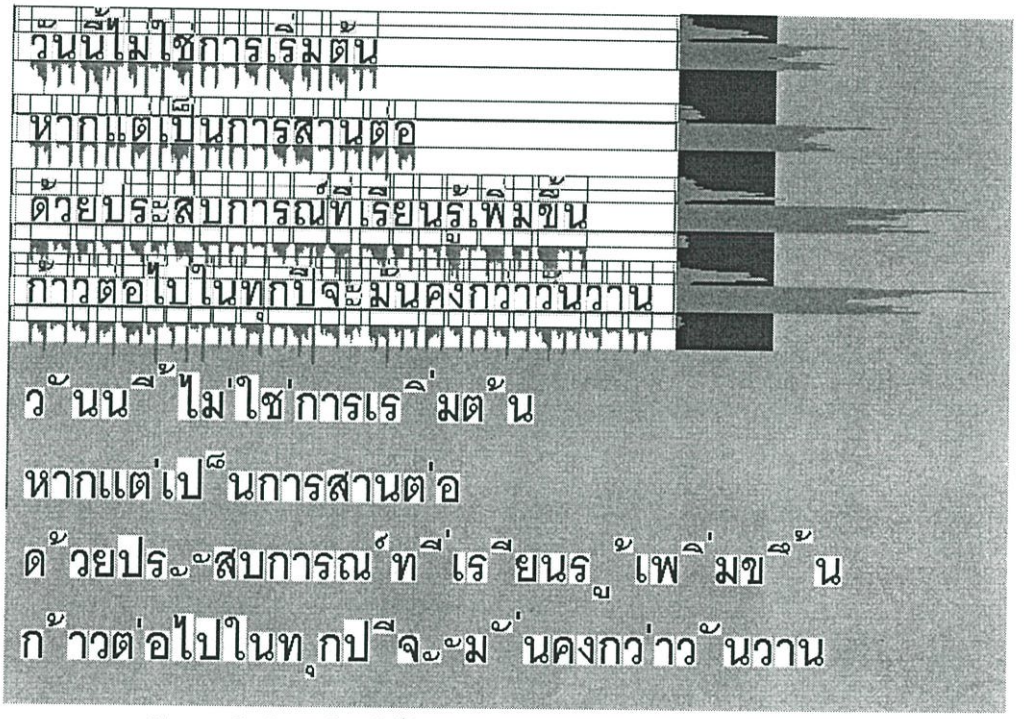


รูปที่ 1.3 แสดงกระบวนการซ่อมแซมตัวอักษรขาด

ทฤษฎีและหลักการที่ใช้ในงานวิจัยจะแบ่งเป็น 3 ส่วนหลัก ๆ คือ

1) การประมวลผลภาพเบื้องต้น

ภาพเอกสารที่เป็นอินพุตของระบบจะประกอบด้วยบรรทัดหลายบรรทัด ดังนั้นในขั้นตอนนี้จะทำการแยกบรรทัดและแยกตัวอักษรแต่ละตัวออกจากบรรทัด โดยใช้ฮิสโตแกรม (Histogram) และการหาขอบภาพ (Contour Algorithm) ซึ่งในขั้นตอนนี้จะได้ตำแหน่งและขนาดของภาพตัวอักษรแต่ละตัว เรียกว่า กรอบตัวอักษร แล้วจัดเรียงลำดับภาพตัวอักษร



รูปที่ 1.4 การแยกบรรทัดและตัวอักษรโดยใช้ฮีสโตแกรม และการหาขอบภาพ

2) การวิเคราะห์การขาดของตัวอักษร

การวิเคราะห์การขาดของตัวอักษรใช้หลักในการพิจารณา 2 หลักการ

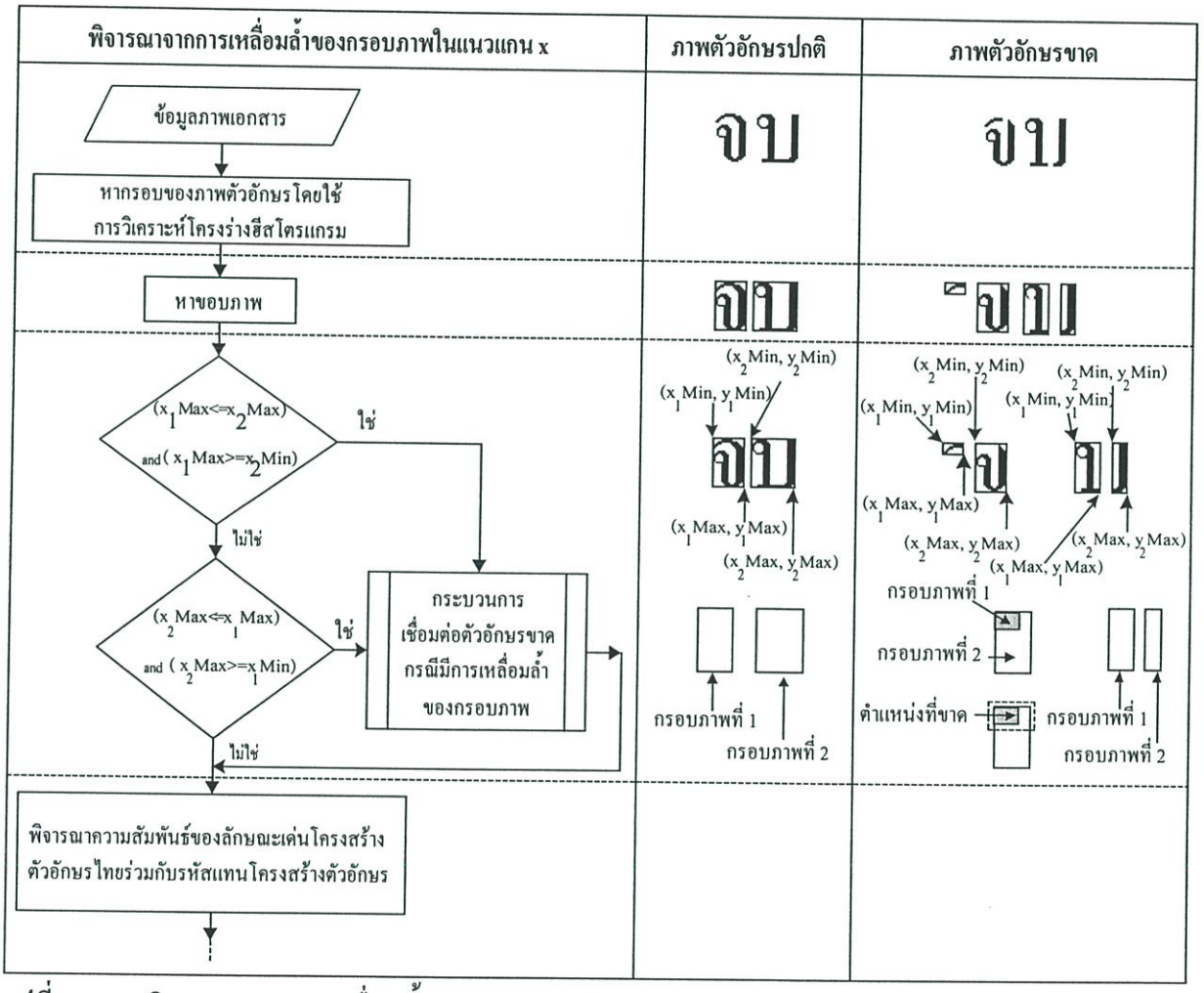
1) พิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

2) พิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

2.1) พิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

ขั้นตอนนี้จะนำกรอบตัวอักษรที่ได้จากการประมวลผลภาพเบื้องต้นมาผ่านกระบวนการหาขอบภาพ หลักเกณฑ์ในการพิจารณาว่าภาพตัวอักษรขาดหรือไม่จะใช้วิธีตรวจสอบค่า $x_i Max$ ของจุดคู่ลำดับของกรอบภาพที่ 1 มีค่าอยู่ระหว่างขอบเขต

ค่า x_2Min และ x_2Max ของอีกกรอบภาพที่ 2 หรือไม่ ถ้าค่า x_1Max ที่พิจารณาอยู่ในขอบเขตของค่า x_2Min และ x_2Max แสดงว่ากรอบภาพที่ 1 มีการเหลื่อมล้ำกรอบภาพที่ 2 ในแนวแกน x นั่นคือ กรอบภาพทั้งสองเป็นกรอบภาพของตัวอักษรเดียวกันและเป็นตัวอักษรขาด แต่ถ้าค่าของ x_1Max ไม่อยู่ในขอบเขตดังกล่าว แสดงว่าไม่มีการเหลื่อมล้ำในแนวแกน x และถือว่าการตรวจสอบในขั้นตอนนี้ไม่พบกรอบภาพขาด



รูปที่ 1.5 การพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

จากรูปที่ 1.5 พบว่าภาพตัวอักษร จ ที่ขาด เมื่อนำมาพิจารณาการเหลื่อมล้ำของกรอบภาพในแนวแกน x จะได้ว่าเงื่อนไข ($x_1Max <= x_2Max$) และ ($x_1Max >= x_2Min$) เป็นจริง นั่นคือ กรอบภาพทั้งสองมีการเหลื่อมล้ำในแนวแกน x และเป็นกรอบภาพตัวอักษรขาด

การพิจารณาการเหลื่อมล้ำของกรอบภาพในแนวแกน x จะมี 2 ลักษณะ ได้แก่

1) กรอบภาพที่เหลื่อมล้ำมีการซ้อนทับกัน ดังรูปที่ 1.6 ก

2) กรอบภาพที่เหลื่อมล้ำไม่มีการซ้อนทับกัน ดังรูปที่ 1.6 ข

ก) กรอบภาพที่มีการซ้อนทับกัน

ข) กรอบภาพที่ไม่มีการซ้อนทับกัน

รูปที่ 1.6 การพิจารณาการเหลื่อมล้ำของกรอบภาพในแนวแกน x

ผลลัพธ์ที่ได้จากขั้นตอนนี้ทำให้ทราบว่ากรอบภาพใดขาดและจะต้องจับคู่กับกรอบภาพใด

2.2) พิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

ขั้นตอนนี้จะนำกรอบภาพตัวอักษรที่ผ่านการตรวจสอบการเหลื่อมล้ำของกรอบภาพในแนวแกน x มาผ่านกระบวนการหาลักษณะเด่นโครงสร้างตัวอักษรไทย ได้แก่ การทำให้บาง (Thinning) การหาตำแหน่งหัวของตัวอักษร การหาจุดปลายของตัวอักษร การหาขาของตัวอักษร จากนั้นจะนำลักษณะเด่นดังกล่าวมาใช้ร่วมกับรหัสแทนโครงสร้างตัวอักษร เพื่อประกอบการตรวจสอบตัวอักษรขาด

ตำแหน่งหัวของตัวอักษร พิจารณาจากการแบ่งพื้นที่กรอบภาพออกเป็น 4 ส่วน ดังรูปที่ 1.7

ก) กำหนดตำแหน่งหัวของตัวอักษร

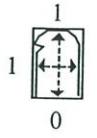
ข) ตัวอักษรที่ไม่มีหัว

ค) ตัวอักษรที่มีหัวในตำแหน่งที่ 1

ง) ตัวอักษรที่มีหัวในตำแหน่งที่ 1 และ 3

รูปที่ 1.7 ตำแหน่งหัวของตัวอักษร

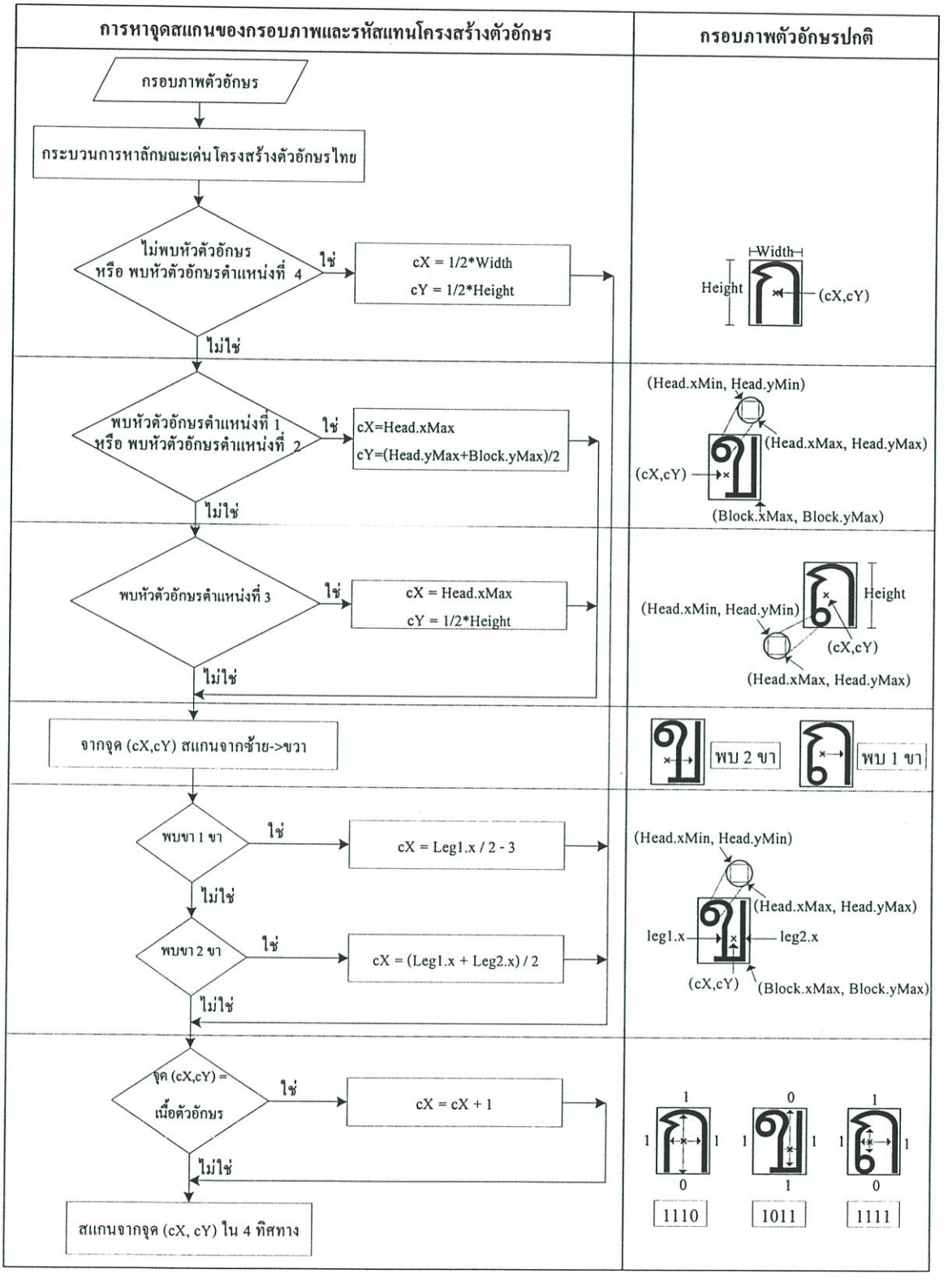
รหัสแทนโครงสร้างตัวอักษร ประกอบด้วยตัวเลข 4 หลัก โดยหาได้จากการนำกรอบตัวอักษร (ที่ผ่านกระบวนการทำให้บาง) 1 กรอบ มาหาจุดสแกนและสแกนจากจุดนี้ 4 ทิศทาง ได้แก่ ซ้าย บน ขวา และล่าง ตามลำดับ ถ้าสแกนแล้วพบจุดที่เป็นโครงสร้างของตัวอักษรจะแทนด้วยเลข 1 และถ้าไม่พบจุดจะแทนด้วยเลข 0 ดังรูปที่ 1.8



รหัสแทนโครงสร้างตัวอักษร คือ 1110

รูปที่ 1.8 การสแกนจากจุดสแกน 4 ทิศทาง

การหาจุดสแกนมีการเปลี่ยนแปลงเลื่อนจุด เพื่อให้ได้รหัสแทนโครงสร้างที่เหมาะสมกับตัวอักษรนั้น ๆ ซึ่งมีวิธีในการเปลี่ยนแปลงจุดสแกน ดังรูปที่ 1.9



รูปที่ 1.9 วิธีหาจุดสแกนของกรอบภาพและหารหัสแทนโครงสร้างตัวอักษร

3) กระบวนการเชื่อมตัวอักษรขาด

ในการเชื่อมต่อจะนำข้อมูลที่ได้จากการวิเคราะห์การขาดของตัวอักษรมาพิจารณาการเชื่อมต่อ ซึ่งจากลักษณะการขาดของตัวอักษรทั้งสองแบบดังที่กล่าวมาในการวิเคราะห์การขาดของตัวอักษร จะทำให้เกิดการเชื่อม 2 แบบเช่นกัน คือ

1) การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

2) การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

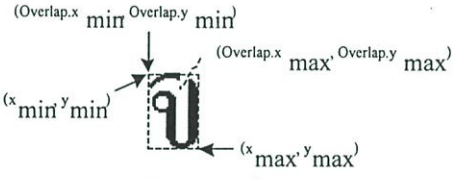
การเชื่อมตัวอักษรขาดจะเริ่มจากการเชื่อมในแบบที่ 1) ก่อน และนำผลลัพธ์ที่ได้ไปทำการเชื่อมต่อในแบบที่ 2) เสมอ

3.1) การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

จากการพิจารณาการเหลื่อมล้ำของกรอบภาพในแนวแกน x จะมี 2 ลักษณะ ได้แก่

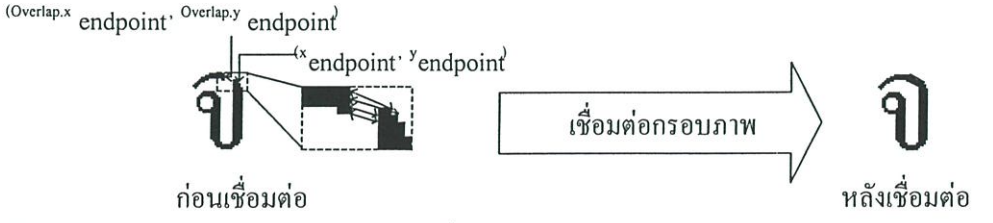
1) กรอบภาพที่เหลื่อมล้ำมีการซ้อนทับกัน

กำหนดพิกัดกรอบตัวอักษรที่ได้จากการประมวลผลภาพเบื้องต้นเป็น $(x_{\min}, y_{\min})$ และ $(x_{\max}, y_{\max})$ ส่วนที่เหลื่อมล้ำที่ได้จากการวิเคราะห์การขาดของตัวอักษรกำหนดเป็น $(Overlap.x_{\min}, Overlap.y_{\min})$ และ $(Overlap.x_{\max}, Overlap.y_{\max})$ ดังรูปที่ 1.10



รูปที่ 1.10 ขอบเขตของตัวอักษรและส่วนที่เหลื่อมล้ำ

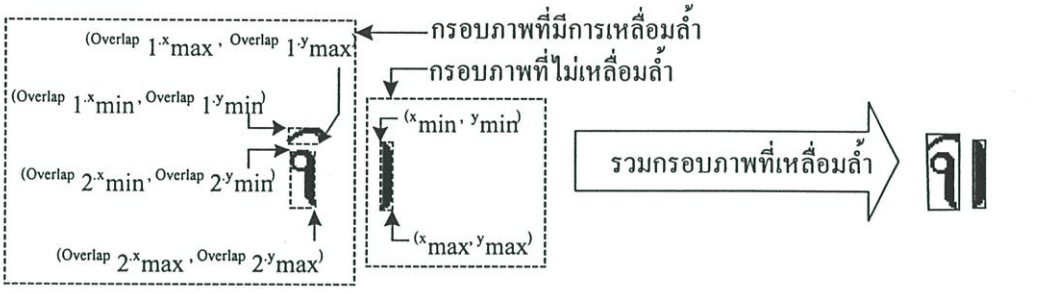
การเชื่อมต่อจะยึดส่วนที่เหลื่อมล้ำเป็นหลัก คือ เริ่มจากจุดปลายของส่วนที่เหลื่อมล้ำ $(Overlap.x_{endpoint}, Overlap.y_{endpoint})$ ลากไปยังจุดปลาย $(x_{endpoint}, y_{endpoint})$ เส้นตรงที่ได้นี้จะเรียก เส้นหลัก จากนั้นจะลากเส้นตรงเหนือและต่ำกว่าเส้นหลัก โดยลากให้เส้นตรงทุกเส้นขนานกัน คือ มีความชัน = $\frac{Overlap.y_{endpoint} - y_{endpoint}}{Overlap.x_{endpoint} - x_{endpoint}}$ ดังรูปที่ 1.11



รูปที่ 1.11 การเชื่อมต่อจุดขาดกรณีเหลื่อมล้ำในแนวแกน x

2) กรอบภาพที่เหลื่อมล้ำไม่มีการซ้อนทับกัน ดังรูปที่ 1.12

ส่วนที่เหลื่อมล้ำที่ได้จากการวิเคราะห์การขาดของตัวอักษรมี 2 ส่วน กำหนดส่วนแรกเป็น $(Overlap_1.x_{\min}, Overlap_1.y_{\min})$ และ $(Overlap_1.x_{\max}, Overlap_1.y_{\max})$ และส่วนที่สอง $(Overlap_2.x_{\min}, Overlap_2.y_{\min})$ และ $(Overlap_2.x_{\max}, Overlap_2.y_{\max})$ ดังรูปที่ 11

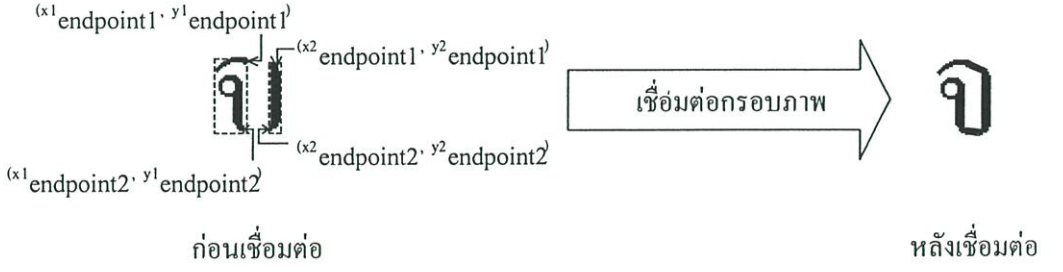


รูปที่ 1.12 การรวมกรอบภาพที่เหลื่อมล้ำแต่ไม่มีการซ้อนทับกัน

กรอบภาพที่เหลื่อมล้ำแต่ไม่มีการซ้อนทับกันจะไม่ทำการลากเส้นเชื่อมต่อ แต่จะทำการรวมให้เป็นกรอบภาพเดียว

3.2) การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

กำหนดพิกัดกรอบตัวอักษรที่ 1 เป็น $(x1_{endpoint1}, y1_{endpoint1})$ และ $(x1_{endpoint2}, y1_{endpoint2})$ และกำหนดพิกัดกรอบตัวอักษรที่ 2 เป็น $(x2_{endpoint1}, y2_{endpoint1})$ และ $(x2_{endpoint2}, y2_{endpoint2})$ ดังรูปที่ 1.13



รูปที่ 1.13 การเชื่อมต่อจุดขาดกรณีพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

การเชื่อมต่อจะทำการลากเส้นด้วยวิธีการเดียวกับการเชื่อมต่อกรอบภาพที่เหลื่อมล้ำที่มีการซ้อนทับกัน

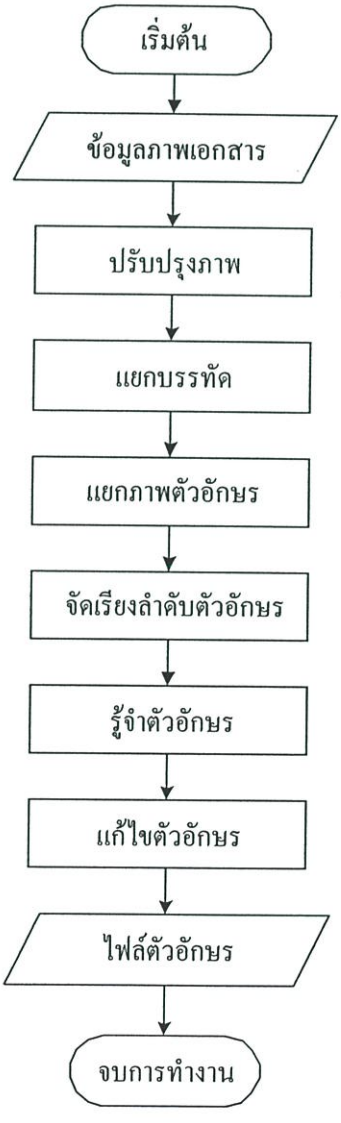
1.4 แผนการดำเนินงาน

1. ศึกษาบทความและผลงานวิจัยต่างๆ ที่มีความเกี่ยวข้องกับงานวิจัยนี้
2. เก็บข้อมูลตัวอย่างของตัวอักษรขาด พร้อมจัดเก็บลงคอมพิวเตอร์
3. ศึกษาลักษณะโครงสร้างของตัวอักษรไทยและรหัสแทนโครงสร้างตัวอักษรไทยเพื่อนำไปวิเคราะห์ลักษณะและตำแหน่งการขาด
4. ออกแบบอัลกอริทึมในการวิเคราะห์ภาพตัวอักษรขาด และทำการเชื่อมต่อตัวอักษรขาด
5. เขียนโปรแกรมเพื่อวิเคราะห์ภาพตัวอักษรขาด และทำการเชื่อมต่อตัวอักษรขาด
6. ทดลองเชื่อมต่อตัวอักษรขาดกับข้อมูลที่จัดเก็บ
7. ทดสอบผลจากการเชื่อมต่อว่าสามารถนำไปใช้งานได้จริงกับซอฟต์แวร์ ThaiOCR โดยทดสอบข้อมูลก่อนและหลังการเชื่อมต่อ
8. สรุปผลการดำเนินการ และรวบรวมนำจัดทำเอกสารนำเสนอเป็นงานวิจัย

บทที่ 2
การประมวลผลภาพเบื้องต้น

2.1 กระบวนการรู้จำตัวอักษร

ในกระบวนการรู้จำตัวอักษรเริ่มจากนำข้อมูลภาพตัวอักษรมาเป็นอินพุตของระบบ ผ่านกระบวนการปรับปรุงภาพในกรณีที่ข้อมูลภาพมีสัญญาณรบกวน จากนั้นทำการแยกบรรทัดและแยกตัวอักษรแต่ละตัวออกจากบรรทัด ในขั้นตอนนี้จะได้ตำแหน่งและขนาดของภาพตัวอักษรแต่ละตัว จากนั้นจัดเรียงลำดับภาพตัวอักษรเพื่อนำไปสู่กระบวนการรู้จำต่อไป ขั้นตอนต่าง ๆ แสดงดังรูปที่ 2.1

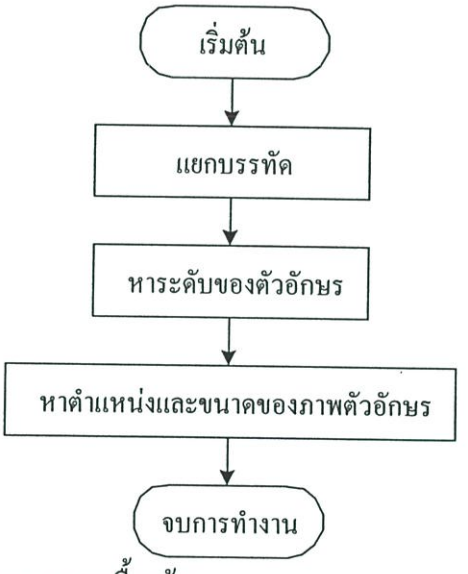


รูปที่ 2.1 กระบวนการรู้จำตัวอักษร

จากรูปที่ 2.1 ขั้นตอนการตรวจสอบและเชื่อมต่อตัวอักษรขาด จะแทรกอยู่ระหว่างกระบวนการจัดเรียงลำดับตัวอักษรและกระบวนการรู้จำตัวอักษร

2.2 การประมวลผลภาพ

อินพุตของระบบเป็นภาพเอกสารที่ประกอบด้วยบรรทัดของตัวอักษรหลายบรรทัด จะต้องทำการแยกบรรทัดและแยกตัวอักษรแต่ละตัวออกจากบรรทัด โดยการหาโครงร่างฮิสโตแกรมและการหาขอบภาพ [10] ผลที่ได้จากขั้นตอนนี้ คือ ได้ตำแหน่งและขนาดของภาพข้อมูลตัวอักษรแต่ละตัว เรียกว่า กรอบตัวอักษร แล้วจัดเรียงลำดับภาพตัวอักษร กระบวนการประมวลผลภาพเบื้องต้นแสดงได้ ดังรูปที่ 2.2



รูปที่ 2.2 กระบวนการประมวลผลภาพเบื้องต้น

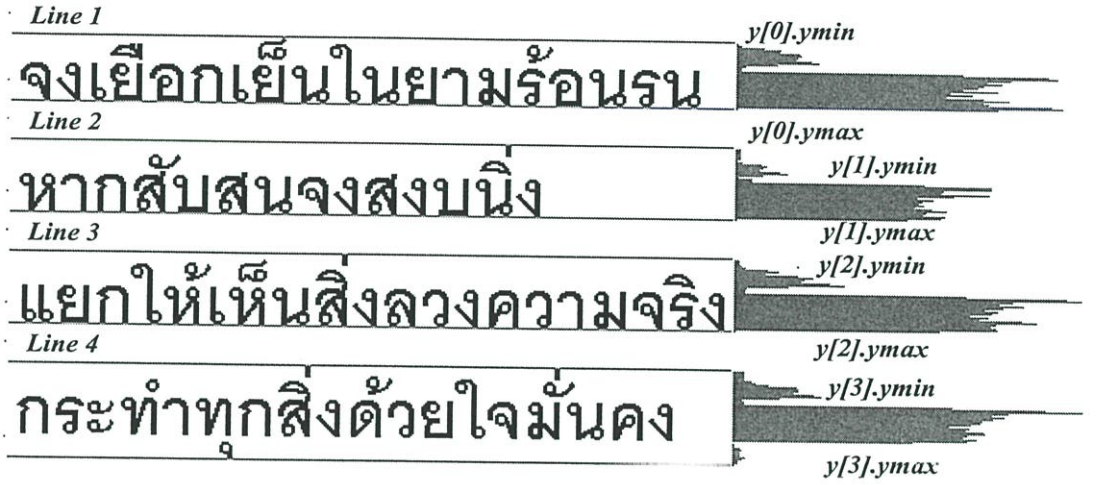
2.2.1 การแยกบรรทัด (Line Separation)

ขั้นตอนนี้นำทฤษฎีฮิสโตแกรมมาใช้ในการวิเคราะห์หาระดับของภาพตัวอักษรในแต่ละบรรทัดที่มีในหน้าเอกสาร โดยใช้ฮิสโตแกรมทางด้านแกน y หรือฮิสโตแกรมในแนวนอน (Horizontal Histogram) ดังสมการ 2.1

$$yHis(y) = \sum_{x=0}^{x=X\,max} P(x,y) \qquad (2.1)$$

เมื่อ $P(x,y)$ เป็นจุดของภาพ และ $X\,max$ เป็นความกว้างของภาพ

ในขั้นตอนนี้จะได้จุดเริ่มต้นและจุดสิ้นสุดในแนวแกน y ของบรรทัด ข้อมูลที่ได้นี้จะเก็บเป็นลิสต์ของการแบ่งที่เกิดขึ้นทั้งหมด เพื่อนำไปวิเคราะห์หาบรรทัด ดังรูปที่ 2.3



รูปที่ 2.3 การแบ่งบรรทัดโดยการวิเคราะห์ค่าฮิสโตแกรมในแนวนอน

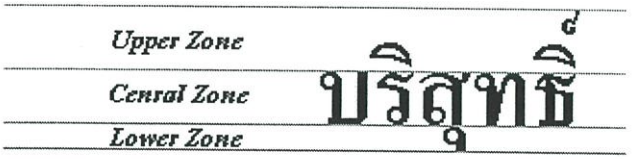
จากรูปที่ 2.3 พบว่าส่วนที่อยู่ระหว่างบรรทัดเป็นส่วนที่มีค่าฮิสโตแกรมเป็นศูนย์และมีระยะที่กว้าง ซึ่งเงื่อนไขนี้ทำให้สามารถแบ่งบรรทัดออกมาได้ จากรูปนี้จะได้ 4 บรรทัด แต่ละบรรทัดจะได้ขอบเขตของตำแหน่ง y เริ่มต้น (*y min*) และตำแหน่ง y สุดท้าย (*y max*) ของบรรทัด ข้อมูลนี้จะถูกจัดเก็บในรูปแบบอาร์เรย์

เมื่อได้ขอบเขตของแต่ละบรรทัดแล้ว จะนำขอบเขตของแต่ละบรรทัดไปทำการแบ่งระดับของตัวอักษรในแต่ละบรรทัดตามลักษณะโครงสร้างของระดับตัวอักษรภาษาไทย [12]

2.2.2 การหาระดับของตัวอักษร (Level Separation)

ขั้นตอนนี้ทำการวิเคราะห์ค่าฮิสโตแกรมในแนวนอนที่ได้จากขั้นตอนก่อนหน้า เพื่อทำการแบ่งระดับของภาพตัวอักษรในแต่ละบรรทัด ดังรูปที่ 2.4 ซึ่งระดับของตัวอักษรแบ่งเป็น 3 ส่วน คือ

- Upper Zone ประกอบด้วย ตัวอักษรในระดับ Tonal Line Level และ Upper Line Level
- Central Zone ประกอบด้วย ตัวอักษรในระดับ Consonant Line Level
- Lower Zone ประกอบด้วย ตัวอักษรในระดับ Lower Vowel Line Level



รูปที่ 2.4 การแบ่งระดับของตัวอักษรทั้ง 3 ส่วน

การแบ่งระดับจะเริ่มด้วยการหาค่าเฉลี่ยของค่าฮิสโตแกรมในแนวนอนที่ได้ในแต่ละบรรทัด ดังสมการ 2.2

$$AverageHist(line) = \frac{\sum_{i=y[line].y\,min}^{y[line].y\,max} yHis(i)}{y[line].y\,max - y[line].y\,min} \qquad (2.2)$$

เมื่อ $yHist$ เป็นค่าฮิสโตแกรมในแนวนอน

$y[line].y\,min$ และ $y[line].y\,max$ เป็นพิกัดขอบบนและขอบล่างของบรรทัด

หลักในการแบ่งระดับในแต่ละบรรทัด คือ ตำแหน่งที่มีค่าฮิสโตแกรมในแนวนอนมากกว่า 90 เปอร์เซ็นต์ของค่าเฉลี่ยฮิสโตแกรมจะถือว่าเป็นส่วน Central Zone ดังรูปที่ 2.5

รูปที่ 2.5 ค่าฮิสโตแกรมในแนวนอนเปรียบเทียบ 90 % ของค่าเฉลี่ยฮิสโตแกรมในแต่ละบรรทัด

จากรูปที่ 2.5 จะพบว่ากลุ่มหรือช่วงที่มีค่าฮิสโตแกรมมากกว่า 90 เปอร์เซ็นต์ของค่าเฉลี่ยฮิสโตแกรม (จากรูปคือแถบสีเทา) จะเป็น Central Zone การเลือกช่วงของฮิตโตแกรม คือ จะเลือกช่วงที่มีค่ากว้างที่สุดให้เป็นส่วนของ Central Zone ดังนั้นส่วนที่อยู่เหนือ Central Zone ก็จะกลายเป็น Upper Zone และส่วนที่อยู่ต่ำกว่า Central Zone ก็จะเป็น Lower Zone

ระดับของตัวอักษรจะมีประโยชน์ในการระบุระดับของภาพตัวอักษรที่ได้จากการหาขอบเขตในขั้นตอนการหาตำแหน่งและขนาดของภาพตัวอักษรในขั้นตอนถัดไป และมีประโยชน์ในการใช้จัดเรียงตัวอักษรด้วย

สำหรับงานวิจัยนี้จะวิเคราะห์การขาดของตัวอักษรเฉพาะตัวอักษรใน Consonant Level หรือส่วนที่เป็น Central Zone ซึ่งประกอบด้วยพยัญชนะไทย 44 ตัว และสระ 6 ตัว คือ า, เ, แ, ไ, ใ และ โ

2.2.3 การหาตำแหน่งและขนาดของภาพตัวอักษร (Character Segmentation)

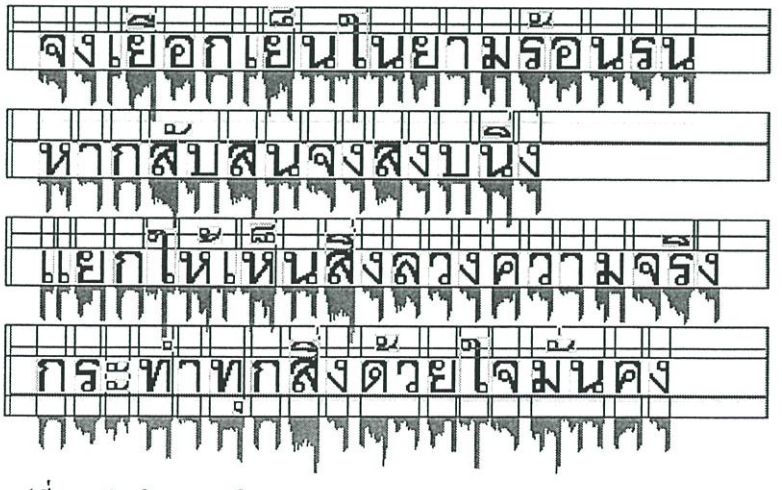
ขั้นตอนนี้จะทำการหาตำแหน่งและขอบเขตของตัวอักษรแต่ละตัวพร้อมระบุระดับของตัวอักษรว่าอยู่ในระดับใดใน 3 ส่วน คือ Upper Zone, Central Zone และ Lower Zone จากนั้นจะทำการจัดเก็บลงลิงค์ลิสต์พร้อมจัดเรียงลำดับให้ถูกหลักตามลักษณะการพิมพ์ภาษาไทย

เพื่อให้การจัดเรียงลำดับทำได้ง่ายขึ้นในขั้นตอนแรกจะทำการแบ่งภาพตัวอักษรแบบหยาบ ๆ ก่อน โดยใช้ Vertical Histogram ดังสมการ 2.3

$$xHis(x) = \sum_{y=0}^{y=y\max} P(x,y) \qquad (2.3)$$

เมื่อ $P(x,y)$ เป็นจุดภาพ และ เป็นความสูงของบรรทัดที่แบ่งได้

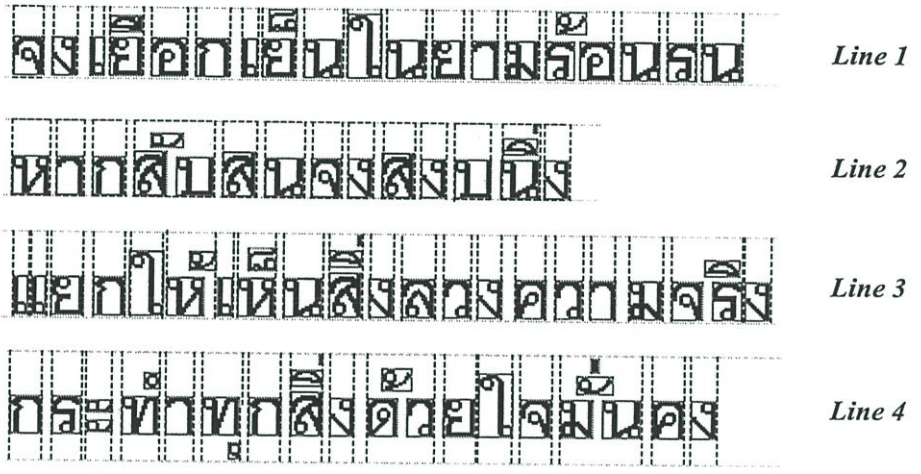
ตำแหน่งที่มีค่า Vertical Histogram มากกว่าศูนย์จะเป็นตำแหน่งที่จะตัดแบ่ง ซึ่งผลของการตัดแบ่งอาจจะได้ตัวอักษรเพียงตัวเดียวหรือเป็นกลุ่มของตัวอักษรก็ได้ ขอบเขตที่ตัดแบ่งได้ในขั้นตอนนี้จะเรียกว่า Character Block ดังรูปที่ 2.6



รูปที่ 2.6 ฮิสโตแกรมในแนวแกน x

ขั้นตอนถัดไปจะทำการไต่ขอบ (Tracing contour) ของภาพตัวอักษรที่อยู่ภายใน Character Block เพื่อหาขอบเขตของตัวอักษรแต่ละตัว แล้วทำการคัดลอกภาพตัวอักษรที่ไต่ขอบแล้วไปใส่ในโครงสร้างข้อมูลแบบลิงค์ลิสส์พร้อมจัดเรียงตามลักษณะการพิมพ์ตัวอักษรในภาษาไทย ซึ่งภาพตัวอักษรแต่ละตัวที่คัดลอกออกมาจะเรียกว่า Character Frame

ผลที่ได้จากการไต่ขอบแสดงดังรูปที่ 2.7 และ Character Frame ที่เก็บในลิงค์ลิสต์แสดงดังรูปที่ 2.8



รูปที่ 2.7 การหา Character Block

รูปที่ 2.8 Character Frame

ผลลัพธ์สุดท้ายที่ได้คือ Character Frame จะนำไปใช้ในขั้นตอนการวิเคราะห์ว่าภาพตัวอักษรที่อยู่ Character Frame ขาดหรือไม่ ซึ่งจะกล่าวรายละเอียดในบทถัดไป

บทที่ 3
การวิเคราะห์ตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด

3.1 นิยามตัวอักษรขาด

ตัวอักษรขาด หมายถึง ภาพตัวอักษรที่ถูกแยกออกเป็นชิ้นส่วนภาพย่อย ๆ ในงานวิจัยนี้เน้นเฉพาะการขาดในส่วนที่เป็นเส้นเชื่อมระหว่างขาของตัวอักษร ตัวอย่างตัวอักษรขาดแสดงดังรูปที่ 3.1

รูปที่ 3.1 ภาพตัวอักษรขาด

เมื่อนำภาพจากรูปที่ 3.1 ที่มีตัวอักษรขาดมาผ่านกระบวนการประมวลผลภาพเบื้องต้น จะได้ผลลัพธ์เป็น Character Frame ดังรูปที่ 3.2

รูปที่ 3.2 Character Frame ที่ได้จากการประมวลผลภาพเบื้องต้น

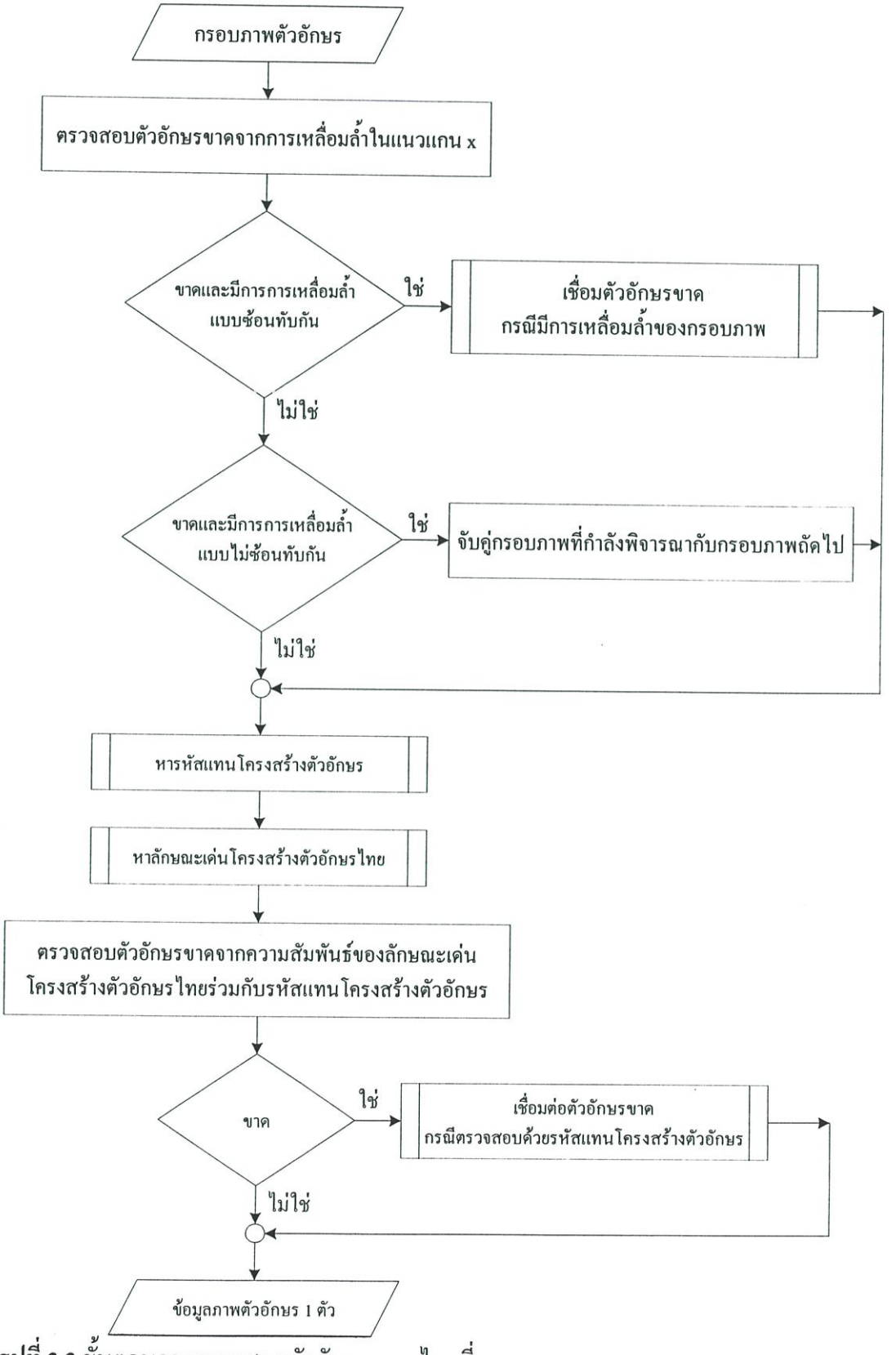
Character Frame ทั้งหมดจะถูกนำมาตรวจสอบว่าเป็นตัวอักษรขาดหรือไม่ และถ้าขาดจะทำการเชื่อมต่ออย่างไร จากการศึกษาพบว่า สามารถตรวจสอบตัวอักษรขาดได้ 2 วิธี คือ

1. การพิจารณาตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x
2. การพิจารณาตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

ขั้นตอนการตรวจสอบตัวอักษรภาษาไทยที่ขาด แสดงได้ดังรูปที่ 3.3 การเชื่อมตัวอักษรที่ได้จากการตรวจสอบทั้งสองวิธี มี 3 ขั้นตอนย่อย คือ

1. การจับคู่กรอบภาพ
2. การหาบริเวณที่จะทำการเชื่อมต่อ
3. การจับคู่จุดปลายและลากเส้นเชื่อมจุด

ขั้นตอนย่อยที่ 1 ของทั้งสองวิธีจะมีขั้นตอนแตกต่าง แต่ในขั้นตอนย่อยที่ 2 และ 3 จะมีวิธีการเหมือนกัน



รูปที่ 3.3 ขั้นตอนการตรวจสอบตัวอักษรภาษาไทยที่ขาด

3.2 การพิจารณาตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

วิธีนี้จะตรวจสอบกรอบภาพ 2 กรอบภาพ โดยคู่ลำดับของจุดปลายกรอบภาพที่ 1 อยู่ภายในกรอบภาพที่ 2 ดังสมการ (3.1) หรือพิจารณาในกรณีกลับกัน

$$overlap = \left(x_1 \max \leq x_2 \max\right) and \left(x_1 \max \geq x_2 mix\right) \tag{3.1}$$

overlap : เป็นค่าความจริงทางตรรกศาสตร์ โดยมีค่าเป็น True หรือ False

$\left(x_1 \min, x_1 \max\right)$: เป็นคู่ลำดับของจุดปลายกรอบภาพที่ 1

$\left(x_2 \min, x_2 \max\right)$: เป็นคู่ลำดับของจุดปลายกรอบภาพที่ 2

จากสมการที่ 3.1 $overlap = True$ แสดงว่ากรอบภาพทั้งสองมีการเหลื่อมล้ำกันในแนวแกน x และเป็นกรอบภาพของตัวอักษรขาด หาก $overlap = False$ แสดงว่ากรอบภาพทั้งสองไม่เหลื่อมล้ำกันและในขั้นตอนนี้สรุปได้ว่าเป็นตัวอักษรปกติ (ไม่ขาด)

ผลลัพธ์ที่ได้จากการตรวจสอบการเหลื่อมล้ำของกรอบภาพในแนวแกน x มี 2 ลักษณะ ดังนี้

3.2.1 เหลื่อมล้ำแบบซ้อนทับกัน ดังรูปที่ 3.3 สามารถส่งเข้าสู่กระบวนการซ่อมตัวอักษรได้ทันที

ตัวอักษรขาด กรอบภาพ พิจารณาการเหลื่อมล้ำ

รูปที่ 3.4 กรอบภาพที่ซ้อนทับกัน

3.2.2 เหลื่อมล้ำแบบไม่ซ้อนทับกัน ดังรูปที่ 3.4 นำเข้าสู่กระบวนการวิเคราะห์ตัวอักษรขาดในขั้นตอนถัดไป

ตัวอักษรขาด กรอบภาพ พิจารณาการเหลื่อมล้ำ

รูปที่ 3.5 กรอบภาพที่ไม่ซ้อนทับกัน

3.3 การพิจารณาตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

ขั้นตอนนี้จะนำกรอบภาพตัวอักษรที่ผ่านการซ่อมแซมตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x มาผ่านกระบวนการหาลักษณะเด่นของโครงสร้างตัวอักษรไทย ได้แก่ การหาตำแหน่งหัวของตัวอักษร การหาจุดปลายของตัวอักษร การหาขาของตัวอักษร จากนั้นจะนำลักษณะเด่นดังกล่าวมาพิจารณาตรวจสอบตัวอักษรขาดร่วมกับรหัสแทนโครงสร้างตัวอักษร

3.3.1 การหาตำแหน่งหัวของตัวอักษร

ตำแหน่งหัวของตัวอักษรหาได้จากการแบ่งพื้นที่กรอบภาพออกเป็น 4 ส่วน โดยการแบ่งจากจุดกึ่งกลางด้านกว้างและจุดกึ่งกลางด้านยาวของกรอบภาพ ดังรูปที่ 3.6

ก) ตำแหน่งหัวของตัวอักษร

ข) ตัวอักษรมีหัวในตำแหน่ง 1 และ 4

รูปที่ 3.6 การหาตำแหน่งหัวของตัวอักษร

3.3.2 การหารหัสแทนโครงสร้างตัวอักษร

การหารหัสแทนโครงสร้างตัวอักษร เริ่มจากการนำตัวอักษร 1 กรอบภาพไปผ่านกระบวนการทำให้บาง จากนั้นทำการหาจุดสแกน (x, y) หาได้จากการพิจารณาตำแหน่งหัวของตัวอักษร คือ Head(0,4) ใช้สมการ (3.2) Head(1,2) ใช้สมการ (3.3) และ Head(3) ใช้สมการ (3.4) สแกนจากจุดนี้ในทิศทางซ้าย บน ขวาและล่าง ตามลำดับ หากสแกนแล้วพบส่วนที่เป็นเนื้อตัวอักษรจะแทนด้วยเลข 1 ถ้าไม่พบจะแทนด้วยเลข 0

$$(x,y)=\left(\frac{1}{2}-Block.width,\frac{1}{2}-Block.height\right) \tag{3.2}$$

$$(x,y)=\left(Head.xMax,\frac{(Head.yMax+Block.yMax)}{2}\right) \tag{3.3}$$

$$(x,y)=\left(Head.xMax,\frac{1}{2}-Block.height\right) \tag{3.4}$$

$Head(0,4)$: ตัวอักษรที่ไม่พบหัวทุกตำแหน่งหรือพบหัวในตำแหน่งที่ 4

$Head(1,2)$: ตัวอักษรที่พบหัวในตำแหน่งที่ 1 หรือตำแหน่งที่ 2

$Head(3)$: ตัวอักษรที่พบหัวในตำแหน่งที่ 3

$Block.width$: ความกว้างของกรอบภาพ

$Block.height$: ความสูงของกรอบภาพ

$(Head.xMax, Head.yMax)$: คู่ลำดับของจุดปลายของหัวตัวอักษร

$(Block.xMax, Block.yMax)$: คู่ลำดับของจุดปลายของกรอบภาพ

กรณี Head(1,2) จะต้องนำขาของตัวอักษรมาพิจารณาร่วมด้วย กล่าวคือ หากเป็นกรอบภาพของตัวอักษรที่มีขาตั้งแต่สองขาขึ้นไป จะทำการเปลี่ยนจุดสแกน ดังสมการที่ (3.5) และ (3.6) เพื่อให้ได้รหัสแทนตัวอักษรที่เหมาะสมกับโครงสร้างตัวอักษร ดังรูปที่ 3.7

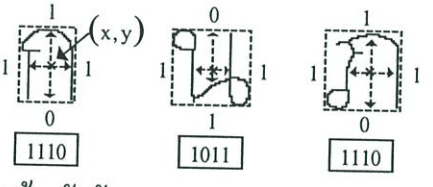
$$(x,y) = \left(Leg_{average}, \frac{(Head.yMax + Block.yMax)}{2} \right) \tag{3.5}$$

$$Leg_{average} = \frac{(Leg_1.x + leg_2.x)}{2} \tag{3.6}$$

$Leg_{average}$: ค่าเฉลี่ยของระยะห่างระหว่างขาตัวอักษร

$Leg1.x$: ค่า x ของตำแหน่งขาที่ 1

$Leg2.x$: ค่า x ของตำแหน่งขาที่ 2



รูปที่ 3.7 การหารหัสแทนโครงสร้างตัวอักษร

3.3.3 การวิเคราะห์รหัสแทนโครงสร้างตัวอักษร

รหัสแทนโครงสร้างตัวอักษรประกอบด้วยตัวเลข 4 หลัก ทำให้มีรหัสแทนโครงสร้างตัวอักษรทั้งหมด 16 รูปแบบ จากการทดลองทำให้ได้ผลดังตารางที่ 3.1

ตารางที่ 3.1 รหัสแทนโครงสร้างตัวอักษร มี 16 รูปแบบ

ลำดับ	โครงสร้าง	รหัสแทนโครงสร้าง	ตัวอักษร
1		1110	ก ข ค ฅ ฆ ฉ ฑ ฒ ด ต ท พ ฟ ภ ศ ส ห ฬ
2		1011	ข ฃ ฆ ช ซ ฌ ญ ฐ ณ น บ ป ผ ฝ ม ษ
3		1111	ง จ ฉ ช ซ ฎ ฏ ฐ ฒ ธ ผ ฝ ย ล ศ ษ ส อ ฮ
4		0110	า
5		1001	เ ไ ใ
6		0111	ฉ ฐ ร ว อ
7		1101	โ
8		1100	ตัวอักษรขาด
9		0011	ตัวอักษรขาด
10		1000	ตัวอักษรขาด
11		0010	ตัวอักษรขาด
12		0100	ตัวอักษรขาด
13		0001	ตัวอักษรขาด
14		0101	ตัวอักษรขาด
15		1010	ไม่มีตัวอักษร
16		0000	ไม่มีตัวอักษร

จากการพิจารณาตารางที่ 3.1 ได้ข้อสรุป ดังนี้

3.3.3.1 การวิเคราะห์ตัวอักษรขาดจากรหัสแทนโครงสร้างตัวอักษร ในขั้นตอนนี้จะพบว่ารหัสแทนโครงสร้างตัวอักษรที่เป็นรหัสของตัวอักษรขาด มี 7 รหัส คือ รหัส “1100” , “0011”, “1000”, “0010”, “0100”, “0001” และ “0101” ดังนั้นหากตรวจสอบพบรหัสเหล่านี้จะสามารถสรุปได้ทันทีว่ากรอบภาพนั้นเป็นภาพตัวอักษรขาด ดังรูปที่ 3.8

รูปที่ 3.8 ตัวอย่างการวิเคราะห์ตัวอักษรขาดด้วยรหัสแทนโครงสร้างตัวอักษร

3.3.3.2 การวิเคราะห์ตัวอักษรขาดจากลักษณะเด่นของตัวอักษรร่วมกับรหัสแทนโครงสร้างตัวอักษร ในขั้นตอนนี้ต้องนำลักษณะเด่นของตัวอักษรไทยมาพิจารณาร่วมด้วย จึงจะสามารถระบุได้ว่ากรอบภาพที่กำลังตรวจสอบเป็นกรอบภาพของตัวอักษรขาด ซึ่งลักษณะเด่นที่นำมาพิจารณาได้แก่ ตำแหน่งหัว จำนวนขา และจำนวนจุดปลายของตัวอักษร ซึ่งจะแยกพิจารณาตามรหัสโครงสร้างตัวอักษรดังนี้

(1) รหัส "1110" เป็นรหัสของตัวอักษรปกติ ได้แก่ ก ข ค ต ฆ ฉ ท ฒ ด ต ท พ ฟ ภ ศ ส ห พ แต่อาจตรวจพบตัวอักษรขาดได้ คือ กรณีที่พบหัวในตำแหน่งที่ 1 และพบหัวในตำแหน่งที่ 4 เช่น ฒ ดังรูปที่ 3.9

รูปที่ 3.9 ตัวอักษร ฒ ขาดจากการวิเคราะห์รหัส "1110"

(2) รหัส "1011" เป็นรหัสของตัวอักษรปกติ ได้แก่ ข ฃ ฆ ช ซ ฌ ญ ณ น บ ป ผ ฝ ม ษ แต่อาจตรวจพบตัวอักษรขาดได้ ดังนี้

รูปที่ 3.10 ตัวอักษรขาดจากการวิเคราะห์รหัส "1011"

(3) รหัส "1111" เป็นรหัสของตัวอักษรปกติ ได้แก่ ง จ ฉ ช ซ ฎ ฏ ฐ ถ ธ ผ ฝ ย ล ศ ษ ส อ ฮ ษ แต่อาจตรวจพบตัวอักษรขาดได้ดังนี้

รูปที่ 3.11 ตัวอักษรขาดจากการวิเคราะห์รหัส "1111"

(4) รหัส "0110" เป็นรหัสของสระอา (า) แต่อาจตรวจพบตัวอักษรขาดได้ ดังนี้

0110 **0011**

รูปที่ 3.12 ตัวอักษรขาดจากการวิเคราะห์รหัส "0110"

(5) รหัส "1001" เป็นรหัสของสระเอ ไม้มลาย ไม้ม้วน (เ ไ ใ) แต่อาจตรวจพบตัวอักษรขาดได้ ดังนี้

1111 **1001**

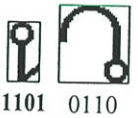
รูปที่ 3.13 ตัวอักษรขาดจากการวิเคราะห์รหัส "1001"

(6) รหัส "0111" เป็นรหัสของตัวอักษรปกติ ได้แก่ ฉ ฐ ร ว อ แต่อาจตรวจพบตัวอักษรขาดได้ ดังนี้

0111 0011

รูปที่ 3.14 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "0111"

(7) รหัส "1101" เป็นรหัสของสระโอ (โ) แต่อาจตรวจพบตัวอักษรขาดได้ ดังนี้



1101 0110

รูปที่ 3.15 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "1101"

(8) รหัส "1100" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ก ค จ ฉ ฒ ด ต ฝ ย ล ศ ส ห พ อ ฮ

1100 0110

รูปที่ 3.16 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "1101"

(9) รหัส "0011" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ข ฆ ฉ ช ญ ฎ ฒ น ผ พ ม ย

0110 **0011**

รูปที่ 3.17 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "0011"

(10) รหัส "1000" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ข ค ด ป ษ อ

0110 **1000**

รูปที่ 3.18 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "1000"

(11) รหัส "0010" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ข ฆ ฌ ญ ฒ ม ว

0110 **0010**

รูปที่ 3.19 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "0010"

(12) รหัส "0100" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ว อ

0100 0111

รูปที่ 3.20 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "0100"

(13) รหัส "0001" เป็นกรณีของตัวอักษรขาดทุกกรณี เช่น ว

0001 0111

รูปที่ 3.21 ตัวอักษรขาดจากการวิเคราะห์ด้วยรหัส "0001"

(14) รหัส "0101" เป็นกรณีของตัวอักษรขาด เช่น จ

(15) รหัส "1010" กรณีนี้ไม่เป็นรหัสของตัวอักษร

บทที่ 4
การซ่อมแซมตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด

การเชื่อมตัวอักษรขาดจะนำข้อมูลจากการวิเคราะห์การขาดของตัวอักษรมาพิจารณาการเชื่อมต่อ ซึ่งจากลักษณะการขาดของตัวอักษรทั้งสองแบบดังที่กล่าวมาในการวิเคราะห์การขาดของตัวอักษร จะทำให้เกิดการเชื่อม 2 แบบเช่นกัน คือ

(1) การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

(2) การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

การเชื่อมตัวอักษรขาดทุกครั้ง จะเริ่มจากการเชื่อมต่อในแบบที่ (1) ก่อน และนำผลลัพธ์ที่ได้ไปทำการเชื่อมต่อในแบบที่ (2)

การเชื่อมตัวอักษรทั้งสองแบบ มีหลักการย่อย ดังนี้

(1) การจับคู่ภาพกรอบตัวอักษร

(2) การพิจารณาหาตำแหน่งเชื่อมต่อ

(3) การจับคู่จุดปลายและลากเส้นเชื่อมจุด

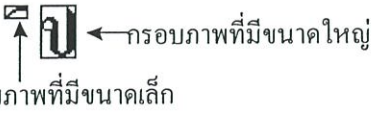
ในหลักการย่อยข้อ (2) และ (3) ของการเชื่อมทั้งสองแบบใช้หลักในการพิจารณาเหมือนกัน ส่วนหลักการย่อยข้อ (1) ของการเชื่อมทั้งสองแบบใช้วิธีการพิจารณาแตกต่างกัน อธิบายได้ดังนี้

4.1 การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

4.1.1 การจับคู่ภาพกรอบตัวอักษร เป็นขั้นตอนที่พิจารณาว่ากรอบภาพปัจจุบันที่กำลังพิจารณาเป็นภาพตัวอักษรที่ขาด เนื่องจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x หรือไม่ ถ้าใช่จะทำการจับคู่กับกรอบภาพถัดไป

4.1.2 การพิจารณาหาตำแหน่งเชื่อมต่อ ในขั้นตอนนี้มีกระบวนการย่อย 2 กระบวนการ ได้แก่

(1) การพิจารณาหาภาพกรอบตัวอักษรที่มีขนาดเล็ก นำภาพกรอบตัวอักษรที่จับคู่ได้ในขั้นตอน 4.1.1 มาหาพื้นที่ ภาพใดมีพื้นขนาดเล็ก จะเป็น**กรอบภาพหลัก**ที่ใช้ในการพิจารณาหาตำแหน่งเชื่อมต่อในขั้นตอนที่ (2) ดังรูปที่ 4.1



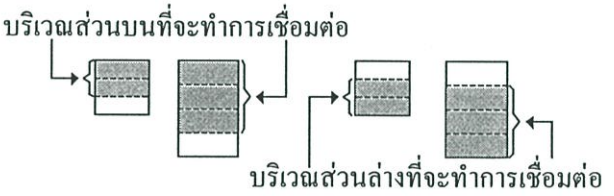
รูปที่ 4.1 แสดงกรอบภาพที่มีขนาดเล็กของตัวอักษร จ ขาด

(2) การพิจารณาหาบริเวณที่จะทำการเชื่อมต่อ จะนำภาพกรอบตัวอักษรที่มีขนาดเล็กไปผ่านกระบวนการหารหัสแทนโครงสร้างตัวอักษร จากนั้นจะนำรหัสแทนโครงสร้างตัวอักษรไปตรวจสอบเงื่อนไข เพื่อให้ทราบว่าจะทำการเชื่อมต่อ ณ บริเวณส่วนบน หรือส่วนล่าง หรือมีการเชื่อมต่อเกิดขึ้นทั้งสองส่วนของกรอบภาพ

การแบ่งกรอบภาพกรณีทั่วไป คือ

2.1 การแบ่งกรอบภาพที่มีขนาดเล็ก จะแบ่งตามแนวนอนออกเป็น 3 ส่วนเท่า ๆ กัน ส่วนบนของกรอบภาพคิดจาก $\frac{2}{3}$ ของกรอบภาพตอนบน และส่วนล่างของกรอบภาพคิดจาก $\frac{2}{3}$ ของกรอบภาพตอนล่าง ดังรูปที่ 4.2

2.2 การแบ่งกรอบภาพที่มีขนาดใหญ่ จะแบ่งตามแนวนอนออกเป็น 4 ส่วนเท่า ๆ กัน ส่วนบนของกรอบภาพคิดจาก $\frac{3}{4}$ ของกรอบภาพตอนบน และส่วนล่างของกรอบภาพคิดจาก $\frac{3}{4}$ ของกรอบภาพตอนล่าง ดังรูปที่ 4.2



รูปที่ 4.2 การพิจารณาส่วนบนและส่วนล่างของตำแหน่งเชื่อมต่อ

การหาบริเวณในการเชื่อมต่อจากรหัสแทนโครงสร้างตัวอักษร มีหลักในการพิจารณาดังนี้

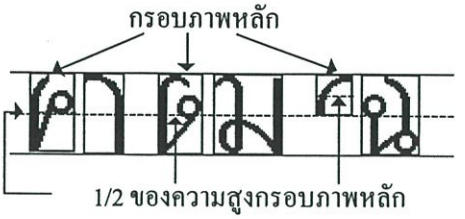
(1) รหัส "1100" ตำแหน่งที่จะทำการเชื่อมต่อ พิจารณาดังนี้

1.1) ตำแหน่งครึ่งหนึ่งของความสูงกรอบภาพหลักอยู่สูงกว่าหรือเท่ากับครึ่งหนึ่งของความสูงระยะบรรทัด

1.1.1) กรอบภาพหลักมีหัวในตำแหน่งที่ 1 บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูป 4.3 ก

รูปที่ 4.3 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 1100 ทำให้ได้การเชื่อมต่อบริเวณส่วนล่าง

1.1.2) กรอบภาพหลักมีหัวในตำแหน่งที่ 2 หรือไม่พบหัวในทุกตำแหน่ง บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนบน ดังรูปที่ 4.4



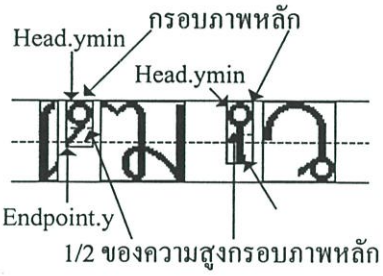
รูปที่ 4.4 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 1100 ทำให้ได้การเชื่อมต่อบริเวณส่วนส่วนบน

1.2) ตำแหน่งครึ่งหนึ่งของความสูงกรอบภาพหลักอยู่ต่ำกว่าครึ่งหนึ่งของความสูงระยะบรรทัด

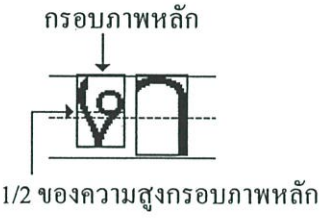
1.2.1) มีหัวในตำแหน่งที่ 1 หรือตำแหน่งที่ 2 และมีจำนวนหัว 1 หัว

1.2.1.1) ค่า ymin ของตำแหน่งหัวอยู่สูงกว่าค่า y ของจุดปลายจุดที่ 1 บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูปที่ 4.5 ก

1.2.1.2) ไม่ตรงกับเงื่อนไข 1.2.1.1 บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนบน ดังรูปที่ 4.5 ข



ก) การเชื่อมต่อบริเวณส่วนล่าง



ข) การเชื่อมต่อบริเวณส่วนบน

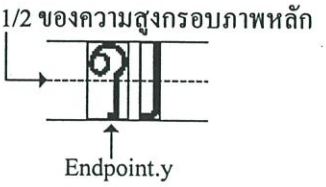
รูปที่ 4.5 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 1100 ในกรณีพบหัวในตำแหน่งที่ 1 หรือ 2

(2) รหัส "1101" พบหัวในตำแหน่งที่ 3 หรือ ตำแหน่งที่ 4 บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนบน ดังรูปที่ 4.6

รูปที่ 4.6 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 1101

(3) รหัส "0110" ตำแหน่งที่จะทำการเชื่อมต่อ พิจารณาดังนี้

3.1) กรณีพบหัวในตำแหน่งที่ 1 และ ค่า y ของจุดปลายจุดที่ 1 อยู่ต่ำกว่าครึ่งหนึ่งของความสูงของกรอบภาพ บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูปที่ 4.7 ก



รูปที่ 4.7 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 0110 กรณีพบหัวในตำแหน่งที่ 1

3.2) กรณีไม่พบหัวในตำแหน่งที่ 1

3.2.1) ความสูงของกรอบภาพหลักน้อยกว่า ¾ ของความสูงระยะบรรทัด บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูปที่ 4.8 ก

3.2.2) ไม่ตรงกับเงื่อนไข 3.2.1บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนบน ดังรูปที่ 4.8 ข

ก) การเชื่อมต่อบริเวณส่วนล่าง

ข) การเชื่อมต่อบริเวณส่วนบน

รูปที่ 4.8 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 0110 กรณีไม่พบหัวในตำแหน่งที่ 1

(4) รหัส "0011" บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูปที่ 4.9

รูปที่ 4.9 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 0011

(5) รหัส "1011" พบขา 1 ขา และมีจำนวนจุดปลาย 3 จุด และค่า y ของจุดปลายที่ 2 อยู่ต่ำกว่า 1/3 ของความสูงกรอบภาพ บริเวณที่จะทำการเชื่อมต่อ คือ ส่วนล่าง ดังรูปที่ 4.10

รูปที่ 4.10 กรอบภาพหลักที่วิเคราะห์ด้วยรหัส 1011

หากกรอบภาพหลักที่นำมาตรวจสอบไม่ตรงกับรหัสทั้งห้ารหัสดังกล่าว บริเวณที่จะทำการเชื่อมต่อ คือ ทั้งบริเวณส่วนบนและส่วนล่าง ซึ่งจะต้องนำไปพิจารณาในขั้นตอนถัดไป

4.1.3 การจับคู่จุดปลายและลากเส้นเชื่อมจุด

ในขั้นตอนนี้จะนำผลที่ได้จากขั้นตอนที่ 4.1.2 ซึ่งก็คือบริเวณที่จะทำการลากเส้นเชื่อมต่อ บริเวณดังกล่าวจะมีจุดปลายของแต่ละกรอบภาพหลายจุด การพิจารณาว่าจะเลือกลากเส้นเชื่อมต่อจากจุดใดไปสิ้นสุดที่จุดใดนั้น จะใช้วิธีการหาระยะห่างที่สั้นที่สุดระหว่างจุด 2 จุด กำหนดให้ $Block1.Endpoint(m)$ เป็นจุดปลายของกรอบภาพที่ 1 ซึ่งมีจุดปลายทั้งหมด m จุด และ $Block2.Endpoint(n)$ เป็นจุดปลายของกรอบภาพที่ 2 และมีจุดปลายทั้งหมด n จุด ดังนั้นจะต้องหาระยะทางทั้งหมด $m \times n$ ครั้ง โดยใช้สมการ

$$\text{distance(i,j)} = \sqrt{(Block2.Endpoint(j).x - Block1.Endpoint(i).x)^2 + (Block2.Endpoint(j).x - Block1.Endpoint(i).x)^2}$$

โดยที่ i = 1, 2, 3, ..., m และ j = 1, 2, 3, ..., n

เมื่อได้ระยะทางที่สั้นที่สุดแล้วจะนำระยะทางมาพิจารณาดังนี้

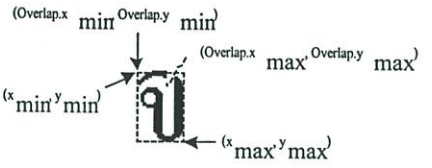
(1) ตำแหน่งที่ลากเส้นเชื่อมต่อเป็นบริเวณส่วนบน จะพิจารณาว่า ระยะห่างระหว่างจุด 2 จุดในแนวแกน x จะต้องน้อยกว่าหรือเท่ากับ 1/2 ของความกว้างของกรอบภาพ และระยะห่างระหว่างจุด 2 จุดในแนวแกน y จะต้องน้อยกว่าหรือเท่ากับ 1/5 ของความสูงกรอบภาพ

(2) ตำแหน่งที่ลากเส้นเชื่อมต่อเป็นบริเวณส่วนล่าง จะพิจารณาว่า ระยะห่างระหว่างจุด 2 จุดในแนวแกน x จะต้องน้อยกว่า 1/2 ของความกว้างกรอบภาพ และระยะห่างระหว่างจุด 2 จุดในแนวแกน y จะต้องน้อยกว่า 1/4 ของความสูงกรอบภาพ

ทั้งสองกรณีจะต้องมีระยะทางมากกว่าศูนย์ หลังจากที่พิจารณาตามเงื่อนไขสองข้อนี้แล้วเป็นจริง จึงจะทำการลากเส้นเชื่อมทั้งสองกรอบภาพ

ก. การเชื่อมต่อที่บริเวณส่วนบน

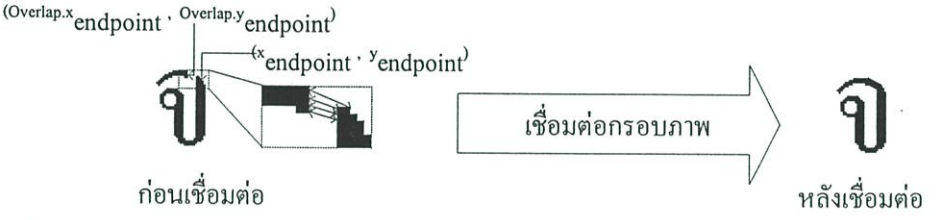
กำหนดพิกัดกรอบตัวอักษรที่ได้จากการประมวลผลภาพเบื้องต้นเป็น $(x_{\min}, y_{\min})$ และ $(x_{\max}, y_{\max})$ ส่วนที่เหลื่อมล้ำที่ได้จากการวิเคราะห์การขาดของตัวอักษรกำหนดเป็น $(Overlap.x_{\min}, Overlap.y_{\min})$ และ $(Overlap.x_{\max}, Overlap.y_{\max})$ ดังรูปที่ 4.11



รูปที่ 4.11 ขอบเขตของตัวอักษรและส่วนที่เหลื่อมล้ำ

การเชื่อมต่อจะยึดส่วนที่เหลื่อมล้ำเป็นหลัก คือ เริ่มจากจุดปลายของส่วนที่เหลื่อมล้ำ $(Overlap.x_{endpoint}, Overlap.y_{endpoint})$ ลากไปยังจุดปลาย $(x_{endpoint}, y_{endpoint})$ เส้นตรงที่ได้นี้จะเรียก เส้นหลัก จากนั้นจะลากเส้นตรงเหนือและต่ำกว่าเส้นหลัก โดยลากให้เส้นตรงทุกเส้นขนานกัน คือ

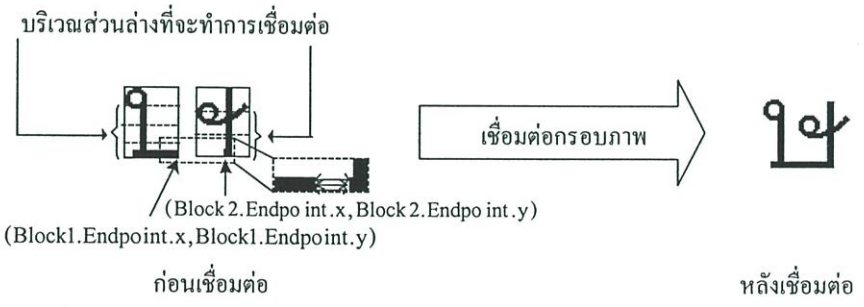
$$\text{ความชัน} = \frac{\text{Overlap.y}_{endpoint} - \text{y}_{endpoint}}{\text{Overlap.x}_{endpoint} - \text{x}_{endpoint}}$$ ดังรูปที่ 4.12



รูปที่ 4.12 การเชื่อมต่อจุดขาดบริเวณส่วนบน

ข. การเชื่อมต่อที่บริเวณส่วนล่าง

การเชื่อมต่อบริเวณส่วนล่างจะทำเช่นเดียวกับการเชื่อมต่อบริเวณส่วนบน ดังรูปที่ 4.13



รูปที่ 4.13 การเชื่อมต่อจุดขาดบริเวณส่วนล่าง

4.2 การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

หลังจากเชื่อมตัวอักษรที่ขาดเนื่องจากพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x แล้ว จะนำกรอบภาพที่ขาด ซึ่งไม่ได้เชื่อมต่อในขั้นตอนก่อนหน้านั้น มาวิเคราะห์ดังนี้

4.2.1. การจับคู่ภาพกรอบตัวอักษร ขั้นตอนนี้จะนำรหัสแทนโครงสร้างตัวอักษรของกรอบภาพปัจจุบันมาวิเคราะห์ เพื่อให้ทราบว่าจะนำกรอบภาพปัจจุบันไปเชื่อมกับกรอบภาพก่อนหน้า หรือกรอบภาพถัดไป การวิเคราะห์รหัสแทนโครงสร้างตัวอักษรพิจารณา ดังนี้

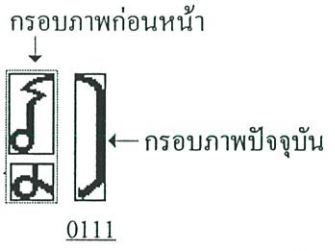
(1) รหัส "1100" การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.14

รูปที่ 4.14 การจับคู่จากการวิเคราะห์ด้วยรหัส "1100"

(2) รหัส "0111" การจับคู่พิจารณา ดังนี้

2.1 ไม่พบหัวในตำแหน่งที่ 1 (ฎ) การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.15 ก

2.2 ไม่ตรงกับเงื่อนไขในข้อ 2.1 การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.15 ข



ก) ไม่พบหัวในตำแหน่งที่ 1

ข) ไม่ตรงเงื่อนไขในข้อ 2.2

รูปที่ 4.15 การจับคู่จากการวิเคราะห์ด้วยรหัส "0111"

(3) รหัส "0110" การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.16

รูปที่ 4.16 การจับคู่จากการวิเคราะห์ด้วยรหัส "0110"

(4) รหัส "1101" การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.17

รูปที่ 4.17 การจับคู่จากการวิเคราะห์ด้วยรหัส "1101"

(5) รหัส "1011" การจับคู่พิจารณาดังนี้

5.1) พบหัวในตำแหน่งที่ 1 และมีขา 1 ขา การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.18 ก

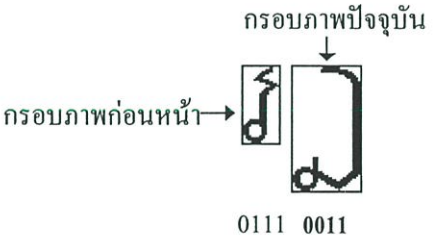
5.2) ไม่ตรงกับเงื่อนไข 5.1 การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.18 ข

ก) พบหัวในตำแหน่งที่ 1 และมี 1 ขา

ข) ไม่ตรงกับเงื่อนไข 5.2

รูปที่ 4.18 การจับคู่จากการวิเคราะห์ด้วยรหัส "1011"

(6) รหัส "0011" การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.19

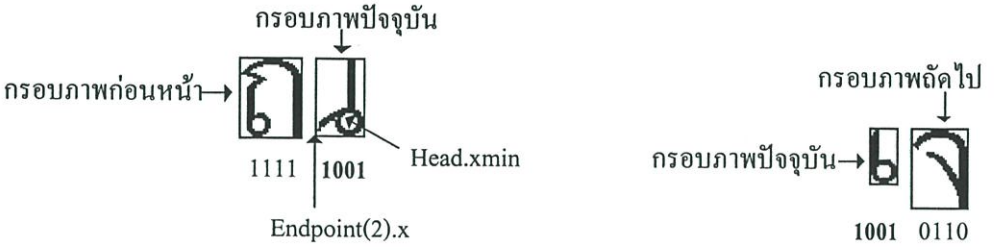


รูปที่ 4.19 การจับคู่จากการวิเคราะห์ด้วยรหัส "0011"

(7) รหัส "1001" การจับคู่พิจารณาดังนี้

7.1) พบหัวในตำแหน่งที่ 4 และเป็นตัวอักษรในระดับ Central Zone และมีค่า x ของจุดปลายจุดที่ 2 น้อยกว่า xmin ของตำแหน่งหัว (Endpoint(2).x < Head.xmin) การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.20 ก

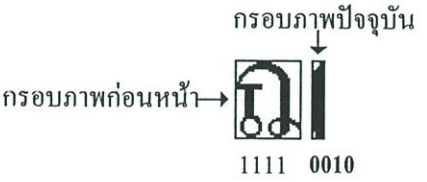
7.2) ไม่ตรงกับเงื่อนไข 7.1 การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.20 ข



ก) พบหัวในตำแหน่งที่ 4 ข) ไม่ตรงกับเงื่อนไข 7.1

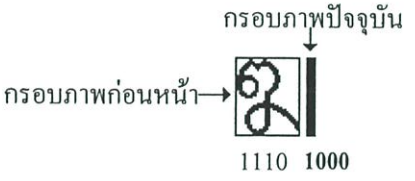
รูปที่ 4.20 การจับคู่จากการวิเคราะห์ด้วยรหัส "1001"

(8) รหัส "0010" การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.21



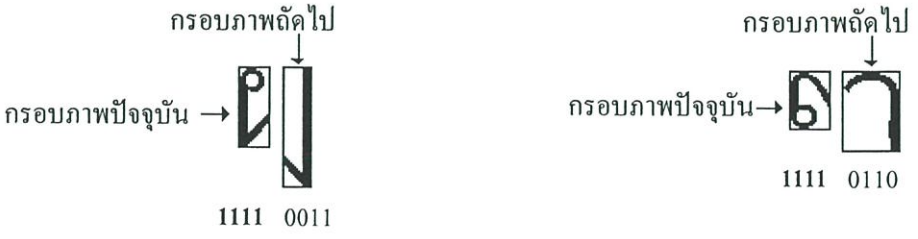
รูปที่ 4.21 การจับคู่จากการวิเคราะห์ด้วยรหัส "0010"

(9) รหัส "1000" การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 4.22



รูปที่ 4.22 การจับคู่จากการวิเคราะห์ด้วยรหัส 0010

(10) รหัส "1111" การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน และกำหนดให้ต่อที่บริเวณส่วนล่าง ดังรูปที่ 4.23

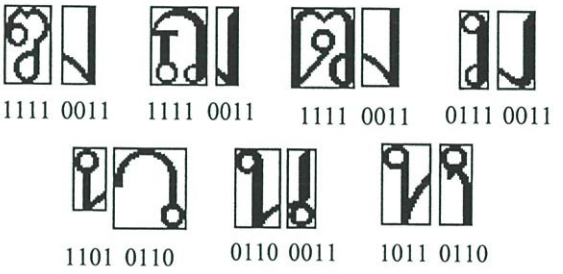


รูปที่ 4.23 การจับคู่จากการวิเคราะห์ด้วยรหัส 1111

4.2.2 การพิจารณาหาตำแหน่งเชื่อมต่อ ใช้วิธีการวิเคราะห์เช่นเดียวกับ 4.1.2 ของการเชื่อมตัวอักษร โดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

4.2.3 การจับคู่จุดปลายและลากเส้นเชื่อมจุด ใช้วิธีการวิเคราะห์เช่นเดียวกับ 4.1.3 ของการเชื่อมตัวอักษร โดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

นอกเหนือจากการเชื่อมต่อกรณีปกติดังที่ได้กล่าวมาแล้ว ยังพบว่าการขาดในบางกรณีที่ไม่พบจุดปลาย แต่ต้องทำการลากเส้นเชื่อมระหว่างสองกรอบภาพ การขาดในกรณีดังกล่าวจะเกิดที่ใกล้บริเวณส่วนหัว เช่น ฆ ฉ ฌ น ม ฒ ห ดังรูปที่ 4.24



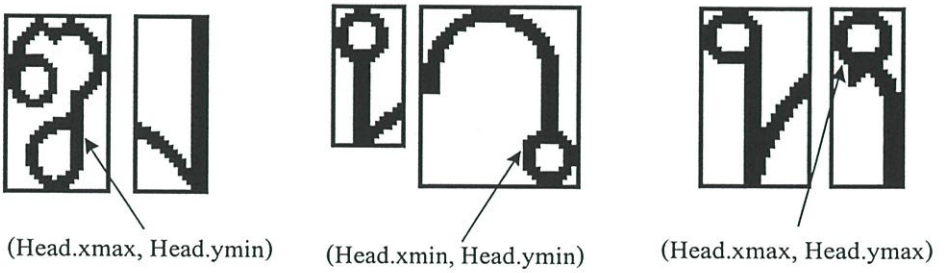
รูปที่ 4.24 การขาดในกรณีไม่พบจุดปลาย

จากการพิจารณาตำแหน่งที่จะทำการลากเส้นเชื่อมต่อ ทำให้สามารถแบ่งเป็น 3 กรณี ดังนี้

(1) การลากเส้นเชื่อมต่อที่ตำแหน่ง (Head.xmax, Head.ymin) ได้แก่ ฆ ฌ ฒ ม ดังรูปที่ 4.25 ก

(2) การลากเส้นเชื่อมต่อที่ตำแหน่ง (Head.xmin, Head.ymin) ได้แก่ ฉ น ดังรูปที่ 4.25 ข

(3) การลากเส้นเชื่อมต่อที่ตำแหน่ง (Head.xmin, Head.ymax) ได้แก่ ห ดังรูปที่ 4.25 ค



รูปที่ 4.25 การลากเส้นเชื่อมต่อกรณีไม่พบจุดปลาย

บทที่ 5

ผลการทดลอง

งานวิจัยนี้พัฒนาโปรแกรมโดยใช้ Microsoft Visual Basic Version 6.0 และได้ทำการออกแบบการทดลองเป็น 3 ส่วน ดังนี้

5.1 การทดลองหารหัสแทนโครงสร้างตัวอักษรปกติ

งานวิจัยนี้ได้ทำการทดลองกับตัวอักษรภาษาไทยที่มีหัว โดยตัวอักษรที่ใช้ในการทดลองนี้มี 7 ชนิด คือ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC, FreesiaUPC และ IrisUPC ดังตารางที่ 5.1 ผลการทดลองที่ได้มี 2 ผลการทดลอง คือ รหัสแทนโครงสร้างอักษรของตัวอักษรปกติ 44 ตัวของทั้ง 7 ชนิดตัวอักษร และสรุปรหัสแทนโครงสร้างตัวอักษรของตัวอักษรปกติทั้ง 7 ชนิดตัวอักษร

ตารางที่ 5.1 ตัวอย่างตัวอักษรทั้ง 7 ชนิด

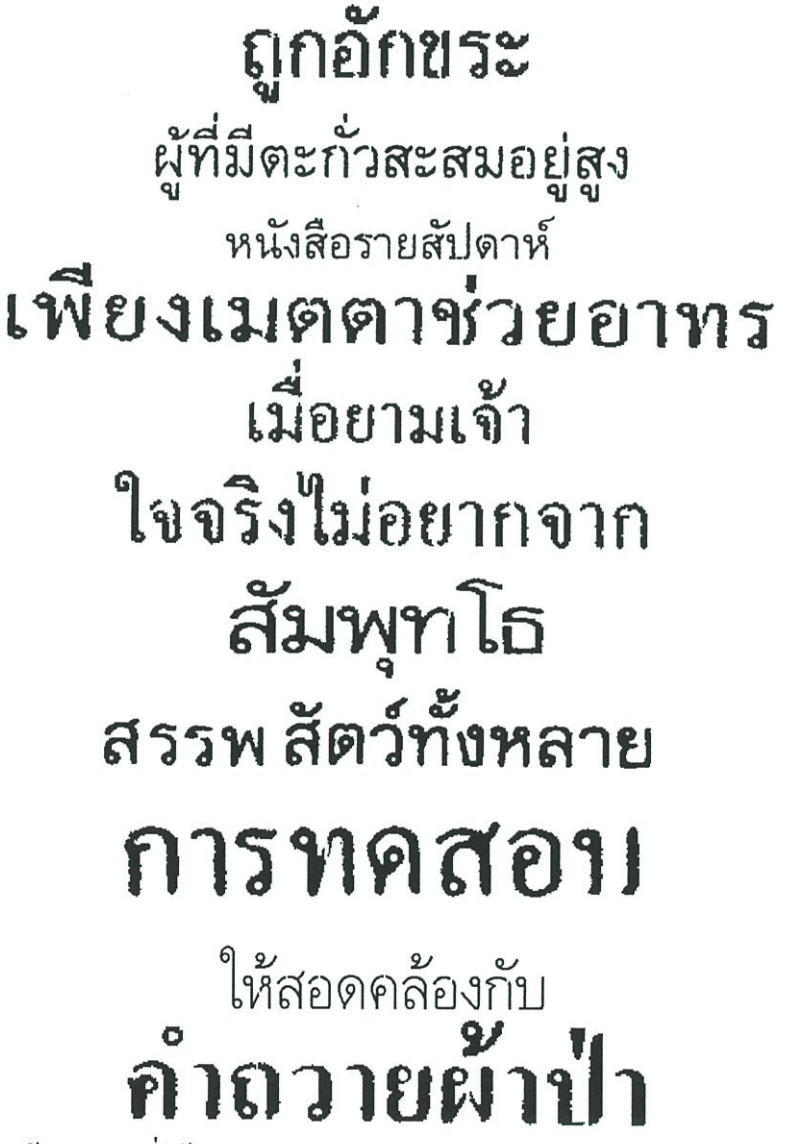
ลำดับ	ชนิดตัวอักษร	ตัวอย่าง
1	AngsanaUPC (A)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
2	BrowalliaUPC (B)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
3	CordiaUPC (C)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
4	DilleniaUPC (D)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
5	EucrosiaUPC (E)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
6	FreesiaUPC (F)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง
7	IrisUPC (I)	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง

5.2 การทดลองหารหัสแทนโครงสร้างตัวอักษรที่ขาด

งานวิจัยนี้ได้ทดลองตัดตัวอักษรขาดในลักษณะต่าง ๆ ที่มักจะพบการขาด และการคาดคะเนโอกาสที่จะเกิดการขาดในลักษณะต่าง ๆ เพื่อให้ได้รหัสแทนโครงสร้างตัวอักษรจากการขาดของตัวอักษรในแบบต่าง ๆ

5.3 การทดลองซ่อมตัวอักษรขาด

งานวิจัยนี้ได้ทำการทดลองกับข้อมูลที่ได้มาจากเอกสารที่มีการสแกนภาพที่ความละเอียด 600 จุด และนำข้อมูลภาพที่ได้รับการซ่อมแซมแล้วมาทำการทดสอบและเปรียบเทียบผลด้วยโปรแกรมการรู้จำภาษาไทยที่มีขายตามต้องตลาด ได้แก่ ThaiOCR Version 1.5b และ ArnThai Version 2.0 เพื่อแสดงให้เห็นว่าวิธีการซ่อมแซมตัวอักษรนี้สามารถนำไปใช้งานได้จริง ข้อมูลที่ใช้ในการทดสอบมีหลายรูปแบบด้วยกัน ได้แก่ ตัวหนา และตัวบาง และมีขนาดที่แตกต่างกัน โดยมีตัวอย่างที่ใช้ในการทดลองแสดงดังรูปที่ 5.1



รูปที่ 5.1 ตัวอย่างข้อมูลภาพที่ใช้ในการทดสอบ

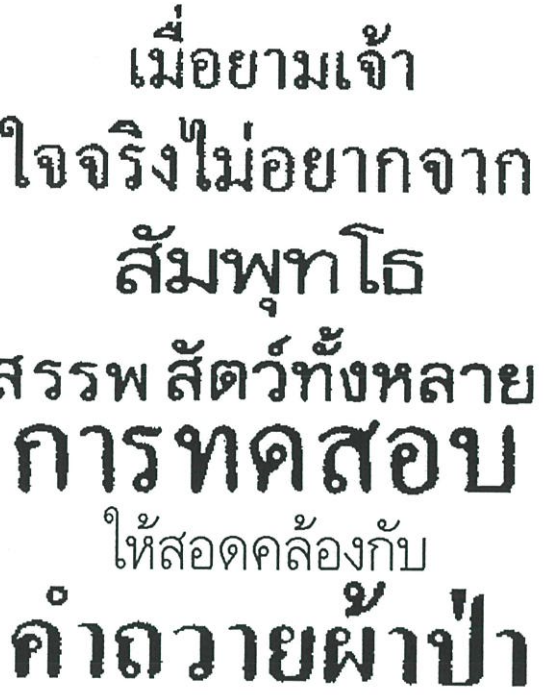
จากตัวอย่างดังรูปที่ 5.1 เมื่อนำภาพข้อความดังกล่าวไปเป็นอินพุตให้กับซอฟต์แวร์ ThaiOCR และ ArnThai เพื่อทำการทดสอบว่ากรณีที่ข้อมูลภาพไม่สมบูรณ์ซอฟต์แวร์ทั้งสองจะสามารถทำการรู้จำได้หรือไม่ มากน้อยเพียงใด ทำให้ได้ผลการทดลองดังตารางที่ 5.2

ตารางที่ 5.2 ผลการทดสอบด้วยโปรแกรม ThaiOCR และ ArnThai ในการรู้จำตัวอักษร

ข้อความทดสอบ	ซอฟต์แวร์ทดสอบ	
	ThaiOCR	**ArnThai**
ถูกอักขระ	ถู่กe▮f1'ระ	[lกE าf)t'ร
ผู้ที่มีตะกั่วสะสมอยู่สูง	4ู้ที่3มีตะr่ะJสะสมPe)I▮สูะร	▮ที่@Jตะโ▮ศ3สะสมM E)Eiสู\'
หนังสือรายสัปดาห์	ร ขัรสือระายะสัรjฬาาข์	'จชL\ๆสือร" ยธ สั]ljจ?าา\{
เพียงเมตตาช่วยอาทร	เ นียงเ IWาตา ̊ 8jย8ยาใาร	เ'▮ยงเ 1?าตาๆf1ศjยภ ดาทร
เมื่อยามเจ้า	เมื่อยามเา้า	เมื่อยามเจ้า
ใจจริงไม่อยากจาก	ใะ จริงไร่อยากจาก	ใธจจริ6รไว 1า 6 fjา ไจาก
สัมพุทโธ	สัมพุทiโธ	สัมพุrlโธ
สรรพสัตว์ทั้งหลาย	สร3พสัดว์ทั้งทลาย	สรภ1พ สัตว์ท ป็หลาย
การทดสอบ	า i1ดสอii	กา>ัฉาดสอlj
ให้สอดคล้องกับ	ใ ▮ส์ อ์ ด าล้ย ง้1บ	ใT▮ L U ตคล้ธoงF1บ
คำถวายผ้าป่า	าง วายฐงาใ า	r▮}วายฒไ้า\11า

จากนั้นจะนำภาพที่ไม่สมบูรณ์เนื่องจากการการขาดของตัวอักษรไปทำการปรับปรุงด้วยกระบวนการเชื่อมตัวอักษรขาด ทำให้ได้ผลการเชื่อมต่อดังรูปที่ 5.2 และทดสอบด้วยโปรแกรม ThaiOCR และ ArnThai อีกครั้ง ผลการรู้จำแสดงดังตารางที่ 5.3

รูปที่ 5.2 ตัวอย่างข้อมูลภาพที่ใช้ในการทดสอบหลังการเชื่อมต่อ



รูปที่ 5.2 (ต่อ)

ตารางที่ 5.3 ผลการทดสอบด้วยโปรแกรม OCR ในการรู้จำตัวอักษรขาดที่ซ่อมแซมแล้ว

ข้อความทดสอบ	ซอฟต์แวร์ทดสอบ	
	ThaiOCR	**ArnThai**
ถูกอักขระ	ถูกอักยระ	ถูกอักๆระ
ผู้ที่มีตะกั่วสะสมอยู่สูง	ผู้ที่มีตะกั่วสะสมอยู่สูง	ผู้ที่มีตะกั่วสะสมอยู่สูง
หนังสือรายสัปดาห์	หนังสือรายสัปดาห์	หนังสือรายสัปดาห์
เพียงเมตตาช่วยอาทร	เพียงเมตตาชวยอาทร	เพียงเมตตาช่วยอาทร
เมื่อยามเจ้า	เมื่อยามเจ้า	เมื่อยามเจ้า
ใจจริงไม่อยากจาก	ใจจริงไม่อยากจาก	ใจจริงไม่อยากจาก
สัมพุทโธ	สัมพุทโธ	สัมพุทโธ
สรรพสัตว์ทั้งหลาย	สรรพสัดว์ทั้งทลาย	สรรพ สัตว์ทั้ปีหลาย
การทดสอบ	การทดสอบ	การทดสอบ
ให้สอดคล้องกับ	ให้สอดคล้องกับ	ให้สอดคล้องกับ
คำถวายผ้าป่า	คงถวายผาปา	คำถวายผ้าป่า

จากการเก็บรวบรวมสถิติของการพบตัวอักษรขาด โดยแยกตามประเภทของเอกสารและสิ่งพิมพ์ ดังตารางที่ 5.4

ตารางที่ 5.4 สถิติการพบตัวอักษรขาดในเอกสารประเภทต่าง ๆ

ประเภทของเอกสาร หรือสิ่งพิมพ์	จำนวนตัวอักษร ทั้งหมด	จำนวน ตัวอักษรขาด	เปอร์เซ็นต์ การพบตัวอักษรขาด
วารสาร	90520	93	0.1027
หนังสือพิมพ์	34001	35	0.0147
ถ่ายเอกสาร	21280	68	0.3195
เอกสารโรเนียว	10320	350	3.3915
เอกสารโทรสาร (FAX)	10684	471	4.4085
เอกสารจากเครื่องพิมพ์ อิงค์เจ็ต	9203	5	0.0543

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้นำเสนอแนวทางเพื่อแก้ปัญหาภาพตัวอักษรขาดอันเป็นสาเหตุหนึ่งที่ทำให้กระบวนการรู้จำตัวอักษรผิดพลาดหรือไม่สามารถรู้จำได้ งานวิจัยนี้อยู่ในส่วนหนึ่งของกระบวนการแยกภาพตัวอักษร โดยขั้นตอนแรกจะต้องทำการหากรอบตัวอักษร โดยใช้วิธีการหาโครงร่างฮิสโตแกรมและการหาขอบภาพ ทำให้ได้ตำแหน่งและขอบเขตของภาพข้อมูลตัวอักษรแต่ละตัว ภายในกรอบตัวอักษรอาจพบทั้งภาพตัวอักษรขาดและไม่ขาด ดังนั้นหลังจากหากรอบตัวอักษรได้แล้ว จะต้องนำกรอบตัวอักษรไปวิเคราะห์ว่าภาพตัวอักษรในกรอบขาดหรือไม่ โดยทำการส่งขอบเขตของภาพตัวอักษรแต่ละตัวไปวิเคราะห์การขาด ซึ่งการวิเคราะห์การขาดของตัวอักษรพิจารณาได้จาก 2 หลักการ ดังนี้

1. พิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

2. พิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

หลังจากการวิเคราะห์จะทำให้ทราบว่ากรอบภาพใดขาด จะนำกรอบภาพดังกล่าวเข้าสู่กระบวนการเชื่อมตัวอักษรขาด ซึ่งจะมีวิธีการเชื่อม 2 วิธีการหลักตามลักษณะของการขาด ดังนี้

1. การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

2. การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัว

การเชื่อมตัวอักษรขาดทุกครั้ง จะเริ่มจากการเชื่อมในแบบที่ 1 ก่อน และนำผลลัพธ์ที่ได้ไปทำการเชื่อมในแบบที่ 2 เสมอ การเชื่อมตัวอักษรทั้งสองแบบ มีหลักการย่อย 3 หลักการ ดังนี้

(1) การจับคู่ภาพกรอบตัวอักษร

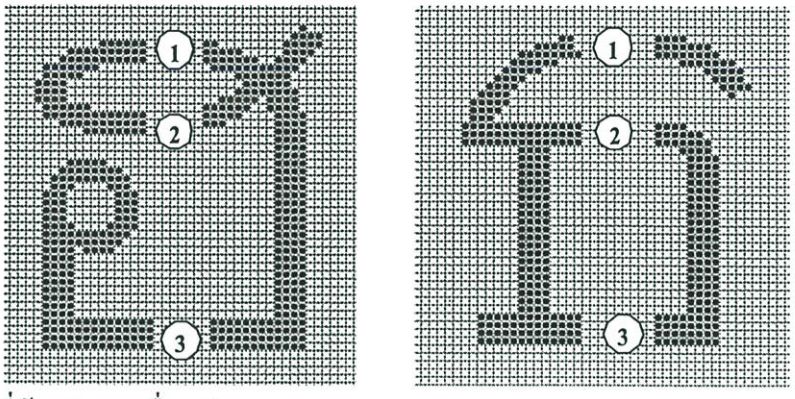
(2) การพิจารณาหาตำแหน่งเชื่อมต่อ

(3) การหาจุดลากเส้นเชื่อมต่อของภาพกรอบตัวอักษรที่จับคู่ได้และการลากเส้นเชื่อมจุดต่อ

ในหลักการย่อยข้อ (2) และ (3) ของการเชื่อมทั้งสองแบบใช้หลักในการพิจารณาเหมือนกัน ส่วนหลักการย่อยข้อ (1) ของการเชื่อมทั้งสองแบบใช้วิธีการพิจารณาแตกต่างกัน

จากผลการทดลองสรุปได้ว่าวิธีการที่นำเสนอเป็นวิธีการที่มีประสิทธิภาพที่ดี และสามารถนำไปใช้กับการเชื่อมต่อตัวอักษรที่ขาดได้จริง เป็นผลทำให้การรู้จำตัวอักษรมีความถูกต้องมากขึ้น

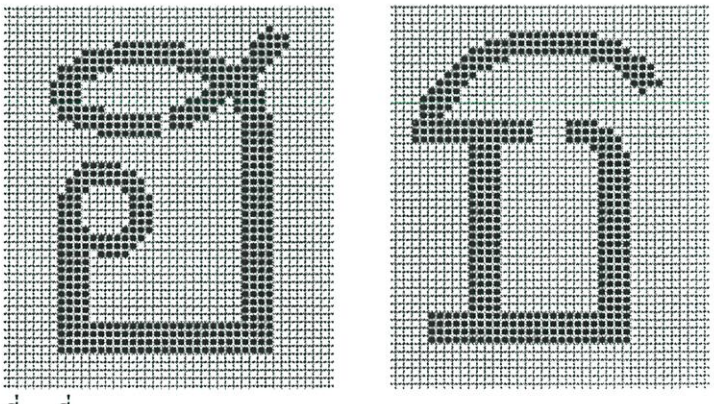
กรณีที่ทำให้การเชื่อมต่อตัวอักษรไม่ครบตามโครงสร้างของตัวอักษรที่พบ คือ กรณีที่ต้องมีการเชื่อมต่อ 3 จุด ดังรูปที่ 6.1 เนื่องจากวิธีการเชื่อมต่อตัวอักษรที่นำเสนอดังกล่าว จะสามารถเชื่อมต่อได้ 2 จุด



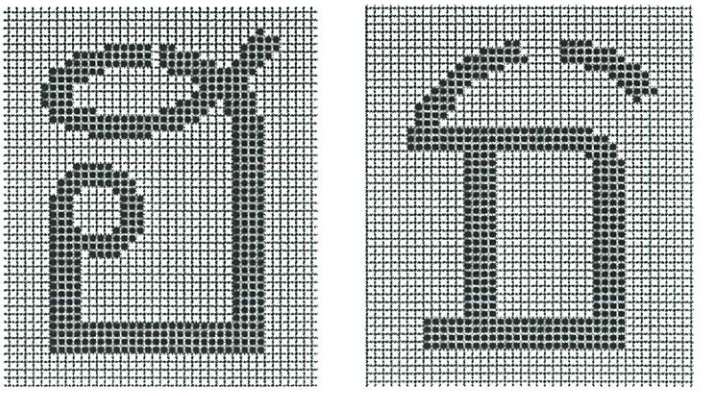
รูปที่ 6.1 ตัวอักษรที่ต้องทำการเชื่อมต่อ 3 จุด

จากรูปที่ 6.1 การเชื่อมต่อจะเกิดขึ้นที่บริเวณส่วนบนและส่วนล่าง การเชื่อมต่อที่บริเวณส่วนบนอาจเกิดการลากเส้นเชื่อมที่บริเวณที่ 1 หรือ 2 เพียงจุดเดียว นั่นคือหากระยะห่างระหว่างจุด 2 จุดของบริเวณใดน้อยกว่า ก็จะทำการลากเส้นเชื่อมต่อในบริเวณนั้น ส่วนการเชื่อมต่อที่บริเวณส่วนล่างจะสามารถลากเส้นเชื่อมต่อได้

การลากเส้นเชื่อมต่อในรูปที่ 6.1 ในกรณีที่ระยะห่างระหว่างจุดที่ 1 น้อยกว่าจุดที่ 2 แสดงดังรูปที่ 6.2 และในกรณีที่ระยะห่างระหว่างจุดที่ 2 น้อยกว่าจุดที่ 1 แสดงดังรูปที่ 6.3



รูปที่ 6.2 กรณีเชื่อมต่อที่จุดที่ 1



รูปที่ 6.3 กรณีเชื่อมต่อที่จุดที่ 2

บรรณานุกรม

[1] อุบลรัตน์ พาชิยานุกูล, วิเชียร เปรมชัยสวัสดิ์, นุชรี เปรมชัยสวัสดิ์. "The 6th National Computer Science and Engineering Conference (NCSEC2002).", 29-31 ตุลาคม 2545. หน้า 19-26.

[2] อุบลรัตน์ พาชิยานุกูล, วิเชียร เปรมชัยสวัสดิ์, นุชรี เปรมชัยสวัสดิ์. "การซ่อมแซมตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด." การประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 25 (EECON-25), 21-22 พฤศจิกายน 2545. หน้า 36-40.

[3] Pongsuree L., Wichian P., Nucharee P. "Reconstruction of Broken Character Images for Thai Character Recognition Systems." International Conference on Digital Image Computing, Techniques and Applications, DICTA'99, Perth, Australia on December, pp. 222-226, 1999

[4] Pongsuree L., Wichian P., Nucharee P. "Repairing Broken Thai Printed Characters Using Feature Extraction", The National computer Science and Engineering Conference, 2000.

[5] Pongsuree L., Wichian P., A. Thammano, S. Narita. "Merged and Broken Printed Thai Character Segmentation." The 1999 International Conference on Artificial Neural Network In Engineering, St. Loius, USA, on November, pp.893-898, 1999.

[6] Nucharee P., Wichian P., and Seinosuke N., "Segmentation of Horizontal and Vertical Touching Thai Character." ITC-CSCC'99 International Technical Conference on Circuit Systems, Computers and Communications, Niigata, Japan.

[7] Wicha P., Somchai J., Prasert C. "Segmentation of Connected Characters Using Distinctive Feature of The Character in Thai Character Recognition System." Electrical Engineering Conference on Circuits and Systems, pp.338-342, 1997.

[8] Shunji, Ching Y. Suen, Kazuhiko Y. "Historical Review of OCR Research and Development" Proceedings of IEEE, vol. 80, 7 Jul 1992.

[9] D.G. Elliman and I.T. Lancaster. "A Review of Segmentation and Contextual Analysis Techniques for Text Recognition." Pattern Recognition, Vol. 23. No. ¾, pp.337-376, 1990.

[10] Rafael C. Gonzalez, and Richard E. Woods, Digital Image Processing, Addison-Wesley, 1993

[11] E. R. Davies, Machine Vision, Academic Press, 1997.

[12] พงษ์สุรีย์ ลิ้มมณีวิจิตร. "การเชื่อมต่อลายเส้นที่ขาดหายไปของอักษรตัวพิมพ์ภาษาไทย." วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2543

ภาคผนวก ก

บทความที่ได้รับการตีพิมพ์ในการประชุมวิชาการทางวิศวกรรมไฟฟ้า ครั้งที่ 25 (EECON-25)

การซ่อมแซมตัวอักษรตัวพิมพ์ภาษาไทยที่ขาด

Repairing Broken of Thai Printed Characters

อุบลรัตน์ พาชิยานุกูล[1] วิเชียร เปรมชัยสวัสดิ์[1] นุชรี เปรมชัยสวัสดิ์[2]
[1]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ถ.ฉลองกรุง เขตลาดกระบัง กรุงเทพฯ 10520
โทร 0-2737-2551-5 ต่อ 530 โทรสาร 0-2326-9074 E-mail:s2067022@kmitl.ac.th, wichian@kmitl.ac.th
[2]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิต
110/1-4 ถ.ประชาชื่น เขตหลักสี่ กรุงเทพฯ 10210
โทร 0-2954-7300-29 ต่อ 215 โทรสาร 0-2954-9605

บทคัดย่อ

ความสมบูรณ์ของตัวอักษรมีผลต่อความถูกต้องของกระบวนการรู้จำตัวอักษร เอกสารที่เป็นอินพุตในกระบวนการรู้จำตัวอักษรอาจพบตัวอักษรขาด ซึ่งเมื่อนำไปประมวลผลแล้วทำให้ไม่สามารถรู้จำได้ ในบทความนี้จึงเสนอวิธีการแก้ไขปัญหาภาพตัวอักษรขาด วิธีการหลักที่ใช้ในการซ่อมแซมตัวอักษรที่ขาดมี 2 วิธีการ คือ (1) ซ่อมตัวอักษรจากการตรวจสอบการเหลื่อมล้ำของกรอบภาพ และ (2) ซ่อมตัวอักษรโดยพิจารณาจากรหัสแทนโครงสร้างตัวอักษรไทยร่วมกับลักษณะเด่นของตัวอักษรไทย ลักษณะเด่นที่พิจารณา คือ ตำแหน่งหัวของตัวอักษร ขาของตัวอักษร และ จุดปลายของตัวอักษร การแก้ปัญหาดังกล่าวจะต้องเริ่มต้นด้วยวิธีที่ (1) และตามด้วยวิธีที่ (2) เสมอ จากการทดลองซ่อมตัวอักษรที่เป็นตัวอักษรปกติและตัวหนา และทดสอบด้วยซอฟต์แวร์ประยุกต์ที่มีขายในปัจจุบัน พบว่าวิธีที่นำเสนอนี้สามารถซ่อมแซมตัวอักษรและรู้จำได้ถูกต้อง

Abstract

Completeness of character affects to the accuracy of character recognition systems. Some input documents may consist of broken characters. When the document is passed to the recognition process, the broken characters cannot be recognized at all. This paper presents a scheme to solve the problem of broken characters. There are two repairing techniques used in the proposed scheme as follows: 1) using overlapping area of the image block in x-axis and 2) using specific characteristics of Thai characters together with character structure code. The specific characteristics of Thai characters that employed in the scheme are position of the head, leg and endpoint of image character. Each character is passed through the two processes. The experiment results that the scheme can repair broken characters both in normal and bold fonts efficiently. The commercial available software can recognized the repaired characters correctly.

Keywords: broken characters, overlapping area, character structure code, specific characteristics of Thai character

1. คำนำ

การขาดของตัวอักษรมีได้หลายสาเหตุด้วยกัน ได้แก่ เอกสารที่เกิดจากการถ่ายเอกสาร เอกสารที่ได้จากเครื่องพิมพ์เลเซอร์ หรือเอกสารที่มีการพับในขณะพิมพ์ เอกสารที่พบการขาดในลักษณะดังกล่าว เป็นผลให้กระบวนการรู้จำเกิดความผิดพลาดหรือไม่สามารถรู้จำได้

ในระบบรู้จำตัวอักษร ความถูกต้องของกระบวนการรู้จำตัวอักษรขึ้นกับข้อมูลภาพตัวอักษรที่ครบถ้วนสมบูรณ์ ในงานวิจัยที่ผ่านมาได้มีการกล่าวถึง การแก้ไขตัวอักษรที่ติดกัน[1] การลดสัญญาณรบกวน[2] การทำขอบตัวอักษรให้เรียบ[3] และการซ่อมแซมตัวอักษรตัวพิมพ์ไทยที่ขาดโดยใช้การพิจารณาลักษณะเด่นของตัวอักษร[4] ในงานวิจัย [4] กล่าวถึงเฉพาะการซ่อมตัวอักษรที่มีการขาดในลักษณะแนวนอน การขาดของตัวอักษรตรวจสอบได้จากการเหลื่อมล้ำกันของกรอบภาพ ซึ่งจากการตรวจสอบด้วยวิธีการดังกล่าว พบว่าอาจเป็นการขาดของตัวอักษรที่เกิดขึ้นในลักษณะแนวตั้งได้เช่นกัน ดังรูปที่ 1 ก กระบวนการซ่อมตัวอักษรในงานวิจัย [4] ไม่สามารถซ่อมตัวอักษรที่ขาดในลักษณะแนวตั้งได้ ดังนั้นวิธีการในงานวิจัย [4] จึงไม่เพียงพอในการเตรียมความพร้อมก่อนที่จะนำไปสู่กระบวนการรู้จำตัวอักษร

บทความนี้ได้เสนอกระบวนการใหม่ในการซ่อมแซมตัวอักษรที่มีการขาดในลักษณะแนวตั้ง โดยนำข้อมูลภาพมาผ่านกระบวนการปรับปรุงภาพในกรณีที่ข้อมูลภาพมีสัญญาณรบกวน จากนั้นทำการแยกบรรทัดและแยกตัวอักษรแต่ละตัวออกจากบรรทัดโดยใช้วิธีการหาโครงร่างฮีสโตรแกรมและการหาขอบภาพ [2] ในขั้นตอนนี้จะได้ตำแหน่งและขอบเขตของภาพข้อมูลตัวอักษรแต่ละตัว เรียกว่า กรอบภาพ ภายใน

กรอบภาพอาจเป็นภาพตัวอักษรที่ขาดหรือไม่ขาด ดังรูปที่ 1 ข ดังนั้น หลังจากหากรอบภาพตัวอักษรได้แล้ว จะนำไปตรวจสอบตัวอักษรขาด และซ่อมแซมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพและซ่อมแซมตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นของโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร และในส่วนสุดท้ายกล่าวถึงผลการทดลองและสรุป

ก) ข้อมูลภาพตัวอักษรขาด ข) กรอบภาพตัวอักษร

รูปที่ 1 ข้อมูลภาพตัวอักษรก่อนและหลังการหาขอบภาพ

2. การซ่อมแซมตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพ

2.1 การตรวจสอบตัวอักษรขาดจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

วิธีนี้จะตรวจสอบกรอบภาพ 2 กรอบภาพ โดยคู่ลำดับของจุดปลายกรอบภาพที่ 1 อยู่ภายในกรอบภาพที่ 2 ดังสมการ (1) หรือพิจารณาในกรณีกลับกัน

$$\text{overlap} = \left(x_1\,\max \le x_2\,\max\right) \text{and} \left(x_1\,\max \ge x_2\,\min\right) \quad (1)$$

overlap : เป็นค่าความจริงทางตรรกศาสตร์ โดยมีค่าเป็น True หรือ False

$\left(x_1\,\min, x_1\,\max\right)$: เป็นจุดปลายในแนวแกน x ของกรอบภาพที่ 1

$\left(x_2\,\min, x_2\,\max\right)$: เป็นจุดปลายในแนวแกน x ของกรอบภาพที่ 2

จากสมการ (1) overlap = True แสดงว่ากรอบภาพทั้งสองมีการเหลื่อมล้ำกันในแนวแกน x และเป็นกรอบภาพของตัวอักษรขาด หาก overlap = False แสดงว่ากรอบภาพทั้งสองไม่เหลื่อมล้ำกันและในขั้นตอนนี้สรุปได้ว่าเป็นตัวอักษรปกติ (ไม่ขาด)

การเหลื่อมล้ำของกรอบภาพในแนวแกน x มี 2 ลักษณะ ดังนี้

2.1.1 กรอบภาพซ้อนทับกัน ดังรูปที่ 2 สามารถส่งเข้าสู่กระบวนการซ่อมตัวอักษรได้ทันที

2.1.2 กรอบภาพไม่ซ้อนทับกัน ดังรูปที่ 3 นำเข้าสู่กระบวนการวิเคราะห์ตัวอักษรขาดในขั้นตอนถัดไป

ตัวอักษรขาด กรอบภาพ พิจารณาการเหลื่อมล้ำ

รูปที่ 2 การเหลื่อมล้ำของกรอบภาพกรณีกรอบภาพซ้อนทับกัน

ตัวอักษรขาด กรอบภาพ พิจารณาการเหลื่อมล้ำ

รูปที่ 3 การเหลื่อมล้ำของกรอบภาพกรณีกรอบภาพไม่ซ้อนทับกัน

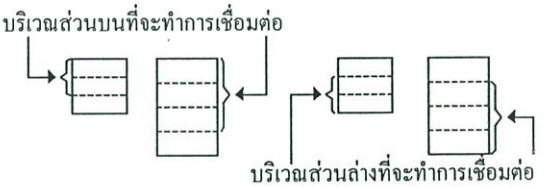
2.2 การเชื่อมตัวอักษรโดยพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x

หลังจากตรวจสอบพบว่าเป็นตัวอักษรขาด จะนำกรอบภาพทั้งสองมาพิจารณาหาตำแหน่งเชื่อมต่อ ซึ่งมี 2 ขั้นตอนย่อย คือ

(1) หากรอบภาพตัวอักษรที่มีขนาดเล็ก ดังรูปที่ 4 กรอบภาพใดมีพื้นที่ขนาดเล็ก จะเป็นกรอบภาพหลัก หากกรอบภาพหลักเป็นส่วนที่ไม่มีตำแหน่งหัว และมีตำแหน่งอยู่เหนือครึ่งหนึ่งของความสูงเฉลี่ยของตัวอักษรในบรรทัดที่กำลังพิจารณา แสดงว่าเป็นการขาดในแนวตั้ง ถ้าไม่ตรงกับเงื่อนไขดังกล่าว จะเป็นการขาดในแนวนอน ทำให้สามารถเชื่อมต่อในแนวนอนตามวิธีการ [4]

(2) หาตำแหน่งเชื่อมต่อในแนวตั้ง โดยนำกรอบภาพหลักไปผ่านกระบวนการหารหัสแทนโครงสร้างตัวอักษร จากนั้นจะนำรหัสแทนโครงสร้างตัวอักษรไปตรวจสอบเงื่อนไข เพื่อให้ทราบว่าจะทำการเชื่อมต่อ ณ บริเวณส่วนบน หรือส่วนล่าง หรือมีการเชื่อมต่อเกิดขึ้นทั้งสองส่วนของกรอบภาพ ดังรูปที่ 5

รูปที่ 4 กรอบภาพที่มีขนาดเล็กของตัวอักษร ณ ขาด



รูปที่ 5 บริเวณส่วนบนและส่วนล่างที่จะทำการเชื่อมต่อ

การลากเส้นเชื่อมต่อภาพกรอบตัวอักษร พิจารณาจากจุดที่อยู่ ณ บริเวณส่วนบน หรือส่วนล่าง บริเวณดังกล่าวจะมีจุดปลายของแต่ละกรอบภาพหลายจุด การพิจารณาว่าจะเลือกลากเส้นเชื่อมต่อจากจุดใดไปสิ้นสุดที่จุดใดนั้น จะใช้วิธีการหาระยะทางที่สั้นที่สุดระหว่างจุด 2 จุด คือ $P_i = \left(x_i, y_i\right)$ และ $P_j = \left(x_j, y_j\right)$ ดังสมการ (2)

$$D_E\left(P_i, P_j\right) = \min\left(\sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2}\right) \quad (2)$$

D_E : ระยะห่างระหว่างจุด 2 จุด

P_i : คู่ลำดับของจุดปลายของกรอบภาพที่ 1

P_j : คู่ลำดับของจุดปลายของกรอบภาพที่ 2

i : ลำดับของจุดปลายของกรอบภาพที่ 1 โดยที่ i = 1, 2, 3,..., m

j : ลำดับของจุดปลายของกรอบภาพที่ 2 โดยที่ j = 1, 2, 3,..., n

m : จำนวนจุดปลายของกรอบภาพที่ 1

n : จำนวนจุดปลายของกรอบภาพที่ 2

ดังนั้นจะต้องหาระยะทางทั้งหมด $m \times n$ ครั้ง เมื่อได้ระยะทางที่สั้นที่สุดแล้ว นำระยะทางมาพิจารณาดังนี้

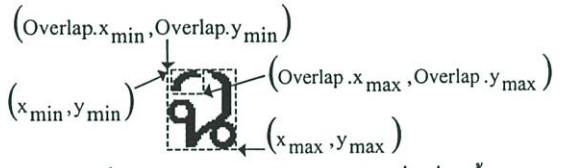
(1) ตำแหน่งที่ลากเส้นเชื่อมต่อเป็นบริเวณส่วนบน จะพิจารณาจากระยะห่างระหว่างจุด 2 จุดในแนวแกน x จะต้องน้อยกว่าหรือเท่ากับ 1/2 ของความกว้างของกรอบภาพ และระยะห่างระหว่างจุด 2 จุดในแนวแกน y จะต้องน้อยกว่าหรือเท่ากับ 1/5 ของความสูงกรอบภาพ

(2) ตำแหน่งที่ลากเส้นเชื่อมต่อเป็นบริเวณส่วนล่าง จะพิจารณาว่า ระยะห่างระหว่างจุด 2 จุดในแนวแกน x จะต้องน้อยกว่า1/2 ของความกว้างกรอบภาพ และระยะห่างระหว่างจุด 2 จุดในแนวแกน y จะต้องน้อยกว่า 1/4 ของความสูงกรอบภาพ

ทั้งสองกรณีจะต้องมีระยะทางมากกว่าศูนย์ หลังจากที่พิจารณาตามเงื่อนไขสองข้อนี้แล้วเป็นจริง จึงจะทำการลากเส้นเชื่อมทั้งสองกรอบภาพ

ก. การเชื่อมต่อที่บริเวณส่วนบน

กำหนดพิกัดกรอบตัวอักษรที่ได้จากการประมวลผลภาพเบื้องต้นเป็น (x_{min}, y_{min}) และ (x_{max}, y_{max}) ส่วนที่เหลื่อมล้ำที่ได้จากการวิเคราะห์การขาดของตัวอักษรกำหนดเป็น $(Overlap.x_{min}, Overlap.y_{min})$ และ $(Overlap.x_{max}, Overlap.y_{max})$ ดังรูปที่ 6



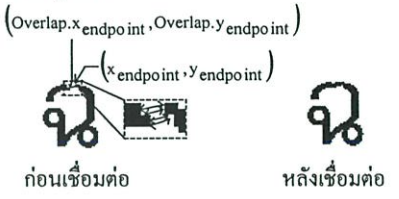
รูปที่ 6 ขอบเขตของตัวอักษรและส่วนที่เหลื่อมล้ำ

การเชื่อมต่อจะยึดส่วนที่เหลื่อมล้ำเป็นหลัก คือ เริ่มจากจุดปลายของส่วนที่เหลื่อมล้ำ $(Overlap.x_{endpoint}, Overlap.y_{endpoint})$ ลากไปยังจุดปลาย $(x_{endpoint}, y_{endpoint})$ เส้นตรงที่ได้นี้จะเรียก เส้นหลัก จากนั้นจะลากเส้นตรงเหนือและต่ำกว่าเส้นหลัก โดยลากให้เส้นตรงทุกเส้นขนานกัน (ความชัน = m) ดังสมการ (3) และดังรูปที่ 7

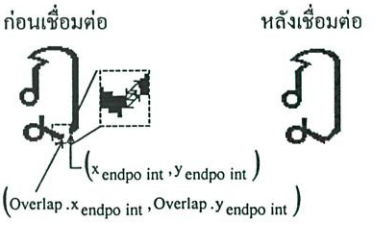
$$m = \frac{Overlap.y_{endpoint} - y_{endpoint}}{Overlap.x_{endpoint} - x_{endpoint}} \qquad (3)$$

ข. การเชื่อมต่อที่บริเวณส่วนล่าง

การเชื่อมต่อบริเวณส่วนล่างจะทำเช่นเดียวกับการเชื่อมต่อบริเวณส่วนบน ดังรูปที่ 8



รูปที่ 7 การเชื่อมต่อจุดขาดบริเวณส่วนบน



รูปที่ 8 การเชื่อมต่อจุดขาดบริเวณส่วนล่าง

3. การซ่อมแซมตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นของโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

3.1 การตรวจสอบตัวอักษรขาดจากความสัมพันธ์ของลักษณะเด่นของโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

ขั้นตอนนี้นำกรอบภาพตัวอักษรที่ผ่านการซ่อมแซมมาผ่านกระบวนการหาลักษณะเด่นของโครงสร้างตัวอักษรไทย ได้แก่ การหาตำแหน่งหัวของตัวอักษร การหาจุดปลายของตัวอักษร การหาขาของตัวอักษร จากนั้นจะนำลักษณะเด่นดังกล่าวมาพิจารณาตรวจสอบตัวอักษรขาดร่วมกับรหัสแทนโครงสร้างตัวอักษร

ตำแหน่งหัวของตัวอักษรหาได้จากการแบ่งพื้นที่กรอบภาพออกเป็น 4 ส่วน โดยการแบ่งจากจุดกึ่งกลางด้านกว้างและจุดกึ่งกลางด้านยาวของกรอบภาพ ดังรูปที่ 9

ก) ตำแหน่งหัวของตัวอักษร ข) ตัวอักษรมีหัวในตำแหน่ง 1 และ 4

รูปที่ 9 การหาตำแหน่งหัวของตัวอักษร

การหารหัสแทนโครงสร้างตัวอักษร เริ่มจากการนำตัวอักษร 1 กรอบภาพไปผ่านกระบวนการทำให้บาง จากนั้นทำการหาจุดสแกน (x, y) ซึ่งหาได้จากการพิจารณาตำแหน่งหัวของตัวอักษร คือ

Head(0,4) ใช้สมการ (4) Head(1,2) ใช้สมการ (5) และ Head(3) ใช้สมการ (6) สแกนจากจุดนี้ในทิศทางซ้าย บน ขวาและล่าง ตามลำดับ หากสแกนแล้วพบส่วนที่เป็นเนื้อตัวอักษรจะแทนด้วยเลข 1 ถ้าไม่พบจะแทนด้วยเลข 0 ดังรูปที่ 10

$$(x,y)=\left(\frac{1}{2}\text{Block.width},\frac{1}{2}\text{Block.height}\right) \quad (4)$$

$$(x,y)=\left(\text{Head.xMax},\frac{(\text{Head.yMax}+\text{Block.yMax})}{2}\right) \quad (5)$$

$$(x,y)=\left(\text{Head.xMax},\frac{1}{2}\text{Block.height}\right) \quad (6)$$

Head(0,4) : ตัวอักษรที่ไม่พบหัวทุกตำแหน่งหรือพบหัวในตำแหน่งที่ 4

Head(1,2) : ตัวอักษรที่พบหัวในตำแหน่งที่ 1 หรือตำแหน่งที่ 2

Head(3) : ตัวอักษรที่พบหัวในตำแหน่งที่ 3

Block.width : ความกว้างของกรอบภาพ

Block.height : ความสูงของกรอบภาพ

(Head.xMax, Head.yMax) : คู่ลำดับของจุดปลายของหัวตัวอักษร

(Block.xMax, Block.yMax) : คู่ลำดับของจุดปลายของกรอบภาพ

กรณี Head(1,2) จะต้องนำขาของตัวอักษรมาพิจารณาร่วมด้วย กล่าวคือ หากเป็นกรอบภาพของตัวอักษรที่มีขาตั้งแต่สองขาขึ้นไป จะทำการเปลี่ยนจุดสแกน ดังสมการที่ (7) และ (8) เพื่อให้ได้รหัสแทนตัวอักษรที่เหมาะสมกับโครงสร้างตัวอักษร

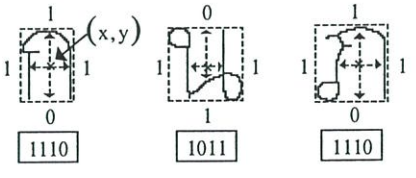
$$(x,y)=\left(\text{Leg}_{\text{average}},\frac{(\text{Head.yMax}+\text{Block.yMax})}{2}\right) \quad (7)$$

$$\text{Leg}_{\text{average}}=\frac{(\text{Leg}_1.x+\text{leg}_2.x)}{2} \quad (8)$$

$\text{Leg}_{\text{average}}$: ค่าเฉลี่ยของระยะห่างระหว่างขาตัวอักษร

$\text{Leg}_1.x$: ค่า x ของตำแหน่งขาที่ 1

$\text{Leg}_2.x$: ค่า x ของตำแหน่งขาที่ 2

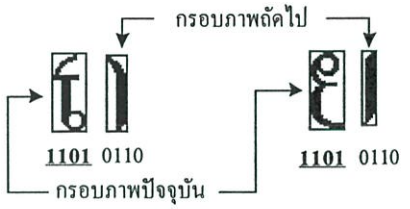


รูปที่ 10 การหารหัสแทนโครงสร้างตัวอักษร

3.2 การเชื่อมตัวอักษรโดยพิจารณาจากความสัมพันธ์ของลักษณะเด่นโครงสร้างตัวอักษรไทยร่วมกับรหัสแทนโครงสร้างตัวอักษร

นำกรอบภาพตัวอักษรขาดมาวิเคราะห์การจับคู่ภาพกรอบตัวอักษร ขั้นตอนนี้จะนำรหัสแทนโครงสร้างตัวอักษรมาวิเคราะห์เพื่อให้ทราบว่าจะนำกรอบภาพปัจจุบันไปเชื่อมกับกรอบภาพก่อนหน้าหรือกรอบภาพถัดไป เช่น รหัส "1101" และรหัส "0111" เป็นต้น

3.2.1 รหัส "1101" ตรวจสอบความสูงของกรอบภาพหากมีค่าน้อยกว่าหรือเท่ากับความสูงเฉลี่ยของบรรทัดที่กำลังพิจารณา ให้ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 11

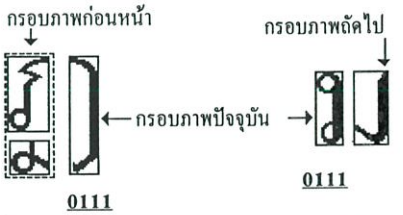


รูปที่ 11 การจับคู่กรอบภาพจากการวิเคราะห์รหัส "1101"

3.2.2 รหัส "0111" ตรวจสอบเงื่อนไข ดังนี้

ก. ไม่พบหัวในตำแหน่งที่ 1 (ฎ) การจับคู่ คือ ดึงกรอบภาพก่อนหน้ามาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 12

ข. ไม่ตรงกับเงื่อนไขในข้อ ก (ม) การจับคู่ คือ ดึงกรอบภาพถัดไปมาต่อกับกรอบภาพปัจจุบัน ดังรูปที่ 12



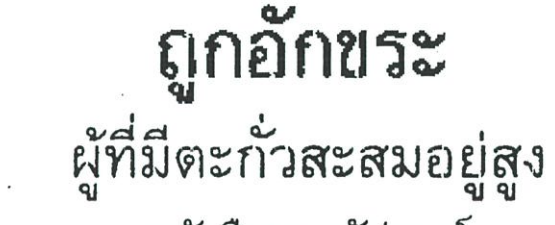
รูปที่ 12 การจับคู่กรอบภาพจากการวิเคราะห์รหัส "0111"

การพิจารณาหาตำแหน่งเชื่อมต่อ และการหาจุดลากเส้นเชื่อมต่อของกรอบภาพตัวอักษรที่จับคู่ได้ ใช้หลักการเดียวกับการเชื่อมตัวอักษรที่ขาดเนื่องจากพิจารณาจากการเหลื่อมล้ำของกรอบภาพในแนวแกน x ดังที่ได้กล่าวมา

4. ผลการทดลอง

วิธีการที่เสนอนี้ได้ทำการทดลองกับข้อมูลตัวอักษรที่ได้มาจากการสแกนภาพที่ความละเอียด 600 จุด ข้อมูลภาพที่ใช้ในการทดสอบ

มีหลายรูปแบบ ได้แก่ ตัวหนา ตัวบาง และขนาดที่แตกต่างกัน หลังจากทำการเชื่อมต่อตัวอักษรแล้ว ได้นำภาพตัวอักษรที่ได้ไปทดสอบกับซอฟต์แวร์รู้จำตัวอักษร ThaiOCR Version 1.5 และ ArnThai Version 2.0 ดังตารางที่ 1 และ 2 ตามลำดับ โดยการทดสอบจะนำทั้งภาพตัวอักษรที่ขาดและภาพตัวอักษรที่ได้หลังจากการซ่อมเป็นอินพุตให้กับซอฟต์แวร์ดังกล่าว เพื่อทำการทดสอบว่ากรณีที่ข้อมูลภาพไม่สมบูรณ์ซอฟต์แวร์จะทำการรู้จำได้หรือไม่มากน้อยเพียงไร และหลังจากข้อมูลภาพผ่านกระบวนการเชื่อมต่อตัวอักษรขาดแล้วจะได้ผลการทดลองที่ดีกว่าเดิมอย่างไร ตัวอย่างข้อมูลภาพที่ใช้ในการทดสอบ ดังรูปที่ 13 และผลที่ได้จากการเชื่อมต่อตัวอักษรแสดงดังรูปที่ 14



รูปที่ 13 ตัวอย่างข้อมูลภาพตัวอักษรขาดที่ใช้ในการทดสอบ

รูปที่ 14 ตัวอย่างข้อมูลภาพตัวอักษรที่ใช้ในการทดสอบหลังการเชื่อมต่อ

ตารางที่ 1 ผลการทดสอบด้วยโปรแกรม ThaiOCR ในการรู้จำตัวอักษร

ข้อความทดสอบ	ผลการทดลองด้วยโปรแกรม ThaiOCR	
	ข้อมูลภาพ ตัวอักษรขาด	หลังซ่อมตัวอักษร
ถูกอักขระ	ถู่กe■fl'ระ	ถูกอักษระ
ผู้ที่มีตะกั่วสะสม อยู่สูง	4ู้ที่3ปีตะกั่วJสะสม Pe)■ู่สูะร	ผู้ที่มีตะกั่วสะสม อยู่สูง
หนังสือราย สัปดาห์	ร ขัรสือระายะสัรj พาาข์	หนังสือราย สัปดาห์

ตารางที่ 2 ผลการทดสอบด้วยโปรแกรม ArnThai ในการรู้จำตัวอักษร

ข้อความทดสอบ	ผลการทดลองด้วยโปรแกรม ArnThai	
	ข้อมูลภาพ ตัวอักษรขาด	หลังซ่อมตัวอักษร
ถูกอักขระ	[ıกE าf)t'ร	ถูกอักๆระ
ผู้ที่มีตะกั่วสะสม อยู่สูง	■ู่ที่@Jตะโ■ศ3สะสม M E)Ei■สู\'	ผู้ที่มีตะกั่วสะสม อยู่สูง
หนังสือราย สัปดาห์	'จชL\ๆสือร" ย๔ สั]lj จ?าา\\	หนังสือราย สัปดาห์

5. สรุป

งานวิจัยนี้ได้นำเสนอแนวทางเพื่อแก้ปัญหาภาพตัวอักษรขาดที่อาจเกิดขึ้นได้ทั้งแนวนอนและแนวตั้ง ซึ่งเป็นสาเหตุหนึ่งที่ทำให้กระบวนการรู้จำตัวอักษรเกิดข้อผิดพลาดหรือไม่สามารถรู้จำได้ หากวิเคราะห์ด้วยบทความนี้พบว่าภาพตัวอักษรมีการขาดในแนวนอนจะซ่อมแซมด้วยงานวิจัย [4] และในแนวตั้งจะใช้วิธีการในบทความนี้ เมื่อทำการทดสอบกับซอฟต์แวร์รู้จำตัวอักษร ซอฟต์แวร์ดังกล่าวสามารถรู้จำตัวอักษรที่ผ่านการซ่อมได้ถูกต้อง จากผลการทดลองสรุปได้ว่า วิธีการที่นำเสนอนี้สามารถนำไปใช้เชื่อมตัวอักษรขาดได้จริงและมีประสิทธิภาพเป็นผลทำให้การรู้จำตัวอักษรมีความถูกต้องมากขึ้น

เอกสารอ้างอิง

[1] Nucharee Premchaiswadi, Wichian Premchaisawadi, and Seinosuke Narita, "Segmentation of Horizontal and Vertical Touching Thai Character", *ITC-CSCC'99 International Technical Conference on Circuit Systems, Computers and Communications*, Niigata, Japan.

[2] Rafael C. Gonzalez, and Richard E. Woods, Digital Image Processing, Addison-Wesley, 1993

[3] E. R. Davies, Machine Vision, Academic Press, 1997.

[4] Pongsuree Limmaneewichid, Wichian Premchaisawadi, and Nucharee Premchaisawadi, "Repairing Broken Thai Printed Characters Using Feature Extraction", *The National computer Science and Engineering Conference*, 2000.

ประวัติผู้เขียน

นางสาวอุบลรัตน์ พาชิยานุกูล เกิดเมื่อวันที่ 3 เมษายน 2518 ที่จังหวัดอุบลราชธานี สำเร็จการศึกษาคณะวิทยาศาสตร์ (วิทยาการคอมพิวเตอร์) จากมหาวิทยาลัยขอนแก่น ปีการศึกษา 2540

ปี พ.ศ. 2540 – 2545 เป็นอาจารย์ประจำภาควิชาคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร

ปี พ.ศ. 2545 – ปัจจุบัน เป็นอาจารย์ประจำภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศศาสตร์ มหาวิทยาลัยเทคโนโลยีมหานคร