การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่สัมผัสกัน

AN ANALYSIS AND SEGMENTATION OF TOUCHING THAI PRINTED CHARACTERS

จักริน สุขสวัสดิ์ชน

JAKKARIN SUKSAWATCHON

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2543

ISBN 974-622-928-1

AN ANALYSIS AND SEGMENTATION OF TOUCHING THAI PRINTED CHARACTERS

JAKKARIN SUKSAWATCHON

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 LADKRABUNG

2000

ISBN 974-622-928-1

หัวข้อวิทยานิพนธ์	การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่สัมผัสกัน
นักศึกษา	นายจักริน สุขสวัสดิ์ชน
รหัสประจำตัว	41067039
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2543
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ. ดร. วิเชียร เปรมชัยสวัสดิ์

บทคัดย่อ

การรู้จำตัวอักษรจะทำกับอักษรเดี่ยวๆ ดังนั้นถ้าตัวอักษรเกิดสัมผัสกันซึ่งหมายถึงว่าจะมีตัวอักษรมากกว่าหนึ่งตัวจะกลายเป็นอินพุทสำหรับกระบวนการรู้จำตัวอักษร จึงเป็นอุปสรรคต่อกระบวนการการรู้จำตัวอักษรไทยด้วยคอมพิวเตอร์ ดังนั้นการแบ่งตัวอักษรไทยที่สัมผัสกันให้เป็นตัวอักษรเดี่ยวๆ เสียก่อนจึงเป็นขั้นตอนที่สำคัญมากขั้นตอนหนึ่งในขบวนการเตรียมข้อมูลก่อนเข้าสู่กระบวนการรู้จำตัวอักษรไทย จึงมีความจำเป็นต้องมีและมีอัลกอริธึมที่สามารถวิเคราะห์และหาจุดที่ใช้แยกตัวอักษรที่สัมผัสกันได้

งานวิจัยนี้จึงต้องการศึกษาและนำเสนอแนวทางในการวิเคราะห์เพื่อหาจุดที่ใช้แบ่งตัวอักษรไทยที่สัมผัสกัน ที่สามารถแยกตัวอักษรที่สัมผัสกันได้ทั้งในแนวตั้งและแนวนอน โดยการนำค่าทางสถิติ คือการหาฐานนิยมของความกว้างตัวอักษรแต่ละตัว เพื่อนำมาใช้เป็นความกว้างมาตรฐาน และใช้ในการตรวจสอบว่าตัวอักษรนั้นเป็นอักษรเดี่ยว หรือตัวอักษรที่สัมผัสกัน ลักษณะเฉพาะของตัวอักษรภาษาไทย และโครงสร้างของประโยคภาษาไทยที่มีหลายระดับซึ่งแตกต่างจากภาษาอังกฤษหรือภาษาอื่น ได้ถูกนำเข้ามาใช้ในวิธีการนี้ เพื่อแยกตัวอักษรออกจากกันเป็นตัวอักษรเดี่ยวๆ โดยวิธีการคำนวณหาจุดที่ใช้แบ่งตัวอักษรโดยการหาค่า Peak-to-Valley โดยใช้สมการที่พัฒนาโดย Kahan และ Pavlidis มาประยุกต์ใช้กับตัวอักษรไทย และพิจารณาร่วมกับค่าโปรเจคชั่น ค่าความกว้างมาตรฐาน และลักษณะเฉพาะของตัวอักษรไทย เพื่อตรวจสอบความถูกต้องของการแบ่งแยกตัวอักษร

Thesis Title	An Analysis and Segmentation of Touching Thai Printed Characters.
Student	Mr. Jakkarin Sukswatchon
Student ID.	41067039
Degree	Master of Science
Programme	Information Technology
Year	2000
Thesis Advisor	Assoc. Prof. Dr. Wichian Premchaiswadi

ABSTRACT

Most of Thai character recognition systems are deal with an isolate single character. So that, touching character means that if is not isolate singled character. It will cause problems in a character recognition process. Thus, the preparation of input data for recognition process is very important. We must have algorithm that could analysis and segment Thai touching characters into isolated characters.

This research will study and presents a new method for analysis and segmentation the touching characters. It could work with both vertically and horizontally touching Thai characters. The proposed segmentation scheme is based on the information of statistical width, the multi-level structure of Thai sentence and characteristics of Thai characters are employed. This method used algorithm for finding the segmentation point proposed by Kahan and Pavlidis (Peak-to-Valley) to developed with Thai characters. It will consider with projection value, standard width, and characteristic of Thai character.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยคำแนะนำและคำปรึกษาเกี่ยวกับระบบการรู้จำตัวอักษร และปัญหาที่เกิดขึ้นกับระบบการรู้จำซึ่งมีผลทำให้ประสิทธิภาพในการรู้จำลดลง จาก รศ.ดร.วิเชียร เปรมชัยสวัสดิ์ ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่านและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณเพื่อนๆ นักศึกษาทุกคน และเพื่อนๆ ในทีมวิทยานิพนธ์ของข้าพเจ้า ที่ได้ช่วยกันแก้ปัญหา ความผิดพลาดด้านต่างๆ โดยเฉพาะอย่างยิ่งนายพงษ์สุรีย์ ลิ้มมณีวิจิตร ที่เป็นผู้นำในการเขียนโปรแกรมและช่วยแก้ปัญหาต่างๆ ได้เป็นอย่างดี จนทำให้โปรแกรมได้เสร็จสมบูรณ์

ขอขอบพระคุณบิดา มารดา และทบวงมหาวิทยาลัย ที่ให้ได้ทุนการศึกษา และทุนสนับสนุนการทำวิทยานิพนธ์ครั้งนี้

สุดท้ายขอขอบพระคุณเจ้าหน้าที่ทุกคนในฝ่ายงานบริการนักศึกษา คณะเทคโนโลยีสารสนเทศ และบัณฑิตวิทยาลัย ที่ได้คอยอำนวยความสะดวกเสมอมา

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

จักริน สุขสวัสดิ์ชน

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง(ต่อ)

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมา และความสำคัญของปัญหา

การเตรียมข้อมูลสำหรับกระบวนรู้จำตัวอักษร (Optical Character Recognition) เป็นขั้นตอนหนึ่งของส่วนจัดการล่วงหน้าของระบบ (Preprocessing Process) ที่มีความสำคัญไปไม่น้อยกว่าส่วนของการรู้จำตัวอักษร จากผลงานวิจัยที่ผ่านๆ มา มุ่งเน้นไปที่การคิดค้นวิธีการรู้จำตัวอักษรให้ได้ประสิทธิภาพและความถูกต้องสูงสุดเป็นส่วนใหญ่ จึงไม่ค่อยได้กล่าวถึงส่วนของการเตรียมข้อมูลสำหรับการรู้จำตัวอักษรมากนัก ซึ่งในการใช้งานจริงของระบบการรู้จำตัวอักษรนั้น มักจะประสบปัญหาในกระบวนการรู้จำตัวอักษร ถ้าข้อมูลที่ส่งเข้ามาเป็นข้อมูลที่ไม่อยู่ในเงื่อนไขที่จะสามารถรู้จำตัวอักษรด้วยระบบการรู้จำที่สร้างขึ้น ดังนั้นเพื่อให้ระบบการรู้จำทำงานได้อย่างมีประสิทธิภาพและถูกต้อง จำเป็นต้องมีขบวนการการทำงานเข้ามาจัดการกับงานในส่วนนี้ โดยจะทำการวิเคราะห์และระบุส่วนประกอบของหน้าเอกสารซึ่งเป็นข้อมูลอินพุท ที่ประกอบด้วยตัวอักษรเรียงต่อกันเป็นคำ ข้อความ ประโยค หรือย่อหน้า และจะต้องมีขบวนการการทำงานที่สามารถแยกตัวอักษรให้ได้เป็นตัวอักษรเดี่ยวๆ (Segmentation) ก่อนส่งต่อไปให้ในขั้นตอนการรู้จำตัวอักษรต่อไป ขั้นตอนการแยกตัวอักษรออกจากภาพของข้อความ เป็นการดึงภาพของตัวอักษรทีละ 1 ตัวอักษรออกจากภาพประโยคของเอกสารในแต่ละบรรทัด ในกรณีของตัวอักษรภาษาไทย จะมีความสลับซับซ้อนมากกว่าภาษาอื่นๆ เนื่องจากแต่ละประโยคของภาษาไทยประกอบด้วย พยัญชนะ สระ และวรรณยุกต์ ถ้าพิจารณาจากตำแหน่งของตัวอักษรเหล่านี้จะเห็นได้ว่ามีระดับต่างกันถึง 4 ระดับ จึงทำให้มีโอกาสที่ตัวอักษรจะสัมผัสหรือซ้อนทับกัน ทั้งในระดับเดียวกันตามแนวนอน (Horizontal) หรือระดับที่ต่างกันตามแนวตั้ง (Vertical) จะเห็นได้ว่าตัวอักษรที่สัมผัสกันนี้เป็นปัญหาสำคัญของกระบวนการรู้จำ ซึ่งจะทำให้ความถูกต้องในการรู้จำลดลง เนื่องจากตัวอักษรมากกว่าหนึ่งตัวที่สัมผัสกัน จะถูกพิจารณาเป็นตัวอักษรเพียงหนึ่งตัวเท่านั้น ซึ่งในปัจจุบันนี้โปรแกรมประยุกต์ที่ใช้ในการรู้จำตัวอักษรไทยนั้น ยังไม่สามารถแก้ปัญหาในส่วนของการสัมผัสกันของตัวอักษรได้ดี ซึ่งจะเห็นได้จากจากผลการทดสอบตัวอย่างของภาพตัวอักษรดังแสดงในรูปที่ 1.1(1)-(4) ซึ่งมีตัวอักษรบางตัวสัมผัสกันอยู่เช่น คำว่า “เพียง” ในรูปที่1.1(1) คำว่า “ปืน” ในรูปที่1.1(3) เป็นต้น ผลการทดสอบกับโปรแกรมอ่านไทย (ArnThai) เวอร์ชัน 1.0 และ ThaiOCR เวอร์ชัน 1.5ได้ผลลัพธ์ดังแสดงในตารางที่ 1.1 จะเห็นได้ว่าโปรแกรมทั้งสองยังไม่สามารถรู้จำตัวอักษรที่สัมผัสกันได้ถูกต้อง

เพียงติดต่อสั่งซื้อสวนกระบองเพชรชุดใดชุด ผู้ร่วมสมทบทุนสามารถตรวจสอบรายชื่อ

(1) (2)

หนังไทยเรื่อง ดอกไม้ในทางปืน ปัญหาราคาน้ำมันขึ้นทำให้คนไทย

(3) (4)

รูปที่ 1.1 ภาพตัวอักษรที่ใช้ทดสอบกับโปรแกรม ArnThai และ ThaiOCR

ตารางที่ 1.1 ผลการทดสอบอักษรไทยที่สัมผัสกันในรูปที่ 1.1 กับโปรแกรม ArnThai และ ThaiOCR

ข้อมูล	Thai OCR 1.5			ArnThai		
รูปที่ 1.1	ผลการรู้จำ	A	B	ผลการรู้จำ	A	B
1	ฒี สั0รี ฒmะ เพ ฑ่ใด9	29	0	ฮี่สั่งง่ฒณะ๖เพฉ,ใด,	29	0
2	ผู ร่ ทบทนmมาร ตรวจ บรายช อ	14	2	ผู้ร่วฒทบฎๆoมารถตรวจ, บรายช่อ	14	2
3	พงไทยเ ง ่กไม1นท	15	0	ฐัง1หยเลง ฯกไม8นทฑฒณ	15	0
4	ปญหาราคาน้ำมันขึ้น ทำให้คนโทย	8	2	ป่หาราคานามันข่นทำให้ คนไทย	8	0

A คือ จำนวนตัวอักษรที่สัมผัสกัน B คือ จำนวนตัวอักษรที่สามารถรู้จำได้อย่างถูกต้อง

จากปัญหาที่กล่าวมาข้างต้น งานวิจัยนี้จึงมุ่งเน้นที่จะทำการศึกษาและวิเคราะห์ถึงลักษณะการสัมผัสกันของตัวอักษรพิมพ์ไทย ตลอดจนหาวิธีการที่ใช้แยกตัวอักษรที่สัมผัสกันออกจากกันให้เป็นตัวอักษรเดี่ยวๆ เพื่อจะได้ข้อมูลอินพุทที่ถูกต้องเข้าสู่กระบวนการรู้จำ อันจะเป็นการเพิ่มประสิทธิภาพให้กับระบบการรู้จำให้มากขึ้นด้วย

1.2 ความมุ่งหมาย และวัตถุประสงค์ของการศึกษา

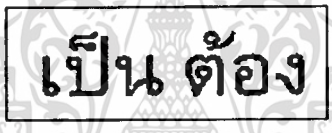
1. เพื่อศึกษาลักษณะการสัมผัสกันของตัวอักษรไทย และวิธีที่ใช้แยกตัวอักษรที่สัมผัสกันเหล่านั้น

2. เพื่อศึกษาปัญหาที่เกิดขึ้นพร้อมทั้งหาแนวทางในการแก้ไข เพื่อนำไปใช้งานได้อย่างมีประสิทธิภาพ
3. เพื่อเป็นแนวทางในการพัฒนาระบบการแยกตัวอักษรที่สัมผัสกันต่อไปในอนาคต
4. เพื่อนำวิธีการที่ศึกษานี้ไปใช้แก้ปัญหาในกระบวนการรู้จำตัวอักษรไทยให้มีประสิทธิภาพมากยิ่งขึ้น

1.3 สมมติฐานของการศึกษา

จากภาพของเอกสารที่ได้จากการสแกน เพื่อที่จะนำเข้าสู่กระบวนการแยกภาพตัวอักษร ให้ได้ตัวอักษรเดี่ยวๆ ออกมานั้นภาพตัวอักษรมีลักษณะต่างๆ กันดังนี้

1. **ตัวอักษรที่มีการเหลื่อมล้ำกัน (Overlap Characters)** คือ การที่บางส่วนของขอบเขตตัวอักษรสองตัวที่อยู่ต่อเนื่องกันมีลักษณะที่เหลื่อมล้ำกัน ดังแสดงในรูปที่ 1.2



รูปที่ 1.2 แสดงตัวอย่างของตัวอักษรที่มีการเหลื่อมล้ำกัน

2. **ตัวอักษรที่มีการสัมผัสกัน (Touching Characters)** คือ การที่บางส่วนของตัวอักษรมีการสัมผัสกัน สามารถเกิดได้ทั้งในระดับเดียวกันและต่างระดับกัน สาเหตุเนื่องมาจากคุณภาพของการสแกน และรูปแบบของตัวอักษร ดังแสดงในรูปที่ 1.3

รูปที่ 1.3 แสดงตัวอย่างของตัวอักษรที่มีการสัมผัสกัน

3. **ตัวอักษรที่มีการซ้อนทับกัน (Crossing Characters)** คือการที่มีการซ้อนทับกันของตัวอักษร ซึ่งในกรณีนี้จะเกิดจากรูปแบบของตัวอักษรเอง ดังแสดงในรูปที่ 1.4

ปัญหา

รูปที่ 1.4 แสดงตัวอย่างของตัวอักษรที่มีการซ้อนทับกัน

สำหรับงานวิจัยนี้ จะพิจารณาเฉพาะการสัมผัสกันของตัวอักษรเท่านั้น ส่วนตัวอักษรที่เหลื่อมล้ำกันจะใช้วิธีการแบ่งระดับของตัวอักษรและการติดตามขอบของตัวอักษรเพื่อแยกออกเป็นตัวอักษรเดี่ยว เมื่อทำการศึกษาในเบื้องต้นพบว่า ความกว้าง และความสูงของตัวอักษรสามารถนำมาใช้พิจารณาระบุถึงการสัมผัสกันของตัวอักษรในรูปแบบต่างๆ ประกอบกับคุณสมบัติทางกายภาพของตัวอักษร ซึ่งน่าจะเป็นแนวทางในการแก้ปัญหาการสัมผัสกันของตัวอักษรได้

1.4 ขอบเขตของการดำเนินงานวิจัย

1. กลุ่มของตัวอักษรที่ใช้เป็นตัวอย่างประกอบด้วยตัวอักษรตัวพิมพ์ภาษาไทย (Thai Printed Character) ที่นิยมใช้กันอยู่ทั่วไป ซึ่งมีแบบอักษร AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC, FreesiaUPC, IrisUPC โดยมีขนาดของตัวอักษรอยู่ระหว่าง 12-16 points
2. การวิจัยนี้ทำการทดสอบกับข้อมูลที่เป็นบทความภาษาไทย เช่น หนังสือพิมพ์, วารสาร, เอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ เป็นต้น
3. ขนาดของตัวอักษรใน 1 บรรทัดที่นำมาวิเคราะห์จะต้องมีขนาดเดียวกัน
4. รูปแบบของตัวอักษรไม่ครอบคลุมถึงตัวเอน (Italic Fonts) และขีดเส้นใต้ ไม่มีภาพ และตาราง
5. ข้อมูลเอกสารไม่ครอบคลุมถึงเอกสารเอียง

1.5 ขั้นตอนของการศึกษา

1. ศึกษาบทความและทฤษฎีต่างๆ ที่มีความเกี่ยวข้องกับงานวิจัยนี้
2. เก็บข้อมูลตัวอย่างจากหนังสือพิมพ์ และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ พร้อมทั้งนำไปทำการสแกนด้วยเครื่องสแกนเนอร์
3. ศึกษาลักษณะการสัมผัสกันของตัวอักษรในแบบต่างๆ

4. เขียนโปรแกรมเพื่อทำการวิเคราะห์การสัมผัสกันของตัวอักษร ว่าเป็นอักษรที่สัมผัสกันหรืออักษรเดี่ยวๆ
5. ทดลองกับข้อมูลตัวอย่าง พร้อมทั้งแก้ไขข้อผิดพลาดของโปรแกรม
6. เขียนโปรแกรมเพื่อทำการหาจุดที่ใช้แบ่งแยกตัวอักษรออกจากกัน ในกรณีที่ตัวอักษรนั้นสัมผัสกัน 2 ตัว หรือมากกว่า
7. ทดลองกับข้อมูลตัวอย่าง พร้อมทั้งแก้ไขข้อผิดพลาดของโปรแกรม
8. รวบรวมผลการทดลองที่ได้จากโปรแกรม
9. วิเคราะห์ข้อผิดพลาดจากการทำงานของอัลกอริธึม และแก้ไขข้อผิดพลาดให้การทำงานมีประสิทธิภาพมากที่สุด
10. สรุปผลที่เป็นไปได้ทั้งหมด พร้อมทั้งทำเอกสารนำเสนอเป็นงานวิจัย

บทที่ 2
งานวิจัยที่เกี่ยวข้อง

ปัญหาการสัมผัสกันของตัวอักษรภาษาไทยยังถือว่าเป็นอุปสรรคต่อกระบวนการรู้จำเป็นอย่างมาก ซึ่งจำเป็นต้องมีกระบวนการแยกตัวอักษรที่สมผัสกันออกจากกันเป็นตัวเดี่ยวๆ เสียก่อน ซึ่งในภาษาอังกฤษมีงานวิจัย [3]-[7] ซึ่งใช้วิธีการต่างๆ เพื่อแยกตัวอักษรภาษาอังกฤษที่สัมผัสกันและได้ผลเป็นอย่างดี แต่ไม่สามารถนำวิธีการที่ใช้สำหรับตัวอักษรภาษาอังกฤษมาใช้กับตัวอักษรภาษาไทยโดยตรงได้ เนื่องจากอักษรภาษาไทยมีลักษณะที่แตกต่าง และซับซ้อนกว่าภาษาอังกฤษ สำหรับภาษาไทยนั้นปัจจุบันนี้ก็มีหลายงานวิจัยที่พัฒนาวิธีการเพื่อแยกตัวอักษรไทยที่สัมผัสกันซึ่งจะกล่าวอธิบายต่อไปในบทนี้ และในท้ายบทที่2 นี้จะสรุปถึงข้อผิดพลาดในแต่ละวิธีการที่นำเสนอมาและแนวทางในการพัฒนาและแก้ไขต่อไป

2.1 การแยกตัวอักษรภาษาไทยที่ติดกันด้วยลักษณะเด่นของตัวอักษร [2]

ในบทความนี้เสนอวิธีการในการตรวจสอบ และแยกตัวอักษรภาษาไทยที่ติดกันโดยใช้ลักษณะเฉพาะของตัวอักษรภาษาไทยเป็นตัวช่วยในการแบ่งกลุ่ม และเป็นเงื่อนไขในการแยกแต่ละกลุ่มออกจากกัน บทความนี้ได้ทำการแบ่งกลุ่มตามลักษณะของตัวอักษรภาษาไทยแต่ละตัวดังแสดงในตารางที่ 2.1 และแยกลักษณะของการติดกันของตัวอักษรออกเป็นอีก 10 กลุ่ม ดังแสดงในตารางที่ 2.2 ซึ่งข้อมูลที่ใช้ในการทดลองได้มาจาก Microsoft Word 6 ซึ่งใช้แบบอักษร 2 แบบ คือ AngsanaUPC และ CordiaUPC ซึ่งมีขนาด 12, 14 และ16 points โดยพิมพ์จาก เครื่องพิมพ์ Laser Jet 5L

ตารางที่ 2.1 แสดงการแบ่งกลุ่มของตัวอักษรไทย

กลุ่ม	ระดับ	ตัวอย่าง
1	ด้านบน	๑ ๘ ๕ ๕ ่ ้ ๗ + ์
2	ด้านบน+ตรงกลาง	ป ฝ ฟ ห ไ ใ ๒ ๖ ๗ ๙
3	ตรงกลาง	ฯ ะ เ า ก ข ฃ ค ต ฆ ง จ ฉ ช ซ ฌ ญ ด ท ธ น บ พ ภ ผ พ ม ย ร ล ว ศ ษ ส ห อ ๐ ๑ ๓ ๔ ๕ ๘
4	ด้านล่าง+ตรงกลาง	ฤ ฤๅ ฯ ญ ฎ ฎ ฐ
5	ด้านล่าง	ุ ู
6	เหนือด้านบน	่ ้ ๗ +

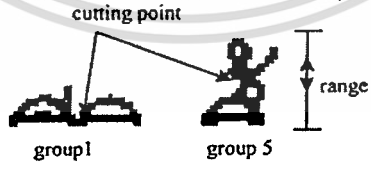
ตารางที่ 2.2 แสดงการแบ่งกลุ่มตามลักษณะการติดกัน

กลุ่ม	กลุ่มอักษรที่ติดกัน	ตัวอย่าง	เปอร์เซ็นต์ที่พบ
1	1 & 1	บัติ ค์กิ	0.2
2	1 & 2 (คอลัมน์เดียวกัน)	ปี ปี	14.5
3	1 & 2 (ต่างคอลัมน์กัน)	ร์ไ ร์ไ	9.2
4	2 & 1 (ต่างคอลัมน์กัน)	ปรี	0.2
5	1 & 6	อี๋	38.1
6	1 & 3	ต้ ศึ ส้	5.9
7	2 & 3	ไม	0.2
8	3 & 5	นุ คู ผุ	31.0
9	3 & 3	รท เภ	0.2
10	เกิน 2 ตัวอักษร	สื่ รูป	0.5

ในบทความนี้ใช้วิธีการวิเคราะห์ลักษณะเด่น (Distinctive Feature) เป็นวิธีการที่ใช้แยกตัวอักษรที่ติดกัน โดยพิจารณาออกเป็น 2 กลุ่ม ดังนี้

1. **กลุ่มที่ 1**

ในกลุ่มนี้มีกลุ่มที่ติดกันเป็นไปได้ 2 กลุ่มย่อย คือ กลุ่มที่ 1 หรือ 5 จากตารางที่ 2.2 เนื่องจากว่าความกว้างของตัวอักษรที่ติดกันในกลุ่มที่ 1 มากกว่าในกลุ่มที่ 5 ซึ่งมีความสูงมากกว่า ดังแสดงในรูปที่ 2.1 ทำให้สามารถใช้ความแตกต่างที่เป็นอัตราส่วนความกว้างต่อความสูงของตัวอักษรเป็นตัวแยกทั้ง 2 กรณีออกจากกัน

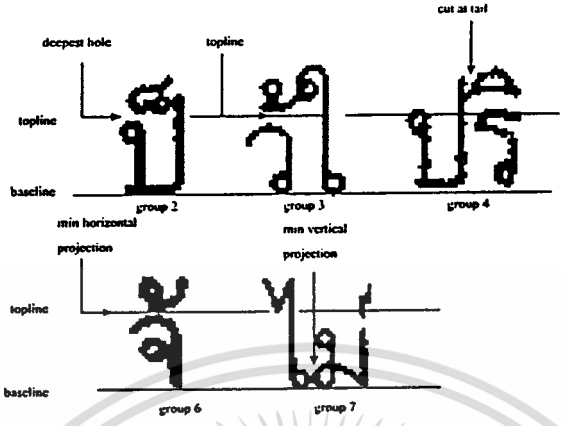


รูปที่ 2.1 แสดงการติดกันของกลุ่มที่ 1 และ 5

จุดตัดของกลุ่มที่ 1 จะอยู่บริเวณตรงกลางระหว่างตัวอักษร และจุดตัดของกลุ่มที่ 5 จะเป็นบริเวณที่น้อยที่สุดของ Horizontal Projection

2. กลุ่มที่ 2

กลุ่มที่ติดกันที่เป็นไปได้คือ กลุ่มที่ 2 3 4 6 หรือ 7 ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 แสดงตัวอย่างการติดกันในกลุ่ม 2, 3, 4, 6 และ 7

ผลการทดลองของวิธีการนี้ได้เปอร์เซ็นต์ความถูกต้องประมาณ 95.6% ซึ่งมีเปอร์เซ็นต์ความผิดพลาดอยู่ 5% เนื่องจากความผิดพลาดดังนี้

1. ความผิดพลาดเนื่องจากความคล้ายคลึงกันจนแยกไม่ออก เช่น ฎ ที่คล้ายกับ ฏ, ฤ กับ ฤ, ฦ กับ ฦ, ป กับ ปั เป็นต้น

2. ความผิดพลาดที่เกิดจากการหาจุดตัดผิดเมื่อมีสัญญาณรบกวน หรือตัวอักษรที่เหลื่อมล้ำกัน (Overlapped Characters) ซึ่งจะทำให้บริเวณที่เป็นจุดตัดนั้นจะไม่ถูกต้อง คือจะมีบางส่วนขาด บางส่วนเกิน

2.2 การวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรพิมพ์ไทยโดยใช้คุณลักษณะทางแนวตั้งและแนวนอนของฮีสโตแกรม [8]

บทความนี้นำเสนอวิธีการวิเคราะห์การติดกันของตัวอักษร และแนวทางการตัดแยกภาพตัวอักษรภาษาไทยในระดับที่นอกเหนือจากระดับกลาง โดยอาศัยระดับของตัวอักษรเพื่อทำการแบ่งประเภทของตัวอักษรที่ติดกัน ทำให้สามารถแบ่งประเภทของตัวอักษรได้ 7 ประเภท จากนั้นใช้คุณสมบัติของฮีสโตแกรมมาวิเคราะห์การติดกัน และการตัดแยกตัวอักษรในแต่ละประเภท

ประเภทที่ 1 หมายถึงตัวอักษรที่มีระดับความสูงอยู่ภายในเส้นแบ่งระดับของอักษรในแต่ละระดับ ดังนั้นจึงไม่มีโอกาสเกิดการติดกันในแนวตั้ง แต่มีโอกาสที่จะติดกันในแนวนอน

ประเภทที่ 2 หมายถึงตัวอักษรที่อยู่ในระดับเหนือบน และระดับบน ดังนั้นประเภทนี้จึงเป็นตัวอักษรที่ติดกันในแนวดิ่ง เพราะไม่มีสระหรือวรรณยุกต์บนที่มีความยาวเกิน 1 ระดับ

ประเภทที่ 3 หมายถึงตัวอักษรที่มีความสูงจากระดับกลางสูงขึ้นไปจนถึงระดับบน ประเภทนี้อาจเป็นได้ทั้งอักษรเดี่ยวเช่น "ป" หรือเป็นพยัญชนะที่ติดกับวรรณยุกต์หรือสระก็ได้

ประเภทที่4 หมายถึงตัวอักษรที่มีระดับความสูงจากระดับกลางยาวลงมาถึงระดับล่าง ประเภทนี้อาจเป็นได้ทั้งอักษรเดี่ยว เช่น "ฦ", "ฦ" หรือเป็นพยัญชนะที่ติดกับสระล่างก็ได้

ประเภทที่ 5 หมายถึงตัวอักษรที่มีความสูงจากระดับกลางสูงขึ้นไปจนถึงระดับเหนือบน ประเภทนี้อาจเป็นอักษรเดี่ยว เช่น "ไ", "โ", "ใ" หรือ อาจเป็นพยัญชนะติดกับสระหรือวรรณยุกต์ก็ได้

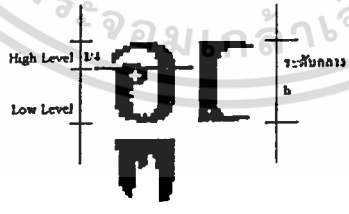
ประเภทที่ 6 หมายถึงตัวอักษรที่มีความสูงตั้งแต่ระดับบนลงมาจนถึงระดับล่าง ประเภทนี้ต้องมีการติดกันของตัวอักษร เพราะไม่มีตัวอักษรที่สูงในระดับนี้

ประเภทที่ 7 หมายถึงตัวอักษรทีมีความสูงจากระดับเหนือบนยาวลงถึงระดับล่าง ประเภทนี้มีการติดกันในระดับเหนือบน เพราะไม่มีตัวอักษรใดๆ ที่มีความสูงในระดับเหนือบน และติดกับระดับล่าง

การกำหนดตัวอักษรที่อยู่ในระดับกลาง เริ่มต้นจากการแบ่งช่วงระดับกลางออกเป็นสองส่วน ดังรูปที่ 2.3 กำหนด h เป็นความสูงของระดับกลาง ช่วงบนมีขนาดความสูงเป็นหนึ่งในสี่ของความสูงระดับกลาง และช่วงล่างมีความสูงเป็นสามในสี่ของความสูงระดับกลาง

HL = 1/4h เมื่อ HL คือระดับความสูง High Level

LL = 3/4h เมื่อ LL คือระดับความสูง Low Level



รูปที่ 2.3 แสดงการแบ่งระดับและกราฟฮีสโตแกรม

จากนั้นทำการหาจำนวนภูเขาของฮีสโตแกรม ที่แสดงถึงแนวเส้นตรงของตัวอักษรทั้งในแนวตั้งและแนวนอนในระดับ HL และ LL ของระดับกลาง ซึ่งได้ผลดังตารางที่ 2.3

ตารางที่ 2.3 แสดงกลุ่มของตัวอักษรระดับกลาง

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

จากนั้น ก็นำเข้าสู่ขั้นตอนการวิเคราะห์การติดกันของตัวอักษร ซึ่งประกอบด้วยขั้นตอนดังนี้

1. ระบุประเภทโอกาสที่จะเกิดการติดกันของตัวอักษรว่าจัดอยู่ในประเภทใดในประเภทที่ 1–7 โดยการเปรียบเทียบตำแหน่งพิกัดของภาพตัวอักษรกับเส้นแบ่งระดับของตัวอักษร
2. ตรวจสอบการติดกันของตัวอักษรแต่ละประเภทว่ามีการติดกันจริงหรือไม่ และกำหนดแนวทางการตัดแยก

การตรวจสอบและแนวทางการตัดแยกตัวอักษร แบ่งได้ตามประเภทดังนี้

- **ประเภทที่ 1** จะไม่มีการติดกันในแนวตั้ง เพราะความสูงไม่เกินกว่าระดับที่อยู่
- **ประเภทที่ 2** ในภาพตัวอักษรที่ได้มีความสูงในระดับบนและเหนือบน จากลักษณะสระและวรรณยุกต์ของภาษาไทยไม่มีตัวใดที่สูงเกินกว่าหนึ่งระดับ ดังนั้นจึงสรุปว่ามีการติดกันในแนวตั้ง และแนวทางการตัดแยกอยู่ที่เส้นแบ่งระดับบนและเหนือบน
- **ประเภทที่ 3** ตัวอักษรนี้จะเป็นตัวอักษรเดี่ยวที่มีความสูงจากระดับกลางถึงระดับบนได้ เช่น “ป” จึงทำการหาจำนวนเส้นทั้งแนวตั้งและแนวนอนของตัวอักษรว่าตกอยู่ในกลุ่มใด ถ้าอยู่ในกลุ่ม V3, V1T3, V2T1, V2T3 ให้ทำการกำหนดจุดตัดในแนวนอน ถ้าตกอยู่ในกลุ่ม V2T2, V2T4 จะต้องนำเข้าสู่กระบวนการรู้จำ ถ้าผลที่ได้เป็น “ข”, “ฃ”, “ย”, “ม”, “น”, “ฆ” ให้กำหนดการตัดตามเส้น Ln2 ข้อผิดพลาดของการวิเคราะห์นี้จะไม่สามารถแยกกรณีที่มีการติดกันของตัวอักษรในกลุ่ม (“บ”, “ป”, “พ”, “ฟ”, “ผ”, “ฝ”) ที่ติดกับไม้เอกได้ ดังรูปที่ 2.4

รูปที่ 2.4 แสดงการวิเคราะห์ที่ผิดพลาดทำให้ได้อักษรเดี่ยว

- **ประเภทที่ 4** ตัวอักษรในระดับกลางมีความสูงลงถึงในระดับล่าง เช่น "ฤ" และ "ฦ" เป็นต้น จึงต้องทำการหาจำนวนแนวเส้นทั้งในแนวตั้งและแนวนอนของภาพตัวอักษรว่าตกอยู่ในกลุ่มใด การระบุกลุ่ม V2T1 ให้นำเข้าสู่กระบวนการรู้จำถ้าผลของการรู้จำได้ตัวอักษรในกลุ่ม ("ถ", "ภ") อยู่ในกลุ่ม V3T3, V1T3, V2T2, V2T3, V2T2 ถ้าผลของการรู้จำอยู่ในกลุ่ม ("ก", "ค", "ศ", "ด", "ต", "ฉ", "ท", "ฑ", "ห") จากนั้นใช้วิธีการสังเกตเช่นเดียวกับข้อ 1, 2 และ 3 ซึ่งการวิเคราะห์นี้ไม่สามารถแยกตัวอักษรระหว่าง "ฦ ฦ" กับ "ภ" ติดกับสระ อู ได้ หรือในกรณี "ฤ" กับ "ถ" สระ อุ ได้
- **ประเภทที่ 5** ในกรณีที่ "ไ", "โ" และ "ใ" มีความสูงถึงระดับเหนือบน ทำให้ไม่สามารถแยกระดับเหนือระดับบนออกไปได้ จึงต้องทำการหาจำนวนแนวเส้นทั้งแนวเส้นตั้งและแนวนอนของภาพว่าตกอยู่ในกลุ่มใด จากนั้นใช้การสังเกตข้อที่ 1 อยู่ในกลุ่ม V3 และข้อที่ 2 อยู่ในกลุ่ม V1T3, V2T2, V2T3 ให้กำหนดแนวเส้นตัดในแนวนอน
- **ประเภทที่** 6 อักษรที่สูงจากระดับบนลงมาจนถึงระดับล่าง เมื่อพิจารณาแล้วไม่พบตัวอักษรในกลุ่มนี้ จึงสรุปได้ว่ามีการติดกันของตัวอักษรในระดับบนหรือระดับล่างอย่างแน่นอน แต่เนื่องจากพยัญชนะไทยมีความสูงเกินระดับกลางได้ทั้งสองระดับ เช่น "ฤ" และ "ป" ทำให้ไม่สามารถระบุแนวการตัดได้ จึงต้องทำการหาจำนวนเส้นในแนวตั้งและแนวนอน และใช้วิธีการสังเกตในข้อที่ 2 ทำการหาเส้นแบ่งแนว
- **ประเภทที่** 7 ตัวอักษรที่มีความสูงถึง 4 ระดับ จากการทดลองไม่พบตัวอักษรในกลุ่มนี้ และไม่มีโอกาสที่ตัวอักษรที่สูงถึง 3 ระดับ จะติดกับสระระดับล่าง เช่น "ไ" ไม่มีโอกาสติดกับสระอุ แสดงว่ามีการติดกันในระดับบนอย่างแน่นอน จึงทำการตัดในระดับเหนือบนได้ และส่วนล่างจะถูกจัดเข้าประเภทที่ 6 เพื่อทำการวิเคราะห์อีกครั้ง

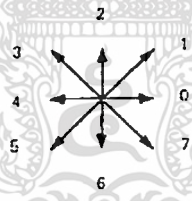
ผลการทดลอง ทดลองกับรูปภาพตัวอักษรจากวารสาร Byte Thailand ผ่านการแสกนด้วยความละเอียด 300 จุดต่อนิ้ว และทำการเลือกเฉพาะบรรทัดที่มีตัวติดกันในระดับบนและล่าง นับจำนวนตัวได้ 4250 ตัว มีตัวอักษรที่ติดกัน 340 ตัวอักษร สามารถวิเคราะห์ได้ถูกต้อง 316 ตัวอักษร คิดเป็น 93% และทำการตัดแยกได้ 292 ตัวอักษร คิดเป็น 86% ซึ่งมีผลของความผิดพลาดแสดงในรูปที่ 2.4

2.3 การแยกสายอักขระตัวพิมพ์ไทยโดยการเข้ารหัสฟรีแมนดัดแปรกับโครงร่างของฮีสโตแกรม [10]

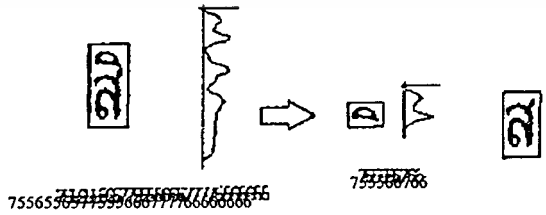
บทความนี้ได้นำเสนอวิธีการที่ทำการแยกแยะตัวอักษรในสายอักขระตัวพิมพ์ไทย โดยการนำโครงร่างของค่าฮีสโตแกรมของกลุ่มตัวอักษรที่อยู่ติดกันมาเข้ารหัสตามทิศทางของฟรีแมน ซึ่งมีวิธีการแก้ปัญหาการติดกันของตัวอักษรดังนี้

บทความนี้ได้ตั้งสมมติฐานที่ว่าตัวอักษรที่ต่างกันควรจะให้ค่าฮีสโตแกรมที่ต่างกันทั้งในแนวตั้งและแนวนอน ดังนั้นตัวอักษร 2 ตัวขึ้นไปที่มาอยู่ติดกันจะทำให้ค่าฮีสโตแกรมที่เป็นลักษณะต่อเนื่องกันของตัวอักษรแต่ละตัวในแนวนอนหรือแนวตั้งขึ้นกับแนวการติดกันของตัวอักษรเหล่านั้น ถ้าทำการเก็บค่าเทมเพลตของตัวอักษรที่ติดกันไว้ในรูปของฮีสโตแกรมโดยตรง จะต้องใช้หน่วยความจำขนาดใหญ่มาก เพราะตัวอักษรไทยมีมากกว่า 90 ตัว ซึ่งมีรูปแบบการติดกันมากมาย จึงต้องทำการเข้ารหัสเพื่อลดขนาดของข้อมูลก่อน

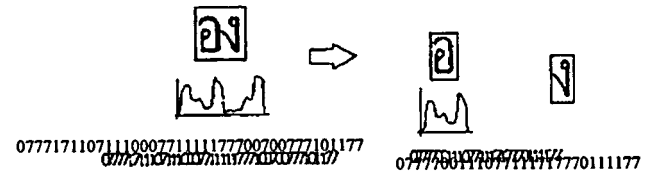
โครงร่างของฮีสโตแกรมจะถูกนำมาปรับให้เป็นการเชื่อมต่อของเส้นตรงสั้นๆ ที่เรียกว่า primitive structure ทิศทางการเชื่อมต่อเหล่านั้นจะเป็นไปตามทิศทางของฟรีแมนดังรูปที่ 2.5 แต่จะใช้เพียง 5 ทิศทางคือ 0, 1, 5, 6 และ 7 ซึ่งจะได้ผลของการเข้ารหัสเป็นดังรูปที่ 2.6 และ 2.7 จากนั้นรหัสโครงร่างของฮีสโตแกรมของกลุ่มตัวอักษรที่จะแยกแยะ จะนำไปเปรียบเทียบกับรหัสโครงร่างต้นแบบโดยให้มีรหัสต่างจากต้นแบบน้อยที่สุด



รูปที่ 2.5 ทิศทางของฟรีแมน



รูปที่ 2.6 ผลการแยกอักษรตามแนวตั้ง



รูปที่ 2.7 ผลการแยกตัวอักษรตามแนวนอน

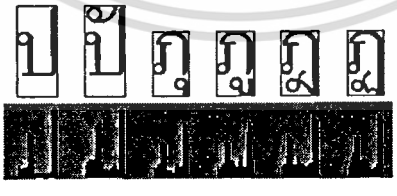
การทดสอบใช้รูปแบบตัวอักษรยูโครเซียยูพีซี ขนาด 18 point และความละเอียด 300 จุดต่อนิ้ว ได้ผลของความถูกต้องประมาณ 92.70 % ซึ่งความผิดพลาดเกิดจาก

1. การที่ตัวอักษรมีลักษณะที่คล้ายคลึงกันเช่น ล กับ ส, ค กับ ศ, ษ กับ น, อิ อี อี อึ เป็นต้น
2. สัญญาณรบกวนจะมีผลทำให้ค่าของรหัสผิดพลาดไป

2.4 ปัญหาที่พบจากงานวิจัยที่ผ่านมา

จากทั้ง 3 งานวิจัยที่นำเสนอมาข้างต้นนั้น เป็นตัวอย่างและแนวทางที่ใช้วิเคราะห์และวิธีการตัดแยกตัวอักษรที่ติดกันได้พอสมควร แต่วิธีการดังกล่าวยังมีข้อผิดพลาดอยู่ ทั้งนี้เนื่องจากทั้ง 3 งานวิจัยนั้น ใช้วิธีการวิเคราะห์โครงร่างของฮีสโตแกรม หรือใช้การพิจารณาลักษณะเด่นของตัวอักษรภาษาไทยเพียงอย่างเดียว จึงทำให้การวิเคราะห์หรือการตัดแยกในบางกรณีนั้นยังทำไม่ได้ จึงสรุปได้ดังนี้

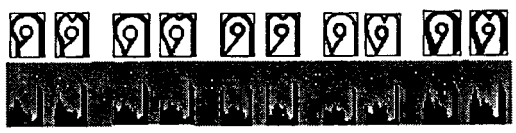
1. การใช้ฮีสโตแกรมวิเคราะห์เพียงอย่างเดียวไม่พอเพียงที่จะใช้ในการตัดสินใจ เนื่องจาก
   - ความคล้ายคลึงกันจากลักษณะของฮีสโตแกรม เช่น "ป" กับ "ป๊" และ "ฎ" กับ "ฏ" หรือ "ฎ" ดังตัวอย่างในรูปที่ 2.8



รูปที่ 2.8 ผลของความผิดพลาดในการใช้ฮีสโตแกรมวิเคราะห์

- ลักษณะโครงสร้างของตัวอักษรภาษาไทยในแต่ละรูปแบบอักษรมีความแตกต่างกัน ซึ่งมีผลทำให้ลักษณะของฮีสโตแกรมที่วิเคราะห์ได้ไม่เหมือนกัน เช่น ในงานวิจัยที่ 2 ใช้ลักษณะของฮีสโตแกรมในการนับจำนวนขาของตัวอักษร ซึ่งจากตัวอย่างรูปที่ 2.9

จะเห็นได้ว่า ด และ ต ในรูปแบบอักษรที่แตกต่างกัน จะมีจำนวนขาของตัวอักษรไม่เท่ากัน บางตัวจะมี 1 ขา และบางตัวจะมี 2 ขา



รูปที่ 2.9 ผลของความผิดพลาดในการใช้ฮีสโตแกรมวิเคราะห์ขาของตัวอักษร

2. การกำหนดจุดตัดของตัวอักษรที่ติดกัน
   - ในงานวิจัยที่ 1 จะใช้แนวที่มีผลรวมของจุดดำในแนวตั้งหรือแนวนอนที่น้อยที่สุด เป็นแนวที่ใช้ตัดตัวอักษรที่ติดกัน ซึ่งแนวคิดจะมีผลความผิดพลาดเกิดขึ้นถ้าจุดที่สัมผัสกันของตัวอักษรไม่เป็นจุดที่มีผลรวมของจุดดำที่น้อยที่สุด หรือจุดที่มีผลรวมของจุดดำที่น้อยที่สุดไม่เป็นจุดที่ตัวอักษรติดกัน
   - ในงานวิจัยที่ 2 ใช้แนวคิดที่คล้ายคลึงกันกับวิธีการที่ 1 แต่มีการนำกระบวนการรู้จำเข้ามาร่วมในการตัดสินใจด้วย ซึ่งอาจทำให้เกิดความผิดพลาดขึ้น และทำให้ยากแก่การตรวจสอบว่าความผิดพลาดที่เกิดขึ้นนั้น เกิดจากกระบวนการรู้จำ หรือเกิดจากวิธีการที่ใช้ในงานวิจัย เพราะว่าระบบการรู้จำแต่ละแบบ ก็ใช้วิธีการในการรู้จำที่แตกต่างกัน
   - ในงานวิจัยที่ 3 ในรหัสโครงร่างต้นแบบมาเปรียบเทียบ เพื่อแยกตัวอักษรแต่ละตัวออกจากกัน ซึ่งเกิดความผิดพลาดมากกว่า 2 งานวิจัยข้างต้น ทั้งนี้เนื่องจาก ไม่สามารถแยกตัวอักษรที่มีลักษณะคล้ายคลึงกันได้ เช่น ล กับ ส เป็นต้น อีกทั้งสัญญาณรบกวนจะมีผลทำให้รหัสที่ใช้ในการวิเคราะห์มีความผิดเพี้ยนไปด้วยเช่นกัน

บทที่ 3
ทฤษฎี และแนวทางที่ใช้

แนวทางการวิเคราะห์ และการหาจุดที่ใช้แบ่งตัวอักษรที่สัมผัสกันแบบต่างๆนั้น จะต้องอาศัยวิธีการหลายวิธีใช้ร่วมกัน เช่น การหาค่าฮีสโตแกรม การวิเคราะห์คุณลักษณะของตัวอักษรภาษาไทย เป็นต้น ซึ่งในบทนี้จะกล่าวถึงทฤษฎี และวิธีการที่ใช้ในงานวิจัยนี้ ดังรายละเอียดต่อไปนี้

3.1 การหาค่าฮีสโตแกรมของภาพตัวอักษร

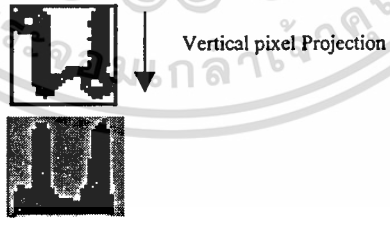
ในงานวิจัยนี้ได้นำวิธีการหาค่าฮีสโตแกรมแบบต่างๆ มาประยุกต์ใช้ให้เหมาะกับตัวอักษรแต่ละแบบของตัวอักษรไทย โดยวิธีการหาค่า Pixel Projection ซึ่งมีวิธีการดังนี้

Pixel Projection [3] เป็นการแสดงค่าจำนวนจุดที่เป็นเนื้อของตัวอักษรในแนวตั้ง (Vertical Projection) และแนวนอน (Horizontal Projection) โดยทำการคำนวณจากสมการ

$$VerticalPXP(x) = \sum_{Y} P(x,y)$$

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

เมื่อ P(x,y) แสดงค่าของจุด ณ.ตำแหน่ง x และ y ผลที่ได้แสดงตัวอย่างดังรูปที่ 3.1



รูปที่ 3.1 แสดงกราฟ Vertical Pixel Projection

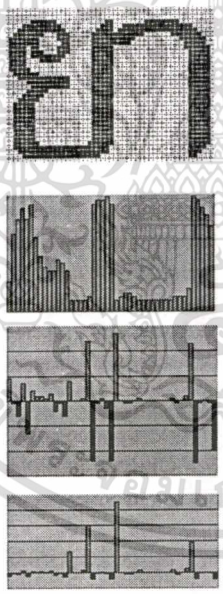
เราสามารถใช้วิธีการหาค่าฮีสโตแกรมตามแนวนอนในการแบ่งแยกบรรทัดของตัวอักษร และใช้ค่าโปรเจคชั่นตามแนวตั้งในการแบ่งตัวอักษรแต่ละตัวออกเป็นตัวเดี่ยวๆ เพื่อนำไปทำการประมวลผลในขั้นตอนต่อไป

3.2 วิธีการวิเคราะห์เพื่อหาจุดที่ใช้แยกตัวอักษรที่ติดกันของตัวอักษรตัวพิมพ์ [1]

สำหรับวิธีการที่เลือกมาใช้ในงานวิจัยนี้ เป็นวิธีการของ Kahan และ Palvidis ซึ่งเขียนเอาไว้ว่า "จุดเชื่อมของ 2 ตัวอักษรจะมีค่าของ Vertical Projection (V(x)) เปลี่ยนแบบ Sharp Minimum และใช้การหาอัตราส่วนระหว่างอนุพันธ์อันดับ 2 คือ V(x-1) – 2V(x) + V(x+1) กับค่าของโปรเจคชั่น เป็นสมการเงื่อนไขในการหาจุดตัด" ดังสมการ

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

จากสมการค่า PV(x) จะมีค่ามาก เมื่อ ค่าโปรเจคชันตามแนวตั้ง V(x) มีค่าน้อย จากตัวอย่างรูปที่ 3.2 (ก) แสดงให้เห็นค่า V(x) ของภาพตัวอักษร รูปที่ 3.2 (ข) แสดงการเปลี่ยนแปลงของ V(x) ในอนุพันธ์อันดับที่ 2 และ รูปที่ 3.2 (ค) แสดงอัตราส่วนระหว่างค่าอนุพันธ์อันดับที่2 กับค่า V(x) จะเห็นได้ว่าค่า PV(x) มีค่ามากบริเวณขาหลังของ ย ซึ่งสัมผัสกับสระอา



(ก)

(ข)

(ค)

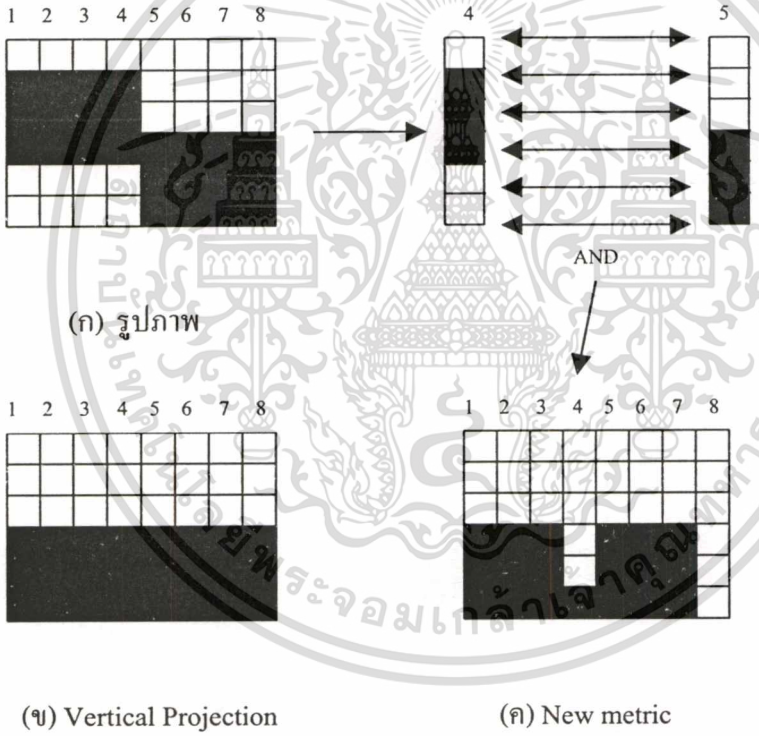
(ง)

รูปที่ 3.2 แสดง (ก) รูปภาพตัวอักษรที่สัมผัสกัน
(ข) ค่าโปรเจคชันตามแนวตั้งของภาพตัวอักษร
(ค) ค่าอนุพันธ์อันดับที่ 2 ของ V(x)
(ง) อัตราส่วนระหว่างอนุพันธ์อันดับที่ 2 กับ V(x)

3.3 การใช้วิธีการ New metric เพื่อกำหนดจุดตัด [1]

โดยทั่วไปการหาจุดตัดของตัวอักษรที่สัมผัสกันอย่างง่ายๆ จะใช้วิธีการหาค่า Pixel Projection เพื่อกำหนดตำแหน่งของจุดตัดจากตำแหน่งที่มีค่าโปรเจคชันที่น้อยที่สุด แต่การใช้วิธีการ New metric ค่าที่คำนวณได้จะบอกถึงนัยสำคัญของการสัมผัสกัน (Degree of contact) ของแต่ละคอลัมน์ที่ติดกัน วิธีการนี้คำนวณโดยการนับจำนวนจุดในแนวตั้งที่ได้ จากการ AND กันของเนื้อภาพในคอลัมน์ที่ติดกัน ดังตัวอย่างในรูปที่ 3.3

จากรูปที่ 3.3 จะเห็นว่าวิธีการ New metric เมื่อแสดงค่าที่ได้จากการ AND กันจะแสดงจุดสัมผัสได้อย่างชัดเจน ในขณะที่วิธีการ Vertical Pixel Projection ไม่สามารถแสดงผลได้ ดังนั้นวิธีการนี้เราสามารถทราบถึงบริเวณที่มีการสัมผัสกันน้อยที่สุด เพื่อกำหนดเป็นตำแหน่งของจุดตัดของตัวอักษรที่ติดกันได้



(ก) รูปภาพ

(ข) Vertical Projection (ค) New metric

รูปที่ 3.3 แสดงนัยสำคัญของการสัมผัสกัน

(ก) รูปภาพต้นแบบ

(ข) ผลจากการทำ Vertical Pixel Projection

(ค) การแก้ปัญหาโดยวิธีการ New metric

บทที่ 4

การวิเคราะห์การสัมผัสกันของตัวอักษรตัวพิมพ์ภาษาไทย และการตัดแยกตัวอักษรที่สัมผัสกัน

การสัมผัสกันของตัวอักษรเกิดจากการที่บางส่วนของตัวอักษรเกิดการสัมผัสกัน ซึ่งอาจจะเกิดจากลักษณะของตัวอักษรเอง หรือเกิดจากสัญญาณรบกวนก็ได้ สามารถเกิดได้หลายรูปแบบ ทั้งในระดับเดียวกัน และต่างระดับกัน และหากพบการสัมผัสกันของตัวอักษรแล้ว วิธีการที่จะใช้หาจุดที่ใช้แบ่งแยกตัวอักษรที่สัมผัสกันออกจากกันนั้นจะต้องใช้วิธีการต่างๆ กัน กับรูปแบบตัวอักษรนั้นๆ ดังนั้นในบทนี้จะกล่าวถึงลักษณะการสัมผัสกันของตัวอักษรในแบบต่างๆ ซึ่งได้จากการสำรวจข้อมูลจากแหล่งข้อมูลต่างๆ เช่น หนังสือพิมพ์ และเอกสารการพิมพ์ เป็นต้น จากนั้นจะกล่าวถึงการวิเคราะห์ว่ามีการสัมผัสกันของตัวอักษรหรือไม่ และสัมผัสกันในรูปแบบใด สุดท้ายจะเป็นวิธีการหาจุดที่ใช้แบ่งแยกด้วยวิธีการต่างๆ ที่กล่าวมาแล้วในบทที่ 3

4.1 ลักษณะของประโยคในภาษาไทย

เมื่อศึกษาลักษณะของประโยคในภาษาไทยพบว่า ประโยคภาษาไทยประกอบด้วยการเรียงกันของพยัญชนะ สระ และวรรณยุกต์ในระดับต่างๆ กัน ซึ่งสามารถแบ่งออกได้เป็น 3 ระดับ ดังต่อไปนี้

- ระดับบน (Upper Zone) ประกอบด้วยสระระดับบน และวรรณยุกต์
- ระดับกลาง (Central Zone) ประกอบด้วย พยัญชนะ และสระในระดับกลาง
- ระดับล่าง (Lower Zone) ประกอบด้วยสระระดับล่าง และบางส่วนของพยัญชนะ

ตัวอย่างแสดงในรูปที่ 4.1

รูปที่ 4.1 แสดงการแบ่งระดับของประโยคในภาษาไทย

เมื่อกำหนดให้

UZ = กลุ่มของตัวอักษรในระดับบน

CZ = กลุ่มตัวอักษรในระดับกลาง

LZ = กลุ่มตัวอักษรในระดับล่าง

to = เส้นบนของระดับบน

up = เส้นแบ่งของระดับกลางกับระดับบน

ba = เส้นแบ่งของระดับกลางกับระดับล่าง

bo = เส้นล่างของระดับล่าง

เมื่อทำการแบ่งระดับของภาพตัวอักษรแล้ว จะต้องทำการแยกภาพของตัวอักษรเพื่อแบ่งแยกออกเป็นตัวอักษรเดี่ยวๆ [12] ดังแสดงตัวอย่างในรูปภาพที่ 4.2

รูปที่ 4.2 กระบวนการแยกภาพตัวอักษรออกเป็นตัวอักษรเดี่ยว

เมื่อทำการแยกเป็นตัวอักษรเดี่ยวๆ แล้ว ภาพของตัวอักษรแต่ละตัวจะมีข้อมูลที่เป็นรายละเอียดของภาพตัวอักษรตัวนั้นๆ ดังต่อไปนี้

- **Xmin, Xmax** เป็นค่าพิกัดตามแนวแกน x ของภาพตัวอักษร
- **Ymin, Ymax** เป็นค่าพิกัดตามแนวแกน y ของภาพตัวอักษร

เมื่อทราบค่าพิกัดของภาพตัวอักษร สามารถจะนำมาวิเคราะห์เพื่อแบ่งกลุ่มของตัวอักษรได้ โดยใช้สมการ

$$X_{cen} = (X_{\min} + X_{\max})/2$$

$$Y_{cen} = (Y_{\min} + Y_{\max})/2$$

และพิจารณาจากเงื่อนไข ต่อไปนี้

CB(i) ∈ upper zone if cp of CB(i) ∈ (to,up-1)

CB(i) ∈ central zone if cp of CB(i) ∈ (up,ba)

CB(i) ∈ lower zone if cp of CB(i) ∈ (ba+1,bo)

เมื่อ

CB คือ Character Block **CP** คือ Central point

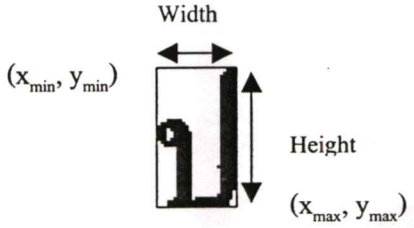
ซึ่งสามารถแบ่งกลุ่มได้ดังตารางที่ 4.1

ตารางที่ 4.1 ตัวอย่างของตัวอักษรภาษาไทยในระดับต่างๆ

กลุ่ม	ระดับ	ตัวอย่าง
1	บน	๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐ ๔
2	กลาง	ก ข ค ต ฆ ง จ ฉ ภ ถ พ ฑ ธ ร น ย บ ล ด เ า ส ว ง ม ท อ ฉ ผ ป ฟ ฝ ฒ พ ฮ ไ โ ใ ฎ ฤ ญ ฐ ฎ ภ
3	ล่าง	ุ ู (ส่วนล่างของ ญ ฐ)

- **Height** เป็นค่าความสูงของภาพตัวอักษร
- **Width** เป็นค่าความกว้างของภาพตัวอักษร
- **Zone** เป็นค่าที่ระบุระดับของตัวอักษร ซึ่งมีค่าเป็น
  - **Upper Zone** คือกลุ่มของตัวอักษรในระดับบน
  - **Central Zone** คือกลุ่มของตัวอักษรในระดับกลาง
  - **Lower Zone** คือกลุ่มของตัวอักษรในระดับล่าง
- **Overlapped Zone** เป็นค่าที่ระบุระดับถ้าตัวอักษรนั้นมีการเหลื่อมล้ำ ซึ่งมีค่าเป็น
  - **Upper Zone** คือการที่ตัวอักษรระดับกลางมีการเหลื่อมล้ำกับระดับบน เช่น ใ
  - **Lower Zone** คือการที่ตัวอักษรระดับกลางมีการเหลื่อมล้ำกับระดับล่าง เช่น ๆ
  - **None** คือ ไม่มีการเหลื่อมล้ำของตัวอักษร

ดังแสดงตัวอย่างในรูปที่ 4.3



รูปที่ 4.3 ตัวอย่างภาพตัวอักษรที่แยกออกเป็นตัวอักษรเดี่ยว

4.2 การกำหนดประเภทการสัมผัสกันของตัวอักษร

จากตารางที่ 4.1 ระดับที่ต่างกันของตัวอักษรภาษาไทย เมื่อเกิดการสัมผัสกันจะเกิดได้ในหลายรูปแบบ ดังแสดงในตารางที่ 4.2 และรูปที่ 4.4

ตารางที่ 4.2 รูปแบบการสัมผัสกันของตัวอักษรในแบบต่างๆ

รูปแบบที่	ลักษณะการสัมผัสกัน	ตัวอย่าง
1	สระระดับบน สัมผัสกับ สระระดับบนในแนวนอน	วิธี
2	สระระดับบน สัมผัสกับ วรรณยุกต์ในแนวตั้ง	ทั้ง ที่
3	สระระดับกลางที่มีความสูงถึงระดับบน สัมผัสกับ สระหรือวรรณยุกต์ระดับบน	นี้ไม่
4	พยัญชนะระดับกลาง สัมผัสกับ พยัญชนะหรือสระระดับกลางในแนวนอน	กลาง
5	พยัญชนะระดับกลาง สัมผัสกับ สระหรือวรรณยุกต์ระดับบนในแนวตั้ง	สีสัน
6	พยัญชนะระดับกลาง สัมผัสกับ สระหรือวรรณยุกต์ระดับล่างในแนวตั้ง	สุข
7	พยัญชนะระดับกลางที่มีความสูงถึงระดับบน สัมผัสกับพยัญชนะหรือวรรณยุกต์ระดับบนในแนวนอน	ปัญหา
8	การสัมผัสกันของตัวอักษรผสมกันระหว่างแนวตั้งและแนวนอน	ฟื้น

รูปที่ 4.4 ตัวอย่างภาพตัวอักษรที่สัมผัสกันในรูปแบบต่างๆ

เมื่อทำการรวบรวมข้อมูลจากเอกสารประเภทหนังสือพิมพ์ต่างๆ และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ในรูปแบบตัวอักษรต่างๆ พบว่ามีจำนวนการติดกันของตัวอักษร แบ่งแยกได้ดังต่อไปนี้

ตารางที่ 4.3 ข้อมูลจากหนังสือพิมพ์

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	5	0.04
2	180	1.26
3	4	0.03
4	388	2.71
5	28	0.19
6	19	0.13
7	51	0.36
8	0	0
รวม	**675**	**4.71**
ตัวอักษรทั้งหมด	14,310	

ตารางที่ 4.4 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร AngsanaUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	8	0.05
2	75	0.49
3	6	0.04
4	126	0.82
5	44	0.29
6	76	0.49
7	72	0.47
8	0	0
รวม	**407**	**2.65**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.5 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	7	0.04
2	216	1.41
3	7	0.04
4	186	1.21
5	0	0
6	0	0
7	50	0.33
8	0	0
รวม	**466**	**3.04**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.6 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร CordiaUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	8	0.05
2	132	0.86
3	7	0.04
4	184	1.20
5	132	0.86
6	170	1.11
7	103	0.67
8	0	0
รวม	**736**	**4.80**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.7 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร DilleniaUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	4	0.03
2	23	0.15
3	42	0.27
4	164	1.06
5	29	0.19
6	46	0.30
7	143	0.93
8	0	0
รวม	**451**	**2.94**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.8 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 12, 14 และ 16 points

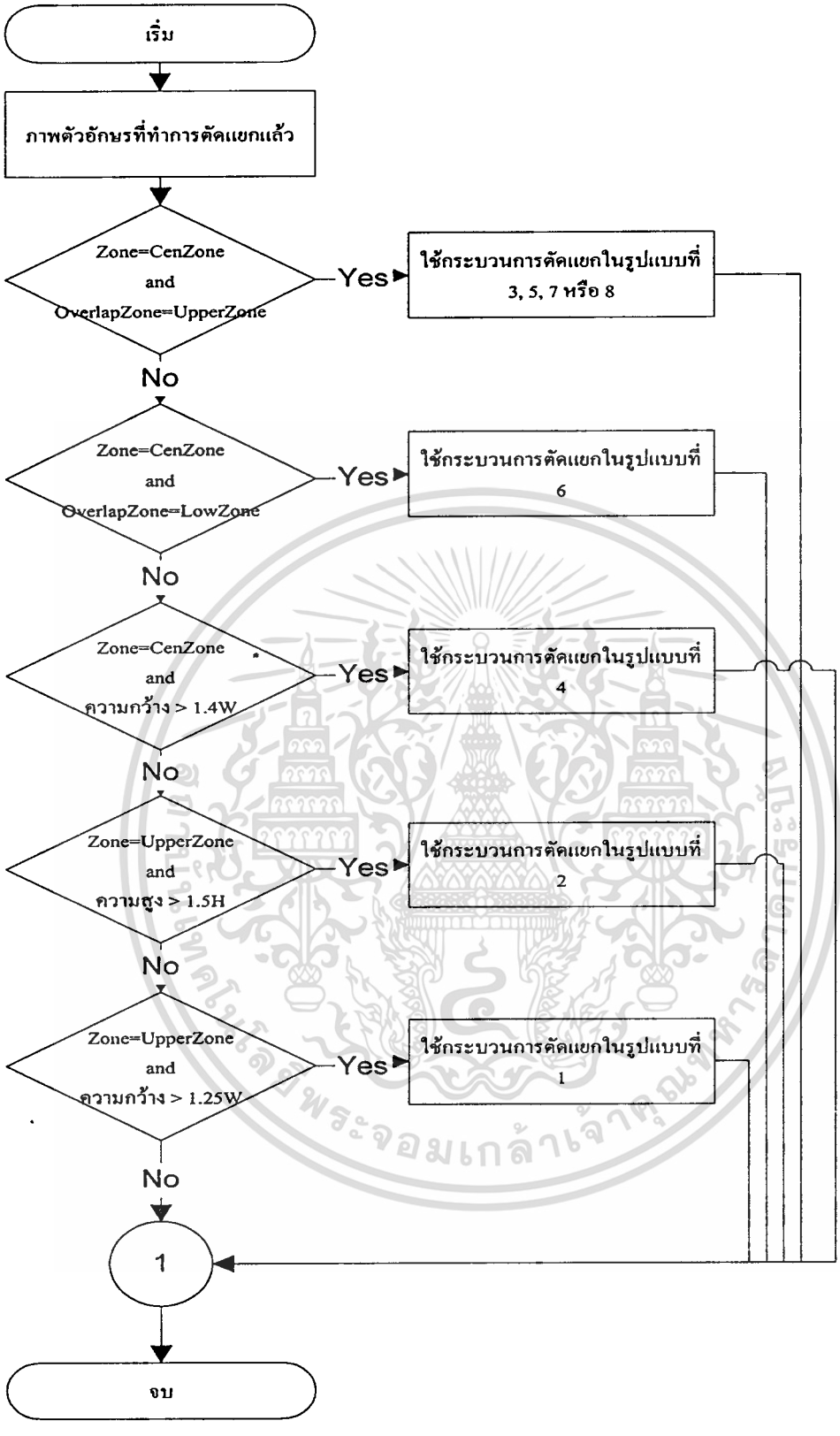
รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	1	0.006
2	204	1.32
3	27	0.17
4	137	0.89
5	4	0.03
6	16	0.10
7	80	0.52
8	4	0.03
รวม	**473**	**3.08**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.9 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร FreesiaUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	1	0.006
2	159	1.04
3	24	0.09
4	29	0.19
5	20	0.13
6	13	0.08
7	70	0.46
8	0	0
รวม	**316**	**2.06**
ตัวอักษรทั้งหมด	15,360	

ตารางที่ 4.10 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร IrisUPC ขนาด 12, 14 และ 16 points

รูปแบบที่	จำนวนตัวอักษรที่สัมผัสกัน	ร้อยละของตัวอักษรที่สัมผัสกัน
1	12	0.07
2	218	1.42
3	35	0.23
4	18	0.12
5	7	0.04
6	79	0.51
7	126	0.82
8	4	0.03
รวม	**499**	**3.25**
ตัวอักษรทั้งหมด	15,360	



รูปที่ 4.5 ผังงานโดยรวมของกระบวนในการวิเคราะห์การสัมผัสกันของตัวอักษร

4.3 การวิเคราะห์การสัมผัสกันของตัวอักษร และการกำหนดจุดตัดแยก

ในหัวข้อ 4.2 กล่าวถึงประเภทการสัมผัสกันของตัวอักษรซึ่งในงานวิจัยนี้แบ่งออกเป็น 8 รูปแบบ ซึ่งการที่จะวิเคราะห์การสัมผัสกันของตัวอักษรในแต่ละแบบจึงต้องใช้วิธีการที่แตกต่างกันไป ซึ่งในงานวิจัยนี้ใช้ค่าความกว้าง และความสูงของตัวอักษรในระดับต่างๆ มาใช้ในการพิจารณา ซึ่งมีวิธีการหาดังต่อไปนี้

4.3.1 การหาค่าความกว้างเฉลี่ยของตัวอักษรในระดับกลาง

เนื่องจากความแตกต่างของความกว้างของตัวอักษรไทย เช่น "เ" "ก" และ "ฌ" เป็นต้น จึงไม่สามารถใช้ความกว้างของตัวอักษรตัวใดตัวหนึ่งมาเป็นมาตรฐานในการเปรียบเทียบได้ เช่น ถ้าหากใช้ สระเอ ซึ่งเป็นตัวที่กว้างน้อยที่สุดมาใช้เปรียบเทียบ จะทำให้อักษรเดี่ยวๆ ทุกตัวถูกวิเคราะห์ว่าเป็นอักษรที่สัมผัสกันได้ ดังนั้นงานวิจัยนี้จึงได้ทำการทดลองหาความกว้างเฉลี่ยของตัวอักษรไทยในรูปแบบต่างๆ กัน ซึ่งพบว่าอักษรไทยโดยส่วนใหญ่จะมีความกว้างเป็น 0.8 เท่าของความสูงของตัวอักษรในระดับกลาง จึงใช้ค่านี้เป็นค่าความกว้างเฉลี่ย (W) เพื่อใช้ในการวิเคราะห์ต่อไป

4.3.2 การหาค่าความสูงเฉลี่ยของตัวอักษรในระดับบน

ความสูงเฉลี่ยของตัวอักษรในระดับบน คือความสูงของสระ และวรรณยุกต์ในระดับบน ซึ่งจากการทดลองพบว่าสระและวรรณยุกต์ในระดับบน จะมีความสูงประมาณ 0.5 เท่าของความสูงของตัวอักษรในระดับกลาง ในงานวิจัยนี้จึงใช้ค่านี้เป็นค่าความสูงเฉลี่ยของตัวอักษรในระดับบน (H) เพื่อใช้ในการวิเคราะห์ต่อไป

4.3.3 การวิเคราะห์การสัมผัสกัน และการกำหนดจุดตัด

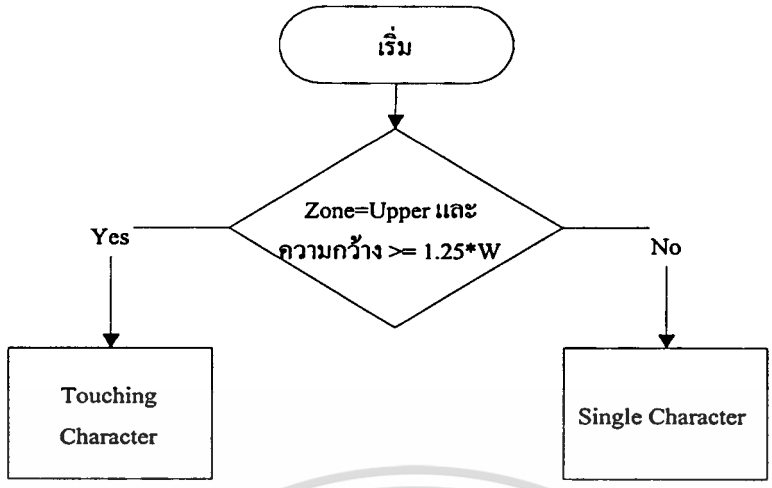
มีวิธีการวิเคราะห์ และการกำหนดจุดตัด ดังต่อไปนี้

1. **รูปแบบที่ 1** สระระดับบน สัมผัสกับ สระระดับบนในแนวนอน

<u>การวิเคราะห์</u>

ถ้าในการสัมผัสกันของสระในระดับบน จะทำให้ความกว้างของตัวอักษรนั้นมีค่ามากกว่าปกติ ซึ่งเราสามารถใช้ความกว้างมาวิเคราะห์ว่ามีการสัมผัสกันของตัวอักษรหรือไม่ ดังต่อไปนี้

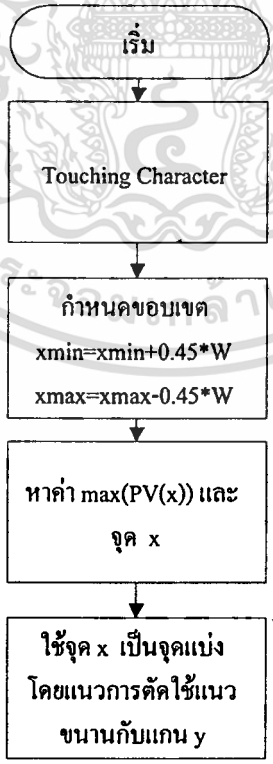
- ถ้าความกว้างของภาพตัวอักษรมีค่ามากกว่า 1.25 เท่าของความกว้างเฉลี่ย จะวิเคราะห์ว่าเป็นตัวอักษรที่สัมผัสกัน ถ้าน้อยกว่า จะวิเคราะห์เป็นตัวอักษรเดี่ยว ดังรูปที่ 4.6



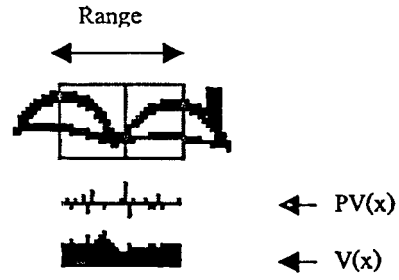
รูปที่ 4.6 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 1

<u>การกำหนดจุดตัดแยก</u>

การกำหนดจุดตัดแยกของตัวอักษรที่สัมผัสกันในรูปแบบนี้ จะใช้วิธีการหาค่า Peak to valley (PV) ของภาพตัวอักษรเพื่อหาตำแหน่งที่มีค่า PV สูง เป็นจุดที่ใช้พิจารณาแบ่งแยก ดังขั้นตอนในรูปที่ 4.7



รูปที่ 4.7 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 1

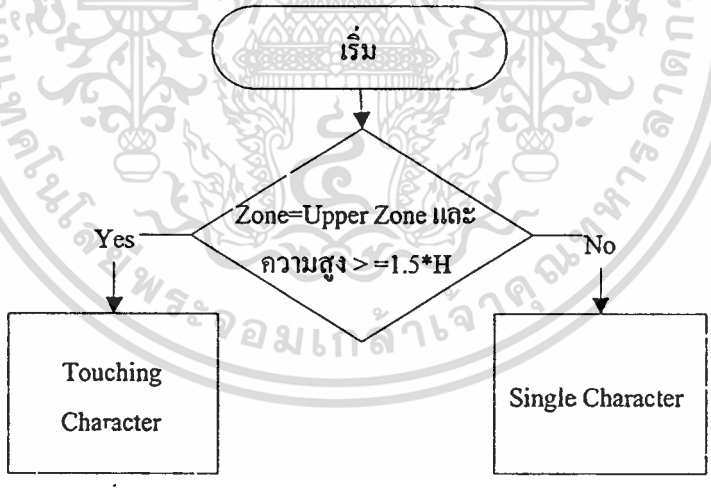


รูปที่ 4.8 ตัวอย่างการตัดแยกในรูปแบบที่ 1

2. **รูปแบบที่ 2 สระระดับบน สัมผัสกับ วรรณยุกต์ในแนวตั้ง**

<u>การวิเคราะห์</u>

การสัมผัสกันของสระระดับบนในแนวตั้ง จะทำให้ค่าความสูงของภาพตัวอักษรผิดปกติ คือ จะมีค่ามากกว่าความสูงเฉลี่ยของสระในระดับบน ซึ่งเราสามารถใช้ค่าความสูงมาวิเคราะห์การสัมผัสกัน ดังขั้นตอนในรูปที่ 4.9



รูปที่ 4.9 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 2

การกำหนดจุดตัดแยก

การกำหนดจุดตัดแยกของตัวอักษรที่สัมผัสกันในรูปแบบนี้ มีขั้นตอนดังนี้

1. หาค่าโปรเจคชันตามแนวแกน y (V(y))

2. ใช้สมการการหาค่า PV(y) เพื่อหาตำแหน่งที่มีค่า PV(y) สูงที่สุด

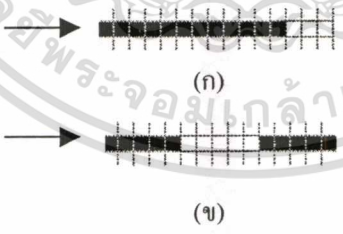
3. ใช้คุณลักษณะของตัวอักษรภาษาไทยร่วมในการพิจารณา คือ ส่วนบนและส่วนล่างของตัวอักษรจากตำแหน่งที่พิจารณาเป็นจุดตัด จะต้องมีความสูงเพียงพอที่จะเป็นสระในระดับบน เพื่อป้องกันข้อผิดพลาดที่เกิดขึ้นจากสัญญาณรบกวน ในงานวิจัยนี้ใช้ 0.5 เท่าของความสูงเฉลี่ยเป็นค่าในการพิจารณา

4. ตรวจสอบแนวการตัดผ่านเนื้อของตัวอักษรว่ามีการตัดผ่านเนื้อของตัวอักษรตลอดทั้งแนวหรือไม่ ดังตัวอย่างรูปที่ 4.10 เพื่อป้องกันข้อผิดพลาดในกรณีของสระ อี อิ เป็นต้น

5. ถ้าเงื่อนไขที่ใช้ในการพิจารณาเป็นจริง ก็จะใช้จุดนั้นเป็นจุดที่ใช้แบ่งแยกตัวอักษร และจบขั้นตอนการทำงาน ถ้าไม่เป็นจริงจะทำในข้อที่ 6 ต่อไป

6. ใช้วิธีการ New Metric ที่กล่าวถึงในบทที่ 3 มาใช้แทนการหาค่า V(y) และทำเช่นเดียวกับข้อ 1-3 อีกครั้ง ถ้าเงื่อนไขทั้งหมดเป็นจริง ก็จะใช้จุดนี้เป็นจุดที่ใช้แบ่งแยกตัวอักษร และจบขั้นตอนการทำงาน ถ้าไม่จริงจะทำในข้อที่ 7 ต่อไป

7. ใช้วิธีการแบ่งตัวอักษรที่สัมผัสกันโดยใช้การประมาณความสูงของตัวอักษรเท่ากับค่าความสูงเฉลี่ย ซึ่งโดยปกติแล้วภาพของตัวอักษรที่สัมผัสกันจะสามารถแยกได้โดย 2 วิธีแรก



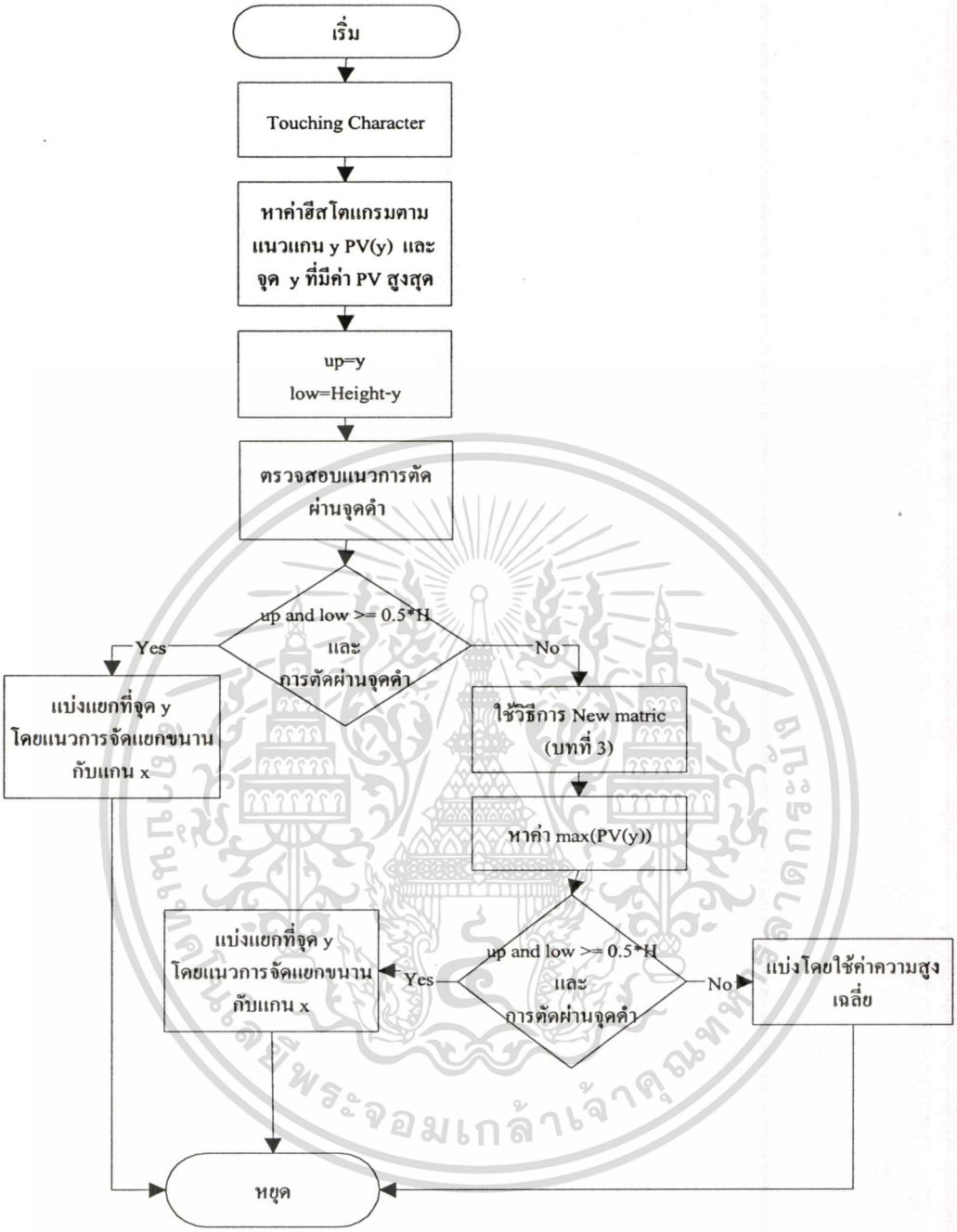
(ก)

(ข)

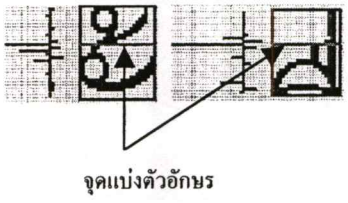
รูปที่ 4.10 ลักษณะการตรวจสอบการตัดผ่านเนื้อของตัวอักษร

(ก) ลักษณะของการตัดผ่านเนื้อที่ถูกต้อง

(ข) ลักษณะของการตัดผ่านเนื้อที่ใช้ไม่ได้



รูปที่ 4.11 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 2



จุดแบ่งตัวอักษร

รูปที่ 4.12 ตัวอย่างการตัดแยกในรูปแบบที่ 2

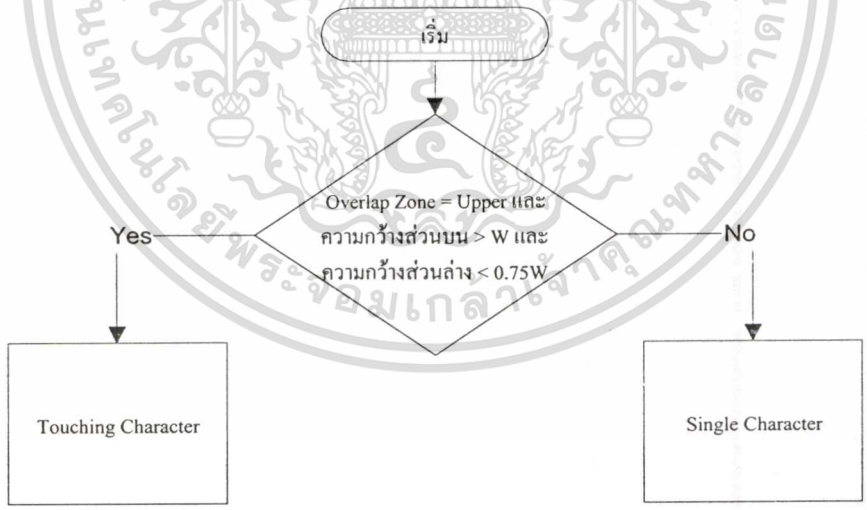
3. **รูปแบบที่ 3** สระระดับกลางที่มีความสูงถึงระดับบน สัมผัสกับ สระหรือวรรณยุกต์ระดับบน

<u>การวิเคราะห์</u>

วิธีการตรวจสอบการสัมผัสกันในรูปแบบนี้ จะใช้การตรวจสอบส่วนของการเหลื่อมล้ำ (Overlap Zone) ของตัวอักษรว่ามีการเหลื่อมล้ำถึงระดับบนหรือไม่ ถ้ามีการเหลื่อมล้ำกับระดับบน จะทำการตรวจสอบความกว้างของส่วนที่มีการเหลื่อมล้ำ และความกว้างส่วนล่างในระดับกลาง ถ้า

- ความกว้างส่วนบน **มากกว่า** ความกว้างเฉลี่ย (W) และ ความกว้างส่วนล่าง **น้อยกว่า** 0.75 เท่าของ W จะวิเคราะห์ว่าเป็นการสัมผัสของตัวอักษรในรูปแบบที่ 3 ดังตัวอย่างในรูปที่ 4.13

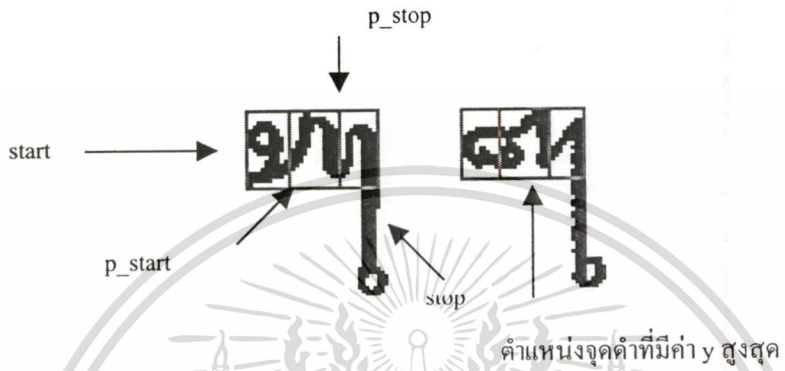
รูปที่ 4.13 ตัวอย่างภาพตัวอักษรที่สัมผัสกันในรูปแบบที่ 3



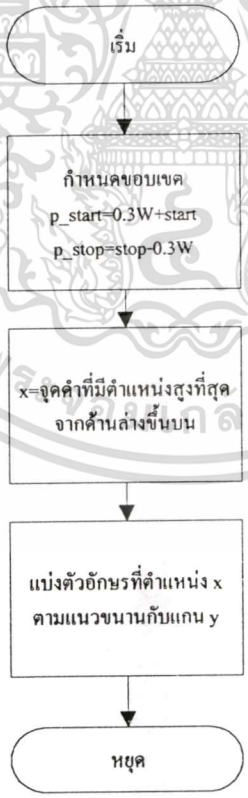
รูปที่ 4.14 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 3

การกำหนดจุดตัดแยก

การสัมผัสกันของตัวอักษรในรูปแบบนี้ จะเกิดที่ส่วนปลายของสระระดับบน เช่น สระ อิ อี ั สัมผัสกับส่วนปลายทางด้านซ้ายของสระระดับกลาง (ไ โ ใ) บริเวณส่วนบน การกำหนดจุดตัดของรูปแบบนี้ จะใช้วิธีการตรวจสอบจุดดำที่มีตำแหน่งในแนวแกน y สูงสุด พิจารณาจากด้านล่างขึ้นบน เป็นจุดตัด ดังตัวอย่างในรูปที่ 4.15 p_start และ p_stop คือ ขอบเขตที่ใช้ในการตรวจสอบเพื่อหาจุดแบ่งแยกตัวอักษรที่สัมผัสกัน ซึ่งมีขั้นตอนดังรูปที่ 4.16



รูปที่ 4.15 การกำหนดขอบเขตเพื่อใช้หาจุดแบ่งของตัวอักษรที่สัมผัสกันในรูปแบบที่ 3

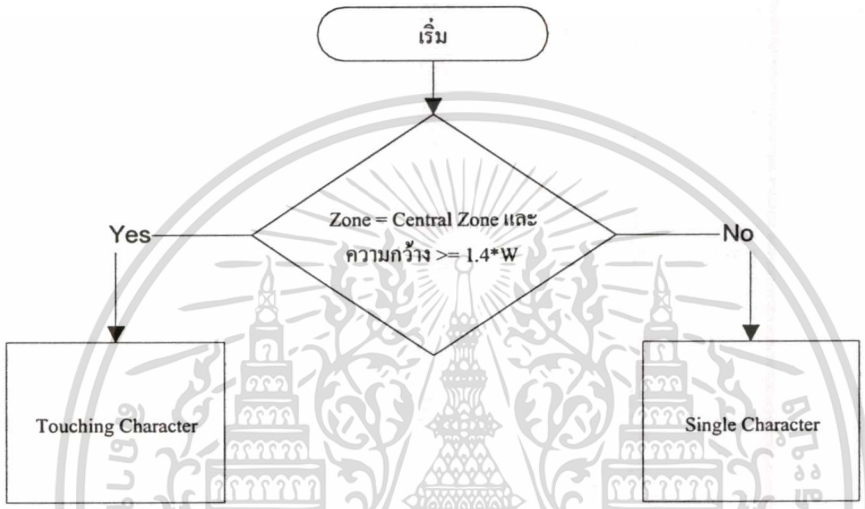


รูปที่ 4.16 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 3

4. **รูปแบบที่ 4** พยัญชนะระดับกลาง สัมผัสกับ พยัญชนะหรือสระระดับกลางในแนวนอน

การวิเคราะห์

การสัมผัสกันของตัวอักษรในรูปแบบที่ 4 นี้จะเกิดขึ้นในระดับกลางของประโยคเท่านั้น การตรวจสอบค่าความกว้างของภาพตัวอักษรซึ่งจะมีค่ามากกว่าความกว้างเฉลี่ย ในงานวิจัยนี้ใช้ค่า 1.4 เท่าของ W เป็นค่าที่ใช้ในการตรวจสอบ ดังขั้นตอนในรูปที่ 4.17



รูปที่ 4.17 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 4

การกำหนดจุดตัดแยก

การหาจุดแบ่งแยกตัวอักษรในรูปแบบนี้ จะมีความซับซ้อนมากกว่ารูปแบบอื่น เนื่องจากตัวอักษรภาษาไทยในระดับกลางมีจำนวนมาก และมีขนาดแตกต่างกัน เช่น เ ก ณ เป็นต้น จะใช้การหาค่า PV(x) เพียงอย่างเดียวจะให้ค่าความถูกต้องไม่ครบทุกกรณี ดังนั้นจึงจะต้องใช้ค่าอื่นๆ พิจารณาร่วมด้วย เช่น ค่าความกว้าง ค่าโปรเจคชัน ซึ่งมีขั้นตอนดังต่อไปนี้

1. ถ้าความกว้างของตัวอักษรมากกว่า 2.2 เท่าของ W ซึ่งหมายถึงตัวอักษรอาจมีการสัมผัสกันมากกว่า 2 ตัว ดังนั้นจึงต้องกำหนดขอบเขตให้ค่า $x_{max}=1.55W+x_{min}$ **ถ้าไม่** x_{max}= ความกว้างของตัวอักษร
2. หาค่าโปรเจคชันตามแนวแกน x (V(x)) จาก x_{min} ถึง x_{max}
3. หาค่า PV(x) จาก x_{min} ถึง x_{max}

4. หาค่า x ที่มีค่า PV สูงสุด และหาค่าความกว้างส่วนหน้า (front) และส่วนหลัง (back) ของจุด x ถ้า

   4.1 front และ back **มากกว่า** 0.6 เท่าของ W ทำขั้นตอนในข้อ 5 ต่อไป

   4.2 front **น้อยกว่า** 0.6W และ back **มากกว่า** 0.6W

   ให้หาค่า PV ที่มีค่าสูงสุดใหม่ตั้งแต่ตำแหน่ง x+1 ถึง x_{max} และกลับไปทำขั้นตอนในข้อ 4 ใหม่

   4.3 front **มากกว่า** 0.6W และ back **น้อยกว่า** 0.6W

   ให้หาค่า PV ที่มีค่าสูงสุดใหม่ตั้งแต่ตำแหน่ง x_{min} ถึง x-1 และกลับไปทำขั้นตอนในข้อ 4 ใหม่

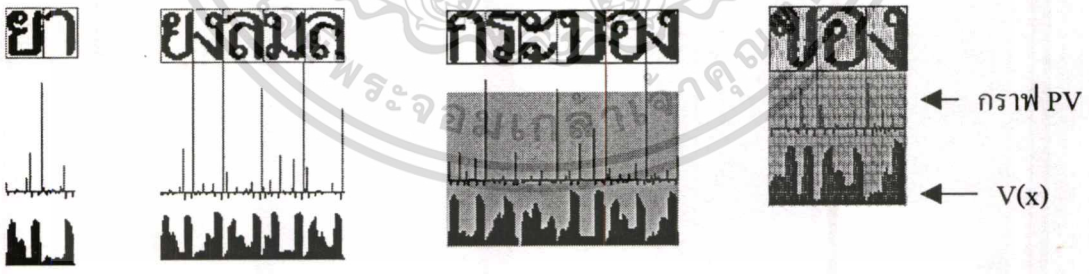
5. พิจารณาเงื่อนไขอื่นๆ

   5.1 ค่าโปรเจคชันที่ตำแหน่ง x จะต้องมีค่าน้อยที่สุด เมื่อเทียบกับจุดใกล้เคียง

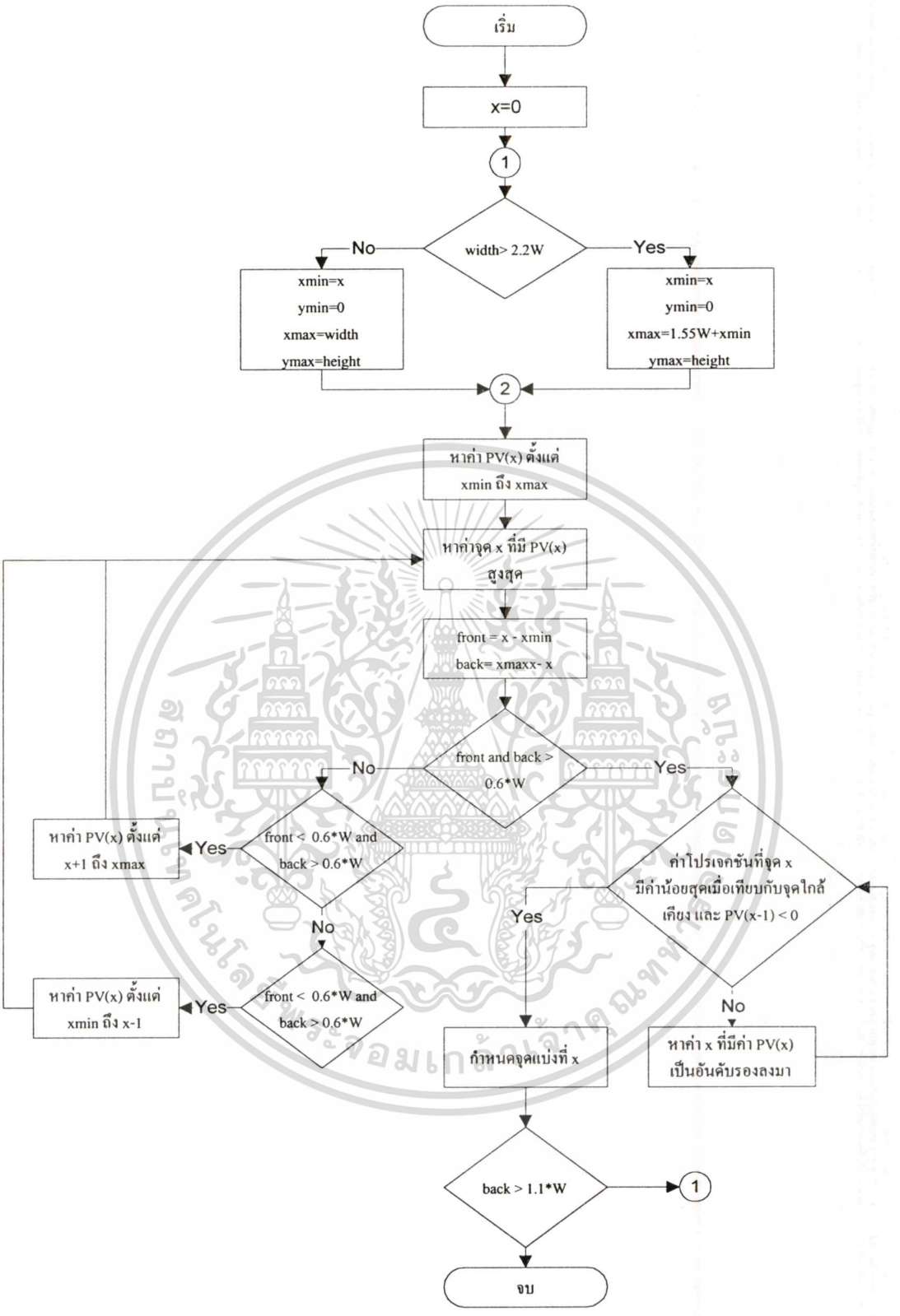
   5.2 ค่า PV ที่ตำแหน่งก่อนหน้าจุด x จะต้องมีค่าลบ

6. ถ้าเงื่อนไขในข้อ 5.1 และ 5.2 เป็นจริง ให้ใช้จุด x เป็นจุดที่ใช้แบ่งตัวอักษร <u>ถ้าไม่</u> ให้ทำการหาค่า x ที่มีค่า PV อันดับรองลงมา จาก x ถึง x_{max} และพิจารณาเงื่อนไขในข้อ 5 อีกครั้ง

7. ถ้า ความกว้างส่วนที่เหลือหลังจุด x มีค่ามากกว่า 1.1เท่าของW ให้กลับไปทำขั้นตอนที่ 1 อีกครั้ง <u>ถ้าไม่</u> ให้หยุดการทำงาน



รูปที่ 4.18 ตัวอย่างภาพที่ทำการแบ่งแยกในรูปแบบที่ 4



รูปที่ 4.19 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 4

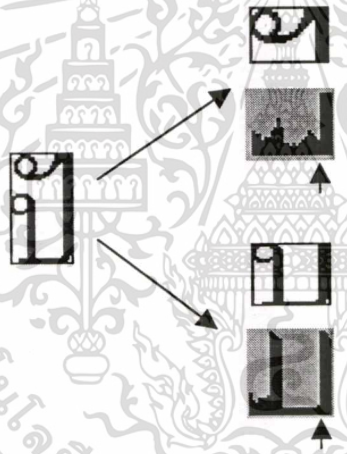
5. **รูปแบบที่ 5 และ รูปแบบที่ 7**

รูปแบบที่ 5 พยัญชนะระดับกลาง สัมผัสกับ สระหรือวรรณยุกต์ระดับบนในแนวตั้ง

รูปแบบที่ 7 พยัญชนะระดับกลางที่มีความสูงถึงระดับบน สัมผัสกับพยัญชนะหรือวรรณยุกต์ระดับบนในแนวนอน

การวิเคราะห์

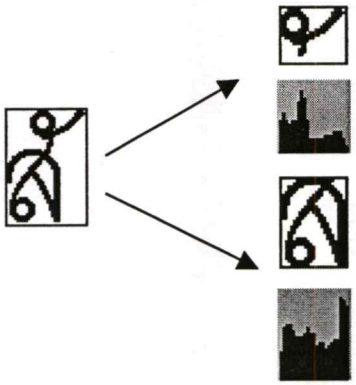
เนื่องจากการสัมผัสกันของตัวอักษรในรูปแบบที่ 5 และ 7 มีความใกล้เคียงกัน คือ การใช้การวิเคราะห์ส่วนของการเหลื่อมล้ำของตัวอักษร ร่วมกับลักษณะเด่นของตัวอักษรและขอบเขตของฮีสโตแกรม ซึ่งจะแบ่งแยกความแตกต่างการสัมผัสกันของทั้งสองรูปแบบนี้ได้ ดังตัวอย่างในรูปที่ 4.20



รูปที่ 4.20 ภาพตัวอย่างในการวิเคราะห์ลักษณะของฮีสโตแกรมในรูปแบบที่ 7

จากรูป 4.19(ก) ส่วนปลายของขาของตัวอักษร “ป” ซึ่งเหลื่อมล้ำกับระดับบน จะมีลักษณะเป็นแท่งสูงขึ้นมา และจากรูป 4.20(ข) จะมีแท่งฮีสโตแกรมถึง 2 แท่ง ถ้าพิจารณายอดที่ 2 จะเห็นได้ว่าตำแหน่งเริ่มต้นของความสูงในยอดที่ 2 ของรูป (ก) และ (ข) มีแนวที่ตรงกัน ซึ่งจะวิเคราะห์เป็นการสัมผัสกันในรูปแบบที่ 7 และจะต้องเป็นตัวอักษรที่มีลักษณะเปิดด้านบนด้วย (ป, ฟ, ฝ)

ถ้าแนวของแท่งฮีสโตแกรมไม่ตรงกัน หรือส่วนบนมีความสูงของแท่งฮีสโตแกรมที่ไม่มีความสูงเพียงพอ ดังตัวอย่างในรูปที่ 4.21 และลักษณะตัวอักษรเป็นตัวอักษรที่ปิดด้านบน เช่น ส ล จะถูกวิเคราะห์เป็นการสัมผัสกันในรูปแบบที่ 5



รูปที่ 4.21 ลักษณะภาพตัวอักษรที่ใช้ฮีสโตแกรมวิเคราะห์ในรูปแบบที่ 5

รูปที่ 4.22 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 5 และ 7

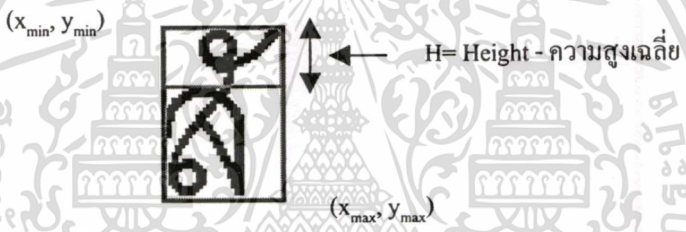
<u>การกำหนดจุดตัดแยก</u> เนื่องจากการกำหนดจุดตัดแยกในแบบที่ 5 และ 7 มีความแตกต่างกัน คือ แบบที่ 5 จะตัดแยกในแนวขนานกับแกน x แต่แบบที่ 7 จะตัดแยกในแนวขนานกับแกน y ดังนั้นจึงต้องแยกอธิบายในแต่ละวิธีดังต่อไปนี้

<u>รูปแบบที่ 5</u>

มีวิธีการดังต่อไปนี้

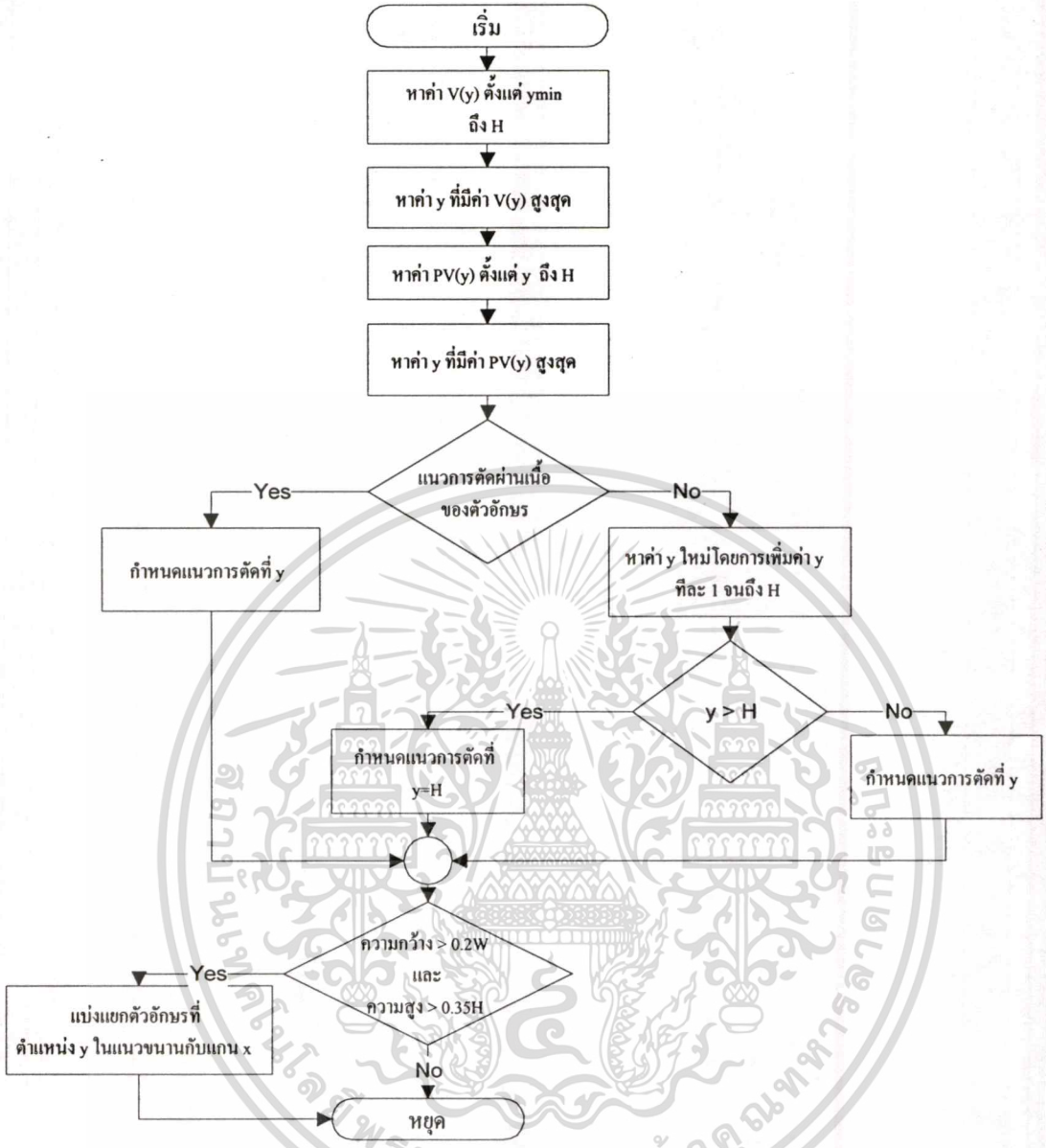
1. หาค่า V(y) ตั้งแต่ y_{min} จนถึง H ดังตัวอย่างรูปที่ 4.22

2. หาค่า y ที่มีค่า V(y) สูงสุด
3. หาค่า PV(y) ตั้งแต่ y ถึง H
4. หาค่า y ที่มีค่า PV(y) สูงสุด
5. ตรวจสอบแนวการตัดผ่านเนื้อของตัวอักษร เช่นเดียวกับรูปแบบที่ 2 ถ้าเป็น **จริง** ให้ทำขั้นตอนที่ 4 **ถ้าไม่จริง** ให้ทำการเลื่อนจุด y ลงมาด้านล่างทีละหนึ่งพิกเซล พร้อมทั้งทำการตรวจสอบอีกครั้งจนกว่าจะได้จุด y ที่มีแนวการตัดผ่านเนื้อที่เป็นจริง ถ้า y มากกว่าค่า H ให้กำหนดแนวการตัดที่ y เท่ากับ H
6. ตรวจสอบค่าความกว้างและความสูงของส่วนที่แบ่งโดยจุด y ถ้าความกว้างมากกว่า 0.2 เท่าของ W และ ความสูง มากกว่า 0.35 เท่าของ H **เป็นจริง** ให้ทำขั้นตอนที่ 7 **ถ้าไม่จริง** ให้หยุดการทำงาน
7. แบ่งแยกตัวอักษร ณ ตำแหน่ง y ในแนวขนานแกน x



รูปที่ 4.23 ตัวอย่างภาพและการกำหนดขอบเขตเพื่อหาจุดแบ่งในรูปแบบที่ 5

รูปที่ 4.24 ตัวอย่างภาพที่ทำการแบ่งแยกในรูปแบบที่ 5



รูปที่ 4.25 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 5

<u>รูปแบบที่ 7</u>

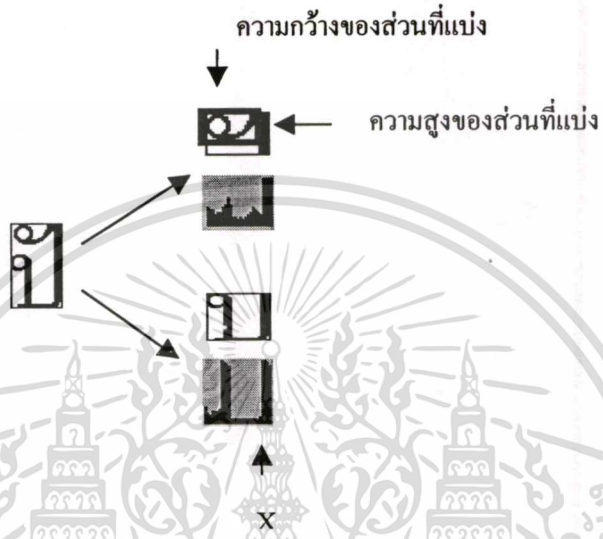
มีวิธีการกำหนดจุดตัดแยกดังนี้

1. กำหนดแนวการตัดที่ตำแหน่ง x เมื่อ x คือตำแหน่งของแท่งฮีสโตแกรมแท่งที่ 2 หรือส่วนที่เป็นขาหลังของตัวอักษร ดังรูปที่ 4.26
2. ตรวจสอบความกว้าง และความสูงของส่วนที่แบ่งออกมาถ้า
   - ความกว้าง มากกว่า 0.2 เท่าของ W และ
   - ความสูง มากกว่า 0.35 เท่าของ H

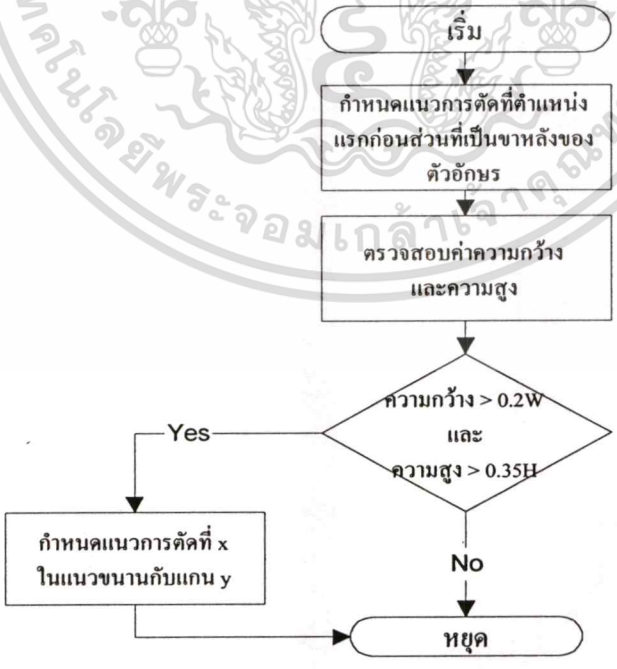
ถ้าเงื่อนไข**เป็นจริง** ให้ทำขั้นตอนที่ 3

ถ้าไม่จริง ให้หยุดการทำงาน

3. แบ่งแยกตัวอักษร ณ ตำแหน่ง x โดยกำหนดแนวการตัดแยกในแนวขนานกับแกน y และรวมส่วนที่เหลือจากส่วนบนเข้ากับส่วนล่างเพื่อให้ได้พยัญชนะในระดับกลาง



รูปที่ 4.26 ตัวอย่างภาพและการกำหนดจุดตัดในรูปแบบที่ 7

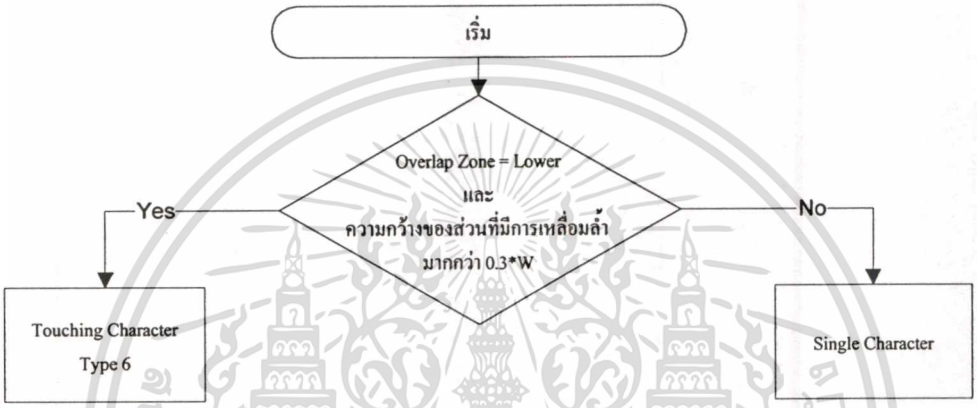


รูปที่ 4.27 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 7

6. **รูปแบบที่ 6** พยัญชนะระดับกลาง สัมผัสกับ สระหรือวรรณยุกต์ระดับล่างในแนวตั้ง

การวิเคราะห์

การวิเคราะห์การสัมผัสกันในรูปแบบนี้ ใช้การตรวจสอบการเหลื่อมล้ำของขอบเขตตัวอักษรว่ามีการเหลื่อมล้ำในระดับล่างหรือไม่ **ถ้ามี** และ ความกว้างของส่วนที่มีการเหลื่อมล้ำมีความกว้างมากกว่า 0.5 เท่าของความกว้างเฉลี่ย ภาพของตัวอักษรนี้จะถูกวิเคราะห์เป็นตัวอักษรที่สัมผัสกันในรูปแบบที่ 6 ดังขั้นตอนในรูปที่ 4.28



รูปที่ 4.28 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 6

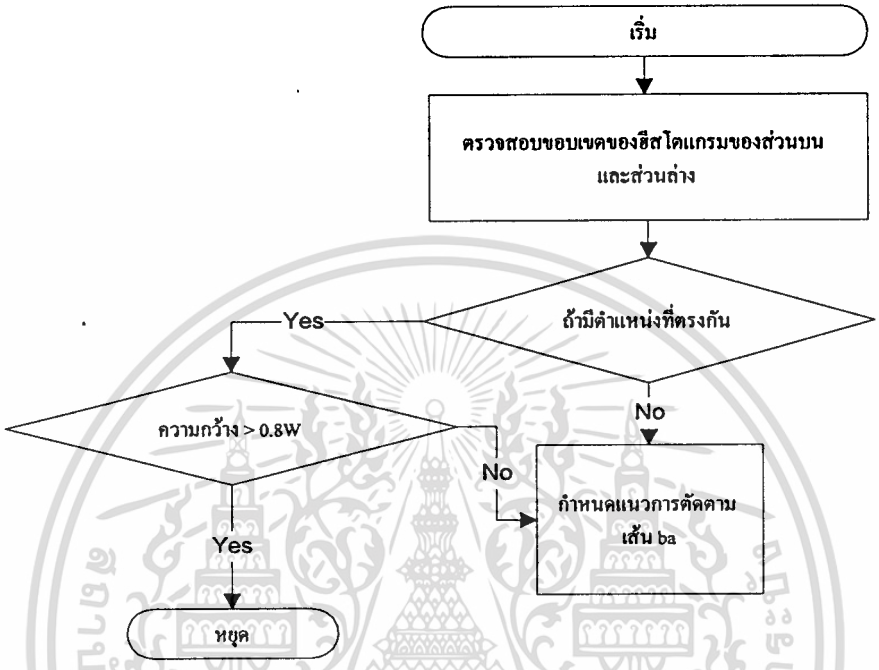
รูปที่ 4.29 ภาพตัวอย่างการสัมผัสกันของตัวอักษรรูปแบบที่ 6

การกำหนดจุดตัดแยก

มีขั้นตอนดังต่อไปนี้

1. ตรวจสอบขอบเขตของฮีสโตแกรมส่วนเป็นขาของส่วนล่างและส่วนบน เช่นเดียวกับรูปแบบที่ 7 ถ้าขอบเขตของแท่งฮีสโตแกรมในส่วนบนและส่วนล่างมีแนวที่ตรงกัน ให้ทำขั้นตอนที่ 2 **ถ้าไม่** ให้ทำขั้นตอนที่ 3

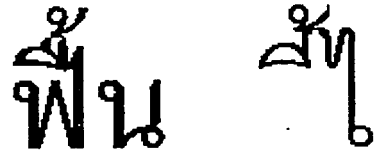
2. ถ้าความกว้างของส่วนที่มีการเหลื่อมล้ำ มากกว่า 0.8เท่าของ W (กรณีของ ฎ ฎ) ให้หยุดการทำงาน **ถ้าไม่** ให้ทำขั้นตอนที่ 3 ต่อไป
3. กำหนดแนวการตัดในแนวของเส้นแบ่งระดับกลางกับระดับล่าง ดังรูป 4.1 คือแนวเส้น ba



รูปที่ 4.30 ขั้นตอนการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบที่ 6

7. **รูปแบบที่ 8** การสัมผัสกันของตัวอักษรผสมกันระหว่างแนวตั้งและแนวนอน

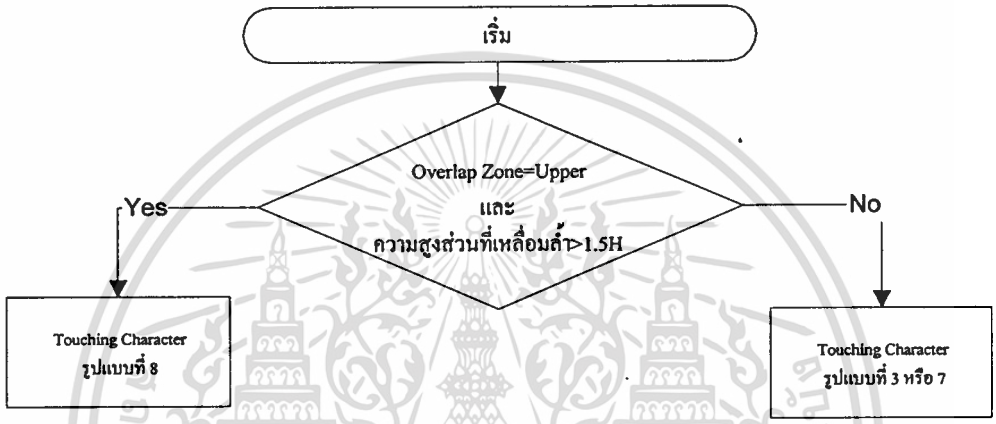
สำหรับรูปแบบที่ 8 นี้ การสัมผัสกันของตัวอักษรอาจจะเกิดการผสมกันระหว่างแบบที่ 2 กับ 7 หรือ แบบที่ 2 กับ 3 ดังตัวอย่างในรูปที่ 4.31 ซึ่งมีขั้นตอนการวิเคราะห์และการกำหนดจุดตัดดังต่อไปนี้



รูปที่ 4.31 ตัวอย่างภาพการสัมผัสกันในรูปแบบที่ 8

การวิเคราะห์

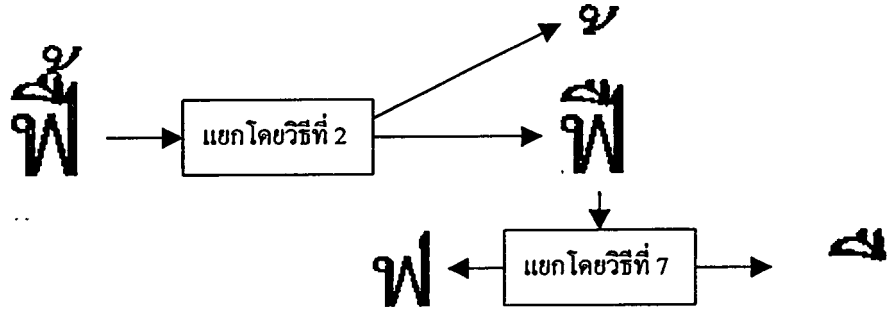
จะเห็นได้ว่าอักษรที่สัมผัสกันในรูปแบบที่ 8 นี้ สามารถใช้ความสูงของตัวอักษรและส่วนของตัวอักษรที่เหลื่อมล้ำจากระดับกลางกับระดับบนเพื่อวิเคราะห์การสัมผัสกัน ซึ่งถ้าความสูงของส่วนที่เหลื่อมล้ำในระดับบนมีความสูงมากกว่า 1.5 เท่าของความสูงเฉลี่ย ภาพของตัวอักษรนี้จะถูกวิเคราะห์เป็นการสัมผัสกันในรูปแบบที่ 8 แต่ถ้าไม่ ภาพของตัวอักษรนี้อาจจะเป็นการสัมผัสกันในรูปแบบที่ 3 หรือ 7 ก็ได้ ดังขั้นตอนในรูปที่ 4.32



รูปที่ 4.32 ขั้นตอนการวิเคราะห์การสัมผัสกันของภาพตัวอักษรรูปแบบที่ 8

การกำหนดจุดตัดแยก

1. ใช้วิธีการกำหนดจุดตัดแยกของรูปแบบที่ 2 เพื่อแยกสระและวรรณยุกต์ที่สัมผัสกันในแนวตั้งออกเสียก่อน
2. ใช้วิธีการกำหนดจุดตัดแยกของรูปแบบที่ 3 หรือ 7 แยกตัวอักษรที่สัมผัสกันที่เหลือ แล้วแต่กรณีที่สัมผัสกัน ดังรูปที่ 4.33



รูปที่ 4.33 ขั้นตอนการกำหนดจุดตัดแยกของภาพตัวอักษรที่สัมผัสกันในรูปแบบที่ 8

บทที่ 5

ผลการทดลอง

จากบทที่ 4 ได้กล่าวถึงการวิเคราะห์การสัมผัสกันของตัวอักษร และการกำหนดจุดตัดแยกในรูปแบบต่างๆ ทั้ง 8 แบบมาแล้ว ในงานวิจัยนี้ได้ทำการทดลองวิธีการดังกล่าวกับข้อมูลจากแหล่งต่างๆ เช่น หนังสือพิมพ์ เอกสารที่ได้จากการพิมพ์ในรูปแบบอักษรที่แตกต่างกัน เป็นต้น ซึ่งในบทนี้จะแสดงรายละเอียดของผลการทดลองแยกตามประเภทของเอกสาร และความถูกต้องในการกำหนดจุดตัดแยกของการสัมผัสกันในรูปแบบต่างๆ ดังตารางด้านล่าง เมื่อ A คือ จำนวนตัวอักษรที่สัมผัสกันทั้งหมด B คือ จำนวนตัวอักษรที่กำหนดจุดตัดแยกได้ถูกต้อง และ C คือ เปอร์เซ็นต์ความถูกต้อง

5.1 ตารางแสดงผลการทดลอง

ตารางที่ 5.1 ข้อมูลจากหนังสือพิมพ์ จำนวน 14,310 ตัวอักษร แสกนที่ความละเอียด 300 dpi

รูปแบบที่	หนังสือพิมพ์								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	4	3	75	1	1	100	-	-	-
2	85	85	100	60	60	100	35	35	100
3	2	2	100	2	2	100	-	-	-
4	173	165	95.37	77	72	93.50	135	132	97.77
5	15	15	100	2	2	100	4	4	100
6	10	9	90.00	4	4	100	5	5	100
7	20	20	100	26	26	100	15	15	100
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	309	299	96.76	172	167	97.09	194	191	98.45
เฉลี่ยรวม	(657/675) 97.33%								

ตารางที่ 5.2 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร AngsanaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	AngsanaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	3	3	**100**	3	3	**100**	2	2	**100**
2	30	30	**100**	26	26	**100**	19	19	**100**
3	2	2	**100**	2	2	**100**	2	2	**100**
4	40	37	**92.50**	47	46	**97.87**	39	38	**97.43**
5	16	15	**93.75**	14	13	**92.87**	14	14	**100**
6	32	32	**100**	20	20	**100**	24	24	**100**
7	32	32	**100**	20	20	**100**	20	20	**100**
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	155	151	**97.41**	132	130	**98.48**	120	119	**99.16**
เฉลี่ยรวม	**(400/407) 98.28%**								

ตารางที่ 5.3 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร BrowalliaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	BrowalliaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	3	3	**100**	2	1	**50.00**	2	2	**100**
2	92	92	**100**	78	78	**100**	46	46	**100**
3	3	3	**100**	2	2	**100**	2	2	**100**
4	72	70	**97.22**	68	66	**97.05**	46	44	**95.65**
5	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-
7	20	20	**100**	16	16	**100**	14	14	**100**
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	190	188	**98.94**	166	163	**98.19**	110	108	**98.18**
เฉลี่ยรวม	**(459/466) 98.49%**								

ตารางที่ 5.4 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร CordiaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	CordiaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	4	4	**100**	2	2	**100**	2	2	**100**
2	56	56	**100**	40	40	**100**	36	36	**100**
3	2	2	**100**	3	3	**100**	2	2	**100**
4	72	70	**97.22**	58	56	**96.55**	54	53	**98.14**
5	52	49	**94.23**	42	39	**92.85**	38	35	**92.10**
6	60	60	**100**	58	58	**100**	52	52	**100**
7	44	44	**100**	29	29	**100**	30	30	**100**
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	290	285	**98.27**	232	227	**97.84**	214	210	**98.13**
เฉลี่ยรวม	**(722/736) 98.06%**								

ตารางที่ 5.5 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร DilleniaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	DilleniaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	2	2	**100**	1	1	**100**	1	1	**100**
2	10	10	**100**	7	7	**100**	6	6	**100**
3	16	16	**100**	14	14	**100**	12	12	**100**
4	85	84	**98.82**	49	49	**100**	30	29	**96.67**
5	9	9	**100**	10	10	**100**	10	10	**100**
6	33	33	**100**	8	8	**100**	5	5	**100**
7	49	49	**100**	47	47	**100**	47	47	**100**
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	204	203	**99.50**	136	136	**100**	111	110	**99.09**
เฉลี่ยรวม	**(449/451) 99.55%**								

ตารางที่ 5.6 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร EucrosiaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	EucrosiaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	-	-	-	-	-	-	1	1	100
2	64	64	100	72	72	100	68	68	100
3	8	6	75	9	7	77.77	10	8	80.00
4	33	33	100	56	56	100	48	48	100
5	-	-	-	2	2	100	2	2	100
6	4	4	100	5	5	100	7	7	100
7	29	29	100	26	26	100	25	25	100
8	-	-	-	2	2	100	2	2	100
เฉลี่ย	138	136	98.55	172	170	98.83	163	161	98.77
เฉลี่ยรวม	(467/473) 98.73%								

ตารางที่ 5.7 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร FreesiaUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	FreesiaUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	-	-	-	-	-	-	1	1	100
2	96	94	97.91	39	39	100	24	24	100
3	13	12	92.30	7	6	85.71	4	3	75
4	15	15	100	8	8	100	9	9	100
5	14	14	100	4	4	100	2	2	100
6	11	11	100	1	1	100	1	1	100
7	36	36	100	20	20	100	14	14	100
8	-	-	-	-	-	-	-	-	-
เฉลี่ย	185	182	98.37	79	78	98.73	55	54	98.18
เฉลี่ยรวม	(314/319) 98.43%								

ตารางที่ 5.8 ข้อมูลจากเอกสารที่พิมพ์จากเครื่องพิมพ์เลเซอร์ ที่มีรูปแบบอักษร IrisUPC ขนาด 12, 14 และ 16 points จำนวน 15,360 ตัวอักษร ที่ความละเอียด 300 dpi

รูปแบบที่	IrisUPC								
	12			14			16		
	(A)	(B)	(C)	(A)	(B)	(C)	(A)	(B)	(C)
1	6	6	**100**	4	4	**100**	2	2	**100**
2	80	80	**100**	78	78	**100**	60	60	**100**
3	14	13	**92.85**	12	11	**91.66**	9	8	**88.88**
4	7	7	**100**	5	5	**100**	6	6	**100**
5	2	2	**100**	3	3	**100**	2	2	**100**
6	31	31	**100**	28	28	**100**	20	20	**100**
7	48	46	**100**	40	38	**95.00**	38	37	**97.36**
8	2	2	**100**	1	1	**100**	1	1	**100**
เฉลี่ย	190	187	**98.42**	171	168	**98.24**	138	136	**98.55**
เฉลี่ยรวม	**(491/499) 98.39%**								

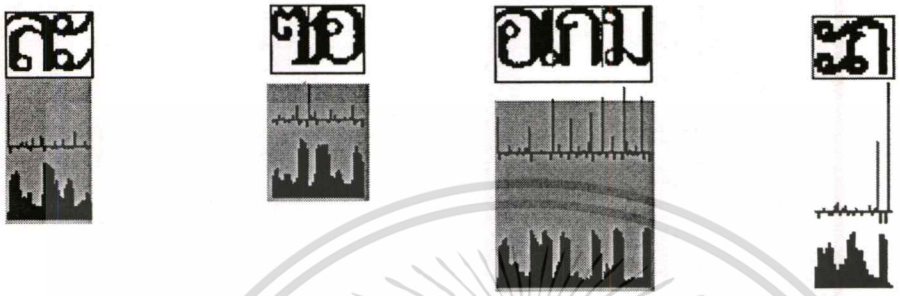
ตารางที่ 5.9 ข้อมูลรวมทั้งหมดจากตารางที่ 5.1-5.8 จำนวน 121,830 ตัวอักษร

รูปแบบที่	ข้อมูลรวมจากตารางที่ 5.1-5.8		
	A	B	C
1	46	44	**95.65**
2	1,207	1,207	**100.00**
3	152	140	**92.10**
4	1,232	1,211	**98.29**
5	257	246	**95.71**
6	422	422	**100.00**
7	695	690	**99.28**
8	8	8	**100.00**
เฉลี่ย	4,019	3,968	**98.73**
เฉลี่ยรวม	**98.73%**		

5.2 สาเหตุของการผิดพลาดจากการกำหนดจุดตัดแยก

ความผิดพลาดที่เกิดขึ้นสามารถสรุปได้ดังต่อไปนี้

1. เกิดจากการพิมพ์ที่ไม่ได้คุณภาพ ตัวอักษรอาจจะมีการซ้อนทับ หรือเหลื่อมล้ำกันจนทำให้ไม่สามารถใช้ สมการ PV ในการหาค่าความแตกต่างของจุดต่างๆ ออกมาได้ ดังรูปที่ 5.1

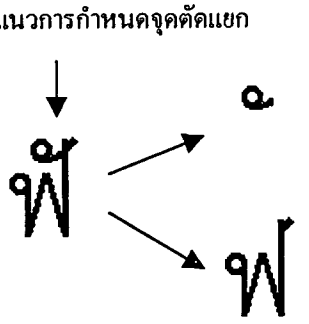


รูปที่ 5.1 ภาพตัวอย่างที่ผิดพลาดจากการกำหนดจุดตัดแยกโดยใช้สมการ PV

2. ความคล้ายคลึงกันของตัวอักษรเดี่ยว กับตัวอักษรที่มีการสัมผัสกัน มีความคล้ายกันมากซึ่งไม่สามารถใช้เงื่อนไขใดๆ ของโปรแกรมทำการแยกได้ เช่น บ่ กับ ป หรือ บี กับ ปี เป็นต้น ดังตัวอย่างในรูปที่ 5.2

รูปที่ 5.2 ภาพตัวอย่างที่ผิดพลาดจากการกำหนดจุดตัดแยกเนื่องจากความคล้ายกันของตัวอักษร

3. การซ้อนทับกันของตัวอักษรในรูปแบบที่ 1 และ 7 ซึ่งสามารถกำหนดจุดตัดแยกได้ถูกต้อง แต่ในงานวิจัยนี้ไม่มีวิธีการในการซ่อมแซมความเสียหายเนื่องจากการซ้อนทับกันนี้ ซึ่งในกรณีของความผิดพลาดนี้ได้กล่าวถึงและนำเสนอวิธีการแก้ไขไว้ในวิทยานิพนธ์ของนายวรวิทย์ เปรมรัตนชัย



รูปที่ 5.3 ความผิดพลาดเนื่องจากการซ้อนทับกันของตัวอักษร

บทที่ 6
สรุปผลงานวิจัยและข้อเสนอแนะ

ในระบบการรู้จำตัวอักษรไทย (Thai Optical Character Recognition) ปัญหาการสัมผัสกันของตัวอักษรมีผลทำให้ประสิทธิภาพของการรู้จำลดลง ซึ่งรูปแบบการสัมผัสกันของตัวอักษรก็มีอยู่ด้วยกันหลายแบบ แต่ละแบบมีวิธีการวิเคราะห์ และการกำหนดจุดตัดแยกที่แตกต่างกัน ซึ่งงานวิจัยนี้แบ่งรูปแบบการสัมผัสกันของตัวอักษรออกเป็น 8 รูปแบบทั้งแนวตั้งและแนวนอน และใช้ค่าความกว้าง ความสูง และการเหลื่อมล้ำของตัวอักษรในการวิเคราะห์การสัมผัสกันของภาพตัวอักษรในแต่ละรูปแบบ ส่วนในการกำหนดจุดตัดแยกนั้นได้ใช้สมการ Peak to valley (PV), การหาค่าฮีสโตแกรมทั้งแนวตั้งและแนวนอน รวมถึงค่าคุณลักษณะของตัวอักษรภาษาไทยมาร่วมในการพิจารณาหาจุดแบ่งแยกด้วย ซึ่งผลของความถูกต้องในการตัดแยกใช้สายตาคนในการแยกแยะ เนื่องจากงานวิจัยนี้ไม่ต้องการนำกระบวนการรู้จำเข้ามาเกี่ยวข้องด้วย เพราะวิธีการที่ใช้ในกระบวนการรู้จำแต่ละวิธีนั้นไม่เหมือนกัน และผลของความถูกต้องยังไม่เท่ากันอีกด้วยซึ่งจะทำให้ไม่สามารถวิเคราะห์ความผิดพลาดได้ว่าเกิดจากส่วนขั้นตอนใด จากผลการทดลองสามารถสรุปถึงข้อดี และข้อผิดพลาดต่างๆที่เกิดขึ้นจากวิธีการที่ใช้ในงานวิจัยนี้ได้ดังนี้

สรุปผลของวิธีการต่างๆ ที่ใช้ในงานวิจัย

1. การวิเคราะห์และการกำหนดจุดตัดแยกใช้วิธีการหลายๆ อย่างร่วมกัน เช่น ฮีสโตแกรม, สมการ PV, ค่าความกว้าง และความสูงของตัวอักษรเข้ามาร่วมประกอบในการพิจารณา ซึ่งได้ผลเป็นอย่างดี
2. สมการที่ใช้ในการคำนวณค่า PV สามารถใช้ในการหาจุดที่ใช้เป็นจุดแบ่งแยกตัวอักษรที่สัมผัสกันได้เป็นอย่างดี โดยเฉพาะอย่างยิ่งถ้าการสัมผัสนั้นเกิดจากส่วนที่เป็นขาของตัวอักษร กับส่วนปลายของอักษรอีกด้วยไม่ว่าจุดนั้นจะเป็นจุดที่มีเนื้อของตัวอักษรที่น้อยที่สุดหรือไม่ ซึ่งแตกต่างกับงานวิจัยที่ผ่านมา ซึ่งใช้วิธีการหาจุดที่มีเนื้อของตัวอักษรที่น้อยที่สุดในนอนหรือแนวตั้ง ทำให้มีข้อผิดพลาดได้สูง
3. การใช้ค่าความกว้าง และความสูงของตัวอักษรเข้ามาร่วมประกอบในการพิจารณาทำให้ความถูกต้องของการแบ่งแยกตัวอักษรมีความถูกต้องมากขึ้น เพราะตัวอักษรภาษาไทยมีแตกต่างกันมากโดยเฉพาะอักษรไทยในระดับกลาง
4. การทดสอบทำกับตัวอักษรในหลายแบบอักษรถึง 7 แบบ และในขนาดต่างๆ กัน คือ 12, 14 และ 16 point ซึ่งพบว่ามีการสัมผัสกันของตัวอักษรสูง และเป็นขนาดที่ใช้พิมพ์กันเป็นมาตรฐานอยู่ทั่วไป และในงานวิจัยนี้ยังได้นำข้อมูลภาพจากหนังสือพิมพ์ต่างๆ ซึ่ง

เป็นปัญหาสำคัญของระบบการรู้จำ และมีการสัมผัสกันของตัวอักษรเป็นจำนวนมากพอสมควรมาทำการทดสอบร่วมด้วยซึ่งผลการทดสอบอยู่ในขั้นที่น่าพอใจ

5. การแบ่งระดับของตัวอักษรไทยออกเป็น 3 ระดับ ทำให้การวิเคราะห์การสัมผัสกัน และการกำหนดจุดตัดแยกมีความง่าย และความถูกต้องมากยิ่งขึ้นเพราะในแต่ละระดับมีรูปแบบการสัมผัสกันที่แตกต่างกัน
6. การกำหนดจุดตัดของตัวอักษรที่สัมผัสกันในระดับกลาง (รูปแบบที่ 4) อาจจะมีการสัมผัสกันมากกว่า 2 ตัว งานวิจัยนี้สามารถกำหนดจุดตัดในกรณีที่ตัวอักษรสัมผัสกันมากกว่า 2 ตัวได้ผลถูกต้องน่าพอใจ

สรุปข้อผิดพลาดที่เกิดขึ้นในงานวิจัยนี้

1. ผลของความเอียงของเอกสารมีผลต่อการแบ่งระดับของตัวอักษร และการหาค่าความสูงเฉลี่ย
2. การผิดพลาดจากการพิมพ์เนื่องจากเครื่องพิมพ์ ซึ่งอาจมีผลทำให้ตัวอักษรมีการเหลื่อมล้ำหรือซ้อนทับกันแบบผิดปรกติ ดังรูป 5.1 ในบทที่ 5 จะทำให้ไม่สามารถกำหนดจุดตัดได้โดยใช้สมการ PV ได้
3. วิธีการในงานวิจัยนี้ไม่ได้ทดสอบกับตัวอักษรไทยที่เป็นเลขไทย

แนวทางการพัฒนาต่อไปในอนาคต

1. ปรับปรุงสมการ PV เพื่อให้ใช้งานได้ดีกับตัวอักษรภาษาไทยให้มากยิ่งขึ้น
2. ทำการสำรวจข้อมูล และรวบรวมให้มากขึ้นทั้งในรูปแบบและขนาดของตัวอักษรต่างๆ
3. นำวิธีการเหล่านี้ไปใช้ร่วมกับกระบวนการรู้จำตัวอักษรไทย โดยเป็นกระบวนการจัดการล่วงหน้า (preprocessing) ก่อนส่งเข้าสู่กระบวนการรู้จำซึ่งจะทำให้ผลของความถูกต้องมีมากยิ่งขึ้น

เอกสารอ้างอิง

[1] Yí Lu. "Machine Printed Character Recognition An Overview." Pattern Recognition, Vol.28, No.4, 1994. pp. 67-79.

[2] Wicha Panich, Somchai Jitapunkul, Prasert Choruengwiwat. "Segmentation of Connected Characters Using Distinctive Feature Of The Character in Thai Character Recognition System." Electrical Engineering Conference on Circuits and Systems, 1997. pp. 338-342.

[3] Su Liang, M. Shridhar, M Ahmadi. "Efficient Algorithm For Segmentation and Recognition of Printed Characters in Document Processing." Proceeding of the 2nd ICDAR, 1993. pp. 240-244.

[4] D.G. Elliman. "A Review of segmentation and contextual analysis techniques for text recognition." Pattern Recognition, Vol.23, No.3/4,1990. pp. 337-346.

[5] Sargur N. Srihari. "High-Performance Reading Machine." Proceedings of the IEEE, Vol.80, No.7, July 1992. pp.1120-1132.

[6] George Nagy. "At the frontiers of OCR." Proceedings of the IEEE, Vol.80, No.7, July 1992. pp.1109-1110.

[7] Shunji Mori, Ching Y. Suen, Kazuhiko Yamamoto. "Historical Review of OCR Research and Development." Proceedings of the IEEE, Vol.80, No.7, July 1992. pp. 1109-1156.

[8] ศุภกร รัตนปราการ, บุญธีร์ เครือตราชู. "การวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรพิมพ์ไทยโดยใช้คุณลักษณะทางแนวตั้งและแนวนอนของฮีสโตแกรม." วารสารสารสนเทศลาดกระบัง, ปีที่ 4, ฉบับที่ 1, กรกฎาคม 2542. หน้า 21-30.

[9] Nucharee Premchaiswadi, Wichian Premchaiswadi, Seinosuke Narita. "Segmentation Of Horizontal and Vertical Touching Thai Character." *ITC-CSCC'99 International Technical Conference on Circuit Systems, Computers and Communications*, 1999. pp.338-342.

[10] สมศักดิ์ คงถาวรวัฒนา, สมชาย จิตะพันธ์กุล. "การแยกแยะสายอักขระตัวพิมพ์ไทยโดยการเข้ารหัสฟรีแมนดัดแปรกับโครงร่างของฮีสโตแกรม." การประชุมวิชาการทางวิศวกรรมไฟฟ้า. ครั้งที่ 19, 2539. หน้า 73-77.

[11] จักริน สุขสวัสดิ์ชน, วิเชียร เปรมชัยสวัสดิ์, นุชรี เปรมชัยสวัสดิ์. "การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่ติดกันในแนวตั้งและแนวนอน." สารเนคเทค, ปีที่ 7, ฉบับที่ 32, มกราคม – กุมภาพันธ์ 2543. หน้า 36-43.

[12] W.Premchaiswadi, N.Premchaiswadi, S.Werawattanakorn. "The data structure for Thai Character Recognition." Proceeding of The First HAS Symposium on Science and Technology, Chiang Mai, Thailand, October 1999, pp.71-74.

[13] N.Premchaiswadi, W.Premchaiswadi, S.Narita. "Segmentation of Horisontal and Vertical Touching Thai Characters." IEICE Trans. Fundamental, Vol.E83-A, No.6, June 2000.

[14] N.Premchaiswadi, W.Premchaiswadi, A.Thammano, S.Narita. "Merged and Broken Printed Thai Characters Segmentation." the 1999 International Conference on Artifical Neural Network In Engineering, St. Loius, USA, Noverber 1999, pp. 893-898.

[15] N.Premchaiswadi, W.Premchaiswadi, S.Narita. "Segementation of Horizontal and Vertical Touching Thai Characters." Proceeding of the 1999 International Technical Conference on Circuit/Systems, July 1999, pp. 25-28.

[16] W.Premchaiswadi, N.Premchaiswadi, J.Suksawatchon. "An analysis and Segmentation of Thai Touching Characters." Proceeding of The First HAS Symposium on Science and Technology, October 1999, pp. 62-65.

ภาคผนวก

การวิเคราะห์และแบ่งตัวพิมพ์อักษรไทยที่ติดกันในแนวตั้งและแนวนอน

AN ANALYSIS AND SEGMENTATION OF HORIZONTAL AND VERTICAL TOUCHING THAI PRINTED CHARACTERS

จักริน สุขสวัสดิ์ชน*, วิเชียร เปรมชัยสวัสดิ์**, นุชรี เปรมชัยสวัสดิ์**

บทคัดย่อ

บทความนี้นำเสนอแนวทางในการแยกตัวพิมพ์อักษรไทยที่ติดกัน ซึ่งแตกต่างจากภาษาอื่น คือ ตัวอักษรภาษาไทยสามารถพบการติดกันได้ทั้งแนวตั้งและแนวนอนขณะที่ภาษาอื่น เช่น ภาษาอังกฤษนั้นจะติดกันในเฉพาะแนวนอนเท่านั้น วิธีการนี้ใช้ลักษณะของประโยคภาษาไทย ซึ่งสามารถแบ่งออกเป็น 3 ระดับเพื่อใช้สำหรับการตรวจสอบชนิดของการติดกันของตัวอักษรไทย ซึ่งวิธีในการแบ่งตัวอักษรที่ติดกันจะขึ้นอยู่กับชนิดของการติดกันของตัวอักษรนั้นๆ วิธีการนี้สามารถแยกตัวอักษรที่สัมผัสกันได้ทั้งในแนวตั้งและแนวนอน จากผลการทดลองโดยการใช้เทคนิคนี้กับข้อมูลที่ได้จากหนังสือพิมพ์และเอกสารจากการพิมพ์ด้วยเครื่องพิมพ์เลเซอร์อย่างละ 10,000 ตัวอักษร มีความถูกต้องถึง 98 เปอร์เซ็นต์ และ 99 เปอร์เซ็นต์ตามลำดับ

Abstract

This paper presents the scheme of segmentation of printed touching Thai characters. Unlike other languages, touching of Thai characters can occurs both horizontally and vertically while other language usually occur only horizontally. The scheme is based on the multi-level structure of a Thai sentence and distinctive features of Thai characters. The multi-level structure is employed to classify characters into three zones, which are used for determining types of touching characters. The proposed method can be applied to both vertically and horizontally touching characters. The efficiency and effectiveness of the proposed method has been tested with scanned data from newspapers and laser printed documents with 10,000 characters each. The percent of correctness are 98 percent respectively.

1. บทนำ

ในกระบวนการรู้จำตัวอักษรด้วยคอมพิวเตอร์นั้นความถูกต้องและประสิทธิภาพของระบบการรู้จำขึ้นอยู่กับวิธีการที่ใช้ในการรู้จำ และข้อมูลนำเข้า (Input data) หากข้อมูลที่นำเข้ามีความสมบูรณ์ กล่าวคือ ไม่มีสัญญาณรบกวน เอกสารไม่มีความเอียง ภาพของตัวอักษรไม่มีการขาดของตัวอักษร แล้วนั้น ความถูกต้องของระบบการรู้จำย่อมมีค่าสูง แต่ในทางกลับกันหากข้อมูลที่นำเข้าเป็นข้อมูลที่ไม่สมบูรณ์ ประสิทธิภาพการรู้จำอาจจะลดลงต่ำมากหรือไม่สามารถทำการรู้จำตัวอักษรได้เลย การติดกันของตัวอักษรก็เป็นอีกปัญหาหนึ่ง ซึ่งโปรแกรมประยุกต์ที่ใช้ในการรู้จำตัวอักษรไทยที่มีอยู่ในท้องตลาดเช่น ArnThai และ ThaiOCR ในทุกวันนี้ ก็ยังไม่สามารถแก้ปัญหาเหล่านี้ได้อย่างสมบูรณ์ จากเหตุผลข้างต้นจึงสรุปได้ว่าข้อมูลที่

เป็นอินพุตมีผลต่อความถูกต้องของการรู้จำเป็นอย่างมาก ดังนั้นเราจะมาพัฒนาแต่ระบบการรู้จำตัวอักษรอย่างเดียวไม่ได้ จะต้องพัฒนาและหาวิธีทางทำให้ข้อมูลอินพุตที่จะส่งเข้าไปในระบบนั้นเป็นข้อมูลที่สมบูรณ์เสียก่อน เช่น การปรับความเอียงของเอกสาร หรือแยกตัวอักษรที่ติดกันให้เป็นตัวอักษรเดี่ยวๆ เป็นต้น

ปัญหาการติดกันของตัวอักษรมีงานวิจัยที่ทำการแก้ปัญหาแล้ว เช่น งานวิจัย [1] ซึ่งใช้คุณลักษณะทางแนวตั้งและแนวนอนของฮีสโตแกรมเพื่อวิเคราะห์การติดกันในระดับบน และระดับล่าง ซึ่งผลการทดลองมีความถูกต้องดี แต่ยังพบปัญหากับตัวอักษรที่มีความคล้ายคลึงกัน เช่น "บ่" กับ "ป" อักษรในกลุ่ม (บ, ป, พ, ฟ, ผ, ฝ) เป็นต้น และอักษรที่มีการเหลื่อมล้ำกัน เช่น ป, ฟ, ฝ, ป, ฝ ทำให้ไม่สามารถใช้ฮีสโตแกรมในการแยกได้ อีกงานวิจัยหนึ่ง [6] ใช้วิธีการ Distinctive Feature แบ่งกลุ่มและชนิดของการติดกัน ซึ่งในแต่ละกลุ่มจะใช้วิธีการในการจุดแยกตัวอักษรที่แตกต่างกัน เช่น ใช้การหาตำแหน่งที่มีค่าโปรเจคชันตามแนวตั้ง หรือแนวนอนที่น้อยที่สุด หรือใช้การหาจุดกึ่งกลางของตัวอักษรเพื่อใช้แบ่งตัวอักษรที่สัมผัสกัน ซึ่งสามารถใช้ได้ดี แต่ไม่สามารถใช้ได้ในทุกกรณี ซึ่งในบางกรณีนั้นจุดที่สัมผัสกันไม่จำเป็นต้องมีค่าโปรเจคชั่นที่น้อยที่สุดเสมอไป และตำแหน่งกลางของตัวอักษรนั้นในการแบ่งบางครั้งแล้วอาจจะทำให้เกิดความเสียหายแก่ตัวอักษรได้ ซึ่งการวิเคราะห์ก็ยังผิดพลาดอยู่เช่นกัน เช่น ("ฎ", "ฏ", "ฏ") และอักษรที่มีการเหลื่อมล้ำกัน

ดังนั้นงานวิจัยนี้จึงนำเสนอแนวทางในการหาจุดแยกอีกวิธีทางหนึ่งเพื่อใช้แยกตัวอักษรที่ติดกัน[2] และแก้ปัญหาที่เกิดขึ้นกับงานวิจัยก่อนๆ โดยใช้ค่าทางสถิติ ค่าความกว้าง และความสูง ค่าโปรเจคชั่น โครงสร้างของประโยคภาษาไทย และคุณลักษณะของตัวอักษรไทยมาพิจารณาร่วมกันเพื่อวิเคราะห์การติดกัน และหาจุดที่ใช้แยกตัวอักษรที่ติดกันนั้น

2. ทฤษฎี

2.1 โครงสร้างของประโยคภาษาไทย

ลักษณะประโยคภาษาไทยตามรูปที่ 1 ประกอบด้วย สระ, พยัญชนะ และวรรณยุกต์เรียงอยู่ในระดับที่แตกต่างกัน ซึ่งสามารถแบ่งได้เป็น 3 ระดับ คือ

- ระดับบน (UZ) ประกอบด้วย สระระดับบน, วรรณยุกต์ และตัวการันต์
- ระดับกลาง (CZ) ประกอบด้วยพยัญชนะ, สระระดับกลาง
- ระดับล่าง (LZ) ประกอบด้วย สระระดับล่าง

รูปที่ 1 โครงสร้างของประโยคในภาษาไทย

เมื่อ to = เส้นบนของระดับบน
up = เส้นแบ่งของระดับกลางกับระดับบน
ba = เส้นแบ่งของระดับกลางกับระดับล่าง
bo = เส้นล่างของระดับล่าง

เราสามารถทำการแยกตัวอักษรออกเป็นตัวอักษรเดี่ยวได้โดยมีกรอบของตัวอักษรเป็น (X_{min},Y_{min}) และ (X_{max},Y_{max}) โดยการทำโปรเจคชั่นตามแนวตั้ง (Vertical Projection) และใช้ช่องว่างของตัวอักษร เพื่อแยกตัวอักษรแต่ละตัวออกเป็นอักษรเดี่ยวๆ และทำการกำหนดกลุ่มของตัวอักษรนั้นๆ ว่าอยู่ในส่วนใด โดยการหาจุดกึ่งกลาง (Central point) ของกรอบตัวอักษร (X_{cen},Y_{cen}) ได้จาก

$$X_{cen} = (X_{min}+X_{max})/2$$

$$Y_{cen} = (Y_{min}+Y_{max})/2$$

และพิจารณาจากเงื่อนไขต่อไปนี้

CB(i) ∈ upper zone if cp of CB(i) ∈ (to,up-1)

CB(i) ∈ central zone if cp of CB(i) ∈ (up,ba)

CB(i) ∈ lower zone if cp of CB(i) ∈ (ba+1,bo)

เมื่อ

CB คือ Character Block

CP คือ Central point

เมื่อพิจารณาตำแหน่งของพยัญชนะและสระ สามารถที่จะแบ่งแยกกลุ่มของตัวอักษรได้ตามตารางที่ 1

ตารางที่ 1 การแบ่งกลุ่มของตัวอักษรไทยตามตำแหน่งในประโยค

กลุ่ม	ตัวอย่าง
1	ิ ี ึ ื ่ ้ ๊ ๋ ็ ั ์ ๎
2	ก ข ค ต ฆ ง จ ฉ ภ ถ พ ฑ ธ ร น ย บ ล ด เ า ส ว ง ม ท อ ฉ ผ ป ฟ ฝ ไ ไ ใ ฏ ฤ ญ ฐ ฎ ฦ
3	ุ ู

1 คือ ระดับบน 2 คือ ระดับกลาง 3 คือ ระดับล่าง

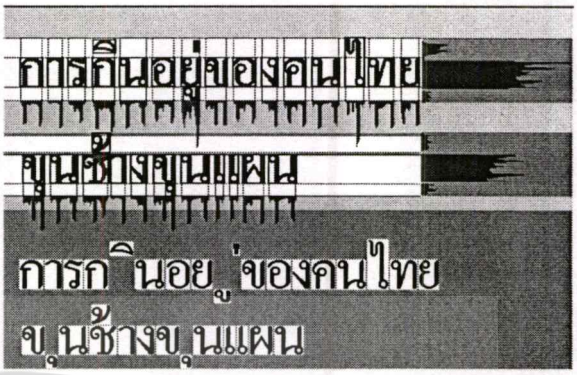
2.2 การหาค่าโปรเจคชั่น

Pixel Projection Profile [7] เป็นการแสดงค่าจำนวนจุดที่เป็นเนื้อของตัวอักษรในแนวตั้ง (Vertical Projection) และแนวนอน (Horizontal Projection) โดยทำการคำนวณจากสมการ

$$VerticalPXP(x) = \sum_{Y} P(x,y)$$

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

เมื่อ P(x,y) เป็นค่าของจุด ณ.ตำแหน่ง x และ y เราสามารถใช้การทำโปรเจคชั่นในการแยกบรรทัดของประโยค และแยกตัวอักษรออกจากประโยคได้ ดังแสดงในรูปที่ 2



รูปที่ 2 การทำ Horizontal และ Vertical Projection Profile

2.3 การหาค่าความแตกต่างของจุดข้างเคียง

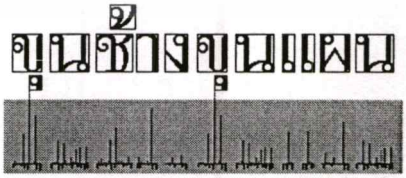
การหาค่าความแตกต่างของจุดข้างเคียง หรือ Peak to valley เป็นอัลกอริธึมที่พัฒนาโดย Kahan และ Pavlidis [4] มีสมการเป็นดังนี้

$$PV(x) = (V(x-1) - 2 \times V(x) + V(x+1)) / (V(x))$$

เมื่อ $V(x-1)$ คือจำนวนบิตที่เป็น 1 (จุดดำ) ที่สดมภ์ x-1

$V(x+1)$ คือจำนวนบิตที่เป็น 1 (จุดดำ) ที่สดมภ์ x+1

จากสมการข้างต้นเป็นการหาค่าเปรียบเทียบระหว่าง ตำแหน่ง x กับตำแหน่ง x+1 และ x-1 หากตำแหน่งที่เปรียบเทียบมีความแตกต่างกันมาก ค่า PV ที่ได้จะมีค่ามากกว่าศูนย์ ในทำนองเดียวกันถ้าค่าโปรเจคชั่นมีความแตกต่างกันน้อย ค่า PV ที่ได้อาจจะเป็นค่าลบหรือศูนย์ ดังแสดงตัวอย่างในรูปที่ 3



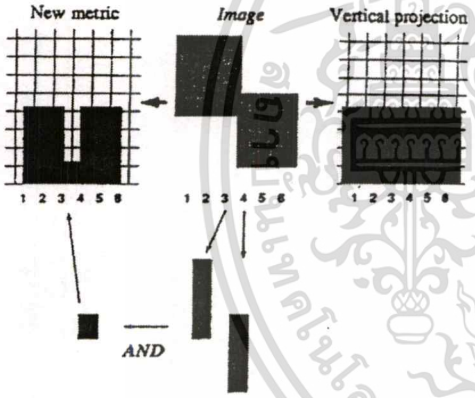
รูปที่ 3 ผลของการหาค่า Peak to valley ของตัวอักษรไทย

2.4 การหาค่า Break Cost เพื่อกำหนดจุดตัด

โดยทั่วไปการหาจุดตัดของตัวอักษรที่ติดกันอย่างง่ายๆ จะใช้วิธีการหาค่า Pixel Projection เพื่อกำหนด

ตำแหน่งของจุดตัดจากตำแหน่งที่มีค่าโปรเจคชั่นที่น้อยที่สุด แต่การใช้วิธีการ Break cost [3] ค่าที่คำนวณได้จะบอกถึงนัยสำคัญของการสัมผัสกัน (Degree of contact) ของแต่ละคอลัมน์ที่ติดกัน วิธีการนี้คำนวณโดยการนับจำนวนจุดในแนวตั้งที่ได้ จากการ AND กันของเนื้อภาพในคอลัมน์ที่ติดกันดังตัวอย่างในรูปที่ 4

จากรูปที่ 4 จะเห็นว่าวิธีการ Break cost เมื่อแสดงค่าที่ได้จากการ AND กันจะแสดงจุดสัมผัสได้อย่างชัดเจน ในขณะที่วิธีการ Vertical Pixel Projection ไม่สามารถแสดงผลได้ ดังนั้นวิธีการนี้เราสามารถทราบถึงบริเวณที่มีการสัมผัสกันน้อยที่สุด เพื่อกำหนดเป็นตำแหน่งของจุดตัดของตัวอักษรที่ติดกันได้



รูปที่ 4 แสดงนัยสำคัญของการสัมผัสกัน

3. ประเภทของตัวอักษรที่สัมผัสกัน

คือ การที่บางส่วนของตัวอักษรมากกว่า 1 ตัวมีการสัมผัสกัน ซึ่งอาจเกิดขึ้นได้ทั้งในระดับเดียวกันและต่างระดับกัน สาเหตุอาจเกิดจากการแสกน หรือเกิดจากรูปแบบของตัวอักษรเองก็ได้ การวิเคราะห์รูปแบบของการสัมผัสกันของตัวอักษรสามารถแบ่งได้ใน 2 ลักษณะคือ

3.1 อักษรที่สัมผัสกันในแนวตั้ง

ลักษณะการสัมผัสกันอาจจะเกิดจากสระระดับบนติดกับวรรณยุกต์ พยัญชนะติดกับสระระดับบน พยัญชนะติดกับวรรณยุกต์ หรือพยัญชนะติดกับสระระดับล่างก็ได้

3.2 อักษรที่สัมผัสกันในแนวนอน

ลักษณะการสัมผัสกันอาจจะเกิดจากการสัมผัสกันของพยัญชนะด้วยกันเอง ส่วนบนของพยัญชนะระดับกลาง+ระดับบนติดกับสระ หรือวรรณยุกต์ก็ได้

ซึ่งในการรวบรวมข้อมูลตัวอย่างจากหนังสือพิมพ์และเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ จำนวน 20,000 ตัวอักษร สามารถทำการแบ่งประเภทการสัมผัสกันของตัวอักษร และจำนวนที่พบการสัมผัสกันได้ดังตารางที่ 2

ตารางที่ 2 ประเภทการสัมผัสกันของตัวอักษรไทยจากแหล่งข้อมูลต่างๆ

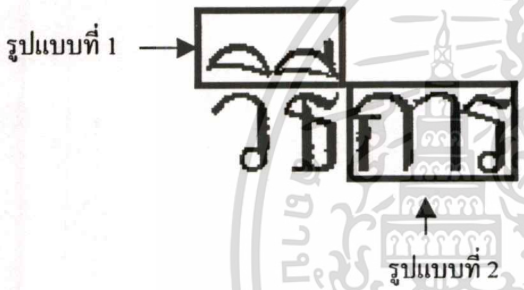
รูปแบบที่	ตัวอย่าง	จำนวนที่พบการติดกัน
1. สระระดับบนติดกันแนวนอน	วิธี	2 (0.01%)
2. อักษรระดับกลางติดกันแนวนอน	กลาง	610 (3.05%)
3. สระระดับบนติดกันในแนวตั้ง	ทั้ง	203 (1.02%)
4. อักษรระดับกลางติดกับสระระดับล่างในแนวตั้ง	ลูก	33 (0.2%)
5. อักษรระดับกลางติดกับสระระดับบนในแนวตั้ง	ส้ม, สั, ปี, ฝี, ปี	18 (0.1%)
6. ติดกันทั้งในแนวตั้งและแนวนอน	นี้ไม่	- -

4. การวิเคราะห์และแยกตัวอักษรที่สัมผัสกัน

จากตารางที่ 2 จะเห็นได้ว่าการติดกันของตัวอักษรไทยเกิดได้ในหลายรูปแบบ ซึ่งในแต่ละแบบจะใช้วิธีในการหาจุดแบ่งตำแหน่งที่ติดกันของตัวอักษรที่แตกต่างกัน ดังนั้นบทความนี้จึงจะนำเสนอวิธีในการหาจุดแบ่งตัวอักษรในแต่ละรูปแบบดังต่อไปนี้

4.1 รูปแบบที่ 1 และ 2

ทั้ง 2 แบบเป็นการติดกันในแนวนอน อาจเกิดจากการการติดกันของสระกับสระ หรือพยัญชนะกับพยัญชนะ ดังตัวอย่างในรูปที่ 5

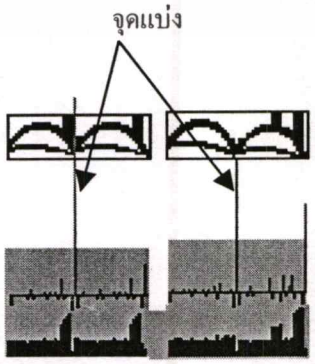


รูปที่ 5 ตัวอย่างของการสัมผัสกันในรูปแบบที่ 1 และ 2

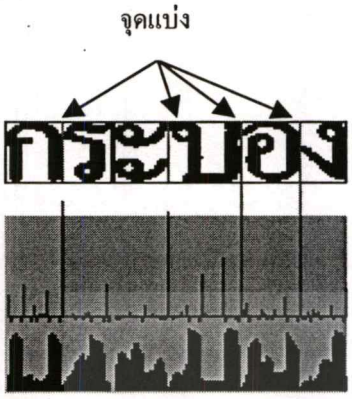
จะเห็นได้ว่าตัวอักษรที่ติดกันในรูปแบบนี้จะมีความกว้างที่ผิดปกติ ดังนั้นจึงสามารถใช้ค่าความกว้างในการตรวจสอบการติดกันได้ แต่ความกว้างมาตรฐานที่จะนำมาใช้ในการเปรียบเทียบนั้นจะใช้ค่าความกว้างของตัวใดตัวหนึ่งไม่ได้ ทั้งนี้เนื่องจากความแตกต่างกันของความกว้างที่มีอยู่หลายขนาด เช่น เ, ก, ญ บทความนี้จึงได้ทำการทดลองเพื่อหาความกว้างมาตรฐานของตัวอักษร ซึ่งพบว่าตัวอักษรไทยโดยส่วนใหญ่จะมีค่าความกว้างประมาณ 0.7-0.9 เท่าของความสูงของตัวอักษร ดังนั้นความกว้างมาตรฐานที่ใช้ในการเปรียบเทียบในงานวิจัยนี้จะใช้ค่าเท่ากับ 0.8 เท่าของความสูงตัวอักษรในระดับกลาง

สำหรับวิธีการหาจุดแบ่งตัวอักษรที่ติดกันจะใช้สมการในการหาค่า PV ในข้อ 2.3 และนำมาพิจารณาร่วมกับความกว้างมาตรฐาน(W) ดังต่อไปนี้

1. ถ้าความกว้างของตัวอักษรมีขนาดมากกว่า 2.2 เท่าของ W แสดงว่าอาจมีการสัมผัสกันของตัวอักษรมากกว่า 2 ตัว จึงต้องทำการแบ่งการหาค่า PV ออกเป็นช่วงๆ โดยมีขนาดเท่ากับ 1.45 เท่าของ W
2. คำนวณค่า PV(x)
3. หาตำแหน่ง x ที่มีค่า PV สูงสุด
4. ถ้าตำแหน่ง x มากกว่า 1.45 เท่าของ W หาค่า PV(x) ใหม่ทางซ้ายของจุด x
5. ถ้าตำแหน่ง x น้อยกว่า 0.6 เท่าของ W หาค่า PV(x) ใหม่ทางขวาของจุด x
6. ถ้า x >= 0.6 เท่าของ W และ (2.2 * W – x) มากกว่า 0.6 เท่าของ W และ PV(x-1) เป็นค่าลบ และค่าโปรเจคชั่นตามแนวตั้งที่ตำแหน่ง x มีค่าน้อยที่สุดเมื่อเทียบกับจุดข้างเคียง
7. ถ้าข้อ 6 เป็นจริง ใช้ตำแหน่ง x นี้เป็นจุดที่ใช้แบ่งตัวอักษรได้
8. ถ้าข้อ 6 เป็นเท็จให้ทำการหาค่า PV(x) ที่มีค่ามากเป็นอันดับ 2, 3, 4 ,5 มาทำการพิจารณาตามข้อ 4, 5 และ 6 ใหม่อีกครั้ง
9. ถ้า (x_{max} – x) < W จะหยุดกระบวนการตัดแยกตัวอักษร
10. กลับไปทำข้อ 1 โดยเลื่อนตำแหน่ง x_{min} ให้เป็นตำแหน่ง x



รูปที่ 6 ตัวอย่างของการหาจุดแบ่งของรูปแบบที่ 1



รูปที่ 6 ตัวอย่างของการหาจุดแบ่งของรูปแบบที่ 2

4.2 รูปแบบที่ 3

ในแบบนี้จะเกิดจากการติดกันของสระและวรรณยุกต์ระดับบนในแนวตั้ง ดังแสดงในรูปที่ 7

รูปที่ 7 ตัวอย่างของการสัมผัสกันในรูปแบบที่ 3

จะเห็นได้ว่าความสูงของสระหรือวรรณยุกต์ที่สัมผัสกันนั้น จะมีความสูงมากกว่าปกติ ซึ่งในกรณีนี้สามารถใช้ค่าความสูงของสระในระดับบนมาใช้เปรียบเทียบเพื่อวิเคราะห์การติดกันของตัวอักษรได้ สำหรับการหาค่าเฉลี่ยของความสูงของสระในระดับบนจะใช้การหาค่าฐานนิยมของความสูงทั้งหมดของสระและวรรณยุกต์ในระดับบนทั้งบรรทัด และทำการวิเคราะห์การติดกันดังนี้

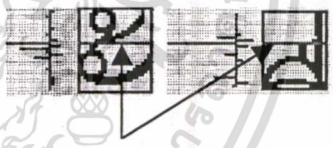
1. เปรียบเทียบความสูงของสระในระดับบนกับค่าความสูงเฉลี่ย ถ้ามากกว่า 1.5 เท่าของความสูงเฉลี่ย จะวิเคราะห์ว่าเป็นตัวอักษรที่ติดกันในรูปแบบที่ 3

2. หาจุดแบ่งตัวอักษรที่ติดกันโดยใช้สมการ PV ในข้อ 2.3 และนำมาพิจารณาร่วมกับความสูงเฉลี่ย ดังต่อไปนี้

- หาตำแหน่ง y ที่มีค่า PV สูงที่สุด
- พิจารณาตำแหน่ง y นั้นกับค่าความสูง คือ จะทำการแบ่งได้ก็ต่อเมื่อ ความสูงของตัวอักษรที่แบ่งแล้วมีความสูงมากกว่า 0.6 เท่าของความสูงเฉลี่ย
- ตรวจสอบการตัดที่ตำแหน่ง y จาก xmin ถึง xmax ซึ่งจะต้องมีการตัดผ่านจุดดำตลอดแนว จะมีการตัดผ่านจุดดำสลับกับจุดขาวไม่ได้ เพื่อป้องกันการตัดผ่านส่วนหัว หรือส่วนในของตัวอักษร

3. หากตำแหน่ง y ในข้อ 2 ใช้เป็นจุดแบ่งไม่ได้ จะนำวิธีการในข้อ 1.4 มาใช้แทนการทำโปรเจคชั่นแนวนอน และทำการหาตำแหน่งที่ใช้แบ่งตัวอักษรเช่นเดียวกับข้อ 2

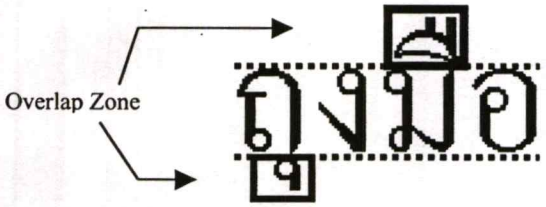
4. หากตำแหน่ง y ในข้อที่ 3 ใช้ไม่ได้ จะใช้วิธีการแบ่งโดยใช้ค่าเฉลี่ยของความสูงระดับบนเป็นตำแหน่งที่ใช้แบ่ง และใช้วิธีการในข้อที่ 2 พิจารณาร่วมด้วย



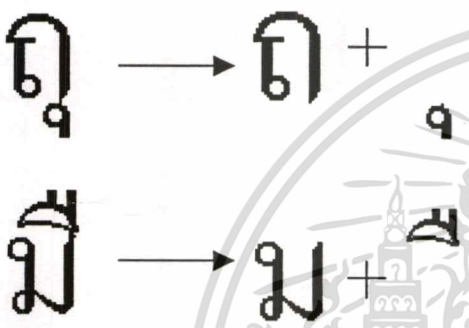
รูปที่ 8 ตัวอย่างของการหาจุดแบ่งของรูปแบบที่ 3

4.3 รูปแบบที่ 4 และ 5

สำหรับการสัมผัสกันในรูปแบบที่ 4 และ 5 จากตารางที่ 2 นั้นสามารถตรวจสอบการเหลื่อมล้ำในระดับบนหรือระดับล่าง ถ้าความสูงของส่วนที่เหลื่อมล้ำ (Overlapped zone) มากกว่า 0.4 เท่าของความสูงในระดับกลาง จะทำการวิเคราะห์ว่ามีการสัมผัสกันในแนวตั้งระหว่างพยัญชนะกับสระระดับบน หรือล่าง และสามารถแบ่งออกเป็น 2 ส่วนโดยการใช้เส้นแบ่ง up และ ba ดังรูปที่ 9



(ก)



(ข)

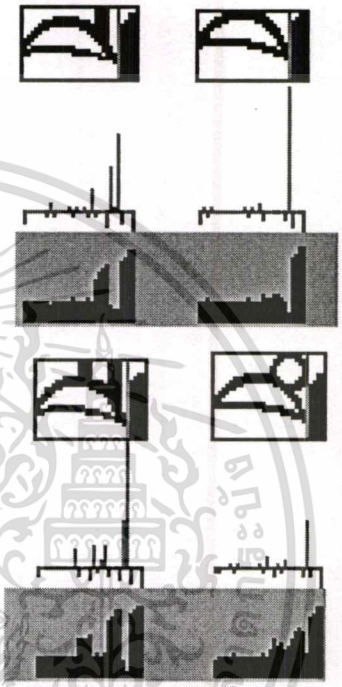
รูปที่ 9 (ก) ตัวอย่างของการสัมผัสกันในรูปแบบที่ 4, 5
(ข) วิธีการแบ่งโดยใช้แนวการแบ่งตามเส้น up และ ba

จากวิธีนี้สามารถแก้ปัญหาของการใช้ฮีสโตแกรมวิเคราะห์ในงานวิจัยก่อนๆ ได้ ซึ่งไม่สามารถวิเคราะห์ตัวอักษรที่มีความคล้ายคลึงกันได้เช่น ถ้าตัวอักษรในกลุ่ม (บ, พ, ผ) ติดกับไม้เอก จะทำให้ไม่สามารถแยกความแตกต่างระหว่าง (ป, ฟ, ฝ) ได้ เป็นต้น

การสัมผัสกันอีกรูปแบบหนึ่งที่ในงานวิจัยก่อนๆ ประสบปัญหาในการวิเคราะห์และการตัดแยกคือการติดกันของตัวอักษรและมีการเหลื่อมล้ำกัน ดังรูปที่ 10 ทำให้ไม่สามารถแบ่งแยกได้อย่างถูกต้อง

รูปที่ 10 การติดกันของตัวอักษรและมีการเหลื่อมล้ำกัน

ในงานวิจัยนี้ใช้การตรวจสอบความสูงของส่วนที่เหลื่อมล้ำ และใช้สมการ PV ในการหาจุดแบ่งแยกตัวอักษร ดังผลในรูปที่ 11



รูปที่ 11 ผลของการหาจุดแบ่งโดยใช้สมการ PV

4.4 รูปแบบที่ 6

การสัมผัสกันในรูปแบบนี้ จะใช้อัลกอริธึมในการแยกของวิธีในข้อ 4.1 และ 4.2

5. ผลการทดลอง

ได้ทำการทดลองวิธีการที่นำเสนอนี้โดยใช้ข้อมูลจากหนังสือพิมพ์มติชนและเอกสารที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์จำนวนทั้งหมด 20,000 ตัว แสกนที่ความละเอียด 300 dpi การทดสอบอัลกอริธึมนี้ใช้โปรแกรม Microsoft Visual C++ 6.0 และทำการนับจำนวนอักษรที่สัมผัสกันตามแนวตั้ง และแนวนอน ได้ผลการทดลองดังตารางที่ 3

ตารางที่ 3 ผลการทดสอบอัลกอริธึมข้างต้นกับข้อมูลตัวอย่าง

แหล่งข้อมูล	ข้อมูลตัวอย่าง	
ลักษณะการสัมผัส	แนวนอน	แนวตั้ง
จำนวนทั้งหมด	20,000	
จำนวนที่พบการสัมผัสกัน	631	252
% การสัมผัสกัน	3.16%	0.13%
จำนวนที่หาจุดตัดได้	611	252
% ความถูกต้อง	96.83%	100%
% ความถูกต้องเฉลี่ย	98.41%	

จากผลการทดลอง แสดงให้เห็นว่าวิธีการที่นำเสนอนี้สามารถแบ่งตัวอักษรที่สัมผัสกันได้อย่างมีประสิทธิภาพและถูกต้องสูง

6. สรุป

จากอัลกอริธึมที่ใช้ในการแบ่งแยกตัวอักษรที่สัมผัสกันในแนวตั้ง และแนวนอนที่ได้นำเสนอมาแล้วนั้น สามารถใช้ได้กับตัวอักษรไทยได้เป็นอย่างดี สามารถที่จะแบ่งตัวอักษรที่สัมผัสกันออกเป็นตัวอักษรเดี่ยวๆ ได้ อีกทั้งยังสามารถแก้ปัญหาที่เกิดจากการวิเคราะห์ด้วยฮีสโตแกรมเพียงอย่างเดียว แต่อย่างไรก็ตามวิธีนี้ยังพบข้อผิดพลาดเช่นกัน เช่น จากตารางที่ 3 มีความผิดพลาดประมาณ 2% ซึ่งมีผลมาจากสัญญาณรบกวน ทำให้การคำนวณค่าผิดพลาด หรือเกิดจากการซ้อนทับกันของตัวอักษรจึงทำให้บริเวณจุดที่สัมผัสกันไม่มีจุดดำรอบข้างที่แตกต่างกัน ทำให้ค่า PV(x) มีค่าน้อยมาก และไม่สามารถใช้ทำการหาจุดที่ใช้แยกได้ ซึ่งได้ทำการศึกษาหาจะต้องมีอัลกอริธึมเพื่อแก้ปัญหาในส่วนนี้ต่อไป

6. เอกสารอ้างอิง

[1] ศุภกร รัตนปราการ, บุญธีร์ เครือตราชู, "การวิเคราะห์การติดกัน และการตัดแยกของตัวอักษรพิมพ์ไทยโดยใช้คุณลักษณะทางแนวตั้งและแนวนอนของฮีสโตแกรม," *วารสารสารสนเทศลาดกระบัง*, ฉบับที่ 1, ปีที่ 1, กรกฎาคม, หน้า 21-30, 2542.

[2] Nucharee Premchaiswadi, Wichian Premchaiswadi, Seinosuke Narita, **"Segmentation Of Horizontal and Vertical Touching Thai Character,"** *ITC-CSCC'99 International Technical Conference on Circuit Systems, Computers and Communications*, 1999.

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

[4] S.Kahan and Pavlids, **"On the Recognition of printed Character of Any Font and Size,"** *IEEE Trans Patt. Anal. Machine Intell*, VolPAMI-9, pp.274-287, March 1987.

[5] N.W. Strathy ANDG, et. al, **"Segmentation Of HandWritting Digit Using Contour Features,"** *Proceeding of the 2nd ICDAR*, pp. 570-580, 1993.

[6] Wicha Panich, Somchai Jitapunkul, Prasert Choruengwiwat, **"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.

[7] Su Liang M. Shridhar, M. Ahmadi, **"Efficient Algorithms for Segmentation and Recognition Of Printed Characters In Document Process,"** *Proceeding of the 2nd ICDAR*, pp. 240-243, 1993.

ประวัติผู้เขียน

ชื่อ	นายจักริน สุขสวัสดิ์ชน
เกิดเมื่อ	วันที่ 6 เมษายน 2519
สถานที่เกิด	จังหวัดชลบุรี
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
สถานที่สำเร็จการศึกษา	คณะวิทยาศาสตร์ มหาวิทยาลัยบูรพา ต.แสนสุข อ.เมือง จ.ชลบุรี
ปี พ.ศ. ที่สำเร็จการศึกษา	2541
ทุนการศึกษาที่ได้รับ	ทุนอุดหนุนการศึกษา โครงการพัฒนาอาจารย์ มหาวิทยาลัยบูรพา จ.ชลบุรี