

การแบ่งตัวอักษรไทยที่สัมพันธ์กัน

นุชรี เปรมชัยสวัสดิ์

ผู้ช่วยศาสตราจารย์

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น

บทคัดย่อ

การแบ่งตัวอักษรไทยที่สัมพันธ์กัน เป็นขั้นตอนที่มีความสำคัญมากขั้นตอนหนึ่งในการเตรียมข้อมูลเพื่อให้ข้อมูลอยู่ในลักษณะที่เหมาะสมกับการนำไปใช้ในขั้นตอนการตรวจรู้ตัวอักษรไทยด้วยคอมพิวเตอร์ บทความนี้ได้เสนอแนวทางในการแบ่งตัวอักษรไทยที่สัมพันธ์กัน โดยใช้คุณสมบัติเฉพาะของตัวอักษรไทยและค่าทางสถิติ เริ่มต้นจากการตรวจสอบว่ามีกลุ่มของตัวอักษรใดบ้างที่สัมพันธ์กันโดยเทียบจากความกว้างมาตรฐาน จากนั้นก็นำกลุ่มตัวอักษรที่สัมพันธ์กันมาแยกออกจากกันให้เป็นอักษรเดี่ยว โดยการคำนวณหาจุดแบ่งตัวอักษร จากค่าโปรเจกชันและความกว้างมาตรฐาน จากผลทดลองพบว่าเทคนิคนี้สามารถแบ่งตัวอักษรไทยที่ติดกันได้อย่างมีประสิทธิภาพ

Thai Touching Character Segmentation

Nucharee Premchaiswadi

Assistance Professor

Department of Computer Engineering

Faculty of Engineering Khon Kaen University

Abstract

Thai touching character segmentation is one of the most importance in the process of input processing for making input data into a suitable form for character recognition. This paper presents the method of touching character segmentation based on characteristic of Thai character and statistical value. The process begins with finding touching character. Then, these touching character will be separated into single character by using projection and standard width of character. The experimentation results shown that the technique can be used with Thai touching character efficiently.

บทนำ

ในระบบการตรวจรู้ตัวอักษรด้วยคอมพิวเตอร์นั้น การเตรียมข้อมูลเพื่อใช้ในการตรวจรู้ตัวอักษรด้วยคอมพิวเตอร์นั้นจัดได้ว่าเป็นสิ่งจำเป็นและที่มีความสำคัญมาก เนื่องจากถ้าขั้นตอนนี้ไม่มีประสิทธิภาพเพียงพอ จะทำให้เกิดความยุ่งยากและต้องใช้เทคนิคที่มีความสลับซับซ้อนมากขึ้นในการตรวจรู้ตัวอักษร ทั้งยังอาจทำให้ได้ผลลัพธ์ที่ไม่ถูกต้องอีกด้วย การแบ่งตัวอักษรที่สัมผัสกันเป็นขั้นตอนสำคัญขั้นตอนหนึ่งในขั้นตอนการเตรียมข้อมูลสำหรับระบบการตรวจรู้ตัวอักษรเนื่องจากข้อมูลที่ได้รับมาจากเครื่องกวาดภาพนั้น ในบางครั้งอาจมีสัญญาณรบกวนทำให้ตัวอักษรจากต้นฉบับที่ไม่สัมผัสกัน กลายเป็นตัวอักษรที่สัมผัสกัน หรือสัมผัสกันมาตั้งแต่ต้นฉบับ จึงมีความจำเป็นที่ระบบจะต้องสามารถตรวจรู้ได้โดยอัตโนมัติว่ามีตัวอักษรที่สัมผัสกันหรือไม่ ถ้ามี ต้องทำการแบ่งตัวอักษรออกให้เป็นอักษรเดี่ยวทั้งหมด เพื่อที่จะส่งให้ระบบการตรวจรู้ตัวอักษรทำการประมวลผลต่อไป

การแบ่งแยกตัวอักษรไทยที่สัมผัสกัน

การแบ่งตัวอักษรที่สัมผัสกันนั้น มีความซับซ้อน เนื่องจากความหลากหลายของรูปแบบตัวอักษร และการเพิ่มขึ้นอย่างรวดเร็วของจำนวนรูปแบบตัวอักษร อีกส่วนหนึ่งก็เนื่องจากคุณภาพของรูปภาพเอง เช่น ภาพที่เกิดจากเครื่องพิมพ์ที่มีคุณภาพต่ำ วิธีการแบ่งตัวอักษรที่สัมผัสกัน มีแนวทางที่แตกต่างกันสองแนวทาง[1] คือ การแบ่งตัวอักษรที่รวมกับการตรวจรู้ตัวอักษร หรืออาศัยลักษณะทางสถิติของตัวอักษร ซึ่งเป็นการพัฒนาและทดลองใช้กับตัวอักษรภาษาอังกฤษเป็นส่วนใหญ่ ในกรณีของการแบ่งตัวอักษรไทยที่สัมผัสกันนั้น ในบทความนี้จะใช้ลักษณะทางสถิติมาใช้ในการแบ่งตัวอักษรที่สัมผัสกัน ซึ่งมีขั้นตอนด้วยกันคือ การตรวจหากลุ่มตัวอักษรที่สัมผัสกัน และการหาตำแหน่งจุดแบ่งตัวอักษร

1. การตรวจหากลุ่มตัวอักษรที่สัมผัสกัน

ตัวอักษรที่สัมผัสกันอาจจะเป็นการติดกันอันเกิดจากการสัมผัสกันของตัวอักษรตั้งแต่สองตัวขึ้นไป ดังนั้นในการตรวจสอบหาว่าตัวอักษรกลุ่มใดบ้างที่สัมผัสกันนั้น จะใช้หลักการของความกว้างตัวอักษร โดยใช้ค่าทางสถิติที่ว่า ตัวอักษรปกตินั้นส่วนมากจะมีความกว้างน้อยกว่าสองเท่าของความกว้างเฉลี่ยของตัวอักษรในบรรทัดนั้น แต่ในความเป็นจริงอาจจะมีตัวอักษรบางตัวที่ติดกันแล้วมีความกว้างน้อยกว่าสองเท่าของความกว้างเฉลี่ยได้ สาเหตุก็เนื่องมาจากความกว้างของตัวอักษรบางตัวนั้นน้อยกว่าความกว้างเฉลี่ยนั่นเอง เช่น การติดกันของตัวอักษร ร กับ ว เป็นต้น ดังนั้นในการพิจารณาจึงใช้ค่า 80 เปอร์เซ็นต์ ของสองเท่าของความกว้างตัวอักษรเฉลี่ยในการเปรียบเทียบแทน

2. การหาตำแหน่งจุดแบ่งตัวอักษร

หลังจากที่สามารถตรวจสอบหากลุ่มของตัวอักษรที่ติดกันได้แล้ว ก็จะทำการหาจุดแบ่งตัวอักษร โดยการนำเอาอัลกอริทึม ที่พัฒนาขึ้นโดย Kahan และ Pavlidis[2] และถูกปรับปรุงเพิ่มเติมโดย Lu[1] ดังสมการ

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

โดยที่

$V(x-1)$ คือ จำนวนบิตที่เป็นหนึ่งที่สดมภ์ $x-1$

$V(x+1)$ คือ จำนวนบิตที่เป็นหนึ่งที่สดมภ์ $x+1$

มาประยุกต์ใช้ เพื่อหาจุดที่มีแนวโน้มว่าจะเป็นจุดแบ่งตัวอักษรได้ อัลกอริทึมที่ใช้ในการหาจุดแบ่งมีการทำงานดังนี้ ในตอนแรกจะเป็นขั้นตอนในการหาช่วงที่จะสามารถใช้เป็นจุดแบ่งได้จากค่าประมาณความกว้างของตัวอักษร ตำแหน่งที่มีค่า $PV(x)$ สูงที่สุดก็คือตำแหน่งที่จะใช้เป็นจุดที่มีแนวโน้มในการแบ่งตัวอักษรที่สัมพันธ์กัน เนื่องจากความกว้างของตัวอักษรที่คำนวณได้ เป็นค่าโดยประมาณ จึงต้องคำนวณหาค่า $PV(x)$ ที่ตำแหน่งข้างเคียงมาเปรียบเทียบเพื่อหาตำแหน่งที่มีค่า PV สูงสุดเป็นจุดแบ่งตัวอักษร หลังจากได้จุดแบ่งตัวอักษร ก็จะนำส่วนย่อยทั้งสองนี้มาพิจารณาความกว้างอีกครั้ง จนกระทั่งความกว้างอยู่ในเกณฑ์ที่ยอมรับให้เป็นตัวอักษรตัวเดียวได้ แต่ถ้าความกว้างของส่วนย่อยนั้นมีแนวโน้มที่จะเป็นตัวอักษรหลายตัวได้ ก็จะต้องทำขั้นตอนของการหาจุดแบ่งต่อไป

รายละเอียดขั้นตอนการทำงานในการตัดตัวอักษรไทยที่สัมพันธ์กันมีดังนี้

- 1 ตัดตัวอักษรในแต่ละบรรทัดออกเป็นกลุ่มหรืออักษรเดี่ยวโดยใช้ช่องว่างระหว่างตัวอักษรหรือกลุ่มของตัวอักษร
- 2 กลุ่มตัวอักษรที่เกิดจากข้อ 1 อาจไม่ใช่ตัวอักษรที่สัมพันธ์กัน ซึ่งสามารถแยกเป็นตัวอักษรเดี่ยวได้โดยเทคนิคการหาขอบภาพ[3]
- 3 หาความกว้างของตัวอักษรแต่ละตัว เมื่อได้แล้วก็จะนำไปหาค่าเฉลี่ย เพื่อที่จะใช้เป็นค่าเริ่มต้นในการหาความกว้างมาตรฐาน
- 4 หาค่าความกว้างมาตรฐานของตัวอักษร โดยนำเอาค่าเฉลี่ยความกว้างของตัวอักษรที่มีความกว้างไม่มากกว่าสองเท่า และไม่น้อยกว่าครึ่งหนึ่งของค่าเฉลี่ยเริ่มต้น มาใช้เป็นค่าความกว้างมาตรฐาน
- 5 ทำการตรวจหาตัวอักษรที่สัมพันธ์กัน
- 6 หาค่าโปรเจกชันในแต่ละสดมภ์ ของกลุ่มตัวอักษรที่สัมพันธ์กัน ซึ่งค่าโปรเจกชันของแต่ละสดมภ์ ก็คือผลรวมของบิตที่เป็น 1 ในสดมภ์นั้นนั่นเอง
- 7 คำนวณหาตำแหน่งจุดแบ่งตัวอักษร
- 8 ทำการตรวจสอบความเป็นไปได้ของตัวอักษรที่จะถูกตัด ว่าน่าจะเป็นการตัดที่ถูกต้องหรือไม่ โดยพิจารณาจากความกว้างของตัวอักษรที่ได้จากการตัดจะต้องไม่น้อยจนเกินไป โดยในที่นี้จะต้องไม่น้อยกว่า 3 ใน 4 ของตัวอักษรมาตรฐาน

9 ตัดตัวอักษร ณ. ตำแหน่งที่ได้ในข้อ 7

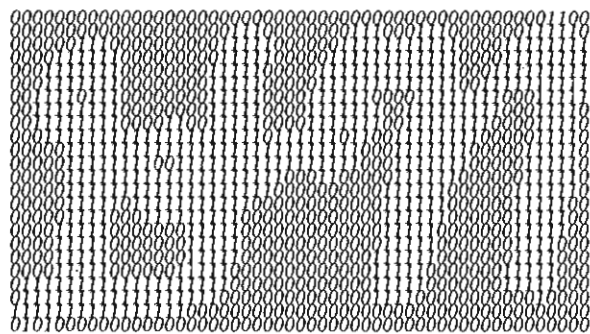
การทดสอบ

วิธีการที่นำเสนอได้ถูกทำการทดสอบการทำงาน โดยอาศัยตัวอย่างจริงของเอกสารที่มีตัวอักษรที่ติดกัน ซึ่งตัวอย่างนี้ไม่ได้เกิดจากการจงใจทำให้ตัวอักษรมาสัมผัสกัน แต่เป็นเอกสารที่มีตัวอักษรสัมผัสกันอยู่แล้ว ดังเช่นตัวอย่างแรกที่เป็นเอกสารในรูปที่ 1

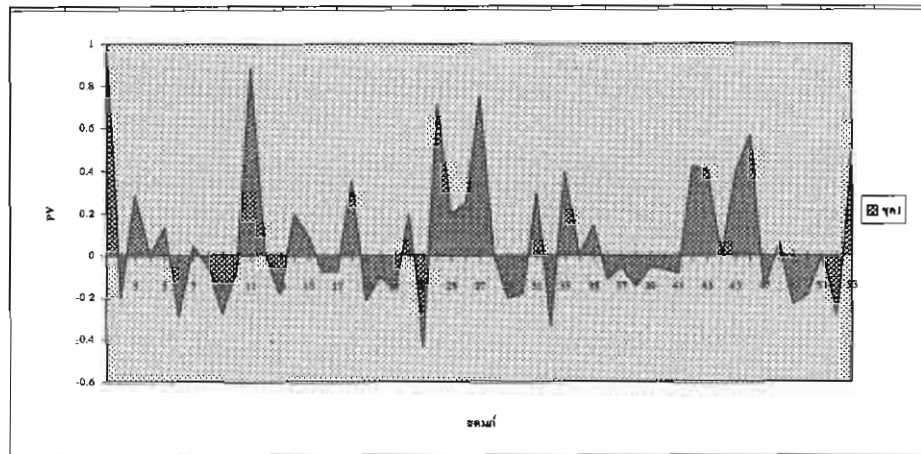
สรุปนิษั ในคัมภีรภควทคิตา คือพรหมพิทยาโยค
ศาสตร ว่าด้วยกถาสวาทะหว่างศรีภุณกับอรุณ ชื่อว่า
ไทวาสุสมบัตวิภาคโยโค อันเป็นโศทโส' ธยายะ
ดังนี้

รูปที่ 1 ตัวอย่างของเอกสารที่ 1

ผลจากการทำงานของโปรแกรมการแบ่งตัวอักษร พบตัวอักษรที่สัมผัสกัน 1 กลุ่ม คือ ข และ ๗ ดังแสดงในรูปที่ 2 และกราฟแสดงค่าของ PV ที่คำนวณได้เพื่อหาตำแหน่งที่เหมาะสมในการแบ่งตัวอักษรได้แสดงไว้ในรูปที่ 3

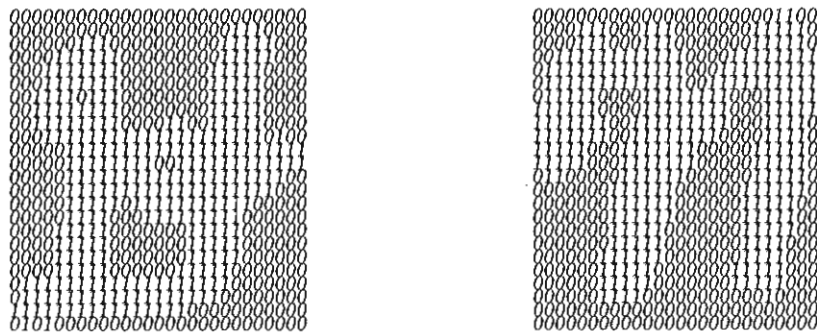


รูปที่ 2 กลุ่มตัวอักษรที่สัมผัสกันของรูปที่ 1



รูปที่ 3 กราฟค่าของ PV ในรูปที่ 2

จากรูปที่ 3 จะเห็นได้ว่ามีค่า PV สูงสุดที่สัดมภ์ที่ 11 แต่เนื่องจากค่าความกว้างที่น้อยเกินไปที่จะเป็นตัวอักษรเดี่ยวได้จึงไม่ถูกเลือก จุดต่อไปที่ให้ค่า PV สูงสุดมีสองจุดคือ ที่สัดมภ์ที่ 24 และ 27 และเมื่อพิจารณาถึงความกว้างของตัวอักษรก็มีความเป็นไปได้ทั้งสองจุด จึงเลือกจุดที่ให้ค่า PV สูงสุดคือที่สัดมภ์ที่ 27 ดังผลลัพธ์ที่แสดงในรูปที่ 4



รูปที่ 4 ผลลัพธ์จากการแบ่งตัวอักษร

สรุปและข้อเสนอแนะ

จากผลการทดลองที่ใช้ตัวอย่างจากเอกสารจริง พบว่าวิธีการนี้สามารถทำการแบ่งตัวอักษรที่สัมผัสกันได้อย่างมีประสิทธิภาพ แต่อย่างไรก็ตามวิธีการนี้ก็ยังมีข้อผิดพลาดได้ ในกรณีที่ตัวอักษรเดี่ยวบางตัวนั้นมีขนาดความกว้างมากผิดปกติ โดยถ้าตัวอักษรนั้นมีความกว้างมากกว่าสองเท่าของความกว้างมาตรฐาน อาจจะถูกมองว่าเป็นตัวอักษรที่สัมผัสกันได้

กิตติกรรมประกาศ

ผู้เขียนขอขอบคุณนักศึกษาชั้นปีที่4 ภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ช่วยเหลือในการทดลอง และรวบรวมข้อมูล

เอกสารอ้างอิง

- 1) Li Lu, "On the Segmentation of Touching Character", IEEE Computers, 1993 pp.440-443
- 2) S.Kahan and Pavlids, "On the Recognition of Printed Characters of Any Font and Size", IEEE Trans. Patt. Anal. Machine Intell., Vol.PAMI-9, pp.274-287, March 1987
- 3) นุชรี เปรมชัยสวัสดิ์ "การเตรียมข้อมูลสำหรับการตรวจรู้ตัวอักษรไทยด้วยคอมพิวเตอร์", วิศวกรรมสาร มข. ปีที่22 ฉบับที่2 (89-107) กค.-ธค., 2538