บทที่ 5
สรุปผลการทดลอง

ระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของอักษรไทยถูกสร้างขึ้นเพื่อทดสอบแนวคิดที่จะใช้ลักษณะพิเศษของภาษาไทยในการรู้จำตัวอักษร การตัดแยกอักษรที่ติดกัน การตัดแบ่งบรรทัดและคอลัมน์

ในส่วนการรู้จำตัวอักษรภาษาไทยใช้การแบ่งกลุ่มโดยใช้ลักษณะของโครงสร้างหลักร่วมกับระดับของอักษร โดยแบ่งเป็นอักษรระดับบน 1 กลุ่ม ระดับล่าง 1 กลุ่ม และระดับกลางอีก 7 กลุ่ม แล้วจึงแยกแยะในกลุ่มย่อยโดยใช้ลักษณะบ่งความต่างของอักษรไทย ในส่วนการตัดแยกอักษรที่ติดกันนั้นใช้ลักษณะบ่งความต่างของอักษรไทยแบ่งประเภทของการติดกันโดยใช้ระดับของอักษรได้เป็น 10 กลุ่มแล้วใช้วิธีเฉพาะของแต่ละกลุ่มในการตัดแยก ในส่วนการวิเคราะห์เอกสารมีการแก้ความเอียงของเอกสาร การตัดคอลัมน์และตัดบรรทัด

จากการทดสอบระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของอักษรไทยยืนยันถึงแนวคิดที่ใช้ลักษณะพิเศษของอักษรไทยในการรู้จำว่าใช้ได้ผล คือได้อัตราความถูกต้องสูงถึง 97.4 % สำหรับอักษรต้นแบบ(AngsanaUPC และ CordiaUPC) และสำหรับอักษรแบบอื่นๆ ยังให้ความถูกต้องสูงมากกว่า 94 % ถ้าแบบอักษรนั้นมีแบบของโครงสร้างหลักและลักษณะบ่งความต่างเหมือนกับอักษรต้นแบบ และยังใช้ได้ดีกับเอกสารจริงอีกด้วยดังจะเห็นจากการทดสอบกับเอกสารจริงที่มีขนาดประมาณ 14 point ให้ความถูกต้องถึง 98% และขนาดประมาณ 13 point ให้ความถูกต้องถึง 96.4 % และเอกสารที่มีสัญญาณรบกวนมากและขนาดเล็กถึง 11 point ให้ความถูกต้องถึง 83.3 %

สำหรับอัตราเร็วในการรู้จำสำหรับอักษรขนาดเล็กที่สุดที่อยู่ในขอบเขตทดสอบ คือ 12 point สูงถึง 67.1 อักษรต่อวินาที สำหรับอักษรขนาดใหญ่คือ 24 point คือ 19.6 อักษรต่อวินาที และเฉลี่ยสำหรับทุกอักษรคือ 36.4 อักษรต่อวินาที

ความผิดพลาดที่เกิดขึ้นเกิดกับทุกส่วนของระบบโดยถ้าระบบตัดคอลัมน์ผิด จะเกิดความผิดพลาดในการเรียงลำดับ ถ้าระบบตัดบรรทัดผิดจะเป็นการผิดติดต่อกันทั้งบรรทัด ถ้าอักษรที่สแกนมีขนาดเล็ก ทำให้ติดกันมากจะเกิดความผิดพลาดในส่วนตัดแยกเพิ่มขึ้นมา แต่ความผิดพลาดที่เกิดขึ้นตลอดเวลาคือในส่วนการรู้จำอักษรเดี่ยวซึ่งเกิดจากสัญญาณรบกวนทำให้ลักษณะบ่งความต่างหายไป และอาจเกิดจากแบบของตัวอักษรมีพื้นฐานของโครงสร้างหลักหรือลักษณะบ่งความต่าง ต่างจากอักษรต้นแบบมาก

ข้อจำกัด

1. การเพิ่มอักษรที่ไม่ได้มีโครงสร้างหลักและลักษณะบ่งความต่างร่วมกับอักษรไทยต้นแบบทำได้ยาก เนื่องจากต้องทำโครงสร้างหลักเพิ่มแล้วให้ผู้ใช้ทำการเลือกว่าจะใช้แบบโครงสร้างชนิดใดซึ่งทำให้ระบบไม่เป็นแบบอัตโนมัติ
2. การตรวจสอบอักษรขาดทำได้ยาก ที่พอทำได้เนื่องจากผิดหลักการผสมคำเช่น เปลี่ยน "เ" เป็น "ถ" เป็นต้น แต่สำหรับอักษรที่ขาดแบบซับซ้อนไม่สามารถตรวจได้
3. เมื่ออักษรที่สแกนได้เอียงมาก เมื่อแก้กลับมาแล้วจะเกิดสัญญาณรบกวนเนื่องจากการแก้ความเอียงมาก มีผลให้ความถูกต้องในการรู้จำลดลง ดังนั้นจึงควรสแกนให้ตรง

งานในอนาคต

1. การมีโครงสร้างหลักหลายแบบให้ผู้ใช้มีทางเลือก ซึ่งมีผลให้ระบบไม่เป็นอัตโนมัติแต่ทำให้ความถูกต้องสำหรับแบบอักษรที่มีโครงสร้างหลักและลักษณะบ่งความต่างต่างออกไปสูงขึ้นได้
2. การตรวจสอบอักษรที่ขาด ซึ่งอาจใช้ฐานข้อมูลทางภาษาร่วมกับการต่อเส้น และการรู้จำแบบมีการตรวจสอบความสัมพันธ์ระหว่างอักษรข้างเคียงเพื่อเชื่อมต่ออักษรที่ขาดนั้น
3. สร้างระบบที่สามารถรู้จำเอกสารที่มีการจัดหน้าซับซ้อนขึ้นเช่น มีตาราง มีจำนวนคอลัมน์ไม่เท่ากันทั้งหน้า มีรูปภาพ
4. เพิ่มให้รู้จำอักษรภาษาอังกฤษได้โดยหาโครงสร้างร่วมที่เหมาะสม

ในระหว่างที่เราพัฒนางานวิจัยนี้ มีระบบรู้จำอักษรที่สามารถรู้จำอักษรไทยได้ 2 ระบบเกิดขึ้นคือ

1. ***"อ่านไทย"*** ของ ห้องปฏิบัติการวิจัยและพัฒนาเทคโนโลยีซอฟต์แวร์ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์ และคอมพิวเตอร์แห่งชาติ (NECTEC : the National Electronics and Computer Technology Center)
2. ***ThaiOCR®*** version 1.5 ของ Atrium Technology Co.,Ltd

โดยทั้ง 2 ระบบนี้พัฒนาอยู่บนพื้นฐานของโครงข่ายประสาทเทียม (ANN : Artificial Neural Networks) ทำให้สามารถเปลี่ยนฐานข้อมูลที่เหมาะสมได้เช่น ฐานข้อมูลแบบอักษรของหนังสือพิมพ์ไทยรัฐ เป็นเอกสารที่มีเฉพาะภาษาไทย หรือเป็นแบบมีทั้งภาษาไทยและภาษาอังกฤษ

จุฬารัตน์ ตันประเสริฐ (2539) ได้กล่าวถึงโปรแกรม "อ่านไทย" ว่าสามารถทำงานได้กับเอกสารที่เอียงจากการสแกน มีความถูกต้องในการรู้จำมากกว่า 94 % สำหรับเอกสารภาษาไทยแบบอักษร AngsanaUPC , BrowalliaUPC , CordiaUPC , EucrosiaUPC , FreesiaUPC , IrisUPC , JasmineUPC และความถูกต้องมากกว่า 90% สำหรับเอกสารที่มีทั้งภาษาไทยและอังกฤษ ที่ขนาดอักษร 8 point ขึ้นไป ได้กล่าวถึงเวลาในการรู้จำว่าประมาณ 1 นาทีต่อ 1 หน้า A4 (ไม่ทราบเงื่อนไข)

สำหรับโปรแกรม ThaiOCR ได้ออก version 1.1 ออกมาก่อนซึ่งมีการแบ่งกรอบคอลัมน์อัตโนมัติมาให้ด้วย แต่เมื่อออก version 1.5 การทำงานส่วนนี้ถูกตัดออกไป จากการทดสอบใช้โปรแกรมกับเอกสารจริงคือ บทความจากหนังสือ Advance Thailand Geographic ได้ความถูกต้อง 96% แบบอักษร AngsanaUPC ขนาด 12 point ได้ความถูกต้อง 81% ขนาด 14 point ได้ความถูกต้อง 93% และ 16 point ได้ความถูกต้อง 96% ความเร็วเฉลี่ย 15 อักษรต่อวินาที(เมื่อทดสอบด้วยเงื่อนไขเดียวกับงานวิจัย) จะเห็นว่าความถูกต้องต่างกันมากซึ่งจริงๆ แล้วความผิดพลาดจากการรู้จำอักษรเดี่ยวเพียง 4 % เห็นได้จากอักษร 16 point ซึ่งมีอักษรติดน้อยมาก แต่เมื่อมีอักษรติดมากขึ้นคือ 14 point และ 12 point ทำให้ความถูกต้องลดลงมากคาดว่าเพราะระบบไม่มีส่วนตรวจสอบและตัดแยกอักษรติด และจากการสังเกตพบว่าไม่ได้นำระดับของอักษรมาช่วยในการรู้จำเช่น อักษรติด “รู” ถูกรู้จำเป็น “โ” เป็นต้น เป็นทั้งข้อดีและเสียคือทำให้ใน 1 บรรทัดมีอักษรได้หลายขนาดแต่ทำให้ความถูกต้องลดลง

จะเห็นได้ว่าข้อได้เปรียบของทั้ง 2 ระบบคือสามารถเพิ่มอักษรที่รู้จำได้จากการเปลี่ยนฐานข้อมูล ข้อได้เปรียบของระบบเราคือการที่ใช้ระดับของอักษรและลักษณะของอักษรไทยช่วยในการแยกอักษรติดและส่วนรู้จำอักษรทำให้ความถูกต้องของระบบสูงขึ้น