บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันได้มีการจัดเก็บเอกสารในรูปแบบสื่ออิเล็กทรอนิกส์กันอย่างแพร่หลาย เนื่องจากมีข้อดีในการค้นหา แก้ไข และปรับปรุงเอกสารได้ง่าย อีกทั้งยังประหยัดที่ในการจัดเก็บอีกด้วย แต่อย่างไรก็ดีการจัดเก็บเอกสารในลักษณะที่เป็นรูปภาพบิทแมปนั้น เป็นการสิ้นเปลืองหน่วยความจำ และนำข้อมูลมาประมวลผลได้ยาก ดังนั้นจึงได้มีการวิจัย คิดค้นพัฒนาวิธีการทำให้เครื่องคอมพิวเตอร์สามารถที่จะอ่านตัวอักษรต่างๆ โดยใช้เครื่องสแกน(Scanner) แล้วทำการวิเคราะห์รูปภาพตัวอักษรตัวพิมพ์ต่างๆ เพื่อแปลงให้เป็นตัวอักษรแบบเท็กซ์ แล้วจัดเก็บในลักษณะเท็กซ์ไฟล์ ซึ่งนิยมเรียกกันว่า การรู้จำตัวอักษรไทย(Thai Character Recognition) วิธีการรู้จำอักษรโดยทั่วไปจะประกอบไปด้วย 3 ขั้นตอนได้แก่ ขั้นตอนการเตรียมการ(Preprocessing) จะทำการแบ่งรูปภาพตัวอักษรที่ได้จากการอ่านด้วยเครื่องสแกนออกเป็น ตัวอักษรภาพเดี่ยว แล้วส่งไปยังขั้นตอนที่สอง เพื่อทำการรู้จำ(Recognition) ให้ได้เป็นตัวอักษรแบบเท็กซ์ จากนั้นก็ส่งตัวอักษรที่ได้จากการรู้จำ ไปยังขั้นตอนที่สามเพื่อทำการแก้ไขตัวอักษรที่ผิดพลาดจากการรู้จำให้ถูกต้อง เรียกว่ากระบวนการปรับปรุง(Postprocessing)

จากการทดสอบโปรแกรม Atrium ThaiOCR รุ่น 1.5 ของ Atrium Technology Co.,Ltd. โดยนำข้อมูลภาพตัวอักษรตัวอย่าง ได้ความถูกต้อง 81% จากแบบตัวอักษร AngsanaUPC ขนาด 12 point ได้ 93% จากตัวอักษร AngsanaUPC ขนาด 14 point และได้ 96% จากตัวอักษร AngsanaUPC ขนาด 16 point พบว่าความถูกต้องต่างกันมาก แท้จริงแล้วความผิดพลาดเพียง 4% จากขนาดตัวอักษร 16 point ที่มีตัวอักษรที่ติดกัน(ตัวอักษรที่มีจุดดำเชื่อมต่อกัน เช่น ปั ตัว "ป" ติดกับ "ั")น้อยมาก แต่เมื่อมีตัวอักษรติดกันมากขึ้นคือ ตัวอักษรขนาด 14 และ 12 point ตามลำดับ ทำให้ความถูกต้องลดลงมามาก นั้นแสดงให้เห็นว่าตัวโปรแกรมรู้จำตัวอักษรไม่มีส่วนตรวจสอบ และตัดแยกตัวอักษรที่ติดกัน และ จากงานวิจัยด้านการรู้จำตัวอักษรเดี่ยวของ สนธยา เมรินทร์ [4] ได้ใช้วิธี Syntactic Pattern Recogntion โดยใช้ Grammar 1 ตัวต่ออักษร 1 ตัว ดังนั้นจะมี Grammar เท่ากับจำนวนอักษร สำหรับ Primitive ที่นำมาสร้างเป็น Grammar นั้นหาได้โดยทำภาพอักษรให้บางแล้วแทนการต่อเชื่อมจุดด้วยรหัสทิศทาง 8 ทิศซึ่งได้ความถูกต้อง 98.5% และงานที่ต่อเนื่องกันของ เดชา รัตนาธาร[3] ใช้วิธีเพิ่มการตัดสินใจด้วย Fuzzy Logic และปรับปรุงส่วนทำให้บาง ของ กิตติพงษ์ เจนวิถีสุข [2] ใช้วิธีนิวรอลเน็ตเวิรก์ ซึ่งทั้งสองได้ความถูกต้องถึง 99.64% และ 99.28% ตามลำดับ ซึ่งพบว่าในกระบวนการเตรียมการ

(Preprocessing) นั้นได้เตรียมข้อมูลตัวอักษรภาพเดี่ยว โดยการตัดภาพตัวอักษรพิมพ์ไทยออกเป็นตัวๆ ที่ถูกต้องไม่มีตัวอักษรตัวใดที่ติดกัน และมีสัญญาณรบกวน

ดังนั้นจึงเห็นว่า ควรจะทำการวิจัยในกระบวนการเตรียมการ เพื่อนำภาพที่ได้จากเครื่องสแกน มาทำการแยกเป็นตัวอักษรเดี่ยวๆ โดยวิเคราะห์ส่วนที่ติดกันของตัวอักษรภาพ เพื่อแยกแยะเป็นอักษรเดี่ยวที่ถูกต้องสมบูรณ์ ก่อนที่จะนำส่งให้กระบวนการรู้จำทำการวิเคราะห์ต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษากระบวนการเตรียมการ (Preprocessing) โดยใช้โปรเจกชั่น นิวเมตริกซ์ อนุพันธ์อันดับสอง และลักษณะบ่งความต่างของตัวอักษรไทย
2. พัฒนาโปรแกรม กระบวนการเตรียมการโดยใช้โปรเจกชั่น นิวเมตริกซ์ อนุพันธ์อันดับสอง และลักษณะบ่งความต่างของตัวอักษรไทยในการแยกตัวอักษรที่ติดกัน
3. เพื่อเปรียบเทียบ ข้อดี ข้อเสีย ของการนำเอากระบวนการเตรียมการมาใช้ รวมถึงคุณสมบัติที่มีผลต่อประสิทธิภาพการรู้จำตัวอักษรพิมพ์ไทย

ขอบเขตของการวิจัย

1. งานวิจัยจะสามารถปรับปรุงกระบวนการเตรียมการ เพื่อให้อัตราการรู้จำมากเพิ่มขึ้น
2. งานวิจัยนี้ได้ทำการพัฒนาโปรแกรม เตรียมการเพื่อให้ได้ภาพอักษรเดี่ยว ก่อนผ่านขบวนการรู้จำตัวอักษรพิมพ์ไทย
3. กำหนดให้ตัวอักษรภาพที่ได้จากเครื่องสแกน (Scanner) ที่ความละเอียดการอ่านภาพ 300 dpi ขึ้นไป
4. ต้นแบบที่ใช้ทำการทดลองจะใช้กับตัวอย่างภาพตัวอักษรพิมพ์ภาษาไทย ที่มีตัวอักษรขนาดระหว่าง 12-36 จุดเท่านั้น

5. พยัญชนะไทยทั้งหมด 44 ตัว ได้แก่ (ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ)
6. สระในภาษาไทยทั้งหมด 21 รูป ได้แก่ (ะ า ิ ี ึ ื ุ ู ั ็ เ แ โ ใ ไ ำ ฤ ฦ ๅ ๆ ฯ)
7. วรรณยุกต์ 4 รูป ได้แก่ (่ ้ ๊ ๋)
8. ตัวเลขไทย ได้แก่ (๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ ๐)
9. ในหนึ่งบรรทัดกำหนดให้มีขนาดตัวอักษรขนาดเดียว และเป็นตัวอักษรแบบเดียวกัน
10. ไม่มีภาพประกอบ หรือตาราง
11. ไม่มีตัวอักษรที่ขีดเส้นใต้ ตัวหนา หรือตัวเอียง
12. คุณภาพของภาพที่ได้จากเครื่องสแกน ต้องไม่เลอะมีรอยเปื้อน และไม่เอียง

ประโยชน์ที่คาดว่าจะได้รับ

1. พัฒนาความรู้ความเข้าใจในเรื่องการตัดบรรทัดอักษรจากภาพ ตัดตัวอักษรจากบรรทัด และ แยกตัวอักษรที่ติดกันเป็นอักษรเดี่ยวๆ
2. ได้โปรแกรมที่ใช้ในการเตรียมการ ก่อนทำการรู้จำตัวอักษรพิมพ์ไทย
3. ได้ทราบถึงข้อจำกัดต่างๆ และผลที่ได้จากการทำกระบวนการเตรียมการก่อนทำการรู้จำ
4. เป็นแนวทางในการพัฒนางานวิจัยทางด้านนี้ต่อไป

วิธีดำเนินการวิจัย

1. ศึกษางานวิจัยในอดีตเกี่ยวกับระบบรู้จำเอกสารภาษาต่างๆ ในส่วนแยกตัวอักษรที่ติดกันเป็นอักษรเดี่ยวๆ
2. ค้นคว้าขั้นตอนวิธีการตัดบรรทัดอักษรจากภาพ ตัดตัวอักษรจากบรรทัด และ แยกตัวอักษรที่ติดกันเป็นอักษรเดี่ยวๆ
3. กำหนดข้อมูลตัวอย่างในการทดลอง

4. พัฒนาโปรแกรมสำหรับกระบวนการเตรียมการ ซึ่งประกอบได้ด้วย การตัดบรรทัด อักษรจากภาพ ตัดตัวอักษรจากบรรทัด และ แยกตัวอักษรที่ติดกันเป็นอักษรเดี่ยวๆ
5. ทดสอบการทำงานของโปรแกรม พร้อมแก้ข้อผิดพลาดของโปรแกรม
6. ทดสอบโปรแกรม กับตัวอย่างข้อมูล เพื่อหาข้อผิดพลาดต่างๆ
7. สรุปผล วิเคราะห์ และนำเสนอ แนวทางการวิจัยเพื่อเป็นประโยชน์ต่อไป