บทที่ 1

บทนำ

ความเบื้องต้น

การใช้คอมพิวเตอร์สำหรับงานเอกสาร และ การประมวลผลข้อมูล เป็นหนึ่งในแนวโน้มหลักของสำนักงานอัตโนมัติ (OA) ในปัจจุบัน ซึ่งการรู้จำตัวอักษรด้วยคอมพิวเตอร์ก็เริ่มเข้ามามีบทบาทในด้านนี้เพิ่มขึ้นอย่างมาก เพื่อจะให้เป็นหนทางเลือกในการป้อนข้อมูลเข้าสู่คอมพิวเตอร์โดยการอ่านข้อความจากเอกสาร หรือ รับรู้จากการเขียนที่จอภาพได้โดยตรง โดยมนุษย์ไม่จำเป็นต้องป้อนข้อมูลผ่านทางแป้นพิมพ์เพียงอย่างเดียว

การรู้จำรูปแบบตัวอักษรได้มีการประยุกต์ใช้กับหลายภาษาเช่น ตัวอักษรภาษาอังกฤษซึ่งได้มีการพัฒนาก้าวหน้าอย่างมากจนกระทั่งสามารถนำมาประยุกต์ใช้ และ จำหน่ายในเชิงพาณิชย์ได้ทั้งแบบตัวพิมพ์และ ตัวเขียน ส่วนภาษาอื่นเช่น ตัวอักษรภาษาอาหรับ , ตัวอักษรภาษาจีน , ตัวอักษรคันจิในภาษาญี่ปุ่น เป็นต้น ก็ได้มีงานวิจัยศึกษาค้นคว้าเพื่อพัฒนาระบบการรู้จำของภาษานั้น ๆ ออกเผยแพร่ และ ตีพิมพ์ หลายงานวิจัย แต่สำหรับตัวอักษรภาษาไทยได้มีการศึกษาค้นคว้ากันมาบางพอสมควร แต่เนื่องจากตัวอักษรภาษาไทยมีโครงสร้างที่ค่อนข้างซับซ้อน ประกอบด้วยลักษณะที่เป็นเส้นตรงผสมเส้นโค้ง และ วงกลม ซึ่งบางครั้งมีการตัดกันของเส้น รวมทั้งตัวอักษรก็มีหลายระดับ ทำให้การศึกษาค้นคว้าไม่แพร่หลาย และ ต่อเนื่องมากนัก ดังเช่นงานวิจัยการรู้จำตัวอักษรไทย และ ตัวเลขไทยเท่าที่พบคือ

สรุพันธ์ เอื้อไพบูลย์ (2531)[1] นำเสนอวิธีการรู้จำตัวอักษรลายมือเขียนภาษาไทย โดยการนำหัวของตัวอักษรนำมาพิจารณาเพื่อทำการจำแนกกลุ่มของตัวอักษรออกเป็นกลุ่มย่อย โดยให้ตัวอักษรที่มีหัวอยู่บริเวณเดียวกันอยู่กลุ่มเดียวกัน และใช้คุณสมบัติทางการจำแนกชนิด (topology) ของตัวอักษร ในการแยกตัวอักษรออกจากกลุ่มย่อย รวมทั้งยังนำเอาเทคนิคหลาย ๆวิธีเพื่อใช้แยกตัวอักษรที่คล้ายคลึงกันออกจากกันด้วย ซึ่งข้อดีคือการเปรียบเทียบแบบนี้ทำให้เปรียบเทียบได้อย่างรวดเร็ว

พิพัฒน์ และ มนลดา[2] นำเสนอการรู้จำตัวอักขระไทยหลายรูปแบบโดยใช้วิธีการไดนามิกโปรแกรมมิ่ง วิธีการนี้อาศัยการพิจารณาเส้นแสดงขอบของอักขระโดยนำรหัสทิศทางแบบลูกโซ่

ของฟรีแมน กับ ความแตกต่างของทิศทางของเส้นแสดงขอบของอักขระ มาใช้ในการตัดแบ่งเส้นแสดงขอบของอักขระออกเป็นส่วนโค้งย่อย คือ ส่วนโค้งเว้า และ ส่วนโค้งนูน โดยจะนำมาเปรียบเทียบหาส่วนโค้งที่คล้ายกันมากที่สุดกับอักขระต้นแบบ โดยใช้เทคนิคของการเปรียบเทียบแบบไดนามิคโปรแกรมมิ่ง ผลอัตราการรู้จำถูกต้อง 94.7 เปอร์เซ็นต์

อนันต์ เอกวงศ์วิริยะ (2537)[3] นำเสนอเรื่องการรู้จำตัวเลขไทยแบบตัวพิมพ์โดยวิธีซินแทกติก โดยวิธีการเริ่มต้นที่ทำการปรับแต่งรูปให้มีคุณภาพดีขึ้น และ เหมาะสมเพื่อที่จะนำไปใช้ในส่วนของการแทนรูปด้วยโครงสร้างของภาษาซึ่งกำหนดโดยไวยากรณ์ต้นไม้ (tree grammar) และจำแนกตัวเลขภาษาไทยโดยใช้เทคนิคในการหาระยะห่างระหว่างต้นไม้ที่เป็นอินพุต กับต้นไม้ต้นแบบ โดยวิธีการซินแทกติกนี้มีจุดเด่นอยู่ที่นำเอาโครงร่างของภาพตัวอักษรมาใช้ประโยชน์ในการรู้จำ ฉะนั้นจึงสามารถรู้จำภาพของตัวอักษรแบบต่าง font และต่างขนาดได้ดี

สนธยา เมรินทร์ (2537)[4] นำเสนอเรื่องการศึกษาการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยวิธีซินแทกติกซึ่งเป็นการพิจารณาที่โครงสร้างของตัวอักษร โดยมีการอธิบายโครงสร้างของตัวอักษรในรูปของประโยคที่ประกอบด้วย primitive ทำให้สามารถจำแนกตัวอักษรที่มีโครงสร้างที่แตกต่างกันออกจากกันได้อย่างง่ายและรวดเร็ว รวมทั้งยังใช้วิธีการเปรียบเทียบทาง feature ควบคู่กันไปด้วยสำหรับตัวอักษรภาษาไทยบางกลุ่มที่มีลักษณะคล้ายคลึงกันมาก

วิทยานิพนธ์ฉบับนี้ได้จัดทำขึ้นเพื่อเสนอแนวทางในการแก้ไขปรับปรุง การรู้จำตัวอักษรไทยด้วยวิธีการซินแทกติกของ สนธยา เมรินทร์[4] โดยได้นำเอาเทคนิคแบบฟัซซีโลจิกเข้ามาช่วยในการปรับปรุงให้มีความแม่นยำในการรู้จำสูงขึ้นโดยยังคงการวิเคราะห์โครงสร้างของตัวอักษรในวิธีทางซินแทกติกไว้ซึ่งจะเป็นการแก้ปัญหาที่มีอยู่คือการไม่สามารถรู้จำตัวอักษร หรือ รู้จำผิดพลาดในกลุ่มพยัญชนะ 20 ตัว ได้แก่ ข 2 ตัว , ฆ 1 ตัว , ฉ 1 ตัว , ช 3 ตัว , ฎ 1 ตัว , ฐ 3 ตัว , ถ 1 ตัว , น 1 ตัว , ศ 5 ตัว , ส 1 ตัว , ห 1 ตัว และกลุ่มสระ 11 ตัวได้แก่ อื 5 ตัว , อี 1 ตัว , อื 2 ตัว , ไม้เอก 1 ตัว , ไม้โท 2 ตัว ใน font แบบ Eucrosia และ แบบ Cordia ขนาด 20,22,24,28,32,36,48

วัตถุประสงค์ของงานวิจัย

ก. เพื่อพัฒนาโปรแกรมสำหรับการรู้จำตัวอักษรพิมพ์ภาษาไทย

ข. เพื่อศึกษาวิธีการปรับปรุงการรู้จำตัวอักษรพิมพ์ภาษาไทย โดยเทคนิคแบบ ฟัซซีโลจิก

ค. เพื่อเป็นแนวทางสำหรับการพัฒนาวิธีการรู้จำในเรื่อง ๆอื่นต่อไป

ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

ก. พัฒนาความรู้ในเรื่องการรู้จำตัวอักษรพิมพ์ภาษาไทย

ข. ผลิตโปรแกรมที่ทำให้คอมพิวเตอร์รู้จำตัวอักษรพิมพ์ภาษาไทยได้

ค. เป็นแนวทางในการพัฒนาเครื่องมืออ่านตัวอักษรภาษาไทยสำหรับผู้พิการทางสายตา และเพิ่มประสิทธิภาพในการทำงานด้านสำนักงาน

ขอบเขตของงานวิจัย

1. ปรับปรุงให้อัตราการรู้จำมากกว่า 98% ขึ้นไป[4]

2. เครื่องคอมพิวเตอร์ที่ใช้ : เครื่อง PC รุ่น 486DX2-66 ใช้ไมโครโปรเซสเซอร์ 80486 ความเร็ว 66 เมกกะเฮิรตซ์ มีหน่วยความจำ 8 เมกกะไบท์ และจานแม่เหล็กแบบแข็งขนาด 330 เมกกะไบท์

3. ภาษาคอมพิวเตอร์ที่ใช้ในการปรับปรุง : ภาษา C

4. อุปกรณ์อ่านเอกสาร : เครื่องสแกนเนอร์ Microtek รุ่น ScanMaker II ใช้ความละเอียดของการอ่านเอกสารที่ 300 dpi

5. รูปแบบตัวอักษรที่ใช้ : ตัวอักษรพิมพ์ภาษาไทยในงานวิทยานิพนธ์ของ สนธยา เมรินทร์ ซึ่งนำรูปแบบตัวอักษรชื่อ EUCROSIA และ CORDIA ที่เป็นตัวธรรมดา ไม่เป็นตัวเอียง หรือมีเส้นใต้ มาใช้เก็บลงไฟล์ 1 ตัวอักษรต่อ 1 ไฟล์ในรูปแบบ BMP หรือ PCX โดยมีตัวอักษรที่สามารถรู้จำได้ 79 ตัวต่อ 1 รูปแบบ ดังนี้

5.1 พยัญชนะ พยัญชนะไทยมีทั้งหมด 44 ตัว คือ

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 ล ว ศ ษ ส ห ฬ อ ฮ

5.2 สระ ในภาษาไทยมีรูปสระอยู่ทั้งหมด 21 รูป คือ

ะ า ิ ี ึ ื ุ ู ็ ั ์ เ โ ใ ไ อำ ฤ ๅ ฦ ๆ ฯ

5.3 วรรณยุกต์ มี 4 รูป คือ ่ ้ ๊ ๋

5.4 ตัวเลขไทย มี 10 รูป คือ ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙

6. ขนาดตัวอักษรที่ใช้ (ดังแสดงในภาคผนวก ง และ จ)

5.1 ขนาดของตัวอักษรที่อ่านโดยเครื่องสแกนเนอร์ 20, 22, 24, 28, 32 และ 36 point นำมาจากโปรแกรม Microsoft Word 2.0c แล้วนำมาสแกนด้วยความละเอียด 300 dpi จุดต่อเนื่อง

5.2 ขนาดของตัวอักษรที่เก็บจากโปรแกรมPaintbrush โดยตรง 48 point