บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การใช้คอมพิวเตอร์สำหรับงานด้านปัญญาประดิษฐ์เป็นการทำให้คอมพิวเตอร์ทำงานง่าย ๆ เป็นธรรมชาติซึ่ง ณ ปัจจุบันมนุษย์สามารถทำได้ดีกว่า การรู้จำตัวอักษรด้วยคอมพิวเตอร์เป็นหนึ่งในงานง่าย ๆ เป็นธรรมชาติซึ่งมนุษย์สามารถทำได้ดีซึ่งมีความสำคัญต่อการทำให้คอมพิวเตอร์ใช้งานง่าย เนื่องจากปัจจุบันมนุษย์จะต้องอาศัยแป้นพิมพ์เป็นอุปกรณ์หลักในการป้อนข้อมูลให้กับคอมพิวเตอร์ นั่นคือมนุษย์จะต้องทำการรู้จำตัวอักษรจากเอกสาร หลังจากนั้นจึงทำการป้อนข้อมูลที่ผ่านการรู้จำแล้วให้กับคอมพิวเตอร์ การรู้จำตัวอักษรด้วยคอมพิวเตอร์จะเป็นหนทางที่ทำให้คอมพิวเตอร์สามารถอ่านข้อมูลจากเอกสารได้โดยตรง ทำให้งานป้อนข้อมูลให้คอมพิวเตอร์จากเอกสารมีความสะดวกยิ่งขึ้น โดยเฉพาะอย่างยิ่ง เมื่อมีเอกสารที่จะต้องป้อนข้อมูลให้แก่คอมพิวเตอร์จำนวนมาก ๆ

การรู้จำตัวอักษรได้มีการวิจัยในหลายภาษาต่าง ๆ กัน ที่ได้ผลที่ก้าวหน้ามากจนถึงขั้นนำมาประยุกต์ใช้ในเชิงพาณิชย์ได้ส่วนใหญ่จะเป็นการรู้จำตัวอักษรภาษาอังกฤษ สำหรับภาษาไทยนั้นมีการวิจัยอยู่บ้างพอสมควร แต่เนื่องจากความซับซ้อนและความหลากหลายของตัวอักษรภาษาไทยทำให้การศึกษาค้นคว้าไม่แพร่หลายและขาดความต่อเนื่อง

การเรียนรู้ของเครื่อง (Machine Learning) เป็นแนวคิดที่จะให้เครื่องคอมพิวเตอร์ทำการเรียนรู้ให้เกิดความรู้ใหม่ ๆ ซึ่งนิวรอลเน็ตเวิร์ก (Neural Networks) ก็เป็นการเรียนรู้รูปแบบหนึ่งซึ่งจะเก็บความรู้ที่ได้ในรูปของเน็ตเวิร์กของค่าน้ำหนักและค่าไบแอส ในงานวิจัยนี้จะเป็นงานวิจัยในด้านการรู้จำที่ประยุกต์การเรียนรู้ในรูปของนิวรอลเน็ตเวิร์กเข้ากับการรู้จำตัวอักษรโดยอาศัยการวิเคราะห์ตัวประกอบสำคัญ (การแปลงข้อมูลภาพที่ได้จากเครื่องสแกนเนอร์ไปยังอีกโดเมน ซึ่งเป็นโดเมนที่จะแสดงค่าตัวประกอบสำคัญ (Principal Component) ของภาพนั้น ๆ เมื่อเทียบกับภาพต้นแบบ) มาร่วมด้วย โดยนิวรอลเน็ตเวิร์กจะใช้สำหรับเรียนรู้ค่าตัวประกอบสำคัญของข้อมูลภาพที่ได้จากการวิเคราะห์ตัวประกอบสำคัญในขั้นตอนของการเรียนรู้ และจะทำการแยกแยะตัวอักษรตามความรู้ที่ได้เรียนรู้มา

งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวเนื่องกับการรู้จำตัวอักษร โดยเฉพาะอย่างยิ่งงานวิจัยการรู้จำตัวอักษรภาษาไทย เท่าที่พบคือ

สุรพันธ์[1] นำเสนอวิธีการรู้จำตัวอักษรลายมือเขียนภาษาไทยโดยการนำหัวของตัวอักษรมาพิจารณาเพื่อทำการจำแนกกลุ่มของตัวอักษรออกเป็นกลุ่มย่อย ๆ โดยให้ตัวอักษรที่มีหัวอยู่บริเวณเดียวกันอยู่กลุ่มเดียวกัน จากนั้นก็จะพิจารณาเปรียบเทียบตัวอักษรที่อยู่ในกลุ่มของตนเองแตกต่างไป ทั้งนี้ขึ้นกับลักษณะเด่นของตัวอักษรที่อยู่ในกลุ่มนั้น ๆ ซึ่งทำให้กลุ่มของตัวอักษรที่ต้องเปรียบเทียบมีจำนวนลดลง ทำให้เปรียบเทียบได้อย่างรวดเร็ว

พิพัฒน์ และ มนลดา [2] นำเสนอวิธีการรู้จำตัวอักขระไทยหลายรูปแบบโดยใช้วิธีการไดนามิคโปรแกรมมิ่ง โดยการพิจารณาเส้นแสดงขอบของอักขระโดยนำรหัสทิศทางแบบลูกโซ่ของฟรีแมน กับความแตกต่างของทิศทางของเส้นแสดงขอบอักขระมาใช้ในการตัดแบ่งเส้นแสดงขอบของอักขระออกเป็นส่วนโค้งเว้าและส่วนโค้งนูน เพื่อนำไปใช้ในการเปรียบเทียบแบบไดนามิคโปรแกรมมิ่ง โดยการคำนวนค่าความคล้ายคลึง (similarity) ของส่วนโค้งเว้าและส่วนโค้งนูนที่ได้กับส่วนโค้งเว้าและส่วนโค้งนูนของตัวอักขระต้นแบบ

สนธยา [3] นำเสนอเรื่องการศึกษาการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยวิธีซินแทกติก[4] ซึ่งเป็นการพิจารณาโครงสร้างของตัวอักษรโดยการเปลี่ยนแปลงเส้นแสดงขอบของตัวอักษรให้อยู่ในรูปรหัสทิศทางแบบลูกโซ่ของฟรีแมน และเปลี่ยนรหัสลูกโซ่เป็นรหัสเวกเตอร์เส้นตรงและวงกลม เพื่อนำมาจัดเก็บเป็นรูปของประโยคที่ประกอบด้วยพิมิทีฟ (primitive) ในลักษณะของโครงสร้างแบบต้นไม้ และอาศัยวิธีการทางการวิเคราะห์ประโยคที่ได้จากการเปลี่ยนเส้นแสดงขอบของตัวอักษรเปรียบเทียบกับประโยคของอักขระต้นแบบ โดยเลือกเปรียบเทียบเฉพาะตัวอักษรต้นแบบที่มีหัวของตัวอักษรอยู่ในบริเวณเดียวกันกับหัวของตัวอักษรที่ต้องการรู้จำ และเปรียบเทียบเฉพาะตัวอักษรที่เป็นตัวอักษรอยู่ในระดับเดียวกันเท่านั้น (โดยการระบุเส้นบอกระดับ) สำหรับตัวอักษรที่แตกต่างกันไม่มากจะถูกนำไปเปรียบเทียบทางฟีเจอร์ (feature) อีกครั้งหนึ่ง โดยการเก็บลักษณะพิเศษของแต่ละตัวอักษรไว้ หากผลการรู้จำในข้างต้นไม่อยู่ในเกณฑ์ที่ยอมรับได้เวกเตอร์ของตัวอักษรจะถูกนำมาปรับปรุงเพื่อตัดส่วนเกินออก หรือเชื่อมเวกเตอร์ที่อยู่ใกล้เคียงกันเข้าด้วยกันแล้วจึงนำมาทำการรู้จำโดยวิธีเดิมอีกจนกว่าผลการรู้จำจะอยู่ในเกณฑ์ที่ยอมรับได้ หรือไม่สามารถทำการปรับปรุงเวกเตอร์ได้อีก

เดชา [5] นำเสนอเรื่องการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยใช้เทคนิคแบบฟัซซีโลจิก และวิธีซินแทกติก โดยทำการปรับปรุงวิธีการซินแทกติกของ สนธยา[3] โดยการนำเทคนิคแบบฟัซซีโลจิกเข้ามาใช้เมื่อการใช้วิธีการทางซินแทกติกไม่สามารถรู้จำตัวอักษร รวมทั้งปรับปรุงวิธีการทำตัวอักษรให้บาง[6] โดยการใช้การทำตัวอักษรให้บางแบบเอสพีทีเอ (SPTA, Save Point Thinning Algorithm)

อภิญญา [7] นำเสนอเรื่องการประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยในการรู้จำตัวพิมพ์อักษรภาษาไทย โดยนำการเรียนรู้โดยการอุปนัยโดยใช้การโปรแกรมตรรกะเชิงอุปนัย (Inductive Logic Programming, ILP) หรือ ไอ แอล พี โดยใช้ความรู้ส่วนหลัง (background knowledge) ในการสร้างสมมติฐานใหม่ที่สอดคล้องกับตัวอย่างที่ได้รับ ซึ่งเทคนิคขั้นต้นจะเป็นการพิจารณาโครงสร้างของตัวอักษรโดยทำการเปลี่ยนขอบของตัวอักษรเป็นรหัสทิศทางแบบลูกโซ่ของฟรีแมน ทำการเปลี่ยนรหัสทิศทางแบบลูกโซ่ของฟรีแมนเป็นเวกเตอร์เส้นตรง และเวกเตอร์วงกลม แล้วทำการเปลี่ยนเวกเตอร์เป็นหน่วยสร้างพื้นฐาน นำการโปรแกรมตรรกะเชิงอุปนัยมาทำการเรียนรู้ลักษณะของหน่วยสร้างพื้นฐานที่ได้จากตัวอักษรต้นแบบ เช่น ระดับของตัวอักษร ขนาดของตัวอักษร ลักษณะส่วนหัวของตัวอักษร ลักษณะส่วนปลายของตัวอักษร เป็นต้น

วิทยานิพนธ์ฉบับนี้ได้จัดทำขึ้นเพื่อเสนอแนวทางการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยการใช้เทคนิคด้านการวิเคราะห์ตัวประกอบสำคัญเพื่อใช้วิเคราะห์คุณลักษณะของตัวอักษร ร่วมกับเทคนิคการเรียนรู้ด้านนิวรอลเน็ตเวิร์กเพื่อใช้แยกแยะคุณลักษณะอีกต่อหนึ่ง เพื่อให้สามารถทำการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยมีรูปแบบตัวอักษรต่าง ๆ กันได้ โดยจะนำผลของการรู้จำตัวอักษรพิมพ์ภาษาไทยที่ได้จากวิทยานิพนธ์ฉบับนี้ไปเปรียบเทียบกับผลของการรู้จำตัวอักษรพิมพ์ภาษาไทยที่ได้จากงานวิจัยอื่น ๆ

วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาโปรแกรมสำหรับรู้จำตัวอักษรพิมพ์ภาษาไทยโดยใช้เทคนิคด้านการวิเคราะห์ตัวประกอบสำคัญควบคู่กับเทคนิคด้านนิวรอลเน็ตเวิร์ก

ขอบเขตของการวิจัย

1. เครื่องคอมพิวเตอร์ที่ใช้: เครื่อง IBM PC ใช้ไมโครโปรเซสเซอร์รุ่น 80486 ความเร็ว 100 MHz ขึ้นไป มีหน่วยความจำ 16 เมกกะไบท์ขึ้นไป และจานแม่เหล็กแบบแข็งขนาด 500 เมกกะไบท์ขึ้นไป
2. ภาษาคอมพิวเตอร์ที่ใช้ : ภาษา C
3. อุปกรณ์อ่านเอกสาร: เครื่องสแกนเนอร์ที่ใช้ความละเอียดได้ตั้งแต่ 200 DPI ขึ้นไป
4. รูปแบบตัวอักษรที่ใช้: ตัวอักษรพิมพ์ภาษาไทยแบบ AngsanaUPC, BrowaliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC, FreesiaUPC โดยการอ่านตัวอักษรแบบ bi-level (ขาวดำ) เก็บลงแฟ้มข้อมูล 1 แฟ้มต่อหนึ่งตัวอักษร
5. ตัวอักษรที่ทำการรู้จำ:

   พยัญชนะ 44 ตัวอักษรคือ

   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

   สระและสัญลักษณ์ 20 ตัวอักษรคือ

   ะ า ิ ี ึ ื ุ ู ็ ั ์ เ แ ไ ใ ๆ ฤ ฦ ฯ

   วรรณยุกต์ 4 ตัวอักษรได้แก่ ่ ้ ๊ ๋

6. ขนาดตัวอักษรที่ใช้: ขนาดตัวอักษรที่อ่านโดยเครื่องสแกนเนอร์ขนาด 16, 18, 20, 22, 24, 26, 28, 36 จุด
7. ตัวอักษรที่ใช้ทดสอบ: ตัวอักษรที่ได้จากการอ่านผ่านเครื่องสแกนเนอร์ในแต่ละรูปแบบและในแต่ละขนาดตัวอักษรจำนวน 68 ตัวอักษรและตัวอักษรที่ได้จากการอ่านผ่านเครื่องสแกนแนอร์และทำการขจัดสัญญาณรบกวนประเภทสัญญาณรบกวนแบบจุดเดี่ยวในแต่ละรูปแบบและในแต่ละขนาดตัวอักษรจำนวน 68 ตัวอักษร (รวมทั้งหมด 6528 ตัวอักษร)

ขั้นตอนและวิธีการดำเนินการวิจัย

1. ศึกษาความเป็นไปได้จากผลงานวิจัยที่ผ่านมา
2. ศึกษาทฤษฎี และเลือกวิธีการที่เหมาะสม
3. พัฒนาโปรแกรม
4. เก็บตัวอย่างตัวอักษรพิมพ์ภาษาไทย
5. ทดสอบโปรแกรมและปรับปรุงโปรแกรม
6. ประเมินผลโดยการเปรียบเทียบกับผลการรู้จำตัวอักษรโดยวิธีอื่น ๆ

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้โปรแกรมคอมพิวเตอร์ที่สามารถทำการรู้จำตัวอักษรพิมพ์ภาษาไทย
2. เป็นแนวทางในการพัฒนาโปรแกรมประยุกต์ที่อาศัยการรู้จำอักษรพิมพ์ภาษาไทย เพื่อทำการรับข้อมูลจากเอกสารโดยตรง
3. เป็นแนวทางในการพัฒนาการรู้จำเรื่องอื่น ๆ