บทที่ 1

บทนำ

1.1. ความเป็นมาของปัญหา

เนื่องจากปัจจุบันเทคโนโลยีทางคอมพิวเตอร์ในการที่จะนำไปใช้เกี่ยวกับทางด้านงานประมวลผลข้อมูล (Data Processing) เป็นไปได้อย่างรวดเร็ว ทำให้มีผู้สนใจในการที่จะนำเครื่องคอมพิวเตอร์ไปใช้ในการที่จะทำการจดจำตัวอักษรและแปลงข้อมูลภาพที่เป็นตัวอักษรให้อยู่ในรูปของ TEXT FILE ตามที่เราต้องการได้ ซึ่งในปัจจุบันการพัฒนาส่วนใหญ่จะเป็นการพัฒนามาจากต่างประเทศทำให้โปรแกรมที่ใช้งานอยู่ สามารถที่จะใช้ได้กับภาษาใดภาษาหนึ่งเท่านั้นหรือเป็นภาษาที่มีการเขียนในลักษณะที่คล้ายคลึงกัน เช่น ภาษาอังกฤษ กับ ภาษาฝรั่งเศส และภาษาอื่น ๆ ในแถบยุโรป ทำให้ไม่สามารถที่จะนำมาใช้กับภาษาไทยซึ่งเป็นภาษาที่มีใช้อยู่เฉพาะในประเทศไทยเท่านั้น การที่จะให้เครื่องคอมพิวเตอร์สามารถเรียนรู้ถึงวิธีการในการจดจำตัวอักษรและแปลงตัวอักษรในปัจจุบันมีเทคโนโลยีที่เรียกว่า OCR (Optical Character Recognition) ซึ่งจะเป็นการอ่านข้อความหรือกลุ่มคำที่จะทำการวิเคราะห์เข้ามาโดยใช้อุปกรณ์ที่ทำหน้าที่ในการอ่านซึ่งได้แก่เครื่อง Scanner เครื่อง Fax กล้อง Video และดิจิไตเซอร์(Digitizer) ซึ่งจะมีประโยชน์ในการนำไปใช้ในงานทางด้านเกี่ยวกับเอกสารหรือการสั่งงานเครื่องคอมพิวเตอร์โดยให้คอมพิวเตอร์เป็นผู้ตรวจสอบคำสั่งนั้น โดยการอ่านข้อมูลภาพที่เป็นตัวอักษรเข้าไป ซึ่งผลลัพธ์ที่ได้จากการรู้จำตัวอักษรนี้จะอยู่ในรูปของ TEXT FILE ทั่วไปและสามารถนำเอาไปใช้งานได้

การรู้จำตัวอักษรไทยนั้น มีผู้ทำการวิจัยไว้แล้วอยู่บ้าง เช่น การรู้จำตัวเลขไทยแบบตัวพิมพ์โดยวิธีซินแทกติด(อนันต์, 2537) การรู้จำตัวอักษรพิมพ์ภาษาไทยโดยวิธีซินแทกติก(สนธยา, 2537) การนำเอาวิธีการรู้จำตัวอักษรไทยไปใช้กับงานทางด้าน OCR จำเป็นต้องมีส่วนการแยกสายอักษรออกเป็นอักษรแต่ละตัวก่อน และทำการจัดเรียงให้ถูกต้องตามมาตรฐานการใช้งานของภาษาไทย จากนั้นจึงนำการรู้จำตัวอักษรมาใช้กับตัวอักษรที่ได้จัดเรียงไว้แล้ว ซึ่งผลลัพธ์ที่ได้จากการรู้จำ เป็น TEXT FILE ที่สามารถนำไปใช้งานกับโปรแกรมทางด้าน word processor ภาษาไทยได้ การศึกษานี้ครั้งนี้จึงมุ่งเน้นที่การแยกสายอักษรออกเป็นตัวอักษรแต่ละตัวเพื่อนำไปใช้ในการรู้จำตัวอักษร

1.2. แนวทางแก้ไขปัญหา

ตัวอักษรภาษาไทย จะประกอบไปด้วย ตัวอักษรต่าง ๆ ซึ่งสามารถแยกออกได้เป็นดังนี้

1.2.1. ตัวพยัญชนะ มีทั้งหมด 44 ตัว แต่ที่ใช้งานมี 42 ตัว และไม่ได้ใช้งาน 2 ตัว คือ ฃ ฅ

1.2.2. ตัวสระ แบ่งเป็น สระระดับบน สระระดับล่าง และสระระดับพยัญชนะ ดังนี้

1.2.2.1. สระระดับบน ได้แก่ ั ิ ี ึ ื ็

1.2.2.2. สระระดับล่าง ได้แก่ ุ ู

1.2.2.3. สระระดับพยัญชนะ ได้แก่ ะ า เ แ ไ โ ใ

1.2.3. ตัววรรณยุกต์ ได้แก่ ่ ้ ๊ และ ๋

1.2.4. ตัวเลขไทย ได้แก่ เลข ๐ - ๙ และตัวเลขอารบิค 0 - 9

1.2.5. ตัวอักขระพิเศษ ได้แก่ ฤ ฦ ๆ ฯ ์ ๎

ลักษณะของตัวอักษรไทยส่วนใหญ่ประกอบด้วย หัวตัวอักษรซึ่งมีลักษณะกลม ส่วนตัวอักษร และ ส่วนปลายตัวอักษร

การรู้จำสายตัวอักษรไทยนั้น จะต้องมีโปรแกรมที่ใช้ในการแยกกลุ่มข้อมูลของตัวอักษรทั้งหมดออกเป็นกลุ่มข้อมูลของตัวอักษรแต่ละตัวเสียก่อน จากนั้นใช้โปรแกรมในด้านการรู้จำตัวอักษร เพื่อให้สามารถรู้ได้ว่าเป็นตัวอักษรอะไร เนื่องจากภาษาไทยเป็นภาษาที่มีความแตกต่างจากภาษาอื่น คือเป็นภาษาที่มีการเขียนในลักษณะที่เป็นระดับโดยสามารถแบ่งออกได้เป็น 4 ระดับ คือ

1. ระดับวรรณยุกต์ ได้แก่ ่ , ้ , ฯลฯ

2. ระดับสระบน ได้แก่ ิ , ี , ฯลฯ

3. ระดับพยัญชนะ แบ่งออกเป็นได้อีก 5 กลุ่มตามลักษณะตัวอักษร ดังนี้

3.1. พยัญชนะทั่วไป ได้แก่ ก , จ , ฯลฯ

3.2. พยัญชนะที่เกินในระดับสระบน ได้แก่ ฟ , ป , ฯลฯ

3.3 พยัญชนะที่เกินในระดับสระล่าง ได้แก่ ฎ , ฏ , ฯลฯ

3.4 สระที่อยู่ในระดับพยัญชนะ ได้แก่ เ , า , ฯลฯ

3.5 พยัญชนะที่แยกเป็น 2 ส่วน ได้แก่ ฐ , ญ

4. ระดับสระล่าง ได้แก่ ุ , ู , ฯลฯ

เพื่อให้การรู้จำสายอักษรเป็นไปอย่างมีประสิทธิภาพเราจึงจะทำการแบ่งพยัญชนะไทย ออกเป็นพยัญชนะในลักษณะตัวเดียวโดด ๆ โดยวิธีการหาขอบของข้อมูลของตัวอักษรแต่ละตัว และทำการจัดเรียงให้อยู่ในลักษณะของพยัญชนะไทย 4 ระดับโดยวิธีการพิจารณาขนาดและตำแหน่งของตัวอักษร จากนั้นจึง

ใช้การรู้จำตัวอักษรภาษาไทยมาใช้กับตัวอักษรที่ได้แยกไว้เป็นตัวอักษรเดี่ยว ๆ โดยวิธีการที่ใช้ในการรู้จำตัวอักษรจะใช้วิธีการซินแทคติก ซึ่งวิธีการซินแทคติกจะใช้โครงร่างของตัวอักษร มาใช้ในการพิจารณา โดยจะอธิบายโครงร่างของตัวอักษรในรูปของประโยคที่ประกอบด้วยโครงสร้าง Primitive ในขั้นตอนต่อไป จึงทำการแก้ไขตัวอักษรบางตัวให้ตรงกับค่าที่ใช้อยู่ในตารางแอสกี เพื่อให้สามารถนำไปใช้งานต่อไปได้

1.3. วัตถุประสงค์

1.3.1. เพื่อหาอัลกอริทึมต้นแบบที่สามารถนำไปใช้ในงานการรู้สายตัวอักษรไทยแบบตัวพิมพ์

1.3.2. ศึกษาและออกแบบการจัดเก็บข้อมูลที่ได้รับเข้ามาในขั้นต้น เพื่อให้สามารถที่จะประมวลผลข้อมูลดังกล่าวข้างต้นได้อย่างสะดวกและรวดเร็ว

1.3.3. ศึกษาการทำงานเทคนิคในการใช้งานของระบบกราฟฟิก

1.3.4. ศึกษาและออกแบบการจัดเก็บข้อมูลของตัวอักษรที่ได้รับ หลังจากที่ได้ผ่านขั้นตอนการรู้จำตัวอักษรซึ่งผลลัพธ์ที่ได้จะเป็นในลักษณะ Text File

1.4. ขอบเขตในการวิจัย

ในงานวิจัยนี้จะเป็นการนำเอาเครื่องคอมพิวเตอร์มาใช้ในการประมวลผลและมีภาษาที่ใช้ในการเขียนโปรแกรมเป็นภาษา C ขอบเขตของงานวิจัยคือความสามารถที่จะนำไปใช้ในการรู้จำตัวอักษรที่มีการเรียงกันเป็นกลุ่มคำ ประโยค หรือข้อความที่เราต้องการ โดยตัวอักษรเหล่านั้นจะประกอบไปด้วย พยัญชนะไทย ตัวสระ วรรณยุกต์ ตัวเลขไทย และตัวเลขอารบิค โดยรูปแบบของตัวอักษรที่ใช้จะเป็นตัวอักษรแบบตัวตรงธรรมดา Font แบบ EucrosiaUPC ขนาด 18 point และสามารถรู้จำตัวอักษรพิมพ์ภาษาไทยที่มีรูปแบบของตัวอักษร ตามที่ได้กำหนดไว้ โดยมีอัตราการรู้จำมากกว่า 80% ขึ้นไป

1.5. ขั้นตอนและวิธีการดำเนินการ

1.5.1. ศึกษาและรวบรวมข้อมูล

1.5.2. ออกแบบโปรแกรมและเลือกอุปกรณ์ที่จำเป็นต้องใช้

1.5.3. พัฒนาโปรแกรมในส่วนการรับข้อมูลและการแยกตัวอักษร ซึ่งโปรแกรมในส่วนนี้จะทำหน้าที่ในการรับข้อมูลที่ได้มาจากการ scan ภาพ มาทำการปรับปรุงภาพและแยกข้อมูลภาพเหล่านั้นออกมาเป็นข้อมูลของตัวอักษรแต่ละตัว และใส่ดัชนีแสดงระดับและตำแหน่งของตัวอักษร เพื่อนำไปใช้ในการรู้จำตัวอักษร

1.5.4. ปรับปรุงโปรแกรมในส่วนการรู้จำตัวอักษร[2] ให้สามารถรับข้อมูลจากขั้นตอนที่ 1.5.3. ได้ ทั้งนี้โปรแกรมในส่วนนี้จะนำเอาข้อมูลที่ได้จากการแยกข้อมูลภาพออกเป็นกลุ่มข้อมูลของตัวอักษรแต่ละตัวมาทำการรู้จำว่าเป็นตัวอักษรอะไร

1.5.5. ทดสอบโปรแกรมทั้งหมดที่ได้พัฒนาขึ้น

1.5.6. สรุปผล

1.6. ประโยชน์ที่คาดว่าจะได้รับ

1.6.1. สามารถใช้โปรแกรมที่ได้พัฒนาขึ้นในการแปลงเอกสารหรือข้อความที่เราต้องการจะทำการเก็บเข้าสู่เครื่องคอมพิวเตอร์แทนการป้อนข้อมูลเข้าทางแป้นพิมพ์หรือเก็บข้อมูลของเอกสารนั้นด้วยวิธีการเก็บในลักษณะของแฟ้มข้อมูลแบบภาพ(Graphic File) ซึ่งการป้อนเข้าสู่เครื่องคอมพิวเตอร์ซึ่งการใช้โปรแกรมที่ได้พัฒนาขึ้นทำให้สามารถป้อนข้อมูลเข้าสู่เครื่องคอมพิวเตอร์โดยวิธีการใช้เครื่อง Scanner แทน

1.6.2. เป็นแนวทางในการพัฒนาระบบ OCR สำหรับการรู้จำคำไทย ซึ่งมีพื้นฐานจากการรู้จำตัวอักษรภาษาไทยต่อไป