บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเครื่องคอมพิวเตอร์มีบทบาทที่สำคัญไม่ว่าจะเป็นการดำเนินชีวิตประจำวันหรือในการทำงาน และมีแนวโน้มที่จะมีความสำคัญเพิ่มขึ้นเรื่อยๆ โดยเฉพาะการนำข้อมูลข่าวสารที่รวบรวมไว้ในระบบคอมพิวเตอร์มาใช้งานได้อย่างถูกต้องและรวดเร็ว และประหยัดเนื้อที่ในการเก็บกระดาษที่เป็นเอกสารต้นฉบับ จึงมีความพยายามที่จะจัดเก็บข้อมูลข่าวสารต่างๆที่มี ให้อยู่ในรูปแบบอิเล็กทรอนิกส์ เพื่อที่จะสามารถนำไปใช้ในระบบคอมพิวเตอร์ได้ เช่น การพิมพ์เอกสารใหม่จากต้นฉบับที่ต้องการ ลงในเครื่องคอมพิวเตอร์แล้วทำการจัดเก็บในฮาร์ดดิสก์ เพื่อการใช้งานในอนาคต แต่การพิมพ์เอกสารใหม่นั้น อาจเกิดข้อผิดพลาดขึ้นได้ ทำให้ข้อมูลข่าวสารนั้นคลาดเคลื่อนจากข้อมูลข่าวสารเดิม และการพิมพ์เอกสารใหม่ยังเสียเวลาและทรัพยากรอย่างมากอีกด้วย จึงมีการพัฒนาเครื่องสแกนเนอร์เพื่อแก้ปัญหาดังกล่าว

การใช้งานเครื่องสแกนเนอร์ เพื่อใช้ในการเก็บรูปภาพและเอกสารต้นฉบับให้อยู่ในรูปแบบอิเล็กทรอนิกส์ ประโยชน์ที่ได้รับก็คือสามารถประหยัดเนื้อที่การเก็บกระดาษที่เป็นเอกสารต้นฉบับและมีความถูกต้องและรวดเร็วกว่าการพิมพ์เอกสารใหม่ อย่างไรก็ตามการเก็บและแก้ไขเอกสารที่ได้จากเครื่องสแกนเนอร์ยังมีปัญหาอยู่มาก เนื่องจากเอกสารที่ได้มานั้น จะอยู่ในแฟ้มข้อมูลแบบรูปภาพ ซึ่งจะเปลืองเนื้อที่ฮาร์ดดิสก์ในการเก็บข้อมูลมากกว่าแฟ้มข้อมูลแบบตัวอักษร และการแก้ไขหรือการค้นหาข้อความในแฟ้มข้อมูลแบบรูปภาพ จะไม่สามารถใช้โปรแกรมประมวลผลคำทั่วๆไปได้

ต่อมาได้มีการพัฒนาโปรแกรมโอซีอาร์ (Optical Character Recognition) ขึ้น เพื่อแก้ปัญหาข้างต้นโดยการทำงานของโปรแกรมจะทำการอ่านแฟ้มข้อมูลแบบรูปภาพแล้วแปลงเป็นตัวอักษรเพื่อจัดเก็บเป็นแฟ้มข้อมูลแบบตัวอักษร ทำให้สามารถลดเนื้อที่ฮาร์ดดิสก์ในการเก็บข้อมูลและสามารถแก้ไขหรือค้นหาข้อความได้ด้วยโปรแกรมประมวลผลคำ แต่การใช้โปรแกรมโอซีอาร์ก็ทำให้เกิดปัญหาอื่นตามมา นั่นคือโปรแกรมโอซีอาร์ไม่สามารถแปลงข้อมูลมาได้ถูกต้องครบถ้วนทั้งหมด โดยเฉพาะโปรแกรมโอซีอาร์ภาษาไทยในปัจจุบันยังมีข้อผิดพลาดอยู่มาก ซึ่งยังไม่เหมาะสมที่จะนำแฟ้มข้อมูลเอกสารที่ได้จากโปรแกรมโอซีอาร์ภาษาไทยนั้นมาใช้งานได้ เพราะต้องทำการแก้ไขข้อมูลในแฟ้มข้อมูลเอกสารให้ถูกต้องเสียก่อน ซึ่งทำให้เสียเวลาและทรัพยากรเป็นจำนวนมาก ซึ่งบางครั้งการพิมพ์เอกสารที่ต้องการขึ้นมาใหม่อาจใช้เวลาน้อยกว่า

ในปัจจุบันได้มีงานวิจัยที่มุ่งแก้ข้อผิดพลาดของโปรแกรมโอซีอาร์ด้วยวิธีการใหม่ๆอยู่เป็นจำนวนมาก ซึ่งโปรแกรมโอซีอาร์ในต่างประเทศได้มีการพัฒนาไปแล้วอย่างมาก เช่นภาษาอังกฤษ ภาษาจีน และ ภาษาญี่ปุ่น ซึ่งในประเทศไทยเองก็มีงานวิจัยที่มุ่งแก้ข้อผิดพลาดของโปรแกรมโอซีอาร์ภาษาไทยด้วยวิธีการใหม่ๆอยู่บ้าง แต่ไม่ได้รับความสนใจเท่าที่ควรในการนำวิธีการใหม่ๆเหล่านั้นมาพัฒนาให้เป็นโปรแกรมโอซีอาร์ภาษาไทย

ในวิทยานิพนธ์ฉบับนี้จะมุ่งเน้นในการออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย โดยนำเอางานวิจัยที่มีอยู่มาเป็นส่วนประกอบในโปรแกรม เพื่อเป็นการนำเอาวิธีการใหม่ๆในงานวิจัยเหล่านั้นมาใช้แก้ปัญหาจริงให้เห็นเป็นรูปธรรม

งานวิจัยและทฤษฎีที่เกี่ยวข้อง

แบบจำลองของโปรแกรมโอซีอาร์ ที่ใช้กันอยู่ทั่วไปมี 3 ขั้นตอน (จุฬารัตน์ [1]) ดังนี้

Pre-Processing → OCR Processing → Post-Processing

รูปที่ 1.1 แสดงแบบจำลองของโปรแกรมโอซีอาร์

1. **ขั้นตอนเตรียมประมวลผล** (Pre-Processing) จะเป็นขั้นตอนที่ทำการปรับลักษณะและคุณสมบัติต่างๆของรูปภาพเอกสารให้เหมาะสมกับขั้นตอนประมวลผลการรู้จำ โดยวิธีการในขั้นตอนนี้มักจะใช้วิธีทางการประมวลผลรูปภาพ (Image Processing)
   - ข้อมูลนำเข้า : รูปภาพของเอกสาร (Document Image) ที่ได้จากการสแกน หรือแฟ้มข้อมูลรูปภาพที่ได้ทำการสแกนเก็บไว้แล้ว
   - ข้อมูลส่งออก : รูปภาพของเอกสารที่เหมาะสมจะนำไปใช้ในขั้นตอนประมวลผลการรู้จำ
2. **ขั้นตอนประมวลผลการรู้จำ** (OCR Processing) เป็นขั้นตอนที่ทำการอ่านตัวอักษรหรือการรู้จำตัวอักษรโดยแบ่งได้เป็นขั้นตอนหลักๆ 2 ขั้นตอน คือ
   1. การตัดแยกตัวอักษร (Character Segmentation)
   2. การรู้จำตัวอักษร (Character Recognition)

   ซึ่งขั้นตอนประมวลผลการรู้จำนี้ถือเป็นขั้นตอนหลักของโปรแกรมโอซีอาร์
   - ข้อมูลนำเข้า : รูปภาพของเอกสารที่ได้จากขั้นตอนเตรียมประมวลผล
   - ข้อมูลส่งออก : ลำดับของรหัสตัวอักษร (Text Stream)

3. ขั้นตอนหลังประมวลผล (Post-Processing) เป็นขั้นตอนที่ช่วยในการปรับปรุงผลลัพธ์ที่ได้จากขั้นตอนประมวลผลการรู้จำ ซึ่งวิธีการที่นำมาใช้ในขั้นตอนนี้ก็คือการแก้คำที่สะกดผิด (Spelling Corrector)
   - ข้อมูลนำเข้า : ลำดับของรหัสตัวอักษร หรือ แฟ้มข้อมูลที่เก็บลำดับของรหัสตัวอักษรที่ได้จากขั้นตอนประมวลผลการรู้จำ
   - ข้อมูลส่งออก : แฟ้มข้อมูลที่เก็บลำดับของรหัสตัวอักษรที่ได้รับการแก้ไขคำที่สะกดให้ถูกต้องแล้ว

วิธีการที่ใช้ในขั้นตอนเตรียมประมวลผล

จากแบบจำลองของโปรแกรมโอซีอาร์ ขั้นตอนเตรียมประมวลผลนอกจากจะใช้ในการปรับลักษณะและคุณสมบัติต่างๆของรูปภาพเอกสารให้เหมาะสมกับขั้นตอนประมวลผลการรู้จำแล้ว ยังมีส่วนช่วยให้การตัดแยกตัวอักษรและการรู้จำตัวอักษรมีความถูกต้องมากยิ่งขึ้น โดยใช้วิธีทางการประมวลผลรูปภาพ ซึ่งบางวิธีสามารถทำได้เองด้วยเครื่องคอมพิวเตอร์ แต่บางวิธีจำเป็นต้องใช้มนุษย์เป็นผู้พิจารณา ขั้นตอนเตรียมประมวลผลโดยส่วนใหญ่จะประกอบด้วยวิธีการดังนี้

- <u>การแปลงระดับความเข้มสีของรูปภาพจากหลายระดับเป็นสองระดับ</u> (Binarization) เนื่องจากการสแกนรูปภาพนั้น ผู้สแกนสามารถกำหนดจำนวนระดับความเข้มสีของรูปภาพได้ เช่น 256 ,16, 4 หรือ 2 ระดับ แต่การแปลงระดับความเข้มสีของรูปภาพนั้นขึ้นอยู่กับว่าขั้นตอนโอซีอาร์โปรเซสซิ่งนั้นต้องการรูปภาพที่มีจำนวนระดับความเข้มเป็นเท่าไร โดยส่วนใหญ่จะใช้รูปที่มีจำนวนระดับความเข้มสี 2 ระดับคือ ขาว และดำ
- <u>การลดจุดภาพรบกวน</u> (Noise Reduction) จุดภาพรบกวน เป็นจุดภาพที่เราไม่ต้องการซึ่งเกิดขึ้นในขั้นตอนการสแกนภาพหรืออาจเกิดจากต้นฉบับที่เป็นเอกสารมีจุดปรากฏอยู่ ดังรูปที่ 1.2

รูปที่ 1.2 แสดงจุดภาพรบกวนที่ปรากฏในภาพอาจทำให้เกิดความผิดพลาดขึ้นได้

จุดภาพรบกวนมีผลทำให้รูปตัวอักษรเพี้ยนไปจากเดิมและจะทำให้การรู้จำผิดพลาดด้วย จึงมีความจำเป็นจะต้องลดจุดภาพรบกวนเพื่อให้ได้รูปภาพที่ใกล้เคียงตามความเป็นจริง

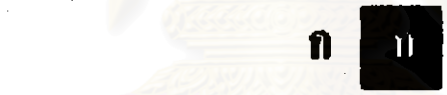
- <u>การหมุนภาพ</u> (Rotation) บางครั้งรูปภาพที่ได้จากการสแกนเอกสารนั้นไม่ตรง อาจเป็นเพราะการสแกนเอียงหรือต้นฉบับเอียง ซึ่งจะทำให้ขั้นตอนประมวลผลการรู้จำนั้น

มีความผิดพลาดเกิดขึ้นได้ จึงต้องทำการหมุนภาพให้กลับมาตรงตามที่ต้องการ ในปัจจุบันมีการนำวิธีการทางสถิติที่เกี่ยวข้องกับการประมาณค่าความเอียงของเอกสารเพื่อทำการหมุนภาพเอกสารเองด้วยเครื่องคอมพิวเตอร์ เช่น การหาองศาความเอียงโดยใช้สมการถดถอย (ยุทธพงษ์ และ กฤษณะ [2]) แต่ในบางครั้งการกำหนดองศาความเอียงของเอกสารโดยผู้ใช้เองนั้น ได้ผลลัพธ์ที่ดีและง่ายกว่ามาก และหลังจากที่ได้ค่าองศาความเอียงของเอกสารแล้วก็จะทำการหมุนภาพตามค่าความเอียงที่ได้ต่อไป

- <u>การกลับภาพ</u> (Flipping, Reflection) เป็นวิธีการกลับภาพให้อยู่ในลักษณะที่ถูกต้อง ดังรูปที่ 1.3 เป็นการกลับภาพในแนวนอน (Flip Horizontal) และ แนวตั้ง (Flip Vertical) ตามลำดับ

รูปที่ 1.3 แสดงการกลับภาพในแนวนอนและการกลับภาพในแนวตั้งตามลำดับ

- <u>การกลับสีของจุดภาพ</u> (Inversion) เป็นการกลับสีของจุดภาพโดยจะแทนจุดสีดำด้วยสีขาว และแทนจุดสีขาวด้วยสีดำตามรูปที่ 1.4



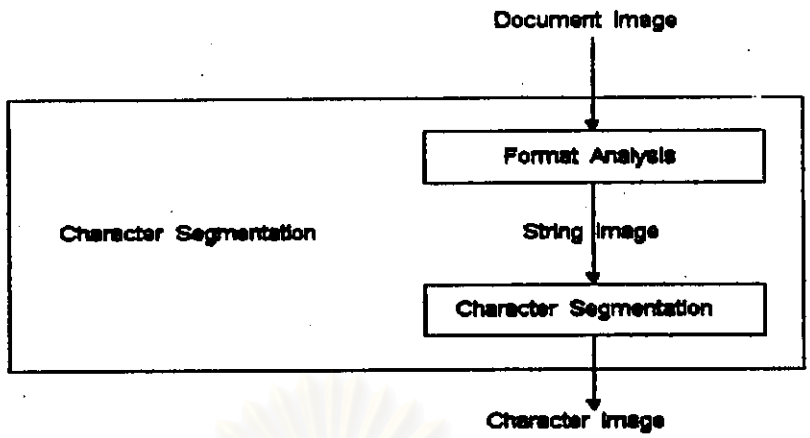
รูปที่ 1.4 แสดงการกลับสีของจุดภาพ

รูปภาพของเอกสารที่จะนำมาใช้ในโปรแกรมโอซีอาร์ จะต้องนำมาพิจารณาก่อนว่าลักษณะของรูปนั้นยังไม่เหมาะสมกับขั้นตอนประมวลผลการรู้จำอย่างไร จึงจะนำวิธีการต่างๆ ในขั้นตอนเตรียมประมวลผลที่เกี่ยวข้องเหล่านี้ มาใช้ปรับลักษณะให้เหมาะสมต่อไป

วิธีการที่ใช้ในขั้นตอนประมวลผลการรู้จำ

จากแบบจำลองของโปรแกรมโอซีอาร์ ขั้นตอนการรู้จำตัวอักษรแบ่งออกเป็น 2 ขั้นตอน คือ

1. *การตัดแยกตัวอักษร* (Character Segmentation) เป็นขั้นตอนที่ใช้แยกรูปของตัวอักษรออกจากรูปของเอกสารเพื่อส่งให้กับขั้นตอนที่ 2 ต่อไป ในงานวิจัยของ Casey and Lecolinet [3] มีแบบจำลองของการตัดแยกตัวอักษรตามรูปที่ 1.5



รูปที่ 1.5 แสดงแบบจำลองของการตัดแยกตัวอักษร

ซึ่งในงานวิจัยของ Casey and Lecolinet [3] สามารถแบ่งการตัดแยกตัวอักษรออกได้เป็น 3 รูปแบบ คือ

- Dissection เป็นรูปแบบที่นิยมใช้ โดยเน้นการพิจารณารูปของตัวอักษร (character-like) คือ ตัวอักษรแต่ละตัวจะต้องสามารถหาขอบเขตได้ชัดเจน และจะได้รูปของตัวอักษรตามขอบเขตนั้น ตามรูปที่ 1.6

รูปที่ 1.6 แสดงรูปที่ได้จากการตัดแยกตัวอักษรแบบ Dissection

- Recognition-based เป็นรูปแบบที่ตรงข้ามกับ Dissection โดยจะไม่เน้นที่การพิจารณารูปของตัวอักษร แต่จะคำนึงถึงความเป็นไปได้ในการรู้จำตัวอักษรนั้น ซึ่งรูปแบบนี้จะรวมขั้นตอนการแยกตัวอักษรกับขั้นตอนการรู้จำตัวอักษรเข้าด้วยกัน เพราะการแยกตัวอักษรต้องใช้ความรู้จากขั้นตอนในการรู้จำมาช่วยในการพิจารณาว่าควรจะต้องแยกตัวอักษรที่ใด ดังรูปที่ 1.7

Input Pattern	Windowed Input	Matching Prototype 1	Residue	Matching Prototype 2
				
				
				

รูปที่ 1.7 แสดงการแยกตัวอักษรโดยใช้ความรู้จากขั้นตอนการรู้จำตัวอักษร

ซึ่งการพิจารณาว่าควรจะแยกตัวอักษรที่ใดอาจใช้ความรู้ทางวากยสัมพันธ์ (syntactic) และ อรรถศาสตร์ (semantics) มาร่วมพิจารณาด้วย

- Holistic เป็นรูปแบบที่จะพิจารณาถึงความเป็นไปได้ในการรู้จำคำทั้งคำ ซึ่งเป็นการหลีกเลี่ยงปัญหาที่จะพบในขั้นตอนการรู้จำตัวอักษร แต่รูปแบบ Holistic หากจะใช้ในงานเอกสารที่มีคำอยู่มากจะต้องใช้ความรู้จากพจนานุกรม (Lexicon) ขนาดใหญ่ จึงนิยมใช้ในงานบางประเภทที่มีคำไม่มาก

จากแบบจำลองของการตัดแยกตัวอักษร (รูปที่ 1.5) จะพบว่าข้อมูลนำเข้าคือรูปภาพของเอกสาร (Document Image) และจะผ่านขั้นตอนย่อยอีก 2 ขั้นตอน คือ

1. Format Analysis เป็นขั้นตอนในการวิเคราะห์โครงสร้างของเอกสาร เช่น ส่วนใดควรจะเป็นบรรทัดเดียวกัน ส่วนใดเป็นย่อหน้า เป็นต้น แล้วส่งผลลัพธ์ที่เป็นรูปของบรรทัด(String Image) ไปยังขั้นตอนต่อไป วิธีการที่ใช้ เช่น การตรวจสอบค่า Histogram ของภาพในแนวนอน เพื่อหาพื้นที่ที่มีความน่าจะเป็นสูงที่จะเป็นช่องว่างระหว่างบรรทัด
2. Character Segmentation เป็นขั้นตอนที่ทำการแยกรูปตัวอักษร (Character Image) ออกจากรูปภาพของบรรทัดที่ได้จากขั้นตอน Format Analysis มักใช้วิธีการล้อมกรอบตัวอักษร (Bounding Box Analysis) เพื่อกำหนดขอบเขตของตัวอักษรแต่ละตัว

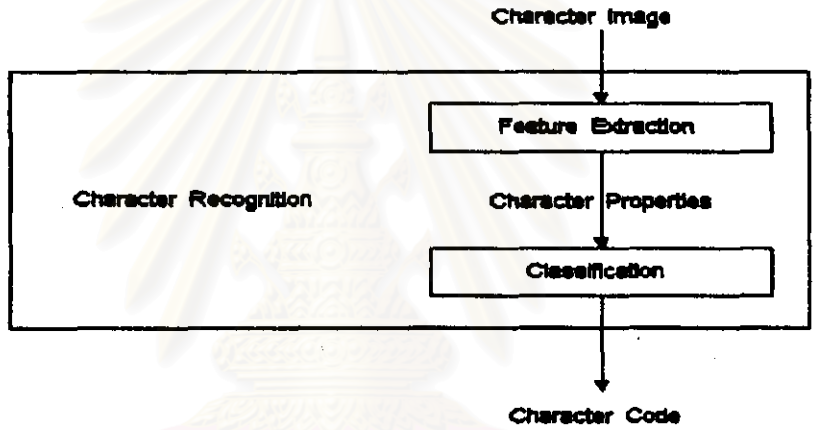
การตัดแยกตัวอักษรโดยเฉพาะตัวอักษรภาษาไทยนั้นจะใช้รูปแบบ Dissection เนื่องจากเหมาะสมกับลักษณะของตัวอักษรและคำในภาษาไทยมากกว่า แต่เนื่องจากตัวอักษรภาษาไทยนั้นมีหลายระดับและมักเกิดการชนกันของตัวอักษร ดังนั้นจึงเกิดปัญหาในการกำหนดขอบเขตที่แน่นอนขึ้น จึงมีงานวิจัยที่เกี่ยวข้องกับการตัดแยกตัวอักษรภาษาไทย เช่น

Premchaiswadi et al. [4] เสนอการตัดแยกตัวอักษรพิมพ์ภาษาไทยที่ติดกันในแนวนอนและแนวตั้ง โดยอาศัยโครงสร้างหลายระดับ (Multi-Level Structure) ของประโยคในภาษาไทย และ Distinctive Features ของตัวอักษรภาษาไทย เพื่อแบ่งระดับของตัวอักษรและใช้ในการพิจารณาประเภทของการติดกันของตัวอักษร ซึ่งการตัดแยกตัวอักษรจะขึ้นอยู่กับประเภทของการติดกันของตัวอักษร

จรรยา และ บุญธีร์ [5] นำเสนอวิธีการแยกตัวอักษรภาษาไทยที่ติดกันโดยวิธี Shortest Path เพื่อมุ่งสร้างระบบการแยกตัวอักษรพิมพ์ภาษาไทย ที่มีลักษณะการติดในแนวนอน และมีจำนวนการติดของตัวอักษร 2 ตัว โดยใช้วิธี Shortest Path ทางตรง และทางเฉียง

Panich et al. [6] เสนอวิธี Segmentation of Connected Character using Distinctive Features of Thai Characters in Thai Character Recognition System โดยแบ่งตัวอักษรไทยออกเป็นกลุ่มต่างๆ 6 กลุ่มตามระดับของตัวอักษรที่ปรากฏ และแบ่งลักษณะของการติดกันของตัวอักษรทั้ง 6 กลุ่มนั้น ออกเป็น 10 กลุ่ม และได้เสนอวิธีการหาการติดกันของตัวอักษรว่าเกิดขึ้นในลักษณะของกลุ่มใด และวิธีการตัดแยกตัวอักษรที่ติดกันนั้น

2. *การรู้จำตัวอักษร* (Character Recognition) เป็นขั้นตอนที่รู้จำตัวอักษรโดยใช้รูปของตัวอักษรที่ได้รับจากขั้นตอนการตัดแยกตัวอักษร เพื่อให้ได้รหัสของตัวอักษร ในงานวิจัยของ Casey and Lecolinet [3] มีแบบจำลองของการรู้จำตัวอักษร ตามรูปที่ 1.8



รูปที่ 1.8 แสดงแบบจำลองของการรู้จำตัวอักษร

จากแบบจำลองของการรู้จำตัวอักษรในรูปที่ 1.8 จะพบว่าข้อมูลนำเข้าคือรูปภาพของตัวอักษร (Character Image) ซึ่งได้รับมาจากการตัดแยกตัวอักษร และจะผ่านขั้นตอนย่อยอีก 2 ขั้นตอน คือ

1. Feature Extraction เป็นขั้นตอนย่อยที่ใช้ในการแยกลักษณะสำคัญของตัวอักษร เพื่อให้ได้คุณสมบัติเฉพาะของตัวอักษรนั้น (Character Properties)
2. Classification เป็นขั้นตอนที่นำเอาคุณสมบัติเฉพาะของตัวอักษรมาทำการพิจารณาและแยกแยะตัวอักษรออกตามประเภทและได้ผลลัพธ์คือรหัสของตัวอักษร (Character Code)

งานวิจัยเกี่ยวกับการรู้จำตัวอักษรภาษาไทย เช่น

สุรพันธ์ [7] นำเสนอวิธีการรู้จำตัวอักษรลายมือเขียนภาษาไทยโดยการนำหัวของตัวอักษรมาพิจารณาเพื่อทำการจำแนกกลุ่มของตัวอักษรออกเป็นกลุ่มย่อยๆ โดยให้ตัวอักษรที่มีหัวอยู่บริเวณเดียวกันอยู่กลุ่มเดียวกัน จากนั้นก็จะพิจารณาเปรียบเทียบตัวอักษรที่อยู่ในกลุ่มของตนเองแตกต่างไป ทั้งนี้ขึ้นกับลักษณะเด่นของตัวอักษรที่อยู่ในกลุ่มนั้นๆ ซึ่งทำให้กลุ่มของตัวอักษรที่ต้องเปรียบเทียบมีจำนวนลดลง ทำให้เปรียบเทียบได้อย่างรวดเร็ว

พิพัฒน์ และ มนลดา [8] นำเสนอวิธีการรู้จำตัวอักขระไทยหลายรูปแบบโดยใช้วิธีการไดนามิคโปรแกรมมิ่ง โดยการพิจารณาเส้นแสดงขอบของอักขระโดยนำรหัสทิศทางแบบลูกโซ่ของฟรีแมน กับความแตกต่างของทิศทางของเส้นแสดงขอบอักขระมาใช้ในการตัดแบ่งเส้นแสดงขอบของอักขระออกเป็นส่วนโค้งเว้าและส่วนโค้งนูน เพื่อนำไปใช้ในการเปรียบเทียบแบบไดนามิคโปรแกรมมิ่ง โดยการคำนวณค่าความคล้ายคลึง (Similarity) ของส่วนโค้งเว้าและส่วนโค้งนูนที่ได้กับส่วนโค้งเว้าและส่วนโค้งนูนของตัวอักขระต้นแบบ

สนธยา [9] นำเสนอเรื่องการศึกษาการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยวิธีซินแทกติก (King Sun Fu [10]) ซึ่งเป็นการพิจารณาโครงสร้างของตัวอักษรโดยการเปลี่ยนแปลงเส้นแสดงขอบของตัวอักษรให้อยู่ในรูปรหัสทิศทางแบบลูกโซ่ของฟรีแมน และเปลี่ยนรหัสลูกโซ่เป็นรหัสเวกเตอร์เส้นตรงและวงกลม เพื่อนำมาจัดเก็บเป็นรูปของประโยคที่ประกอบด้วยพรีมิทีฟ (Primitive) ในลักษณะของโครงสร้างแบบต้นไม้ และอาศัยวิธีการวิเคราะห์ประโยคที่ได้จากการเปลี่ยนเส้นแสดงขอบของตัวอักษรเปรียบเทียบกับประโยคของอักขระต้นแบบ โดยเลือกเปรียบเทียบเฉพาะตัวอักษรต้นแบบที่มีหัวของตัวอักษรอยู่ในบริเวณเดียวกันกับหัวของตัวอักษรที่ต้องการรู้จำ และเปรียบเทียบเฉพาะตัวอักษรที่เป็นตัวอักษรอยู่ในระดับเดียวกันเท่านั้น (โดยการระบุเส้นบอกระดับ) สำหรับตัวอักษรที่แตกต่างกันไม่มากจะถูกนำไปเปรียบเทียบทางฟีเจอร์ (feature) อีกครั้งหนึ่งโดยการเก็บลักษณะพิเศษของแต่ละตัวอักษรไว้ หากผลการรู้จำในช้างต้นไม่อยู่ในเกณฑ์ที่ยอมรับได้ เวกเตอร์ของตัวอักษรจะถูกนำมา ปรับปรุงเพื่อตัดส่วนเกินออก หรือเชื่อมเวกเตอร์ที่อยู่ใกล้เคียงกันเข้าด้วยกันแล้วจึงนำมาทำการรู้จำ โดยวิธีเดิมอีกจนกว่าผลการรู้จำจะอยู่ในเกณฑ์ที่ยอมรับได้หรือไม่สามารถทำการปรับปรุงเวกเตอร์ได้อีก

เดชา [11] นำเสนอเรื่องการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยใช้เทคนิคแบบฟัซซีโลจิก และวิธีซินแทกติก โดยทำการปรับปรุงวิธีการซินแทกติกของ สนธยา [9] โดยการนำเทคนิคแบบฟัซซีโลจิกเข้ามาใช้เมื่อการใช้วิธีการทางซินแทกติกไม่สามารถรู้จำตัวอักษร รวมทั้งปรับปรุงวิธีการทำตัวอักษรให้บาง (Pavlidis [12]) โดยการใช้วิธีทำตัวอักษรให้บางแบบเอสพีทีเอ (SPTA, Save Point Thinning Algorithm)

อภิญญา [13] นำเสนอเรื่องการประยุกต์ใช้การโปรแกรมตรรกะเชิงอุปนัยในการรู้จำตัวอักษรพิมพ์ภาษาไทย โดยนำการเรียนรู้โดยการอุปนัยโดยใช้การโปรแกรมตรรกะเชิงอุปนัย (Inductive Logic Programming, ILP) หรือ ไอ แอล พี โดยใช้ความรู้ส่วนหลัง (background knowledge) ในการสร้างสมมติฐานใหม่ที่สอดคล้องกับตัวอย่างที่ได้รับ ซึ่งเทคนิคขั้นต้นจะเป็นการพิจารณาโครงสร้างของตัวอักษรโดยทำการเปลี่ยนขอบของตัวอักษรเป็นรหัสทิศทางแบบลูกโซ่ของฟรีแมน ทำการเปลี่ยนรหัสทิศทางแบบลูกโซ่ของฟรีแมนเป็นเวกเตอร์เส้นตรง และเวกเตอร์วงกลม แล้วทำการเปลี่ยนเวกเตอร์เป็นหน่วยสร้างพื้นฐาน นำการโปรแกรมตรรกะเชิงอุปนัยมาทำการเรียนรู้ลักษณะของหน่วยสร้างพื้นฐานที่ได้จากตัวอักษรต้นแบบ เช่น ระดับของตัว อักษรขนาดของตัวอักษร ลักษณะส่วนหัวของตัวอักษร ลักษณะส่วนปลายของตัวอักษร เป็นต้น

ธเนศ [14] นำเสนอเรื่องการรู้จำตัวอักษรพิมพ์ภาษาไทยโดยใช้เทคนิคด้านการวิเคราะห์ตัวประกอบสำคัญและนิวรอลเน็ตเวิร์ก ซึ่งใช้วิธีการแปลงแบบเค-แอล (Karhunen Loeve Transform) เป็นขั้นตอนที่ใช้ในการวิเคราะห์ตัวประกอบสำคัญ ซึ่งจะทำการแปลงข้อมูลภาพให้กลายเป็นข้อมูลเวกเตอร์ ซึ่งลักษณะของตัวอักษรต่างๆในภาพนั้นจะมีข้อมูลเวกเตอร์ที่บ่งบอกถึงลักษณะสำคัญและคุณสมบัติเฉพาะของตัวอักษรที่แตกต่างกัน และใช้นิวรอลเน็ตเวิร์กแบบแบคพรอพาเกชัน เป็นขั้นตอนที่นำเอาเวกเตอร์ของตัวอักษรที่ได้ มาทำการพิจารณาและแยกแยะตัวอักษร ในวิทยานิพนธ์ฉบับนี้จะใช้วิธีการของธเนศ [14] เป็นหลักในขั้นตอนการรู้จำ

วิธีการที่ใช้ในขั้นตอนหลังประมวลผล

ผลลัพธ์ที่ได้จากขั้นตอนประมวลผลการรู้จำนั้น อาจไม่ถูกต้องซึ่งจะต้องนำมาแก้ไขก่อนนำไปใช้งานจริง วิธีที่นิยมใช้กันโดยทั่วไปก็คือใช้โปรแกรมตรวจและแก้ไขคำที่สะกดผิด (Spell Corrector) ซึ่งโปรแกรมตรวจและแก้ไขคำที่สะกดผิดโดยทั่วไปจะใช้วิธีค้นหาคำศัพท์ในพจนานุกรมเท่านั้น แต่วิธีการแก้ไขคำผิดที่นำเสนอโดย อนันต์ลดา และ ชลวิช [15] จะพิจารณาถึงความถูกต้องด้านภาษาศาสตร์และความผิดพลาดที่มักเกิดขึ้นจากโปรแกรมโอซีอาร์ ร่วมกันในการแก้ไขด้วย โดยใช้ค่าสถิติของไตรแกรมของประเภทของคำ (Part of speech Tri-gram), ค่าความน่าจะเป็นของคำเมื่อรู้ประเภทของคำ (probability of word given part of speech), ค่าสถิติของความผิดพลาดที่เกิดขึ้นจากโปรแกรมโอซีอาร์ และ ค่าสถิติของไตรแกรมของตัวอักษร (Character Tri-gram) เพื่อใช้ในการตรวจสอบคำผิดรวมถึงการเลือกคำที่เหมาะสมที่สุดที่จะนำมาใช้แทนคำผิดในประโยค และใช้ระยะแก้ไขแบบเคลื่อนที่ (Dynamic Edit Distance) เพื่อความเหมาะสมในการแก้ไขคำที่มีความสั้นยาวแตกต่างกัน

วัตถุประสงค์

เพื่อออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย โดยพัฒนาวิธีการที่ใช้ในขั้นตอนเตรียมประมวลผลและวิธีการตัดแยกตัวอักษร และประยุกต์วิธีการแยกลักษณะสำคัญของตัวอักษร (เค-แอล ทรานส์ฟอร์ม), วิธีการแยกแยะตัวอักษร (นิวรอลเน็ตเวิร์กแบบแบคพรอพาเกชัน) และวิธีการแก้ไขคำที่สะกดผิด (ไตรแกรมของประเภทของคำ) เพื่อให้สามารถทำงานร่วมกับส่วนอื่นๆในโปรแกรม

ขอบเขตของวิทยานิพนธ์

ออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย โดยมีความสามารถดังนี้

1. โปรแกรมจะทำงานภายใต้ระบบปฏิบัติการ Microsoft Windows บนเครื่องคอมพิวเตอร์ที่มีข้อกำหนดขั้นต่ำคือ ใช้หน่วยประมวลผลกลางรุ่น Pentium 233 MMX มีหน่วยความจำ 64 เมกกะไบต์ขึ้นไปและเนื้อที่ฮาร์ดดิสก์ที่เพียงพอต่อการทำงานของโปรแกรม
2. สามารถดำเนินการในขั้นตอนเตรียมประมวลผลได้
3. สามารถดำเนินการในขั้นตอนประมวลผลรู้จำได้ ด้วยวิธีการแยกลักษณะสำคัญของตัวอักษร (เค-แอล ทรานส์ฟอร์ม), วิธีการแยกแยะตัวอักษร (นิวรอลเน็ตเวิร์กแบบแบคพรอพาเกชัน)
4. สามารถดำเนินการในขั้นตอนหลังประมวลผลได้ ด้วยวิธีการไตรแกรมของประเภทของคำ
5. สามารถรับข้อมูลภาพเอกสารจากเครื่องสแกนเนอร์ผ่านทางโปรแกรมโอซีอาร์ภาษาไทยได้โดยตรง
6. สามารถเปิดแฟ้มข้อมูลภาพที่ได้ทำการสแกนเก็บไว้ มาทำการประมวลผลรู้จำได้
7. สามารถเก็บแฟ้มข้อมูลตัวอักษรที่ได้จากการประมวลผลรู้จำ เพื่อนำไปใช้งานในโปรแกรมประมวลผลคำทั่วไปได้
8. แฟ้มข้อมูลรูปภาพเอกสารที่จะใช้ในการทดสอบจะต้องมีลักษณะดังนี้
   - แฟ้มข้อมูลรูปภาพเอกสารจะต้องมีสีขาว-ดำ จัดเก็บแบบ BMP
   - ตัวอักษรที่ปรากฏในเอกสารคือ ตัวอักษรพิมพ์ภาษาไทยและภาษาอังกฤษ ตัวเลขพิมพ์ไทยและอาราบิค แบบ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC และ FreesiaUPC

- ขนาดของตัวอักษรที่จะใช้ทดสอบคือขนาด 14, 16, 18, 20, 22, 24, 28 และ 36 โดยที่ขนาดของตัวอักษรในเอกสารจะต้องมีขนาดเท่ากันทั้งเอกสาร
- รูปแบบของตัวอักษรจะต้องไม่มีรูปแบบเอียง, รูปแบบหนา หรือขีดเส้นใต้

ขั้นตอนการทำวิทยานิพนธ์

1. ศึกษาโปรแกรมโอซีอาร์ภาษาไทย
2. ศึกษางานวิจัยที่เกี่ยวข้องกับการทำงานของโปรแกรมโอซีอาร์ภาษาไทย
3. ออกแบบขั้นตอนการทำงานของโปรแกรมโอซีอาร์ภาษาไทย
4. พัฒนาฐานความรู้ที่เหมาะสมในการรู้จำตัวอักษรที่ใช้กับโปรแกรมโอซีอาร์ภาษาไทย
5. พัฒนาโปรแกรมโอซีอาร์ภาษาไทยตามที่ได้ทำการออกแบบไว้
6. ทดสอบโปรแกรมโอซีอาร์ภาษาไทย และเปรียบเทียบผลกับโปรแกรมโอซีอาร์ภาษาไทยอื่นๆ
7. สรุปผลการทำงาน

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้โปรแกรมโอซีอาร์ภาษาไทยใหม่ที่ใช้วิธีการแยกลักษณะสำคัญของตัวอักษร (เค-แอล ทรานส์ฟอร์ม), วิธีการแยกแยะตัวอักษร (นิวรอลเน็ตเวิร์กแบบแบคพรอพาเกชัน) และวิธีการแก้ไขคำที่สะกดผิด (ไตรแกรมของประเภทของคำ)
2. สามารถนำโปรแกรมโอซีอาร์ภาษาไทยที่ได้ ไปใช้ในการจัดเก็บข้อมูลจากเอกสารให้อยู่ในแฟ้มข้อมูลแบบรหัสตัวอักษรของระบบคอมพิวเตอร์
3. ได้ฐานความรู้ในการรู้จำตัวอักษรและการแก้ไขคำสะกดผิด ที่เหมาะสมกับโปรแกรมโอซีอาร์ภาษาไทย
4. สามารถนำวิธีการที่ได้จากวิทยานิพนธ์นี้ไปประยุกต์ใช้กับงานอื่นๆได้