บทที่ 2
แนวคิด ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

ในการจัดทำโครงการวิจัยโปรแกรมรู้จำและวิเคราะห์ตัวอักษรล้านนา ผู้วิจัยได้ศึกษาแนวคิด ทฤษฎีและผลงานวิจัยที่เกี่ยวข้องที่ใช้เป็นแนวทางในการออกแบบโครงการวิจัย โดยหลักการและทฤษฎีที่เกี่ยวข้องมีดังนี้ OpenCV, OCR , Histogram Matching และภาษาล้านนา

OpenCV (Open Source Computer Vision)

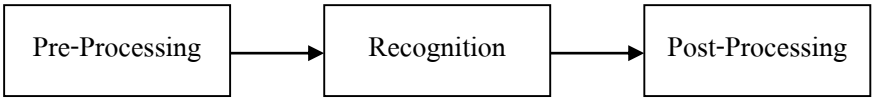
โปรแกรม OpenCV เป็น Library ในภาษา C++ สำหรับการพัฒนาโปรแกรมที่เกี่ยวข้องกับ Image Processing และ Computer Vision โดยสามารถพัฒนาได้ทั้งในระบบปฏิบัติการวินโดว์และ ระบบปฏิบัติการ Linux วิธีใช้งาน OpenCV ใช้ร่วมกับโปรแกรม Visual Studio C++ ในการนำ OpenCV มาใช้งานร่วมกับ Visual Studio C++ เพื่อใช้พัฒนาโปรแกรมนั้น จำเป็นต้องตั้งค่าให้กับโปรแกรม Visual Studio C++ ก่อน เพื่อระบุตำแหน่งของ Library ของ OpenCV ตำแหน่งของไฟล์ที่ต้องใช้ในโปรแกรม และตำแหน่งของ Source File ให้ตัวโปรแกรมทราบ และสามารถดึงมาใช้ได้

OCR (โอซีอาร์)

Optical Character Recognition หรือที่เรียกกันสั้นๆว่า OCR (โอซีอาร์) : คือ การแปลงไฟล์ภาพเอกสาร ให้เป็นไฟล์ข้อความโดยอัตโนมัติ โครงสร้างระบบโอซีอาร์โดยทั่วไปประกอบไปด้วยขั้นตอนการทำงานหลัก 3 ขั้น ซึ่งในแต่ละขั้นตอนประกอบด้วยรายละเอียดดังต่อไปนี้

1. ขบวนการประมวลผลขั้นต้น (Pre-Processing)
    - 1.1 การกรองข้อมูลแทรกซ้อน (Noise Filtering)
    - 1.2 การปรับแต่งข้อมูล (Normalization)
    - 1.3 การตัดแบ่งพื้นที่ใช้งาน (Cropping)
    - 1.4 การสกัดลักษณะสำคัญ (Feature Extraction)
2. การรู้จำ (Recognition)
    - 2.1 วิธีทางการเข้าคู่รูปแบบ (Template Matching)
    - 2.2 วิธีทางสถิติ (Statistical Approach)
    - 2.3 วิธีการวิเคราะห์ทางโครงสร้าง (Structural Analysis)
    - 2.4 วิธีทางโครงข่ายประสาทเทียม (Neural Network)

3. ขบวนการประมวลผลขั้นปลาย (Post-Processing)



ภาพที่ 2.1 โครงสร้างของระบบโอซีอาร์

1.ขบวนการประมวลผลขั้นต้น (Pre-Processing)

ในการทำงานของโปรแกรมโอซีอาร์นั้น ก่อนที่โปรแกรมจะสามารถบอกได้ว่ารูปภาพที่ส่งเข้าไปประกอบด้วยตัวอักษรอะไรบ้าง จำเป็นจะต้องผ่านขั้นตอนที่สำคัญหลายขั้น ขั้นตอนดังกล่าวนี้มักถูกเรียกรวมกันว่า กระบวนการประมวลผลขั้นต้น (Pre-Processing) ซึ่งเป็นขั้นตอนในการปรับแต่งและจัดเตรียมข้อมูลให้เหมาะสมกับขั้นตอนการรู้จำต่อไป ขั้นตอนเหล่านี้มีความสำคัญต่อประสิทธิภาพโดยรวมของระบบ เพราะหากมีความผิดพลาดเกิดขึ้นในส่วนนี้ ก็จะส่งผลกระทบไปยังส่วนถัดไปของระบบด้วย

ขั้นตอนการประมวลผลเบื้องต้นในโปรแกรมโอซีอาร์ที่สำคัญ ได้แก่

1.1 การกรองข้อมูลแทรกซ้อน (Noise Filtering)

การกรองข้อมูลแทรกซ้อนมีจุดประสงค์เพื่อลดทอนส่วนของรูปภาพที่เป็นสิ่งแปลกปลอมอันไม่พึงประสงค์ออกไป โดยข้อมูลแทรกซ้อนที่เกิดขึ้นส่วนใหญ่มักจะมาจากคุณภาพของเอกสารต้นฉบับที่นำมาทำการอ่าน ซึ่งเป็นต้นเหตุสำคัญที่ทำให้ความถูกต้องของโปรแกรมลดลง จึงจำเป็นที่จะต้องจัดการกับส่วนเกินเหล่านี้ออกไปให้ได้มากที่สุดเท่าที่จะเป็นไปได้ แต่เรายังไม่มีวิธีการใดที่รับรองได้ว่าสามารถจัดการกับข้อมูลแทรกซ้อนได้ โดยสมบูรณ์ ดังนั้นส่วนการรู้จำของโอซีอาร์ก็จะต้องมีความทนทานต่อการแทรกซ้อนเหล่านี้ได้พอสมควร

(a) (b)

ภาพที่ 2.2 (a) ภาพที่ไม่ได้ทำการกรองข้อมูลแทรกซ้อน (b) ภาพที่ทำการกรองข้อมูลแทรกซ้อน

1.2 การปรับแต่งข้อมูล (Normalization)

การปรับแต่งข้อมูลเป็นการปรับภาพตัวอักษรให้อยู่ในรูปแบบที่ระบบต้องการเพื่อนำไปใช้ในขั้นต่อไป ตัวอย่างการปรับแต่งข้อมูลในโปรแกรมโอซีอาร์ทั่วๆ ไป อาทิเช่น การปรับขนาดรูปตัวอักษร, การปรับตัวอักษรที่เอียงให้ตรง, การแปลงรูปสีหรือเกรย์สเกลให้เป็นขาวดำ หรือในทางกลับกัน การแปลงรูปขาวดำให้เป็นสีหรือเกรย์สเกล เป็นต้น

(a) (b)

ภาพที่ 2.3 (a) เป็นภาพที่ไม่ได้ทำการปรับแต่งข้อมูล (b) เป็นภาพที่ทำการปรับแต่งข้อมูล

1.3 การตัดแบ่งพื้นที่ใช้งาน (Cropping)

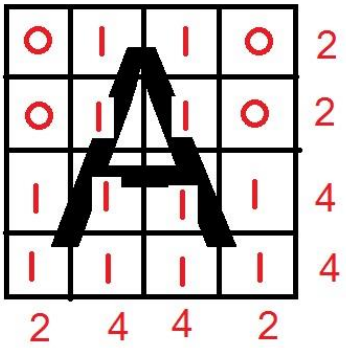
การตัดแบ่งพื้นที่เป็นการตัดแยกเอาเฉพาะรูปตัวอักษรออกมาจากภาพ เพื่อส่งให้ขั้นตอนการรู้จำในการระบุว่ารูปตัวอักษรนั้นเป็นรหัสอักษรอะไร หลักการพอสังเขปที่ใช้สำหรับการตัดรูปตัวอักษรโดยทั่วไปจะใช้พื้นที่สีขาว (สีพื้น) รอบรูปเป็นตัวกำหนดขอบเขตในการตัด ในขั้นตอนนี้มักจะประสบปัญหาที่ส่งผลกระทบต่ออัตราความถูกต้องของระบบโดยรวมอยู่สองปัญหา ปัญหาแรกคือปัญหาตัวติด เกิดจากรูปของตัวอักษรตั้งแต่สองตัวขึ้นไปมีส่วนที่เชื่อมติดกัน ทำให้ไม่สามารถแยกตัวอักษรออกจากกันโดยใช้พื้นที่สีขาวรอบๆ ได้ จำเป็นต้องหาอัลกอริทึมพิเศษมาช่วยในการแยกตัวอักษรออกจากกัน ส่วนปัญหาที่สองในทางตรงกันข้าม เป็นปัญหาตัวขาดที่รูปตัวอักษรหนึ่งๆ ถูกแยกออกเป็นส่วนๆ ทำให้เวลาตัดตัวอักษรจากตัวเดียวจะได้เป็นสองตัว ซึ่งก็ต้องหาวิธีการเฉพาะสำหรับมาจัดการอีกเช่นกัน

(a) (b)

ภาพที่ 2.4 (a) เป็นภาพที่ไม่ได้ทำการตัดแบ่งพื้นที่ใช้งาน (b) เป็นภาพที่ทำการตัดแบ่งพื้นที่ใช้งาน

1.4 การสกัดลักษณะสำคัญ (Feature Extraction)

การสกัดลักษณะสำคัญเป็นอีกขบวนการหนึ่งที่สำคัญมาก ตำราส่วนใหญ่จะแยกส่วนนี้ออกจากการประมวลผลเบื้องต้น คือจะอยู่ระหว่างขั้นตอนการประมวลผลเบื้องต้นกับขั้นตอนการรู้จำ แต่ในที่นี้ขอรวมไว้ในตอนเดียวกัน การสกัดลักษณะสำคัญเป็นการดึงเอาโครงสร้างพื้นฐานที่สำคัญของตัวอักษรนั้นออกมา โดยโครงสร้างพื้นฐานที่ว่าจะต้องมีการกำหนดไว้ก่อนว่าจะมีอะไรบ้าง มีการนิยามอย่างไร ตัวอย่างเช่น สำหรับภาษาไทยเราอาจกำหนดว่าตัวอักษรภาษาไทยทั้งหมดประกอบด้วยโครงสร้างพื้นฐานคือ เส้นตรง (แนวตั้ง/นอน) เส้นเอียง หัว (วงกลม) ส่วนโค้ง ส่วนเว้า จุดแตกกิ่ง จุดตัด เป็นต้น เมื่อเราสามารถแยกเอาองค์ประกอบของตัวอักษรแต่ละตัวออกมาได้แล้ว จากนั้นเราก็นำเสนอรูปภาพของตัวอักษรนั้นในรูปแบบของรายการขององค์ประกอบพื้นฐานต่างๆ แทน ซึ่งจะถูกส่งต่อเป็นอินพุตสำหรับขั้นตอนการรู้จำต่อไป



รูปที่ 2.5 การสกัดตัวอักษร

2.การรู้จำ (Recognition)

ขั้นตอนนี้ถือเป็นหัวใจของระบบรู้จำและวิเคราะห์ตัวอักษร เพราะเป็นส่วนที่จะตัดสินว่ารูปตัวอักษรที่ส่งเข้าไปเป็นตัวอักษรอะไร เช่นเดียวกับส่วนอื่นๆ ที่มีวิธีการหลากหลายซึ่งนำมาใช้เพื่อให้ได้ผลการทำงานที่ดีที่สุด เทคนิคใหม่ๆ ได้ถูกพัฒนาขึ้นมาอย่างต่อเนื่องเพื่อรองรับกับปัญหาที่เกิดจากเอกสารที่มีความสลับซับซ้อนมากขึ้น อย่างไรก็ตามเราพอที่จะจัดแบ่งเทคนิคเหล่านี้ออกเป็นกลุ่มตามแนวทางหลักที่ใช้ในการแก้ปัญหา ถึงแม้บ่อยครั้งที่พบว่ามีความคาบเกี่ยวกันของเทคนิคที่นำมาใช้ระหว่างกลุ่มที่ว่านี้ ทั้งนี้เพราะแต่ละแนวทางก็มีข้อดีข้อเสียที่แตกต่างกัน จึงมีความพยายามที่จะรวมเอาเทคนิคเหล่านี้มาใช้ร่วมกันเพื่อเพิ่มความสามารถของระบบเท่าที่จะเป็นไปได้ ดังนั้นการแบ่งกลุ่มในที่นี้ เป็นการแบ่งที่เน้นความชัดเจนในแง่ของขอบเขตทางทฤษฎีเป็นหลัก โดยแนวทางการรู้จำสามารถแบ่งออกได้เป็น 4 กลุ่ม ดังนี้

2.1 วิธีทางการเข้าคู่รูปแบบ (Template Matching)

วิธีการเข้าคู่รูปแบบเป็นวิธีการแรกๆ ที่มาใช้ในการรู้จำตัวอักษร หลักการโดยทั่วไปคือจะต้องมีรูปแบบ (template) ที่สร้างขึ้นมาสำหรับอ่านตัวอักษร โดยมีการกำหนดเป็นเอกลักษณ์ที่สำคัญของตัวอักษรแต่ละตัว เพื่อใช้แยกแยะความแตกต่างระหว่างตัวอักษรแต่ละตัว เวลาทำงานก็ให้นำรูปภาพที่ต้องการอ่านไปเปรียบเทียบกับรูปแบบ(template) เพื่อวัดความคล้ายคลึงกันของภาพกับตัวแบบ จากนั้นก็ระบุว่าเป็นตัวอักษรอะไร โดยใช้ค่าฮิสโตแกรมแมทชิ่งในการตัดสิน วิธีการนี้จะค่อนข้างอ่อนไหวต่อข้อมูลแทรกซ้อน ขนาด และการเอียงของตัวอักษร จึงจำเป็นต้องมีขั้นตอนการปรับแต่งข้อมูลที่ดี นอกจากนั้นขั้นตอนการเปรียบเทียบก็ไม่ใช่ว่าสามารถเทียบกันแบบจุดต่อจุดได้ เพราะในทางปฏิบัติตัวอักษรที่ส่งเข้าสามารถมีความแปรปรวนได้หลายรูปแบบ ดังนั้นวิธีการเทียบก็ต้องมีประสิทธิภาพเพียงพอที่จะรองรับกับปัญหาดังกล่าวได้

2.2 วิธีทางสถิติ (Statistical Approach)

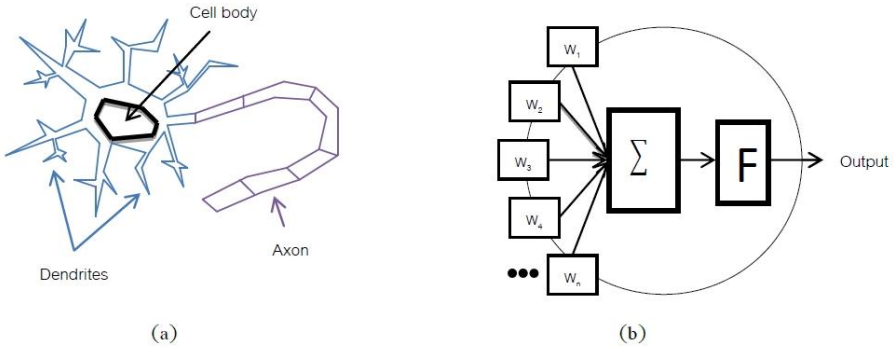
วิธีทางสถิติเป็นวิธีการที่ใช้หลักการทางสถิติ โดยนำค่าความน่าจะเป็น และ/หรือ ฟังก์ชันการแจกแจงความน่าจะเป็นมาใช้ในการตัดสินใจ รูปภาพอินพุตที่ได้มาจากขั้นตอนการสกัดลักษณะสำคัญ จะถูกส่งเข้าไปในส่วนการรู้จำเฉพาะของแต่ละตัวอักษร ซึ่งได้ผลลัพธ์ออกมาเป็นค่าความน่าจะเป็นที่อินพุตเป็นตัวอักษรใด เมื่ออินพุตได้ผ่านส่วนการรู้จำครบทุกตัวแล้ว ก็นำเอาผลลัพธ์ที่ได้ทั้งหมดมาเปรียบเทียบกันว่าได้ค่าความน่าจะเป็นของตัวอักษรใดมากที่สุด ผลลัพธ์จะออกเป็นตัวอักษรนั้น

2.3 วิธีการวิเคราะห์ทางโครงสร้าง (Structural Analysis)

วิธีการวิเคราะห์ทางโครงสร้างคือการวิเคราะห์โครงสร้างตัวอักษร โดยถือว่าตัวอักษรทุกตัวประกอบด้วยองค์ประกอบพื้นฐาน ซึ่งได้มาจากการสกัดลักษณะสำคัญ เช่นเดียวกันกับวิธีการทางสถิติ ต่างกันตรงที่ลักษณะสำคัญ ที่ส่งมาให้กับขั้นตอนการรู้จำแบบการวิเคราะห์ทางโครงสร้างนี้ มักจะใช้เป็นชื่อหรือค่าที่บอกว่าลักษณะโครงสร้างสำคัญนั้นเป็นอะไร เช่น เส้นตรง วงกลม เป็นต้น แทนที่จะเป็นค่าจำนวนจริง ในขั้นตอนการรู้จำลักษณะสำคัญทั้งหลายที่ประกอบเป็นตัวอักษรนั้น จะถูกส่งเข้าไปให้กับส่วนที่ตรวจวิเคราะห์กฎการเขียนตัวอักษร เช่น ฟอร์มอลแกรมมาแมชชีน (formal grammar machine) โครงสร้างกราฟ หรือโครงสร้างต้นไม้เป็นต้น เพื่อระบุว่าเป็นตัวอะไร ซึ่งจะตัดสินโดยการดูที่รูปแบบการเชื่อมต่อขององค์ประกอบต่างๆ เข้าเป็นตัวอักษรนั้น

2.4 วิธีทางโครงข่ายประสาทเทียม (Neural Network)

วิธีทางโครงข่ายประสาทเทียม (Neural Network) เป็นวิธีการที่เป็นที่นิยมใช้ เนื่องจากเป็นวิธีการที่อาศัยการทำงานที่จำลองมาจากการทำงานของสมองมนุษย์ โดยอาศัยการรู้จำของต้นแบบตัวอักษรหลายรูปแบบในการจดจำ เพื่อที่เมื่อนำข้อมูลอินพุตเข้ามา โปรแกรมสามารถที่จะวิเคราะห์และตัดสินใจได้ว่าเป็นตัวอักษรตัวใด ซึ่งต้องอาศัยวิธีการสกัดลักษณะสำคัญเพื่อนำค่าผลลัพธ์ที่ได้มาวิเคราะห์ในขั้นตอนนี้ เครือข่ายใยประสาท (Neural Network) ซึ่งเป็นโมเดลทางคณิตศาสตร์ สำหรับการประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อจำลองการทำงานของสมองมนุษย์ดังแสดงในภาพที่ 2.3 โดยอาศัยการรู้จำของต้นแบบตัวอักษรหลายรูปแบบในการจดจำ



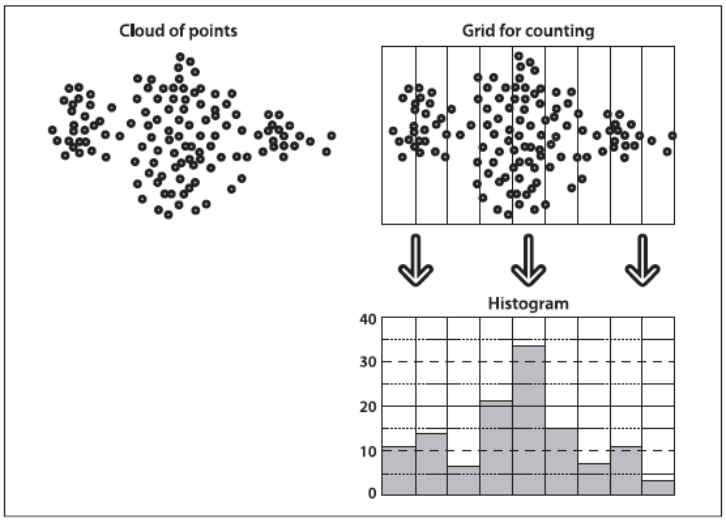
ภาพที่ 2.3 (a) Model ของ Neural ในสมองมนุษย์, (b) Model ของ Neural ในคอมพิวเตอร์
(ที่มา: https://app.enit.kku.ac.th/mis/administrator/doc_upload/20120312103907.pdf)

3.ขบวนการประมวลผลขั้นปลาย (Post-Processing)

หลังจากที่ผ่านขั้นตอนการรู้จำแล้ว รูปตัวอักษรที่ถูกส่งเข้าไปจะได้ผลลัพธ์ออกมาเป็นรหัสตัวอักษร ขั้นตอนนี้เป็นการนำรหัสตัวอักษรมาวิเคราะห์ในโปรแกรมว่ารหัสที่ได้คือตัวอักษรอะไร

Histogram Matching

Histogram Matching เป็นฟังก์ชั่นที่แสดงว่าในแต่ช่องมีจำนวนจุดภาพ (Pixel) ในภาพนี้มีจำนวนเท่าไร เมื่อนำมาเขียนเป็นกราฟ แกนนอนคือช่องของจุดภาพและแกนยืนคือจำนวน พิกเซล (number of Pixels) ฮิสโตแกรมเป็นการนำเสนอภาพในลักษณะแบบรวมๆ ดังรูปภาพที่ 2.4

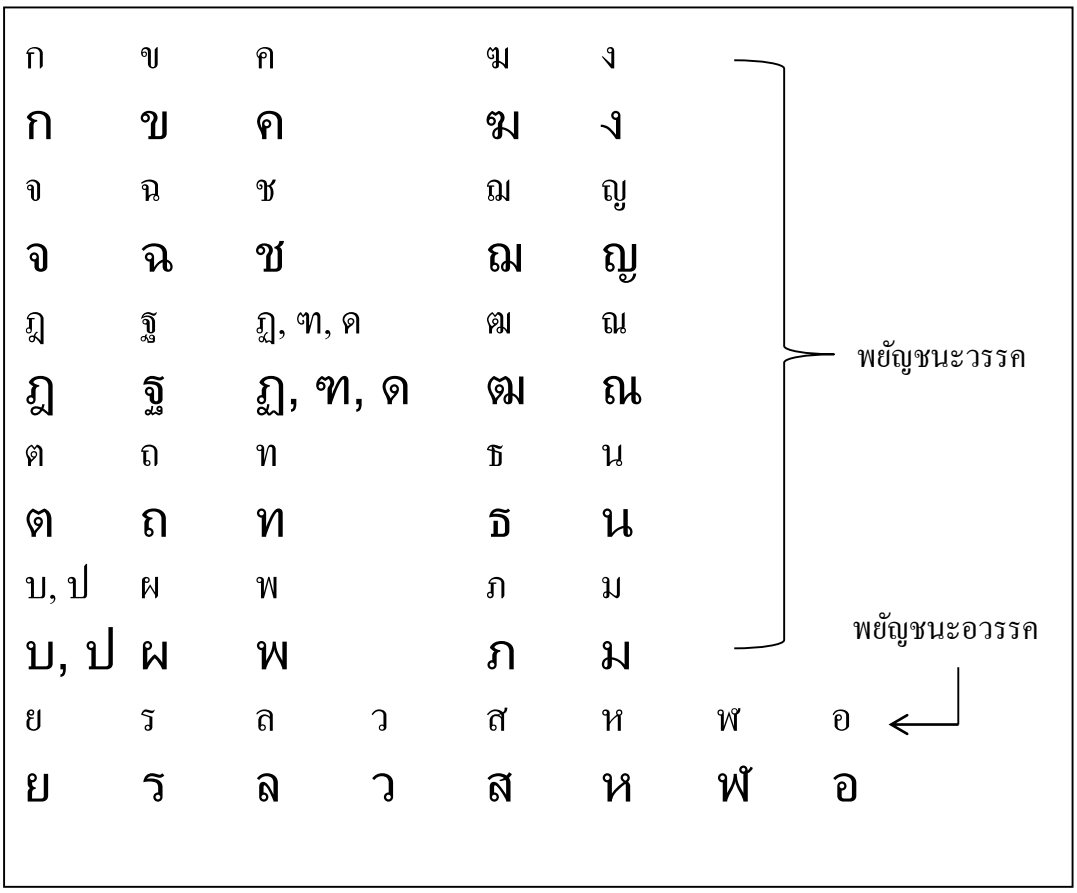


รูปภาพที่ 2.4 ฮิสโตแกรมของกลุ่มจุดภาพ(บนซ้าย) ที่แบ่งออกเป็น 9 ช่อง(บนขวา) และแสดงกราฟฮิสโตแกรม 1 มิติ(ล่างขวา)

(ที่มา : Gary Bradski and Arian Kaebler. 1993 : 13-15)

พยัญชนะภาษาล้านนา

พยัญชนะล้านนาจัดตามกลุ่มพยัญชนะวรรคตามพยัญชนะภาษาบาลี แบ่งออกเป็น 5 วรรค วรรคละ 5 ตัว เรียกว่า “พยัญชนะวรรค” หรือ “พยัญชนะในวรรค” อีก 8 ตัวไม่จัดอยู่ในวรรค เรียกว่า “พยัญชนะอวรรค” หรือ “พยัญชนะนอกวรรค” หรือ “พยัญชนะเศษวรรค” ส่วนการอ่านออกเสียงเรียกพยัญชนะทั้งหมดนั้น จะเรียกว่า“ตั๋ว”เช่น ตั๋ว กะ/ก/ ตั๋ว ขะ/ข/ ตั๋ว จะ/จ/ ดังรูปที่ 2.5



รูปที่ 2.5 พยัญชนะที่ใช้ในการเขียนภาษาล้านนาซึ่งจัดเรียงตามลำดับ

ผลงานวิจัยที่เกี่ยวข้อง

การพัฒนาโปรแกรม OCR สำหรับการอ่านเลขบนมิเตอร์ การพัฒนาโปรแกรม OCR สำหรับการอ่านตัวเลขบนมิเตอร์ เป็นการระบุตัวเลขการใช้น้ำหรือไฟฟ้า โดยใช้การประมวลผลรูปภาพ (Image Processing) โครงการพัฒนาบนพื้นฐานภาษา C++ และ OpenCV (Open Source Computer Vision) ซึ่งเป็น library ของ Programming Functions ในการพัฒนา ขั้นแรกใช้เทคนิคที่เหมาะสมคือการทำ ให้ภาพถ่ายของเลขบนมิเตอร์น้ำหรือไฟฟ้าดีขึ้นเพื่อง่ายต่อการนำ ภาพเข้าสู่ส่วนของ Template Matching เพื่อดึงตัวเลขจากภาพถ่ายของมิเตอร์น้ำหรือไฟฟ้าออกมาแปลเป็น Text File ซึ่งโปรแกรมสามารถที่จะระบุเลขของมิเตอร์น้ำหรือไฟฟ้าได้ถูกต้อง 90 % เครื่องจ่ายของเหลว

การแยกตัวอักษรธรรมอีสานบนหนังสือใบลานด้วยกระบวนการประมวลผลภาพ หนังสือใบลานเป็นมรดกทางวัฒนธรรมและบันทึกแห่งภูมิปัญญาของบรรพชน ที่ควรจะได้รับการจัดทำให้อยู่ในรูปดิจิตอล เพื่อเป็นประโยชน์แก่การศึกษาค้นคว้าของอนุชนรุ่นหลัง วิทยานิพนธ์ฉบับนี้นำเสนอการแยกตัวอักษรธรรมอีสานบนหนังสือใบลานด้วยกระบวนการประมวลผลภาพ โดยมีเป้าหมายว่าจะนำข้อมูลผลลัพธ์ไปใช้ต่อในกระบวนการรู้จำตัวอักษรต่อไป ข้อมูลนำเข้าเป็นภาพถ่ายด้วยกล้องดิจิตอลของหนังสือใบลาน ที่จดจารด้วยภาษาธรรมอีสาน โดยนำเสนอให้ทำการปรับปรุงคุณภาพของภาพด้วยการปรับปรุงความเปรียบเทียบต่าง (Contrast Improvement) ด้วยการตัดค่าทางปลายค่าสูงส่วนที่ไม่กระทบต่อรายละเอียดสำคัญทิ้ง แล้วจึงกระจายค่าฮิสโตแกรม จากนั้นทำการแยกตัวอักษรออกจากพื้นหลัง (Segmentation) โดยเปรียบเทียบสีเทคนิคได้แก่การทำเทรชโฮลด์แบบทั้งภาพ (Global Thresholding) การหาขอบ (Edge Detection) การติดตามขอบวัตถุ (Contour Following) และการทำเทรชโฮลด์แบบเฉพาะที่ (Local Thresholding) ซึ่งจากการทดลองกับภาพถ่ายใบลานจำนวน 10 ภาพ พบว่าการทำเทรชโฮลด์แบบเฉพาะที่ให้ผลดีกว่าวิธีการแยกตัวอักษรโดยการทำเรชโฮลด์แบบทั้งภาพ การหาขอบ และการติดตามขอบวัตถุทุกกรณี จากนั้นได้ทำการทดลองแยกตัวอักษรเดี่ยวๆ โดยทำการหากลุ่มของพิกเซลที่เชื่อมโยงต่อเนื่องกัน (Connected Component) ซึ่งถือว่าแต่ละกลุ่มเป็นวัตถุแต่ละชิ้น แต่วัตถุที่ได้เหล่านั้นมีทั้งส่วนที่เป็นลายเส้นตัวอักษรจริง และส่วนที่เป็นสัญญาณรบกวน จึงต้องทำการคำนวณหาคุณสมบัติต่างๆของวัตถุเหล่านั้นเพื่อขจัดส่วนที่เป็นสัญญาณลบกวนออก ก็จะเหลือส่วนที่เป็นลายเส้นซึ่งได้แยกออกจากกันแล้ว นำตัวอักษรเหล่านั้นให้ผู้เชี่ยวชาญสองท่านตรวจสอบ ผลการทดสอบเบื้องต้นมีความถูกต้องในระดับ 67.41 %