การแยกภาพตัวอักษรลายมือเขียนภาษาไทยแบบอัตโนมัติ โดยใช้การวิเคราะห์ถดถอยเชิงเส้นและการเรียนรู้ด้วยต้นไม้ตัดสินใจ

Automatic Thai Handwritten Characters Segmentation Based on Linear Regression Analysis and Decision Tree Learning

วิเชษฐ์รจน์ เอี่ยมสำอางค์* และรัชฎา คงคะจันทร์

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Wichedroat Iamsamang* and Rachada Kongkachandra

Department of Computer Science, Faculty of Science and Technology, Thammasat University,

Rangsit Centre, Klong Nueng, Khlong Luang, Pathum Thani 12120

บทคัดย่อ

บทความฉบับนี้นำเสนอหลักการสำหรับแยกภาพตัวอักษรลายมือเขียนที่อยู่ติดกันแบบสัมผัสในเอกสารภาพตัวอักษรออกจากกัน ซึ่งเป็นกระบวนการเตรียมพร้อมสำหรับการรู้จำลายมือเขียน เนื่องจากลักษณะของการเขียนภาษาไทยมีความแตกต่างจากภาษาอังกฤษ ซึ่งสามารถแบ่งออกได้เป็น 4 ระดับ โดยสามารถติดกันได้ในระดับเดียวกันและข้ามระดับทั้งในแนวนอนและแนวตั้ง หลักการที่ใช้ในบทความนี้ประกอบด้วยการรับภาพเอกสารลายมือเขียนมาคัดแยกให้เป็นตัวอักษรเดี่ยวและตัวอักษรติดกัน จากนั้นจะวิเคราะห์ตัวอักษรด้วยคุณลักษณะต่าง ๆ ของตัวอักษรไทย เพื่อแยกตัวอักษรที่ติดกันในแนวนอนและแนวตั้ง โดยมีการใช้เส้นการวิเคราะห์การถดถอยสำหรับตัดแบ่งระดับพยัญชนะกับสระ ขั้นตอนนี้ทำการตัดแบ่งตัวอักษรก่อนการรู้จำตัวอักษรตามหลักของการรู้จำตัวอักษรไทย ผลการทดลองพบว่าความถูกต้องของการแยกตัวอักษรลายมือเขียนภาษาไทยเป็นร้อยละ 90.44

คำสำคัญ : การรู้จำตัวอักษร, การวิเคราะห์องค์ประกอบหลัก, ต้นไม้ตัดสินใจ, เส้นการวิเคราะห์การถดถอย, เส้นอ้างอิงกรอบภาพ

Abstract

This paper presents an approach to analyze and segment Thai handwritten characters that are touched with the adjacent characters. In Thai handwriting system, the characters are displayed in four-levels and they can be touched both in vertical and horizontal axis. In this paper, the linear

***Corresponding author:** wichedroat.i@gmail.com*

regression line is used to segment characters in horizontal line. For segmenting characters in vertical axis, Thai character's attributes are learnt for building a classification model i.e. decision tree. The experimental result yields the average accuracy as 90.44 %.

Keywords: optical character recognition (OCR), principal component analysis (PCA), decision tree, linear regression, bounding box

1. ความเป็นมาและความสำคัญของปัญหา

การรู้จำภาพตัวอักษรลายมือเขียน (handwriting recognition) เป็นงานวิจัยขั้นสูงต่อจากการรู้จำภาพตัวอักษรพิมพ์ (optical character recognition) โดยมีวัตถุประสงค์เพื่อช่วยให้ผู้ใช้สามารถจัดเก็บเอกสารที่มีอยู่เข้าสู่ระบบคอมพิวเตอร์ได้สะดวกขึ้น ซึ่งนอกจากจะเป็นเอกสารพิมพ์แล้ว ยังมีเอกสารจำนวนมากที่อยู่ในรูปของเอกสารลายมือเขียน เช่น ข้อมูลที่ผู้ใช้กรอกใบสมัครงาน รายงานการประชุม ข้อความจากการบรรยาย เป็นต้น งานวิจัยทางด้านการรู้จำภาพตัวอักษรลายมือเขียนในภาษาไทย (Phokharatkul and Kimpan, 2002) เสนอวิธีการรู้จำตัวอักษรลายมือเขียนในภาษาไทยโดยใช้ตัวอธิบายรูปร่างแบบฟูริเยียร์ (Fourier descriptor) เป็นตัวอธิบายลักษณะเด่นของตัวอักษร จากนั้นฝึกฝนและรู้จำโดยใช้โครงข่ายประสาทเทียมแบบหลายชั้น ซึ่งค่าน้ำหนักของโหนดถูกพิจารณาโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม (genetic algorithm) ผลการรู้จำเมื่อทดลองกับตัวอักษรลายมือเขียนไทยแบบบรรจง จำนวน 13,500 ตัวอักษร จากผู้ทดลอง 60คน เท่ากับร้อยละ 99.12 แต่อย่างไรก็ตามตัวอักษรลายมือเขียนที่ใช้เป็นแบบเขียนบรรจง ซึ่งไม่สอดคล้องกับการใช้งานจริง ต่อมา Methasate และ Sae-tang (2004) เสนอวิธีการจัดกลุ่มตัวอักษรลายมือเขียนด้วยการพิจารณาโครงสร้างของตัวอักษรที่คล้ายกัน ลักษณะที่นำมาพิจารณา ได้แก่ เส้นตัวอักษรในแนวตั้ง (vertical stroke) และการกระจายของพิกเซลภาพตัวอักษร (pixel distribution) จากนั้นนำมาฝึกฝนและรู้จำโดยใช้โครงข่ายประสาทเทียมแบบแพร่กระจายกลับ (back propagation neural network) ผลการวิจัยสามารถแบ่งกลุ่มตัวอักษรออกได้เป็น 21 กลุ่ม มีระดับความถูกต้องที่ร้อยละ 97.60 นอกจากนี้ Nopsuwanchai และคณะ (2006) ได้ใช้ส่วนประกอบสำคัญ (principal component) ของตัวอักษรแต่ละตัว ร่วมด้วยรูปแบบต่าง ๆ ของตัวอักษร ได้แก่ ภาพกลับขั้วของตัวอักษร (polar transformed image) และภาพตัวอักษรหมุนขวา 90 องศา เป็นลักษณะเด่นของตัวอักษร จากนั้นนำไปฝึกฝนและรู้จำโดยใช้ แบบจำลองฮิดเด้นมาร์คอฟ โดยเปรียบเทียบระหว่างคลังข้อมูล 2 ชุด ได้ผลการรู้จำถูกต้องร้อยละ 95.98 และ 95.13 ตามลำดับ

แต่อย่างไรก็ตาม ถึงแม้งานวิจัยที่ผ่านมาจะให้ความถูกต้องในการรู้จำสูง แต่ภาพตัวอักษรลายมือเขียนที่นำมาทดสอบจะมีการแบ่งช่องว่างระหว่างตัวอักษรที่ชัดเจน ไม่มีการสัมผัส และ ทับซ้อนกัน ซึ่งการนำเอาวิธีการเหล่านี้มาใช้งานจริงอาจจะไม่สามารถรู้จำภาพลายมือเขียนที่มีการสัมผัส (ขวา) และทับซ้อนกัน (ซ้าย) ได้อย่างถูกต้อง รูปที่ 1 แสดงถึงตัวอย่างของลายมือเขียนที่อาจทำให้เกิดข้อผิดพลาดในการรู้จำ

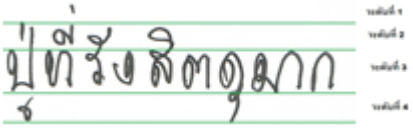
มีนักวิจัยหลายท่านเห็นว่า ถ้ามีการแยกตัวอักษรที่สัมผัสกันได้อย่างถูกต้อง เมื่อนำไปรู้จำแล้วจะได้ผลของการรู้จำที่สูงขึ้น (Chatchinarat, 2009)

รูปที่ 1 รูปลักษณะตัวอักษรติดแบบทับซ้อน (ซ้าย) และติดแบบสัมผัส (ขวา)

โดยเทคนิคของแอ่งน้ำ ซึ่งเป็นการเติมน้ำลงบนตัวอักษร โดยพบว่าตัวอักษรที่เชื่อมติดกันจะมีแอ่งน้ำที่เกิดขึ้นระหว่างตัวอักษรคู่นั้นซึ่งสามารถเป็นจุดตัดได้ สามารถแยกได้ถูกต้องร้อยละ 87.60 ซึ่งข้อผิดพลาดเกิดจากตัวอักษรที่ไม่สามารถหาแอ่งน้ำได้ และตัวอักษรที่มีจำนวนแอ่งน้ำเท่ากัน นอกจากนี้ขนาดของตัวอักษรแต่ละตัวจะต้องมีขนาดเท่ากัน ดังนั้นงานวิจัยนี้จึงเสนอการแยกตัวอักษรลายมือเขียนที่เขียนในลักษณะธรรมชาติคือ มีการสัมผัสกันทั้งในแนวนอนและแนวตั้ง และมีขนาดของตัวอักษรแตกต่างกันได้ โดยใช้การวิเคราะห์ถดถอยเชิงเส้น (linear regression analysis) สำหรับวิเคราะห์หากลุ่มของตัวอักษรที่อยู่ติดกันในแนวตั้ง และใช้ความกว้างมัธยฐานของตัวอักษร (median of character width) ในการวิเคราะห์กลุ่มอักษรที่อยู่ติดกันในแนวนอน จากนั้นแยกตัวอักษรโดยใช้แบบจำลองต้นไม้ตัดสินใจที่ได้มาจากการฝึกฝน

2. หลักการที่เกี่ยวข้อง

เนื่องจากระดับของการเขียนในภาษาไทยมีอยู่ 4 ระดับ ดังแสดงในรูปที่ 2 คือ ระดับวรรณยุกต์ ระดับสระบน ระดับพยัญชนะ/สระกลาง และระดับสระล่าง ดังนั้นโอกาสที่ตัวอักษรจะเขียนสัมผัสกันสามารถเป็นไปได้ทั้งแนวตั้ง เช่น ระหว่างระดับที่ 2 กับระดับที่ 3 ได้แก่ “รั” และ “สิ” และในแนวนอน เช่น “มา” ดังนั้นการแยกตัวอักษรที่สัมผัสกัน เราควรทราบก่อนว่าแนวพยัญชนะกลางอยู่ตำแหน่งใด จากนั้นมาพิจารณาว่าส่วนของตัวอักษรตัวใดสัมผัสกันบ้าง โดยใช้หลักการต่อไปนี้



รูปที่ 2 ระดับการเขียนในภาษาไทย

2.1 การวิเคราะห์ตัวอักษรที่สัมผัสกันในแนวตั้ง

เมื่อเรานำภาพตัวอักษรมาหากรอบของวัตถุเรียบร้อยแล้ว เราจะใช้ตำแหน่งมุมบนซ้ายสุดของกรอบตัวอักษรเป็นตัวแทนจุดสูงสุดของแต่ละส่วน ดังแสดงด้วยรูป “*” สีน้ำเงินในรูป 3ก จากนั้นลากเส้นตรงเชื่อมระหว่างจุดที่ได้ นำมาเข้าสมการการวิเคราะห์การถดถอยเชิงเส้น จากนั้นปรับความชันของเส้นที่ได้ให้เท่ากับ 0 จะได้เป็นเส้นชมพู ดังแสดงในรูป 3ข ซึ่งเส้นดังกล่าวเป็นเส้นบนของระดับกลาง ในทำนองเดียวกัน เราใช้ตำแหน่งมุมล่างขวาสุดของกรอบตัวอักษรเป็นตัวแทนจุดต่ำสุดของแต่ละส่วน ดังแสดงด้วยรูป “*” สีแดงในรูป 3ค แล้วลากเส้นเชื่อมจุด คำนวณหาเส้นตรงที่เป็นค่าเฉลี่ย แล้วปรับความชันของเส้นให้เท่ากับ 0 จะได้เส้นสีชมพู เป็นเส้นล่างของระดับกลาง ดังรูป 3ง

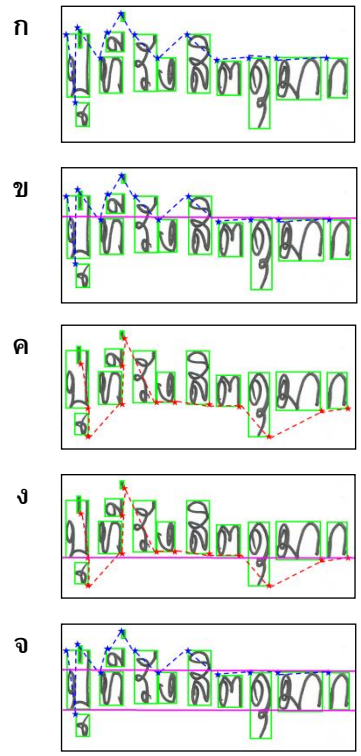
สมการการวิเคราะห์การถดถอยเชิงเส้น

$$y = bx + a \qquad \text{(สมการที่ 1)}$$

หาความสัมพันธ์ระหว่างตัวแปรทั้งสอง ซึ่งมีสูตรในการหา a และ b ดังนี้

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad \text{(สมการที่ 2)}$$

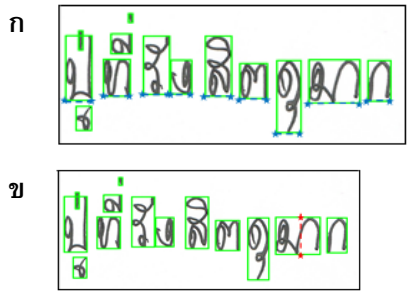
$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad \text{(สมการที่ 3)}$$



รูปที่ 3 การวิเคราะห์ตัวอักษรที่สัมผัสกันในแนวตั้ง

2.2 การวิเคราะห์ตัวอักษรที่สัมผัสกันในแนวนอน (โดยใช้ความกว้างมัธยฐานของตัวอักษร)

นำภาพตัวอักษรที่มีกรอบของวัตถุ เราจะใช้ตำแหน่งมุมล่างซ้ายสุดและขวาสุดของกรอบตัวอักษรเป็นตัวแทนความกว้างของแต่ละวัตถุ ดังแสดงด้วยรูป "*" สีน้ำเงินในรูป 4ก ได้ความกว้างของแต่ละวัตถุจากภาพข้อความ เก็บค่าของแต่ละวัตถุไว้ในอาร์เรย์ แล้วนำค่าที่ได้จากตัวอักษรแต่ละตัวมาเรียงกันแล้วนำเข้าสมการหาค่ามัธยฐาน (median) ของข้อมูลวัตถุบนภาพข้อความ จากนั้นนำค่ามัธยฐานที่ได้มาเทียบตัวอักษรแต่ละตัว ถ้าพบว่ามีความกว้างมากกว่าเท่าครึ่งหรือ 1½ ซึ่งเกินค่ามัธยฐานแสดงว่าวัตถุนั้นเป็นตัวอักษรที่ติดกันในแนวนอน จากนั้นทำการลากเส้นเชื่อมจุดตัดที่มีค่าเท่ากับค่ามัธยฐานของตัวอักษรดังแสดงด้วยรูป "*" สีแดงในรูป 4ข



รูปที่ 4 การวิเคราะห์ตัวอักษรที่สัมผัสกันในแนวนอน

การหาค่ามัธยฐานของข้อมูลที่มีจำนวนคี่ สมการหาตำแหน่งค่ามัธยฐาน

$= (n + 1) \div 2$ (สมการที่ 4)

การหาค่ามัธยฐานของข้อมูลที่มีจำนวนคู่ สมการหาค่าเฉลี่ยของข้อมูลที่อยู่ในตำแหน่งที่ $= n \div 2$ และ $(n + 1) \div 2$ (สมการที่ 5)

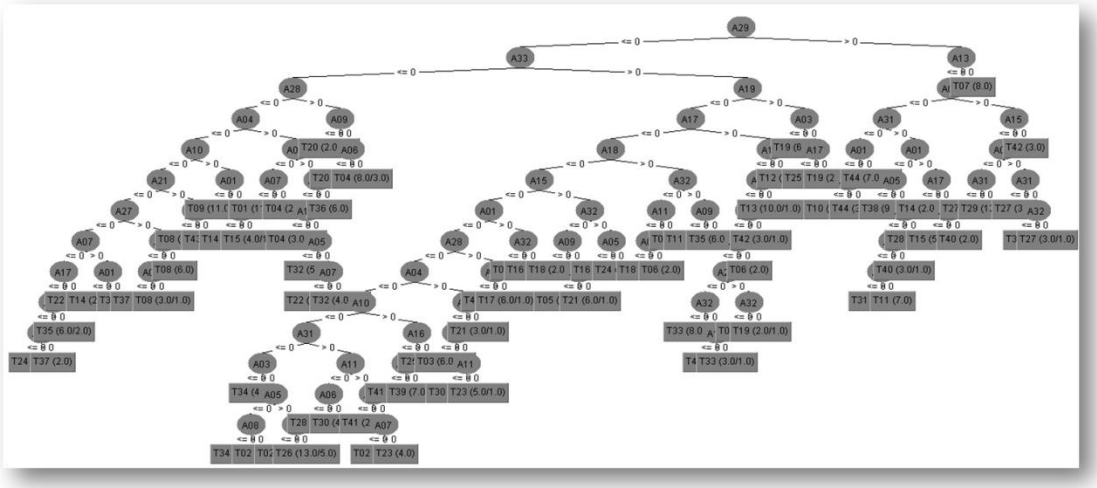
2.3 การสร้างแบบจำลองตัวอักษรด้วยต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (decision tree classifiers, J48) เป็นโครงสร้างต้นไม้ที่ประกอบไปด้วยโหนดตัดสินใจ (decision nodes) แล้วเชื่อมต่อกันด้วยสาขา (branches) โดยขยายจากโหนดราก (root node) มายังโหนดใบ (leaf nodes) ค่าคุณสมบัติ (attribute) ที่ใช้สำหรับเปรียบเทียบเงื่อนไขจะอยู่ในโหนดตัดสินใจ ซึ่งผลลัพธ์การเปรียบเทียบจะอยู่ในสาขา (Branches) และในแต่ละสาขาจะนำไปสู่โหนดตัดสินใจอื่น หรือโหนดใบ ซึ่งโหนดใบจะเก็บผลลัพธ์ของการจำแนกไว้ (Quinlan,1986; Quinlan, 1990) โดยกำหนดให้ภาพตัวอักษร ก-ฮ เป็นภาพจำลองที่ได้จากลายมือเขียนของกลุ่มข้อมูลตัวอย่าง ตัวอย่างละ 10 ตัวอักษร ตรวจเช็คคุณลักษณะเด่นตัวอักษรภาษาไทยแต่ละตัว นำคุณลักษณะตัวอักษรแทนข้อมูลเวกเตอร์ (eigenvector) ทำการแปลงโครงสร้างเมตริกซ์

ข้อมูลไปเป็นเวกเตอร์แถว คำนวณหาไอเกนเวกเตอร์ที่สอดคล้องกันกับค่าไอเกน โดยนำผลรวมของคุณลักษณะต่าง ๆ ของตัวอักษรที่ซ้ำกันนำมาเรียงลำดับไอเกนเวกเตอร์ที่สอดคล้องกับค่าไอเกนจากมากไปน้อย แล้วคัดเอาเฉพาะค่าที่มีคุณลักษณะเด่น เข้าสมการหาค่า eigenvalue ดังตารางที่ 1จากนั้นแล้วนำเอาลักษณะเด่นของภาพ (feature extraction) ทุกตัวอักษรเพื่อเข้าโปรแกรม Weka สร้างต้นไม้ตัดสินใจของตัวอักษรลายมือเขียน ก-ฮ ดังรูปที่ 5 เพื่อทำการรู้จำของตัวอักษรต่อไป

ตารางที่ 1 การหาคุณลักษณะตัวอักษร "ฆ"

คุณลักษณะตัวอักษร ฆ	N Loop ซ้ายบน	N Loop ขวาบน	N Loop ขวาล่าง	ลูป 2	Loop ซ้อง 7	Loop ซ้อง 8	เปิดบน 2	เปิดล่าง 2	เปิดบนขวา	N Loop ซ้ายล่าง	เปิดล่างซ้าย	เปิดล่าง	Loop ซ้อง 4	หัวหน้าบน
code	A05	A06	A08	A10	A17	A18	A33	A31	A26	A07	A27	A04	A14	A30
รวม	10	10	10	10	10	10	10	9	8	7	6	2	1	1
ค่า Eigen vector	0.2688	0.2688	0.2688	0.2688	0.2688	0.2688	0.2688	0.24	0.215	0.188	0.16	0.05	0.027	0.027



รูปที่ 5 ผลลัพธ์ต้นไม้ตัดสินใจของตัวอักษรลายมือเขียน ก-ฮ

สมการหาค่า eigenvalue

$$\text{Eigenvalue} = \sqrt{[\sum(\text{ของน้ำหนักองค์ประกอบของแต่ละตัวแปรในองค์ประกอบนั้น})^2]}$$ สมการที่ 6

3. ภาพรวมของระบบและขั้นตอนการทำงาน

ภาพรวมของระบบและขั้นตอนการทำงานมีวิธีการดำเนินงานวิจัยดังนี้

3.1 รับข้อมูลภาพตัวอักษรเข้าสู่ระบบ

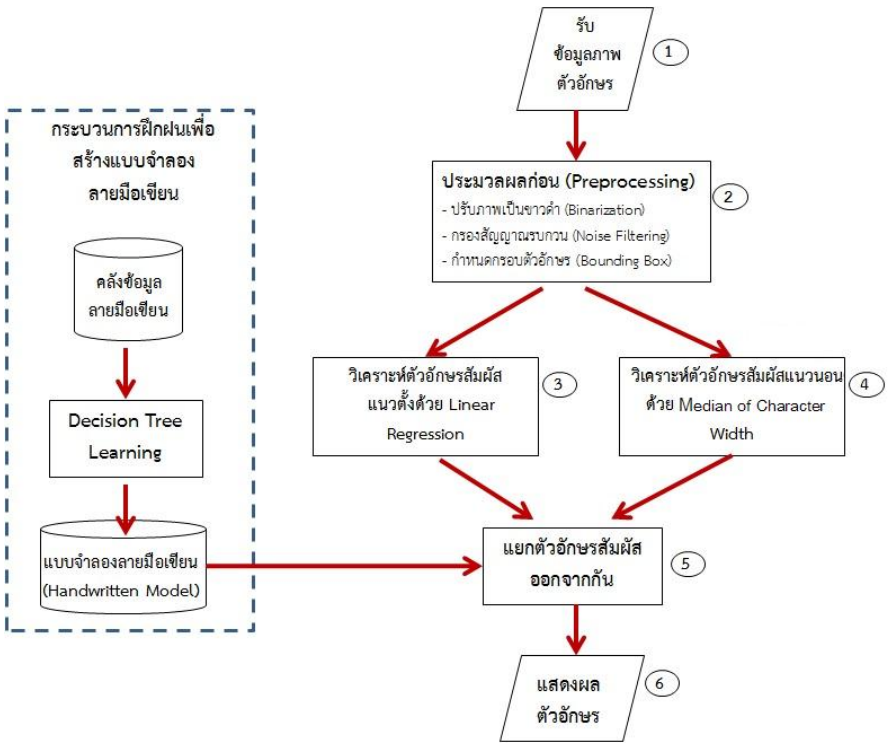
3.2 ประมวลผลก่อนโดยปรับภาพสีของตัวอักษรบนรูปภาพให้เป็นตัวอักษรสีขาวพื้นหลังสีดำ (binarization) นำมากรองสัญญาณรบกวนออกจากภาพขาวดำ (noise filtering) ทำการตีกรอบรอบแต่ละวัตถุบนรูปภาพ (bounding box)

3.3 วิเคราะห์ตัวอักษรที่สัมผัสกันในแนวตั้งด้วยเส้นจากการวิเคราะห์ถดถอยเชิงเส้นตามข้อ 2.1

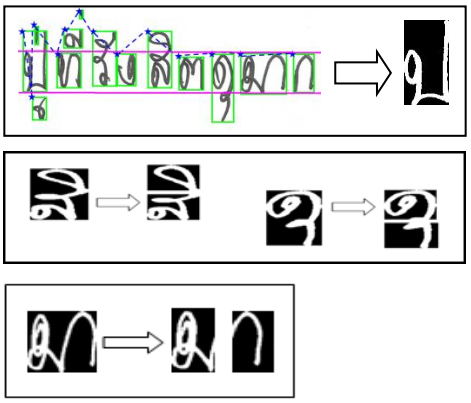
3.4 การวิเคราะห์ตัวอักษรที่สัมผัสกันในแนวนอนด้วยความกว้างมัธยฐานของตัวอักษร (median of character width) ตามข้อ 2.2

3.5 นำแบบจำลองลายมือเขียน (hand written model) มาแยกตัวอักษรที่สัมผัสออกจากกัน ด้วยการใช้กฎที่ได้จากต้นไม้ตัดสินใจ

3.6 แสดงผลตัวอักษร นำวัตถุบนภาพได้จากการ bounding box ไปทำการตรวจสอบระดับพยัญชนะด้วยเส้นจากการวิเคราะห์ถดถอยเชิงเส้นบนพยัญชนะและเส้นจากการวิเคราะห์ถดถอยเชิงเส้นล่างพยัญชนะ ถ้าวัตถุนั้นมีความสูงมากกว่าความสูงของเส้นจากการวิเคราะห์ถดถอยเชิงเส้นทั้งสองเส้นจะได้วัตถุตัวติดหรืออักษรเดี่ยวหางยาวดังรูปที่ 7ก ไม่ต้องนำไปตัด ส่วนอักษรที่สัมผัสกันในแนวตั้ง ทำการตัดด้วยเส้นจากการวิเคราะห์ถดถอยเชิงเส้นดังรูปที่ 7ข และตัวอักษรที่สัมผัสกันในแนวนอน ตัดด้วยความกว้างมัธยฐานของตัวอักษรจะได้อักษรแต่ละตัวออกมาดังรูปที่ 7ค



รูปที่ 6 ขั้นตอนการทำงานของระบบ



รูปที่ 7 แสดงผลตัวอักษร

4. ผลการวิจัยและวิเคราะห์

ในการนำตัวอักษรลายมือเขียนมาเทียบด้วยต้นไม้ตัดสินใจ ซึ่งมีความรู้จำน้อยกว่าตัวอักษรพิมพ์หรือตัวอักษรที่คัดตัวบรรจงทั้งหมด จากการวิเคราะห์พบว่าตัวอักษรลายมือเขียนบางตัวมีคุณลักษณะไม่ครบเหมือนตัวพิมพ์ เช่น การเขียน ร, ล, ย, พ, บ ไม่มีหัว หรือมีหางยาวเกินกำหนด จึงทำให้การรู้จำคลาดเคลื่อนลดลงได้

การนำตัวอักษรลายมือเขียนที่ตัดได้ตรวจสอบคุณลักษณะก่อนว่าตัวอักษรแต่ละตัวมีคุณลักษณะอะไรบ้าง เช่น “ม” เป็นลักษณะตัวอักษรลายมือเขียน (รูปที่ 1) มีคุณลักษณะเปิดบนขวา (รูปที่ 2) ตัวอักษรสั้น ไม่มีหัวแตก (รูปที่ 3) มีลูปช่องที่ 1, 4 และช่องที่ 7 (รูปที่ 4) เมื่อนำมาเทียบด้วยต้นไม้ตัดสินใจของตัวอักษรลายมือเขียน ก-ฮ ที่เตรียมไว้ ผลลัพธ์ที่ได้คือ T33 เท่ากับอักษร “ม” ดังรูปที่ 8ก ส่วนตัวอักษรที่ติดด้านหลังตัวอักษร “ม” นำตัวอักษรนั้นมาตรวจสอบคุณลักษณะด้วยต้นไม้ตัดสินใจตัวติดด้านหลังที่เตรียมไว้เพื่อให้ทราบว่ามีคุณลักษณะเหมือนตัวอักษรใด พบว่าเป็นลักษณะตัวอักษรลายมือเขียน (รูปที่ 1) มีคุณลักษณะเปิดล่าง (รูปที่ 2) ตัวอักษรสั้น ไม่มีหัวแตก (รูปที่ 3) ไม่มีลูป (รูปที่ 4) ผลลัพธ์ที่ได้คือ T46 เท่ากับอักษรสระอา (-า) ดังรูปที่ 8ข

ผลลัพธ์การทดลองตัดตัวอักษรลายมือเขียนที่ผ่านการสแกนมีจำนวนทั้งหมด 1,234 ตัว ใช้ Matlab 2010 ทำการแยกตัวอักษรได้ 1,116 ตัว คิดเป็นร้อยละ 90.44 ตามตารางที่ 2 ดังนี้

ตารางที่ 2 ผลลัพธ์การทดลองตัดตัวอักษรลายมือเขียนจำแนกตามพยัญชนะ สระ และวรรณยุกต์

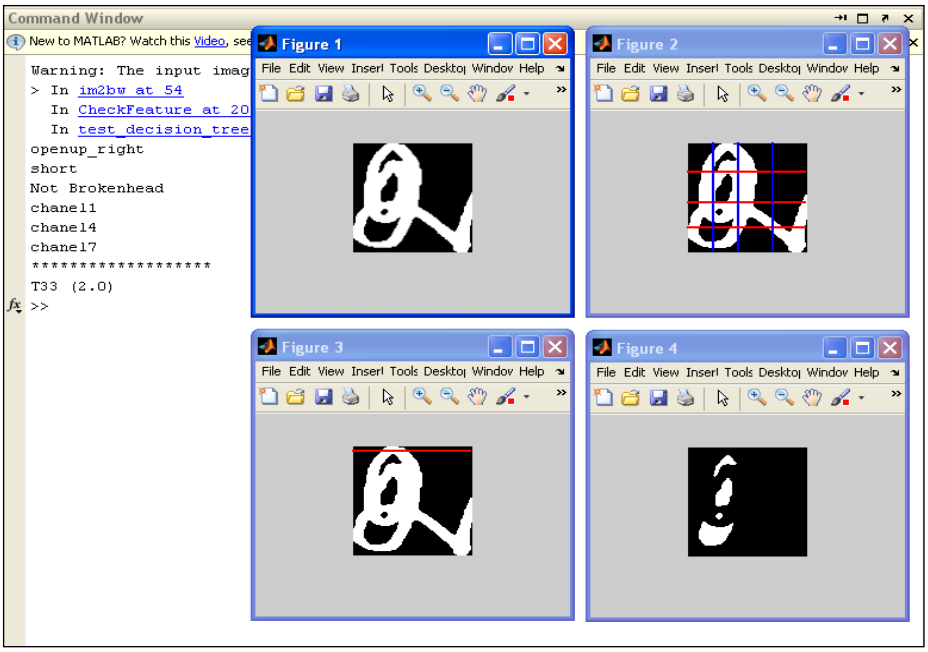
ตัวอักษร	จำนวนทั้งหมด	จำนวนที่แยกได้	คิดเป็นร้อยละ
พยัญชนะ	772	701	90.80
สระล่าง	59	54	91.53
สระข้าง	124	115	92.74
สระบน	225	197	87.56
วรรณยุกต์	54	49	90.74
รวม	1,234	1,116	90.44

5. สรุป

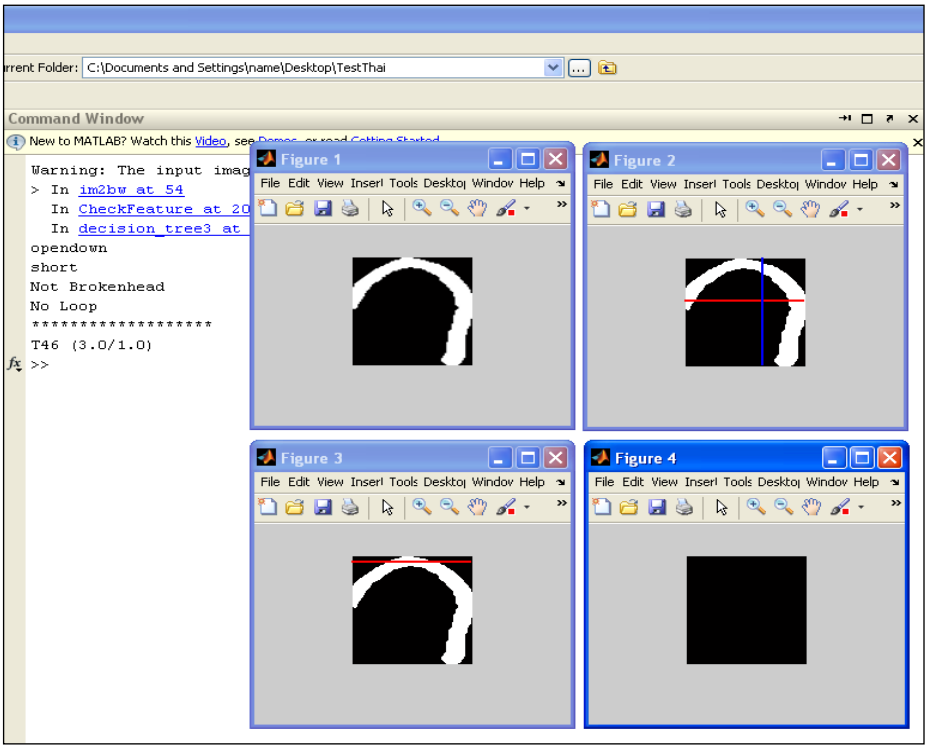
ในปัจจุบันข้อมูลมีหลากหลายรูปแบบ ซึ่งถ้าข้อมูลที่เขียนบนกระดาษ ถ้าใช้กำลังคนในการป้อนข้อมูลเข้าสู่ระบบจะใช้เวลานานและอาจเกิดการผิดพลาดจากการนำเข้าข้อมูลได้ แต่ถ้าเก็บเป็นรูปภาพ ระบบฐานข้อมูลก็จะต้องใช้พื้นที่ในการเก็บข้อมูลนั้นมาก จึงมีระบบการรู้จำภาพตัวอักษรเกิดขึ้นเพื่อความรวดเร็วและประหยัดพื้นที่ในการเก็บข้อมูล ซึ่งขั้นตอนของระบบการรู้จำภาพตัวอักษรจะมีหลายขั้นตอน เริ่มจากรับข้อความเป็นรูปภาพ ทำการตัดแยกตัวอักษรทั้งหมด แล้วนำมาทำการรู้จำตัวอักษร จากนั้นนำผลที่ได้จากการรู้จำมาเก็บลงระบบฐานข้อมูล ข้อมูลที่ได้ในการรู้จำตัวอักษรจะมีความถูกต้องก็ต่อเมื่อมีผลจากการแยกตัวอักษรที่ถูกต้อง ในการตัดแยกตัวอักษรนั้นมีความ

ยากในการตัดรูปแบบตัวเขียน เนื่องจากรูปแบบการเขียนตัวอักษรภาษาไทยมีทั้งตัวอักษรเดี่ยวและตัวอักษรติดระหว่างพยัญชนะกับสระบน และพยัญชนะกับสระล่าง

ก



ข



รูปที่ 8 ผลลัพธ์การตรวจสอบคุณลักษณะตัวอักษร

จากการทดลองการตัดในแนวตั้งจากการใช้เส้นจากการวิเคราะห์ถดถอยเชิงเส้นสามารถตัดแยกระหว่างพยัญชนะกับสระบนและสระล่างได้ แต่บางครั้งเส้นจากการวิเคราะห์ถดถอยเชิงเส้นที่ได้มีความผิดพลาด จึงต้องใช้ค่าเฉลี่ยของความสูงพยัญชนะเพื่อหาเส้นระดับพยัญชนะแทน ส่วนการตัดในแนวนอนด้วยค่าความกว้างมัธยฐานของตัวอักษรบนรูปภาพในการหาค่าความกว้างมัธยฐานของตัวอักษรนั้น จะต้องนำเอาเฉพาะความกว้างของพยัญชนะมาคำนวณเท่านั้น เนื่องจากคำในบางข้อความอาจจะมีสระข้างปะปนอยู่รวมด้วย ซึ่งสระข้างโดยปกติจะมีความกว้างน้อยกว่าพยัญชนะ เช่น สระเอ สระไอ ทำให้การคำนวณค่าเฉลี่ยความกว้างเกิดความผิดพลาดได้ ส่งผลให้การตัดผิดพลาดตามไปด้วย แต่เมื่อนำเอาเฉพาะพยัญชนะมาคำนวณทำให้ผลที่ได้จากการตัดในแนวนอนมีความถูกต้องแม่นยำเพิ่มมากขึ้น

6. เอกสารอ้างอิง

Chatchinarat, A., 2009, Thai handwritten segmentation using proportional invariant recognition technique, International Conference on Future Computer and Communication 2009.

Methasate, I., Sae-tang, S., 2004, The clustering technique for Thai handwritten recognition, IEEE IWFHR-9 2004.

Nopsuwanchai, R., Biem, A., Clocksin, W.F., 2006, Maximization of mutual information for offline Thai handwriting recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 28: 1347-1351.

Phokharatkul, P., Kimpan, C., 2002, Handwritten Thai character recognition using Fourier descriptors and genetic neural networks, Computational Intelligence 18: 270-93.

Quinlan, J.R., 1986, Induction of decision trees, Machine Learning 1: 81-106.

Quinlan, J.R., 1990, Decision trees and decision making IEEE transactions on systems, Man and Cybernatics 20: 339-46.