การค้นคืนคำจากภาพเอกสารภาษาล้านนาโดยการเปรียบเทียบภาพคำ

วิลาวัลย์ ยาทองคำ[a*] ณัฐษิมา สุรเดช[b] และ จีรยุทธ ไชยจารุวณิช[c]

Word Retrieval from Lanna Document Images by Synthetic Word Image Matching

Wilawan Yathongkhum[a*], Natsima Suradet[b] and Jeerayut Chaijaruwanich[c]

[a]โรงเรียนบ้านแม่เทย ต.แม่ตืน อ.ลี้ จ.ลำพูน 51110

[b]คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ตาก 41/1 ถ.พหลโยธิน ต.ไม้งาม อ.เมือง จ.ตาก 63000

[c]ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ 239 ถ.ห้วยแก้ว ต.สุเทพ อ.เมือง จ.เชียงใหม่ 50202

*Corresponding Author. E-mail address: y.wilawan@gmail.com

Received 30 April 2013; accepted 12 July 2013

บทคัดย่อ

บทความนี้นำเสนอการค้นคืนคำจากภาพเอกสารภาษาล้านนา ด้วยการนำคำที่ต้องการค้นหามาสร้างเป็นภาพคำภาษาล้านนา จากนั้นนำภาพคำดังกล่าวไปทำการเปรียบเทียบเพื่อหาขอบเขตของคำในภาพเอกสารด้วยวิธีการ 2 แบบ คือ เลื่อนหน้าต่างเพื่อเปรียบเทียบภาพทีละพิกเซล (PSTM) และพิจารณาจากคุณลักษณะความกว้างและความสูงของตัวอักษรภายในคำ (WFM) จากนั้นสกัดคุณลักษณะของภาพคำที่ได้และหาค่าความคล้ายคลึงโดยพิจารณาค่าความหนาแน่นของพิกเซลในหน้าต่างย่อย (S) เปรียบเทียบกับการหาค่าความคล้ายคลึงโดยพิจารณาจากค่าระยะห่างแบบยูคลิเดียน (E) จากการทดสอบประสิทธิภาพของวิธีการค้นคืน พบว่าการหาขอบเขตของคำในภาพโดยพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ และการหาค่าความคล้ายคลึงของคำโดยพิจารณาค่าระยะห่างแบบยูคลิเดียน (WFM+E) เหมาะสมสำหรับการค้นคืนคำในภาพเอกสารภาษาล้านนาโดยการเปรียบเทียบภาพ ซึ่งวิธีการดังกล่าวมีค่า $F_{measure}$ เท่ากับ 0.82 ค่าเฉลี่ยความแม่นยำ (average precision) เท่ากับ 0.83 และค่าเฉลี่ยความถูกต้อง (average recall) เท่ากับ 0.82 แสดงว่าวิธีการค้นคืนที่นำเสนอมีประสิทธิภาพอยู่ในระดับที่สามารถนำไปใช้งานได้จริง

คำสำคัญ: การค้นคืนคำ ภาพเอกสารภาษาล้านนา การสร้างภาพคำภาษาล้านนา

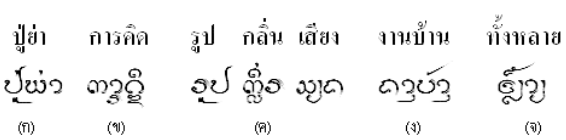
Abstract

In this paper, we propose a method for Lanna word retrieval from Lanna document images. The first step in the proposed method is creation of synthetic keywords image. Then we select the candidate words from the document image by using 2 methods: Pixel-based Sliding Window Template Matching (PSTM) and Word Feature Matching (WFM). Next, feature vector of each word image is extracted by window-based feature extraction. Finally, all relevant words are retrieved by comparing similarity between keyword and word image feature vector. The similarity between two feature vectors is evaluated by using 2 methods: Sub-window similarity (S) and Euclidean distance (E). The experimental results show that the method which combination of Word Feature Matching and Euclidean distance provides best performance. The Fmeasure value of this method reaches 82 percent, the average recall and precision are 82 percent and 83 percent, respectively. It is shown that the proposed method is feasible, valid, and effective for Lanna word image searching.

Keywords: word retrieval, lanna document images, synthetic lanna word image

บทนำ

คำในภาษาล้านนา[1] มีลักษณะการผสมคำคล้ายคลึงกับภาษาไทย คือ จะวางสระไว้รอบ ๆ พยัญชนะต้น และจะวางวรรณยุกต์ไว้บนสระหรือพยัญชนะต้น ดังรูปที่ 1(ก) แต่โครงสร้างคำที่มีตัวสะกดจะแตกต่างกับภาษาไทย คือ ตำแหน่งของตัวสะกดขึ้นอยู่กับสระที่ใช้ในคำนั้น เช่น ถ้าสระนั้นเป็นสระอะ, อา, อิ, อี, อึ, อือ, เอะ, เอ, แอะ, แอ, โอะ, โอ และเออ ตัวสะกดจะถูกวางไว้ด้านล่างของพยัญชนะต้น ดังรูปที่ 1(ข) เมื่อผสมกับสระอุ, อู, ออ, อัว, เอีย, และ เอือ หรือมีพยัญชนะควบกล้ำอยู่ข้างล่างพยัญชนะต้น ตัวสะกดจะถูกวางไว้ด้านหลังของพยัญชนะต้น ดังรูปที่ 1(ค) นอกจากนี้ยังมีการเปลี่ยนรูปของตัวสะกดที่เรียกว่า “ตัวสะกดหาง” ดังรูปที่ 1(ง) และมีการใช้งานพยัญชนะพิเศษและสัญลักษณ์พิเศษต่าง ๆ ดังรูปที่1 (จ)



รูปที่ 1 เปรียบเทียบลักษณะของคำในภาษาล้านนากับคำในภาษาไทย

เอกสารภาษาล้านนาในอดีตจะถูกบันทึกไว้ในสื่อต่าง ๆ เช่น ใบลาน พับสา หรือศิลาจารึก แต่เมื่อเวลาผ่านไปเอกสารเหล่านี้ก็ผุพังเสียหายไปตามกาลเวลา ทำให้ข้อมูลซึ่งถูกบันทึกไว้ไม่ว่าจะเป็นข้อมูลทางประวัติศาสตร์ ประเพณี ศิลปวัฒนธรรม พิธีกรรม หลักคำสอนทางพระพุทธศาสนา กฎหมาย วรรณกรรมพื้นบ้าน ตำรายาสมุนไพร

ตำราโหราศาสตร์ ฯลฯ เกิดความเสียหายและสูญหายไปด้วย เพื่อเป็นการอนุรักษ์และป้องกันข้อมูลอันมีค่าเหล่านั้น จึงได้มีการเก็บรักษาและบันทึกข้อมูลเหล่านั้นไว้ในสื่อรูปแบบอื่น[2] เช่น ภาพถ่าย หรือ หนังสืออิเล็กทรอนิกส์ เป็นต้น

ในปัจจุบันการค้นคืนสารสนเทศจากรูปภาพเอกสาร [3] สามารถแบ่งออกได้เป็น 2 แบบคือ การแปลงรูปภาพเอกสารให้เป็นข้อความในลักษณะของรหัสแอสกีโดยใช้การรู้จำภาษาเป็นเครื่องมือในการแปลง จากนั้นจึงทำการค้นคืนข้อความหรือสารสนเทศที่ต้องการ ส่วนในแบบที่ 2 เป็นการดำเนินการกับรูปภาพเอกสารโดยไม่ใช้เทคนิคการรู้จำภาษา ซึ่งจะพิจารณาจากลักษณะต่าง ๆ ที่มีอยู่ในรูปภาพเอกสารเป็นหลัก เช่น โครงสร้างของข้อความ รูปร่างของคำหรือตัวอักษร เป็นต้น แล้วใช้ลักษณะเฉพาะเหล่านั้นเพื่อค้นคืนข้อความหรือสารสนเทศที่ต้องการ

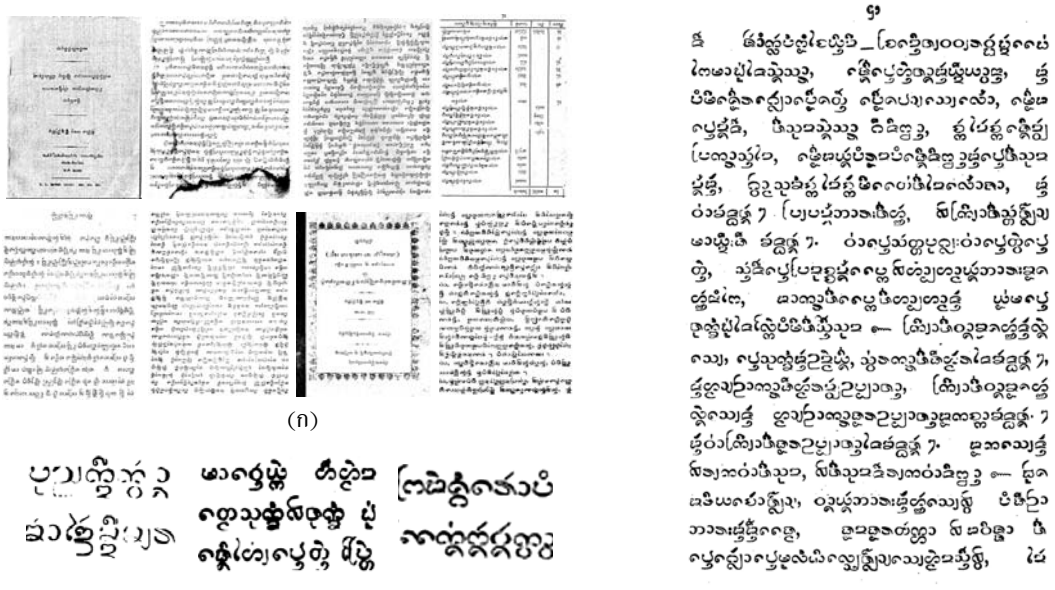
เนื่องจากการพัฒนาระบบการรู้จำภาษาสำหรับภาษาล้านนายังไม่สามารถนำมาใช้เป็นเครื่องมือในการแปลงข้อความจากภาพเอกสารภาษาล้านนาให้เป็นข้อความในลักษณะของรหัสแอสกีได้ ดังนั้น การดำเนินการกับภาพเอกสารภาษาล้านนาจึงต้องพิจารณาถึงลักษณะต่าง ๆ ที่มีอยู่ในภาพเอกสาร จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่ามีการใช้เทคนิคในการค้นคืนภาพเอกสารด้วยการดำเนินการกับภาพเอกสารเป็นหลัก[4-8] โดยเทคนิคดังกล่าวเป็นการเปรียบเทียบความคล้ายคลึงระหว่างคำที่ต้องการค้นหากับภาพคำในเอกสาร โดยมีขั้นตอนหลักในการดำเนินการ คือ การประมวลผลภาพเพื่อแปลงภาพเอกสารให้อยู่ในรูปของลักษณะเฉพาะของภาพตัวอักษรหรือภาพคำในเอกสาร แล้วจัดเก็บลักษณะเฉพาะเหล่านั้นเป็นฐานข้อมูลไว้เพื่อทำการค้นคืน และขั้นตอนต่อไป คือ ทำการค้นคืนโดยคำนวณค่าความคล้ายคลึงระหว่างคำที่ต้องการค้นหากับคำในภาพเอกสารด้วยการเปรียบเทียบลักษณะเฉพาะดังกล่าว

จากปัญหาเกี่ยวกับการเข้าถึงข้อมูล หรือเอกสารภาษาล้านนาของผู้ที่สนใจแต่ขาดความชำนาญในการอ่านการเขียนภาษาล้านนา รวมทั้งข้อจำกัดของลักษณะเฉพาะของภาษาที่ส่งผลต่อวิธีการค้นคืนข้อความจากภาพเอกสารภาษาล้านนา ทำให้ผู้วิจัยมีแนวคิดที่จะพัฒนารูปแบบการค้นคืนคำจากภาพเอกสารภาษาล้านนา เพื่อให้ผู้ที่มีความสนใจเกี่ยวกับภาษาล้านนาสามารถค้นหาและเข้าถึงเอกสารหรือข้อมูลที่ตรงตามความสนใจได้เร็วขึ้น รวมทั้งเป็นประโยชน์ในการศึกษาและวิจัยของผู้ที่มีความสนใจเกี่ยวกับภาษาล้านนาต่อไป

เนื้อหาภายในบทความต่อจากนี้ ประกอบไปด้วยส่วนต่าง ๆ ดังนี้ ส่วนที่ 2 จะกล่าวถึงรายละเอียดเกี่ยวกับวิธีการดำเนินการวิจัย ส่วนรายละเอียดเกี่ยวกับการวัดประสิทธิภาพและผลการทดลองจะแสดงในส่วนที่ 3 และสุดท้ายจะกล่าวถึงบทสรุปและข้อเสนอแนะในส่วนที่ 4

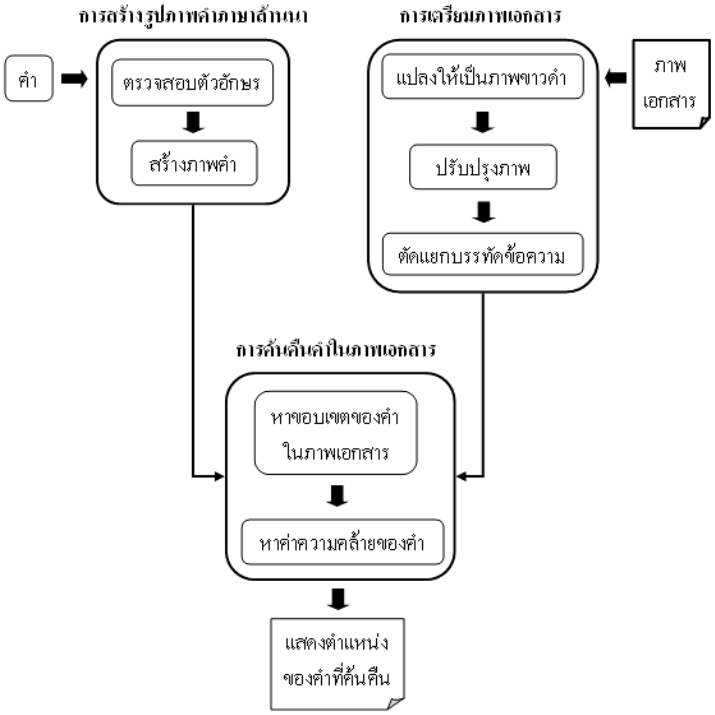
วิธีการ

การค้นคืนคำจากภาพเอกสารภาษาล้านนา เริ่มด้วยการสร้างภาพของคำภาษาล้านนาที่ต้องการค้นหา แล้วนำภาพคำดังกล่าวไปทำการเปรียบเทียบเพื่อหาตำแหน่งของคำจากนั้นแสดงตำแหน่งของคำดังกล่าวที่ปรากฏในภาพเอกสาร ซึ่งขั้นตอนในการดำเนินการแสดงได้ดังรูปที่ 3



(ก) (ข) (ค)

รูปที่ 2 ตัวอย่างภาพเอกสารภาษาล้านนา (ก) ภาพเอกสารที่ได้รับความเสียหายในลักษณะต่าง ๆ (ข) ลักษณะของข้อความที่ชำรุดไม่สมบูรณ์ (ค) ลักษณะของภาพเอกสารที่ถูกนำมาใช้ในการวิจัยนี้



รูปที่ 3 การค้นคืนคำจากภาพเอกสารภาษาล้านนา

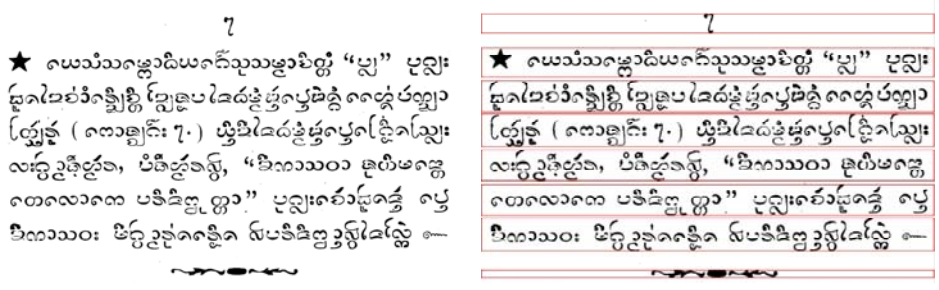
จากรูปที่ 3 จะเห็นว่าในกระบวนการของการค้นคืนคำจากภาพเอกสารภาษาล้านนาประกอบไปด้วยขั้นตอนทั้งหมด 3 ขั้นตอน คือ การเตรียมภาพเอกสารภาษาล้านนา การสร้างไฟล์ภาพคำภาษาล้านนา และการค้นคืนคำในภาพเอกสาร ซึ่งมีรายละเอียดดังต่อไปนี้

2.1 การเตรียมรูปภาพเอกสารภาษาล้านนา

การเตรียมรูปภาพเอกสารภาษาล้านนาแบ่งออกได้เป็น 3 ขั้นตอนได้แก่ ขั้นตอนที่ 1 เป็นขั้นตอนของการแปลงภาพ ซึ่งจะทำการแปลงภาพจากภาพสีเทามาเป็นภาพขาวดำ จากนั้นในขั้นตอนที่ 2 จะเป็นขั้นตอนของการปรับปรุงภาพ เพื่อกำจัดสัญญาณรบกวนในภาพ ทำให้ภาพมีความชัดเจนขึ้น และขั้นตอนที่ 3 จะเป็นขั้นตอนของการตัดบรรทัดเพื่อแยกตัวอักษรภายในภาพ ตัวอย่างภาพเอกสารที่ผ่านการเตรียมในขั้นตอนของการประมวลผลภาพแสดงดังรูปที่ 4

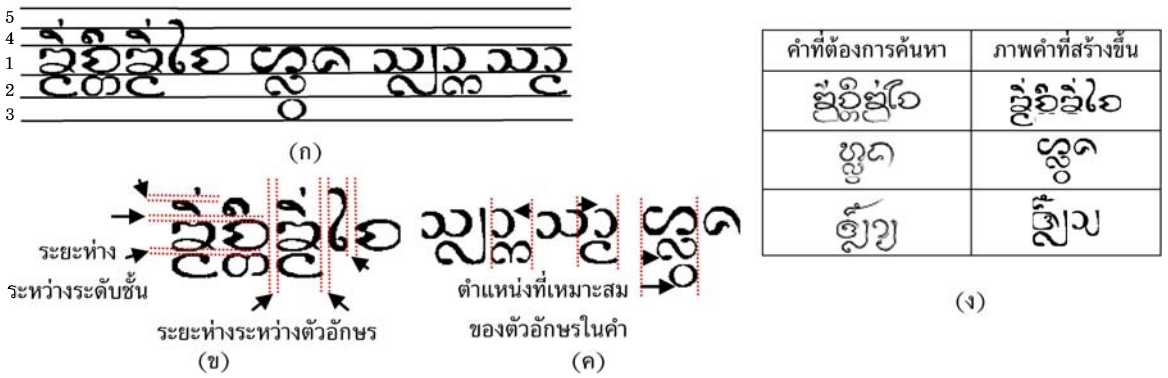
2.2 การสร้างไฟล์รูปภาพคำภาษาล้านนา

จากการวิเคราะห์ลักษณะการวางตัวอักษรในการเขียนภาษาล้านนาพบว่า มีรูปแบบการวางตัวอักษรแบ่งออกเป็น 5 ระดับ ดังรูปที่ 5(ก) บรรทัดในระดับที่ 1 คือ บรรทัดหลักซึ่งใช้เขียนตัวอักษรที่เป็นพยัญชนะต้น, สระ หรือตัวสะกดที่วางระดับเดียวกับพยัญชนะต้น และตัวสะกดที่มีการเปลี่ยนรูปเป็นตัวสะกดหาง บรรทัดในระดับที่ 2 จะใช้เขียนตัวสะกดหรือสระที่อยู่ใต้พยัญชนะต้น ส่วนสระที่อยู่ใต้พยัญชนะควบกล้ำจะเขียนไว้ในบรรทัดระดับที่ 3 ส่วนบรรทัดในระดับที่ 4 จะใช้เขียนสระ, วรรณยุกต์หรือตัวอักษรพิเศษที่เรียกว่า “ไม” และบรรทัดในระดับที่ 5 จะเป็นส่วนที่ใช้เขียนวรรณยุกต์ซึ่งจะวางอยู่บนสระที่อยู่เหนือพยัญชนะต้น [9]



(ก) (ข)

รูปที่ 4 ตัวอย่างรูปภาพเอกสารภาษาล้านนา (ก) ภาพต้นฉบับ และ (ข) รูปภาพที่ผ่านขั้นตอนของการตัดแยกบรรทัดแล้ว



รูปที่ 5 ตัวอย่างลักษณะของคำที่ต้องพิจารณาในขั้นตอนการสร้างภาพคำภาษาล้านนา (ก) การวางตัวอักษรในคำซึ่งแบ่งเป็น 5 ระดับ (ข) ระยะห่างในคำ (ค) ตำแหน่งที่เหมาะสมในการวางตัวอักษรในคำ (ง) ภาพคำภาษาล้านนาที่สร้างขึ้นจากคำที่ต้องการค้นหา

เมื่อพิจารณาตำแหน่งในการจัดวางตัวอักษรทั้ง 5 ระดับ พบว่าระยะห่างระหว่างตัวอักษรและตำแหน่งการวางตัวอักษรในแต่ละระดับดังรูปที่ 5(ข) ก็เป็นเรื่องที่ต้องพิจารณา นอกจากนี้การจัดวางตำแหน่งตัวอักษรในแต่ละคำก็แตกต่างกันขึ้นอยู่กับลักษณะและหน้าที่ของตัวอักษรนั้นว่าเป็น พยัญชนะ, ตัวสะกด, สระ, วรรณยุกต์ หรือสัญลักษณ์พิเศษ ซึ่งตัวอักษรบางตัวมีตำแหน่งในการจัดวางแตกต่างไปจากตัวอื่นๆ ดังรูปที่ 5(ค) จึงต้องมีการกำหนดเงื่อนไขในการจัดวางตำแหน่งตัวอักษร เพื่อให้สามารถสร้างภาพคำภาษาล้านนาได้ถูกต้อง ดังรูปที่ 5(ง)

จากรายละเอียดต่างๆ ที่ต้องนำมาพิจารณาก่อนการสร้างภาพคำภาษาล้านนาดังที่กล่าวมาข้างต้นนั้น ทำให้สามารถสรุปแผนภาพการทำงานของขั้นตอนการสร้างไฟล์ภาพคำภาษาล้านนาทั้งหมดได้ดังรูปที่ 6 โดยเริ่มจากการเตรียมภาพตัวอักษรภาษาล้านนาเพื่อใช้ในการสร้างภาพคำ ในขั้นตอนถัดไปจะเป็นการวิเคราะห์ข้อมูลเข้าที่อยู่ในรูปแบบอักษรตัวพิมพ์ ซึ่งเป็นรูปแบบการพิมพ์โดยใช้ฟอนต์ TILOK จากนั้นทำการตรวจสอบเพื่อหาตำแหน่งที่เหมาะสมของตัวอักษรในภาพคำที่จะสร้าง แล้วนำภาพของตัวอักษรภาษาล้านนาที่ตรงกับตัวอักษรที่วิเคราะห์มาสร้างเป็นภาพของคำภาษาล้านนา

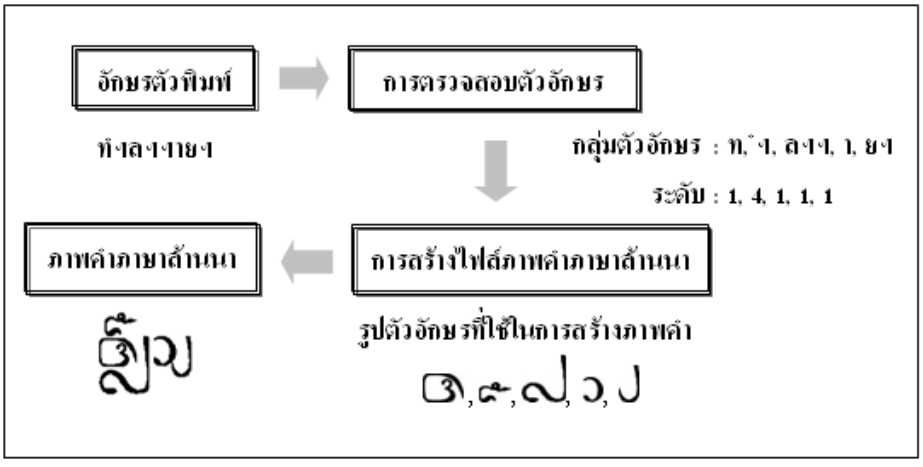
2.3 การค้นคืนคำในภาพเอกสาร

การค้นคืนคำจากรูปภาพเอกสารภาษาล้านนามีขั้นตอนในการดำเนินการ 2 ขั้นตอน โดยขั้นตอนแรกเป็นการหาขอบเขตของคำในภาพเอกสารที่มีความคล้ายคลึงกับภาพคำภาษาล้านนาที่สร้างขึ้น จากนั้นจะนำภาพคำทั้งหมดที่ได้จากขั้นตอนแรกมาสกัดคุณลักษณะโดยแบ่งภาพคำออกเป็นหน้าต่างย่อยๆ แล้วทำการคำนวณหาค่าความคล้ายคลึงและแสดงตำแหน่งของคำที่ค้นคืนได้

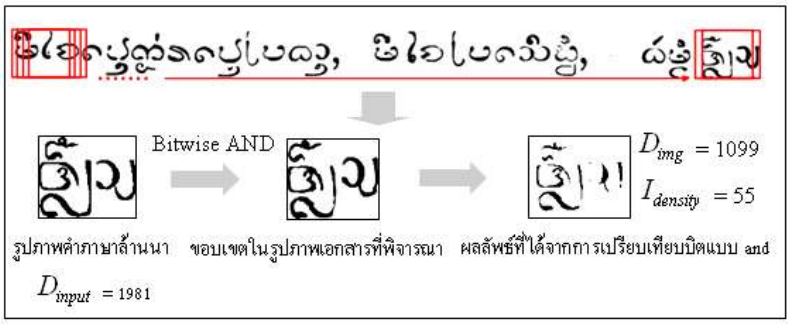
2.3.1 การหาขอบเขตของคำในภาพเอกสารโดยการเปรียบเทียบภาพ

(1) การหาขอบเขตของคำโดยเลื่อนหน้าต่างไปตามบรรทัดข้อความทีละพิกเซล (PSTM)

ทำได้โดยการสร้างหน้าต่างให้มีขนาดเท่ากับภาพคำที่ต้องการค้นหา และเลื่อนหน้าต่างดังกล่าวไปตามบรรทัดข้อความในภาพเอกสาร จากนั้นนำเวคเตอร์ของภาพคำที่ต้องการค้นหาและภาพเอกสาร ณ ตำแหน่งที่พิจารณามาเปรียบเทียบกันโดยใช้การเปรียบเทียบบิตแบบ and แล้วคำนวณค่าความหนาแน่นของพิกเซลในภาพผลลัพธ์ที่ได้ การพิจารณาตำแหน่งในภาพเอกสารที่มีความคล้ายคลึงกับภาพคำที่ต้องการค้นหา จะพิจารณาเปรียบเทียบจากค่า $I_{density}$ ซึ่งคำนวณได้จากสมการที่ (1) หากมีค่าอยู่ในช่วง 50 – 120 ขอบเขตของภาพที่พิจารณาดังกล่าวจะถูกจัดเก็บไว้ ขั้นตอนในการหาขอบเขตของคำแสดงดังรูปที่ 7



รูปที่ 6 การสร้างภาพคำภาษาล้านนา



รูปที่ 7 การหาขอบเขตของคำโดยการเลื่อนหน้าต่างไปตามบรรทัดข้อความทีละพิกเซล

$$I_{density} = \frac{D_{img}}{D_{input}} \times 100 \qquad (1)$$

โดยที่ D_{img} คือ ค่าความหนาแน่นของพิกเซลภาพผลลัพธ์ในเอกสาร และ D_{input} คือ ค่าความหนาแน่นของพิกเซลในภาพคำที่ต้องการค้นหา

(2) การหาขอบเขตของคำในภาพโดยพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ (WFM)

ทำการสกัดคุณลักษณะของภาพคำที่ต้องการค้นหาเมื่อได้คุณลักษณะที่เป็นตัวแทนของภาพคำแล้ว ทำการตัดแยกตัวอักษรในบรรทัดข้อความโดยใช้เทคนิคการโปรเจคชันตามแนวนอน จากนั้นพิจารณากลุ่มภาพตัวอักษรในบรรทัดข้อความที่มีคุณสมบัติสอดคล้องกับคุณลักษณะที่สกัดมาได้โดยเลื่อนพิจารณาไปทีละภาพตัวอักษรจนหมดบรรทัดข้อความดังแสดงในรูปที่ 8 จากนั้นพิจารณาค่าที่ได้จากการคำนวณตามสมการที่ (2) หากมีค่าเท่ากับ 1 ขอบเขตในภาพเอกสารที่พิจารณาจะถูกจัดเก็บไว้

โดยที่ F_q คือ คุณลักษณะที่เป็นตัวแทนของคำที่ต้องการค้นหาประกอบไปด้วย H_q และ W_q ซึ่งคือ ค่าความสูงของภาพตัวอักษรภายในคำ และค่าผลรวมความกว้างของภาพตัวอักษรภายในคำตามลำดับ

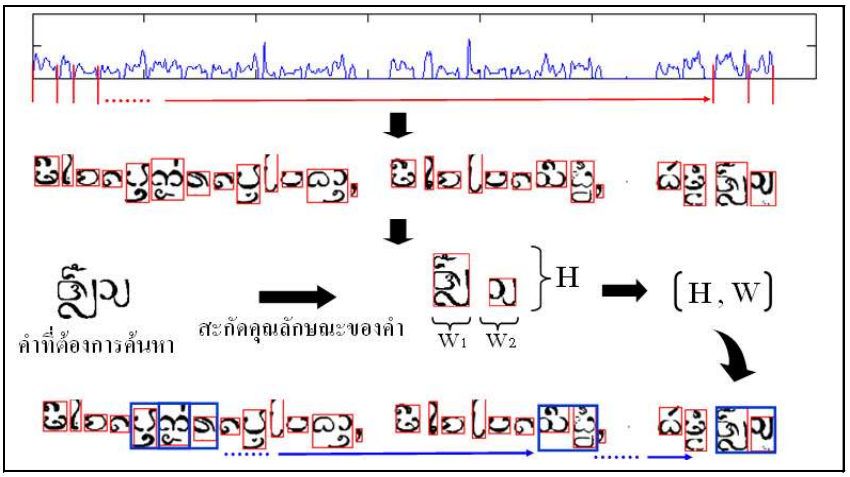
$b_{i\to j}$ คือ ภาพตัวอักษรในบรรทัดข้อความที่พิจารณาลำดับที่ i ถึงลำดับที่ j

w_k คือ ความกว้างของภาพตัวอักษรภายในคำลำดับที่ k

$upper(b_{i\to j})$ คือ ขอบเขตบนของภาพตัวอักษรในบรรทัดข้อความที่พิจารณาลำดับที่ i ถึงลำดับที่ j

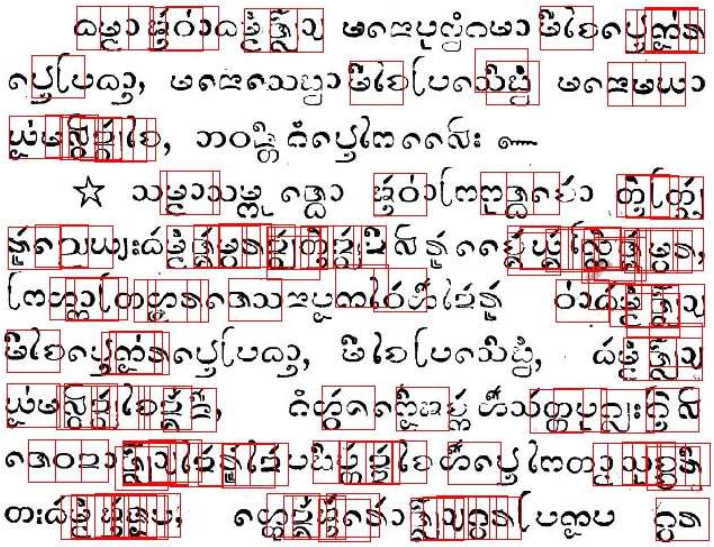
$lower(b_{i\to j})$ คือ ขอบเขตล่างของภาพตัวอักษรในบรรทัดข้อความที่พิจารณาลำดับที่ i ถึงลำดับที่ j

ขอบเขตของคำในภาพเอกสารที่ได้มาจากขั้นตอนการเปรียบเทียบภาพ เพื่อหาขอบเขตของคำในภาพเอกสารดังที่ได้อธิบายมาแล้วข้างต้นนั้นแสดงได้ดังรูปที่ 9 จากรูปดังกล่าวจะเห็นว่าผลลัพธ์ที่ได้จากขั้นตอนนี้ประกอบไปด้วยคำเป็นจำนวนมาก ซึ่งเป็นคำที่ไม่ถูกต้องและไม่ต้องการให้ปรากฏในการค้นคืน ดังนั้น จึงต้องมีการคัดกรองหาตำแหน่งที่ถูกต้องของคำที่ค้นคืนมาได้ โดยผ่านกระบวนการของการหาค่าความคล้ายคลึงของคำดังจะอธิบายในขั้นตอนต่อไป



รูปที่ 8 การหาขอบเขตของคำโดยพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ

$$Match(F_q, b_{i\to j}) = \begin{cases} 1 & \text{if } max(upper(b_{i\to j})) - min(lower(b_{i\to j})) = H_q \quad \text{and} \quad \sum_{k=1}^{j} w_k = W_q \\ 0 & otherwise \end{cases} \qquad (2)$$



รูปที่ 9 ตำแหน่งของคำในภาพเอกสารที่ได้มาจากการเปรียบเทียบภาพคำที่ต้องการค้นหา

2.3.2 การหาความคล้ายคลึงของคำ

(1) การหาความคล้ายคลึงของคำด้วยการพิจารณาค่าความหนาแน่นในแต่ละหน้าต่างย่อย (S)

แบ่งภาพที่พิจารณาออกเป็นหน้าต่างย่อยขนาด m x n จากนั้นคำนวณค่าความหนาแน่นของพิกเซลที่มีสีดำ หรือมีค่าเป็น 1 ในแต่ละหน้าต่างย่อย โดยกำหนดให้ เวคเตอร์ W เป็นเวคเตอร์ความหนาแน่นในแต่ละหน้าต่างย่อยของภาพผลลัพธ์ในภาพเอกสาร และเวคเตอร์ Q เป็นเวคเตอร์ความหนาแน่นในแต่ละหน้าต่างย่อยของภาพคำที่ต้องการค้นหา จากนั้นทำการเปรียบเทียบเวคเตอร์ W และเวคเตอร์ Q โดยการนับจำนวนหน้าต่างย่อย $Count(k)$ ตามสมการที่ (3) หากมีค่ามากกว่าหรือเท่ากับค่าเธรดโฮลด์ที่กำหนดไว้ก็จะถือว่าคำที่พิจารณาดังกล่าวมีความคล้ายคลึงกัน

$$Count(k) = \begin{cases} 1 \text{ if } \left(\frac{w_k}{q_k} \times 100\right) \geq 50 \text{ or } w_k = q_k = 0 \\ 0 \text{ otherwise} \end{cases} \quad (3)$$

โดยที่ w_k คือ จำนวนพิกเซลที่มีสีดำในแต่ละหน้าต่างย่อยของภาพผลลัพธ์ที่พิจารณาในภาพเอกสาร โดย k มีค่าตั้งแต่ 1 ถึง m x n

q_k คือ จำนวนพิกเซลที่มีสีดำในแต่ละหน้าต่างย่อยของภาพคำที่ต้องการค้นหา โดย k มีค่าตั้งแต่ 1 ถึง m x n

(2) การหาความคล้ายคลึงของคำด้วยการพิจารณาค่าระยะห่างแบบยูคลิเดียน (E)

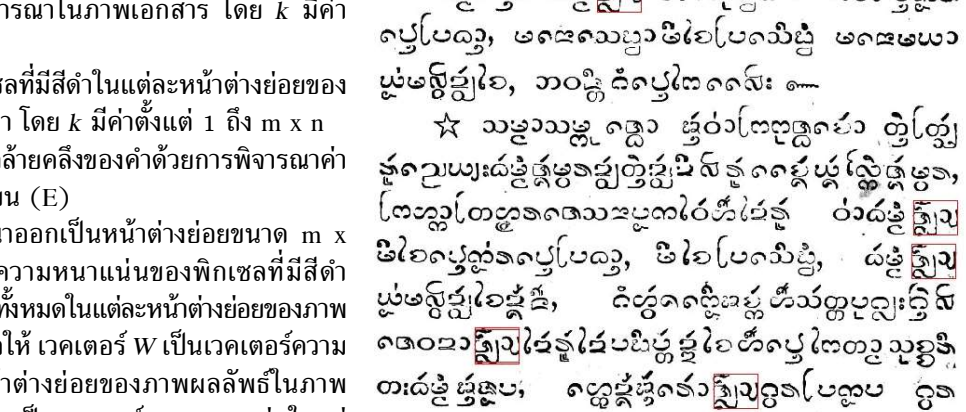
แบ่งภาพที่พิจารณาออกเป็นหน้าต่างย่อยขนาด m x n จากนั้นคำนวณค่าความหนาแน่นของพิกเซลที่มีสีดำเทียบกับจำนวนพิกเซลทั้งหมดในแต่ละหน้าต่างย่อยของภาพที่พิจารณา โดยกำหนดให้ เวคเตอร์ W เป็นเวคเตอร์ความหนาแน่นในแต่ละหน้าต่างย่อยของภาพผลลัพธ์ในภาพเอกสารและเวคเตอร์ Q เป็นเวคเตอร์ความหนาแน่นในแต่ละหน้าต่างย่อยของภาพคำที่ต้องการค้นหา จากนั้นทำการเปรียบเทียบเวคเตอร์ W และเวคเตอร์ Q โดยการพิจารณาค่าระยะห่างแบบยูคลิเดียน $Edist(W,Q)$ ตามสมการที่ (4) หากมีค่ามากกว่าหรือเท่ากับค่าเธรดโฮลด์ที่กำหนดไว้ก็จะถือว่าคำที่พิจารณาดังกล่าวมีความคล้ายคลึงกัน

$$Edist(W,Q) = \left(\sum_{k=1}^{mxn} (w_k - q_k)^2\right)^{\frac{1}{2}} \quad (4)$$

โดยที่ w_k คือ จำนวนพิกเซลที่มีสีดำในแต่ละหน้าต่างย่อยของภาพผลลัพธ์ที่พิจารณาในภาพเอกสาร โดย k มีค่าตั้งแต่ 1 ถึง m x n

q_k คือ จำนวนพิกเซลที่มีสีดำในแต่ละหน้าต่างย่อยของภาพคำที่ต้องการค้นหา โดย k มีค่าตั้งแต่ 1 ถึง m x n

เมื่อนำขอบเขตของคำในภาพเอกสารที่ได้มาจากขั้นตอนการเปรียบเทียบภาพมาหาค่าความคล้ายคลึงของคำดังที่ได้อธิบายมาแล้วข้างต้น ผลลัพธ์สุดท้ายที่ได้จากขั้นตอนการหาค่าความคล้ายคลึง คือ ตำแหน่งของคำที่ค้นคืนมาได้ในภาพเอกสาร ดังแสดงในรูปที่ 10



รูปที่ 10 ตำแหน่งของคำที่ค้นคืนได้ในภาพเอกสารภาษาล้านนา

ผลและอภิปราย

ในงานวิจัยนี้ได้ทำการทดสอบประสิทธิภาพของการค้นคืนคำจากภาพเอกสารภาษาล้านนาโดยใช้คำในการทดสอบจำนวนทั้งหมด 187 คำ และกำหนดค่าเธรดโฮลด์ที่ใช้ในการเปรียบเทียบความคล้ายคลึงของคำเท่ากับ 0.7 ซึ่งประสิทธิภาพของการค้นคืนคำจากภาพเอกสารวัดจากค่าความแม่นยำ (precision) ความถูกต้อง (recall) และค่า Fmeasure ของการค้นคืนดังสมการที่ (5), (6) และ (7) ตามลำดับ

$$\text{ความแม่นยำ} = \frac{\text{ตำแหน่งของคำที่ค้นคืนได้ถูกต้อง}}{\text{ตำแหน่งของคำที่ค้นคืนมาได้ทั้งหมด}} \quad (5)$$

$$\text{ความถูกต้อง} = \frac{\text{ตำแหน่งของคำที่ค้นคืนได้ถูกต้อง}}{\text{ตำแหน่งของคำที่ต้องการค้นคืนที่ปรากฏในรูปภาพเอกสาร}} \quad (6)$$

$$F_{measure} = 2 \times \frac{precision \times recall}{precision + recall} \quad (7)$$

ในการทดสอบประสิทธิภาพการค้นคืนคำจากภาพเอกสารภาษาล้านนา มีวิธีการในการทดสอบทั้งหมดจำนวน 4 วิธี ซึ่งอธิบายได้ดังต่อไปนี้

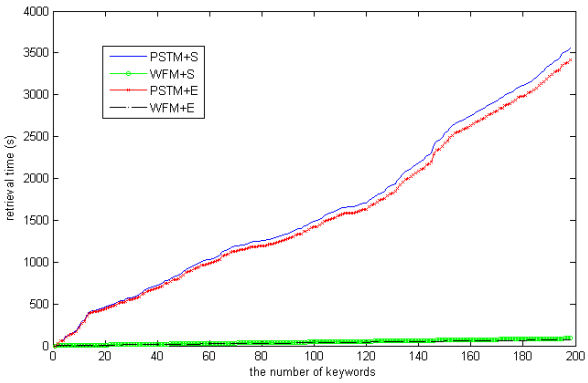
1. PSTM+S คือ วิธีการหาขอบเขตของคำในภาพเอกสารด้วยการเปรียบเทียบภาพโดยเลื่อนหน้าต่างไปตามบรรทัดข้อความทีละพิกเซล และหาค่าความคล้ายคลึงของคำด้วยการพิจารณาค่าความหนาแน่นในแต่ละหน้าต่างย่อย
2. WFM+S คือ วิธีการหาขอบเขตของคำในภาพโดยพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ และหาค่าความคล้ายคลึงของคำด้วยการพิจารณาค่าความหนาแน่นในแต่ละหน้าต่างย่อย
3. PSTM+E คือ วิธีการหาขอบเขตของคำในภาพเอกสารด้วยการเปรียบเทียบภาพโดยเลื่อนหน้าต่างไปตามบรรทัดข้อความทีละพิกเซล และหาค่าความคล้ายคลึงของคำด้วยการพิจารณาค่าระยะห่างแบบยูคลิเดียน
4. WFM+E คือ วิธีการหาขอบเขตของคำในภาพโดยพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ และหาค่าความคล้ายคลึงของคำด้วยการพิจารณาค่าระยะห่างแบบยูคลิเดียน

ผลการทดสอบประสิทธิภาพของวิธีการค้นคืนคำในภาพเอกสารแสดงในตารางที่ 1 และการเปรียบเทียบเวลาที่ใช้ในการทดสอบแสดงในรูปที่ 11

ตารางที่ 1 ประสิทธิภาพของวิธีการที่ใช้ทดสอบการค้นคืนคำในภาพเอกสารภาษาล้านนา

วิธีการค้นคืน	ค่าเฉลี่ยความแม่นยำ (Average precision)	ค่าเฉลี่ยความถูกต้อง (Average recall)	$F_{measure}$	เวลาเฉลี่ยในการค้นคืนคำ (วินาที)
PSTM+S	0.65	0.91	0.76	17.98
WFM+S	0.73	0.73	0.73	0.45
PSTM+E	0.85	0.79	0.82	17.25
WFM+E	0.83	0.82	0.82	0.37

จากการเปรียบเทียบประสิทธิภาพของวิธีการค้นคืนในตารางที่ 1 เมื่อพิจารณาประสิทธิภาพด้านความถูกต้องในการค้นคืนของวิธีการที่ทดสอบ พบว่าวิธีการค้นคืนแบบ PSTM+S มีค่าเฉลี่ยความถูกต้องสูงที่สุด คือ 0.91 เมื่อพิจารณาประสิทธิภาพด้านความแม่นยำในการค้นคืน พบว่าวิธีการค้นคืนแบบ PSTM+E มีค่าเฉลี่ยความแม่นยำสูงที่สุด คือ 0.85 เมื่อพิจารณาประสิทธิภาพโดยรวมของการค้นคืน พบว่าวิธีการค้นคืนแบบ PSTM+E และ WFM+E มีค่า Fmeasure สูงที่สุด คือ 0.82 และเมื่อพิจารณาประสิทธิภาพในการค้นคืนร่วมกับเวลาที่ใช้ในการค้นคืนคำ พบว่าวิธีการค้นคืนแบบ WFM+E มีความเหมาะสมมากที่สุดในการค้นคืนคำจากภาพเอกสารภาษาล้านนา นอกจากนี้หากพิจารณาจากรูปที่ 7 ซึ่งเป็นการเปรียบเทียบเวลาที่ใช้ในการค้นคืนคำในภาพเอกสารทั้งหมด 187 คำ จะเห็นว่าวิธีการค้นคืนแบบ PSTM+S และ PSTM+E ใช้เวลาในการค้นคืนคำมากกว่าวิธีการค้นคืนแบบอื่น ๆ อย่างเห็นได้ชัดเจน ดังนั้น จึงแสดงให้เห็นว่าวิธีการหาขอบเขตของคำในภาพเอกสารด้วยการเปรียบเทียบภาพโดยเลื่อนหน้าต่างไปตามบรรทัดข้อความทีละพิกเซลจะใช้เวลาในการประมวลผลมากที่สุด



รูปที่ 11 เปรียบเทียบเวลาที่ใช้ในการทดสอบการค้นคืนคำในภาพเอกสาร

บทสรุป

การวิจัยนี้นำเสนอการค้นคืนคำจากภาพเอกสารภาษาล้านนา ด้วยการสร้างภาพคำที่ต้องการค้นหา แล้วนำภาพดังกล่าวไปทำการเปรียบเทียบเพื่อหาตำแหน่งของคำที่ปรากฏในภาพเอกสาร โดยใช้ค่าความแม่นยำ ค่าความถูกต้อง และค่า $F_{measure}$ เป็นตัววัดประสิทธิภาพของการ

ค้นคืน จากการทดสอบพบว่าวิธีเหมาะสมในการค้นคืนคำจากภาพเอกสารภาษาล้านนา คือ การหาขอบเขตคำในภาพเอกสารโดยการพิจารณาจากความกว้างและความสูงของตัวอักษรภายในคำ และหาค่าความคล้ายคลึงของคำโดยพิจารณาจากค่าระยะห่างแบบยูคลิเดียน เมื่อนำวิธีการค้นคืนคำดังกล่าวมาพัฒนาต้นแบบเครื่องมือที่ใช้ในการค้นคืนคำจากภาพเอกสารภาษาล้านนาและทดสอบการใช้งาน พบว่าวิธีการค้นคืนคำจากภาพเอกสารภาษาล้านนามีประสิทธิภาพอยู่ในระดับที่สามารถนำไปใช้งานได้จริง ซึ่งในอนาคตสามารถนำแนวคิดในการค้นคืนคำจากภาพเอกสารภาษาล้านนาโดยการเปรียบเทียบภาพไปพัฒนาเครื่องมือเพื่อใช้ในการช่วยคัดกรองภาพเอกสารภาษาล้านนาเบื้องต้นสำหรับผู้ที่มีความสนใจเกี่ยวกับภาษาล้านนา ทำให้สามารถค้นหาและเข้าถึงเอกสารหรือข้อมูลที่ตรงตามความสนใจได้เร็วขึ้น และเพื่อเป็นประโยชน์ในการศึกษาและวิจัยของผู้ที่มีความสนใจเกี่ยวกับภาษาล้านนาต่อไป

อย่างไรก็ตามวิธีการค้นคืนคำจากภาพเอกสารภาษาล้านนาในการวิจัยนี้นำไปใช้ได้ผลดีกับภาพเอกสารที่มีลักษณะของตัวอักษรและรูปแบบการจัดวางตัวอักษรในภาพเอกสารใกล้เคียงกับภาพคำภาษาล้านนาที่สร้างขึ้น เมื่อพิจารณาถึงปัจจัยที่ส่งผลต่อประสิทธิภาพของการค้นคืนพบว่า ลักษณะของตัวอักษรภายในคำจำนวนพยางค์ในคำและความซับซ้อนของรูปแบบโครงสร้างคำเป็นสิ่งที่ส่งผลต่อความแม่นยำและความถูกต้องของวิธีการค้นคืนที่นำเสนอในงานวิจัยนี้

เอกสารอ้างอิง

[1] เกษมศิริรัตน์พิริยะ. (2548). *ตั๋วเมือง: การเรียนภาษาล้านนาผ่านโครงสร้างคำ.* โรงพิมพ์มหาวิทยาลัยสุโขทัยธรรมาธิราช. นนทบุรี.

[2] จีรยุทธ ไชยจารุวณิช และคณะ. (2552). *วรรณพิมพ์ล้านนา: วรรณกรรมที่ตีพิมพ์ด้วยอักษรธรรมล้านนา 60 เล่ม.* มหาวิทยาลัยเชียงใหม่. เชียงใหม่, สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ.

[3] Doermann, D. (1998). "The Indexing and Retrieval of Document Images: A Survey." *Computer Vision and Image Understanding, 70*(3), 287-298.

[4] Chew Lim Tan et al. (2002). "Imaged Document Text Retrieval without OCR." *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(6), 838-844.

[5] Chew Lim Tan et al. (2003). "Text Retrieval from Document Images Based on Word Shape Analysis." *Applied Intelligence, 18*(3), 257-270.

[6] Yue, L & Chew Lim Tan. (2002). "Word Searching in Document Images using Word Portion Matching." Fifth IAPR International Workshop on Document Analysis Systems, USA. 319-328.

[7] Soo Hyung Kimet al. (2005). "Keyword Spotting on Korean Document Images by Matching the Keyword Image." Lecture Notes In Computer Science, Digital Libraries: Implementing Strategies and Sharing Experiences, Heidelberg: Springer Berlin, 3815, 158-166.

[8] Seema, Y. & Sudhir, S. (2009). "Retrieval Of Information In Document Image Databases Using Partial Word Image Matching Technique." *In Proceeding of IMECS 2009*, Vol. I, Hong Kong, 902-907.

[9] Yathongkhum, W. et al. (2010). "Word Retrieval from Lanna Document Images," in *Proc. The 14th National Computer Science and Engineering Conference (NCSEC).* Thailand, 307-312.