บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

จากการศึกษาขั้นตอนกระบวนการรู้จำตัวอักษรที่ผ่านมาในงานวิจัยที่เกี่ยวข้องนั้นได้มีการศึกษาวิจัยกันมาอย่างต่อเนื่องและกว้างขวางเพื่อเพิ่มประสิทธิภาพความสามารถในการรู้จำตัวอักษรให้มีความถูกต้องมากขึ้นอาทิเช่น การรู้จำตัวอักษรโดยอาศัยโครงข่ายประสาทเทียมรูปแบบต่าง ๆ ได้แก่ Back-Propagation, BAM และ CPN หรือจะเป็นการรู้จำโดยอาศัยการใช้ลักษณะเด่นของตัวอักษรซึ่งอาจจะมีการนำเอาทฤษฎีฟัซซี่ลอจิกหรือทฤษฎีอื่น ๆ เข้ามาช่วยในกระบวนการรู้จำ ทั้งนี้ความสามารถในการรู้จำนั้นไม่ได้ขึ้นกับวิธีที่นำมาใช้ในขั้นตอนของกระบวนการรู้จำตัวอักษรเท่านั้น แต่ยังต้องอาศัยกระบวนการก่อนการรู้จำที่ดีเพื่อให้ได้ภาพตัวอักษรที่มีคุณภาพดีด้วย ด้วยเหตุนี้จึงทำให้มีข้อดีและข้อเสียแตกต่างกันไป โดยเฉพาะถ้าภาพตัวอักษรก่อนการรู้จำมีคุณภาพที่ไม่ดีก็จะส่งผลให้ประสิทธิภาพในการรู้จำภาพตัวอักษรนั้นไม่มีความถูกต้องเท่าที่ควร[1] ดังนั้นในการรู้จำตัวอักษรก็จะขึ้นอยู่กับกระบวนการก่อนการรู้จำที่จะจัดการข้อมูลภาพตัวอักษรให้อยู่ในรูปแบบที่เหมาะสมก่อนจะเข้าสู่กระบวนการรู้จำ เช่นการปรับให้คุณภาพของข้อมูลภาพตัวอักษรก่อนการรู้จำมีคุณภาพที่ดีโดยจะต้องไม่มีสัญญาณรบกวน และรวมถึงปัญหาต่าง ๆ ที่พบเจอในงานวิจัยที่ผ่านมาเช่น การแก้ไขตัวอักษรที่ติดกัน [2] [3] [4] การลดสัญญาณรบกวน [4] การทำขอบตัวอักษรให้เรียบ [5] การเชื่อมต่อลายเส้นที่ขาดหายไปของภาพอักษรตัวพิมพ์ภาษาไทยในแนวนอน [6] และการเชื่อมต่อลายเส้นที่ขาดหายไปของภาพอักษรตัวพิมพ์ภาษาไทยในแนวตั้ง [7] เหล่านี้ช่วยให้กระบวนการก่อนการรู้จำได้ข้อมูลภาพตัวอักษรที่ดีขึ้น จากนั้นจะพิจารณากระบวนการรู้จำที่มีการออกแบบอัลกอริทึมสำหรับการรู้จำภาษาต่าง ๆ โดยแต่ละรูปแบบภาพตัวอักษรของแต่ละภาษาก็จะมีคุณลักษณะโครงสร้างที่แตกต่างกันไป สำหรับงานวิจัยที่ผ่านมานั้นจะเป็นการศึกษาในการรู้จำตัวอักษรของแต่ละภาษาโดยเฉพาะ ตัวอย่างเช่นการรู้จำตัวอักษรภาษาอังกฤษ การรู้จำตัวอักษรภาษาจีน การรู้จำตัวอักษรภาษาฝรั่งเศส หรือแม้กระทั่งการรู้จำตัวอักษรภาษาไทย

สำหรับรูปแบบโครงสร้างของตัวอักษรภาษาไทยนั้นปกติแล้วจะประกอบไปด้วยเส้นตัวอักษรที่เป็นเส้นตรง เส้นโค้ง เส้นซิกแซก หรือลูป โดยส่วนมากแล้วตัวอักษรภาษาไทยจะมีจุดเริ่มต้นเป็นลูป หรือที่เราเรียกว่าส่วนที่เป็นหัวตัวอักษร แต่ในปัจจุบันมีการพัฒนารูปแบบตัวอักษรไทยที่เปลี่ยนไปจากเดิมคือที่ส่วนหัวของตัวอักษรไม่ได้เป็นลูปเราเรียกว่ารูปแบบตัวอักษรที่ไม่มีหัว ตามตารางที่ 1.1 แสดงการเปรียบเทียบลักษณะของตัวอักษรภาษาไทยในรูปแบบที่มีหัว

และไม่มีหัว ดังนั้นเมื่อมีการพิจารณาลักษณะเด่นของโครงสร้างตัวอักษรภาษาไทย สิ่งแรกที่จะใช้พิจารณาคือลักษณะของตัวอักษรภาษาไทยในรูปแบบที่มีหัว ซึ่งงานวิจัยที่เกี่ยวกับการรู้จำตัวอักษรภาษาไทยส่วนมากนั้นไม่ได้เน้นถึงลักษณะของตัวอักษรภาษาไทยในรูปแบบที่ไม่มีหัว ซึ่งสามารถพบเจอได้ตามเอกสารต่าง ๆ เช่นหนังสือพิมพ์ วารสาร สื่อแผ่นพับโฆษณาต่าง ๆ ที่เป็นตัวอักษรภาษาไทยที่มีรูปแบบการเขียนหรือพิมพ์ออกมาแล้วเป็นรูปแบบที่ไม่มีหัวเป็นส่วนใหญ่

ตัวอย่างข้อมูลภาพตัวอักษรตามรูปที่ 1.1 เมื่อผ่านการทดสอบด้วยโปรแกรมอ่านไทย (ArnThai 2.5 Lite) พบว่าผลที่ได้ออกมามีความถูกต้องเท่ากับ 34.85% แสดงผลลัพธ์ที่ได้จากการรู้จำตามรูปที่ 1.2 ซึ่งแสดงให้เห็นว่าผลของการรู้จำที่ได้ออกมายังคงมีปัญหาในเรื่องของตัวอักษรพิมพ์ภาษาไทยในรูปแบบที่ไม่มีหัว

ในงานวิจัยครั้งนี้จึงมุ่งเน้นที่กลุ่มข้อมูลตัวอย่างลักษณะของตัวอักษรภาษาไทยในรูปแบบที่ไม่มีหัวซึ่งจะมีทั้งหมด 39 รูปแบบ แต่ละรูปแบบจะประกอบด้วยพยัญชนะ สระ วรรณยุกต์ ตัวเลขอารบิก และสัญลักษณ์อื่น ๆ รวมทั้งหมดรูปแบบละ 80 ตัวอักขระ นอกจากนี้ยังพิจารณากลุ่มข้อมูลภาพตัวอักษรที่อยู่บนป้ายทะเบียนรถที่เป็นภาษาไทยและตัวเลขอารบิก

แต่ทั้งนี้จะเห็นว่าตั้งแต่อดีตมาจนถึงปัจจุบันได้มีนักวิจัยทำการพัฒนาเทคโนโลยีทางด้านการรู้จำตัวอักษรกันมากขึ้น แต่ในการพัฒนาจะมุ่งเน้นกลุ่มทดลองตามภาษาที่ใช้ของแต่ละประเทศ ด้วยเหตุนี้การรู้จำของแต่ละภาษาก็มีวิธีในการรู้จำที่ต่างกันออกไปตามโครงสร้างของแต่ละภาษาเนื่องมาจากตัวอักษรของแต่ละภาษาที่มีรูปแบบต่างกันออกไป และสำหรับงานวิจัยที่เกี่ยวกับการรู้จำภาษาไทยนั้น จะรู้จำในกลุ่มของตัวอักษรมีหัว แต่ส่วนมากในการทำสื่อโฆษณา เผยแพร่ หรือหนังสือต่าง ๆ จะประกอบไปด้วยตัวอักษรภาษาไทยไม่มีหัว ซึ่งปัจจุบันนี้พบอยู่อย่างแพร่หลาย

ดังนั้นงานวิจัยครั้งนี้ได้เล็งเห็นถึงความสำคัญของการรู้จำตัวอักษรตัวพิมพ์ภาษาไทยไม่มีหัว เพื่อเพิ่มประสิทธิภาพในการรู้จำตัวอักษรภาษาไทยให้ครอบคลุมไปถึงตัวอักษรภาษาไทยไม่มีหัว ให้สามารถได้ผลลัพธ์เอกสารสื่อต่าง ๆ ออกมาในรูปแบบที่ยอมรับได้ และมีประสิทธิภาพยิ่งขึ้น

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1.2.1 เพื่อศึกษาลักษณะพิเศษของตัวพิมพ์อักษรไทยในกลุ่มตัวอักษรไม่มีหัว ที่จะนำมาใช้ในวิธีการรู้จำ โดยศึกษาเปรียบเทียบจากวิธีต่าง ๆ ที่ผ่านการวิจัยมาแล้ว

1.2.2 เพื่อศึกษาวิธีการที่เหมาะสมของการวิเคราะห์หาความสัมพันธ์ของโครงสร้างทางกายภาพมาใช้กับการรู้จำตัวพิมพ์อักษรไทยในกลุ่มตัวอักษรไม่มีหัว

1.2.3 เพื่อศึกษาวิธีการที่เหมาะสมของการนำโครงข่ายประสาทเทียมมาใช้กับการรู้จำตัวพิมพ์อักษรไทย

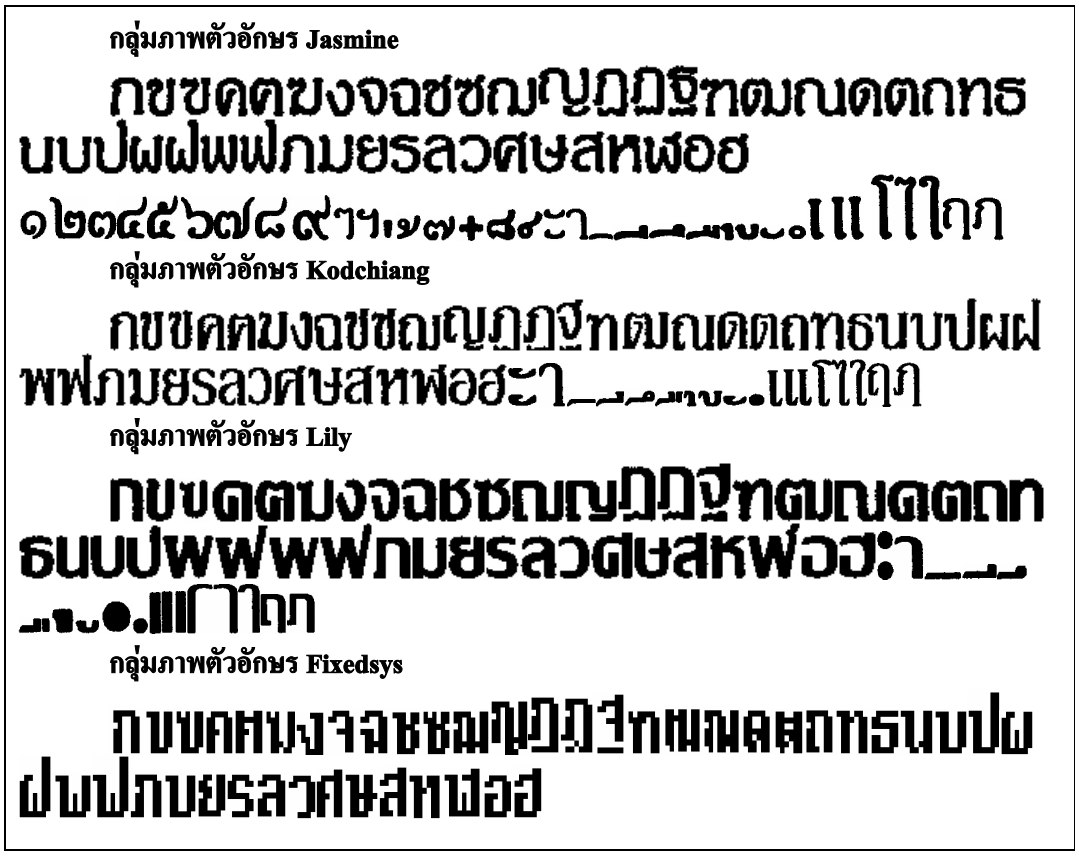
1.2.4 เพื่อเพิ่มประสิทธิภาพในระบบการรู้จำของตัวอักษรภาษาไทยครอบคลุมในรูปแบบตัวอักษรไม่มีหัว

ตารางที่ 1.1 การเปรียบเทียบลักษณะของตัวอักษรภาษาไทยในรูปแบบที่มีหัวและไม่มีหัว

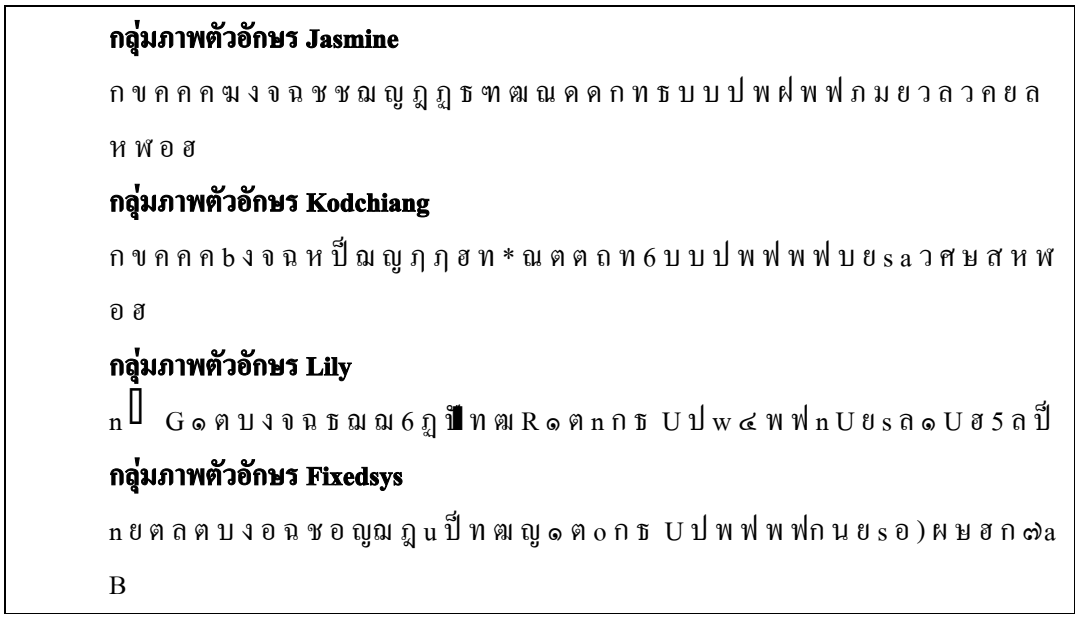
ตัวอักษรที่	ตัวอักษรภาษาไทยรูปแบบที่มีหัว	ตัวอักษรภาษาไทยรูปแบบที่ไม่มีหัว
1. กอไก่	ก	ก
2. ขอไข่	ข	ข
3. ฃอฃวด	ฃ	ฃ
4. คอควาย	ค	ค
5. ฅอฅน	ฅ	ฅ
6. ฆอระฆัง	ฆ	ฆ
7. งองู	ง	ง
8. จอจาน	จ	จ
9. ฉอฉิ่ง	ฉ	ฉ
10. ชอช้าง	ช	ช
11. ซอโซ่	ซ	ซ
12. ฌอกะเฌอ	ฌ	ฌ
13. ญอผู้หญิง	ญ	ญ
14. ฎอปฏัก	ฎ	ฏ
15. ฏอชฏา	ฏ	ฏ
16. ฐอฐาน	ฐ	ฐ
17. ฑอมณโฑ	ฑ	ฑ
18. ฒอผู้เฒ่า	ฒ	ฒ
19. ณอเณร	ณ	ณ
20. ดอเด็ก	ด	ด
21. ตอเต่า	ต	ต

ตารางที่ 1.1 (ต่อ)

ตัวอักษรที่	ตัวอักษรภาษาไทยรูปแบบที่มีหัว	ตัวอักษรภาษาไทยรูปแบบที่ไม่มีหัว
22. ถอถุง	ถ	ถ
23. ทอทหาร	ท	ท
24. ธอธง	ธ	ธ
25. นอหนู	น	น
26. บอใบไม้	บ	บ
27. ปอปลา	ป	ป
28. ผอผึ้ง	ผ	ผ
29. ฝอฝา	ฝ	ฝ
30. พอพาน	พ	พ
31. ฟอฟัน	ฟ	ฟ
32. ภอสำเภา	ภ	ภ
33. มอม้า	ม	ม
34. ยอยักษ์	ย	ย
35. รอเรือ	ร	ร
36. ลอลิง	ล	ล
37. วอแหวน	ว	ว
38. ศอศาลา	ศ	ศ
39. ษอบอฤษี	ษ	ษ
40. สอเสือ	ส	ส
41. หอหีบ	ห	ห
42. ฬอจุฬา	ฬ	ฬ
43. ออ่าง	อ	อ
44. ฮอนกฮูก	ฮ	ฮ



รูปที่ 1.1 ตัวอย่างรูปแบบข้อมูลภาพตัวอักษรที่ใช้ในการทดลอง



รูปที่ 1.2 ตัวอย่างแสดงผลการรู้จำเมื่อผ่านโปรแกรมการรู้จำ ArnThai 2.5 Lite

1.3 ทฤษฎีหรือแนวความคิดที่ใช้ในการวิจัย

แนวความคิดที่ใช้ในงานวิจัยนี้ประกอบไปด้วย 2 กระบวนการคือกระบวนการก่อนการรู้จำและกระบวนการรู้จำ ในขั้นตอนของกระบวนการก่อนการรู้จำนั้นเมื่อรับเอาภาพตัวอักษรในเอกสารเข้ามาแล้วจะทำการแบ่งแต่ละบรรทัดออกจากกันด้วยวิธีของการทำ Histogram จากนั้นในแต่ละบรรทัดจะแบ่งออกเป็น 3 ระดับซึ่งจะทำการจัดเข้ากลุ่มแต่ละกลุ่มตามระดับของภาพตัวอักษรนั้น ๆ ได้แก่ กลุ่มตัวอักษรระดับบน กลางและล่างตามลำดับ แล้วใช้วิธี Histogram อีกเช่นกันในการแบ่งภาพตัวอักษรให้ได้ออกมาเป็นกรอบภาพตัวอักษรแต่ละตัว จากนั้นจะผ่านกระบวนการ Thinning กรอบภาพตัวอักษร ขั้นตอนต่อไปคือการแบ่งภาพตัวอักษรที่ Skeleton แล้วออกเป็น 9 ส่วนย่อยที่เท่า ๆ กัน แล้วทำการพิจารณาหาจุดปลายของภาพตัวอักษรและมีการเก็บข้อมูลโดยเริ่มจากจุดปลายที่ 1 ไปจนถึงจุดปลายที่ 2 และหากกรอบส่วนย่อยที่ถูกแบ่งนั้นมีจุดปลายที่สามก็จะอยู่ในกลุ่มของตัวอักษรที่มีจุดปลายมากกว่า 2 จุด ซึ่งค่าที่เก็บนั้นจะมีการแทนรหัสแทนทิศทาง 9 รหัสตามทิศทางที่ทำมุมของเส้นตัวอักษร โดยที่แต่ละกรอบที่ถูกแบ่งย่อย 9 กรอบจะเก็บค่าที่อยู่ในช่วงกรอบของตนเอง และทำการเรียงเก็บข้อมูลต่อไป กล่าวคือรูปแบบ Input Vector จะเก็บเฉพาะส่วนข้อมูลที่เป็นลายเส้นภาพตัวอักษร จากนั้นแต่ละกรอบที่ถูกแบ่งออกจะส่งต่อเข้าสู่กระบวนการรู้จำโดยอาศัยโครงข่ายประสาทเทียมแบบใหม่ที่พัฒนาขึ้นมาโดยเน้นที่บริเวณส่วนไม่มีหัวของตัวอักษรภาษาไทยมาเป็นหลักการเบื้องต้นของแนวคิดในการพัฒนากระบวนการเรียนรู้ทั้งหมดและพยายามที่จะสร้างกระบวนการรู้จำใหม่ที่มีการปรับเปลี่ยนเฉพาะบางส่วนย่อยเพื่อให้เกิดประสิทธิภาพในการเรียนรู้บริเวณตัวอักษรไม่มีหัวที่เป็นจุดเริ่มต้นของการเขียนตัวอักษรไม่มีหัว นอกจากนี้งานวิจัยนี้ได้นำเอาวิธีการวัดความเหมือนหรือความคล้ายคลึงกันที่อาศัยฟังก์ชันครอสคอรีเลชั่น (Cross-correlation) ที่มีประสิทธิภาพและสามารถนำมาเป็นตัววัดค่าความเหมือนของรูปแบบ ส่วนแนวทางในการรู้จำ Fuzzy Adaptive Resonance Theory (Fuzzy ARTMAP) หรือที่เรียกว่าฟัซซี่อาร์ทแมพนั้น มีความสามารถในการเรียนรู้สิ่งใหม่ได้พร้อมกับยังคงความรู้จำเก่าไว้ได้ ดังนั้นงานวิจัยนี้จึงใช้แนวทางฟัซซี่อาร์ทแมพเป็นพื้นฐานของกระบวนการเรียนรู้และพยายามหาวิธีการปรับเอาครอสคอรีเลชั่นมาใช้ร่วมกับแนวทางฟัซซี่อาร์ทแมพให้ได้

1.4 ขอบเขตของการวิจัย

งานวิจัยนี้เน้นการพัฒนากระบวนการรู้จำสำหรับรูปแบบภาพตัวอักษรในเอกสารที่มีขอบเขตต่อไปนี้

1.4.1 ภาพตัวอักษรเอกสารที่ได้มาจากการสแกนเข้าสู่เครื่องคอมพิวเตอร์ ภาพตัวอักษรจะต้องไม่ติดกันทั้งภายใน-นอก ตัวอักษร รวมไปถึงไม่มีการซ้อนทับกันของภาพตัวอักษร และภาพลายเส้นของตัวอักษรจะต้องไม่ขาด

1.4.2 ภาพตัวอักษรเอกสารที่ได้มาจากการสแกนเข้าสู่เครื่องคอมพิวเตอร์ จะต้องประกอบตัวอักษรภาษาไทยไม่มีหัว โดยครอบคลุมตัวอักษรตามตารางที่ 1.2 ต่อไปนี้

ตารางที่ 1.2 ตัวอักษรภาษาไทยฟอนต์ไม่มีหัว

พยัญชนะ 46 ตัว	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ ฤ ฦ
สระ 15 ตัว	ะ ั า ำ ิ ี ึ ื ุ ู เ แ โ ใ ไ
วรรณยุกต์ 7 ตัว	่ ้ ๊ ๋ ็ ์ ๎ ๐
เลขอารบิก 10 ตัว	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
สัญลักษณ์ 2 ตัว	ๆ ฯ

1.5 ขั้นตอนของการศึกษา

1.5.1 ศึกษาบทความและผลงานวิจัยต่างๆ ที่มีความเกี่ยวข้องกับงานวิจัยนี้

1.5.2 เก็บข้อมูลตัวอย่างของตัวอักษร พร้อมจัดเก็บลงคอมพิวเตอร์

1.5.3 ศึกษาลักษณะโครงสร้างของตัวอักษรภาษาไทยเพื่อนำไปวิเคราะห์

1.5.4 ออกแบบอัลกอริทึมในการรู้จำโดยใช้ Hierarchical Cross-Correlation Neural Network

1.5.5 เขียนโปรแกรมเพื่อวิเคราะห์ลักษณะของภาพตัวอักษร และทำการจัดกลุ่มของตัวอักษรในส่วนของกระบวนการก่อนการรู้จำ

1.5.6 เขียนโปรแกรมเพื่อรู้จำภาพตัวอักษร โดยใช้ Hierarchical Cross-Correlation Neural Network

1.5.7 ทดลองรู้จำตัวอักษรกับข้อมูลที่จัดเก็บ

1.5.8 ทดสอบผลจากการรู้จำตัวอักษรว่ามีความถูกต้องเมื่อเทียบกับซอฟต์แวร์ที่ใช้งานจริง ArnThai version 2.5 Lite และ Thai OCR โดยทดสอบข้อมูลก่อนและหลังการรู้จำตัวอักษร

1.5.9 สรุปผลการดำเนินการ และรวบรวมนำจัดทำเอกสารนำเสนอเป็นงานวิจัย

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ทราบถึงบทความงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับงานวิจัยครั้งนี้

1.6.2 ทราบถึงลักษณะเด่นของตัวพิมพ์อักษรภาษาไทยแบบฟอนต์ไม่มีส่วนหัวของตัวอักษร

1.6.3 เป็นพื้นฐานในการรู้จำตัวอักษรภาษาไทยที่สมบูรณ์แบบ เพื่อใช้ในการพัฒนาต่อไปในอนาคตให้เหมาะกับตัวอักษรภาษาไทยทุกรูปแบบ