บทที่ 4
ผลการวิจัย

แหล่งที่มาของรูปภาพเอกสาร

รูปภาพเอกสารที่ใช้ในการทำวิทยานิพนธ์ฉบับนี้มีอยู่ 2 ชุด คือ

1. รูปภาพในชุดเรียนรู้ เป็นรูปภาพของตัวอักษรโดยเก็บไว้ในแฟ้มข้อมูล 1 ตัวต่อหนึ่งแฟ้ม การจัดเก็บรูปภาพของตัวอักษรจัดเก็บโดยใช้รูปแบบ Windows Bitmap (BMP) โดยตัวอักษรที่ใช้ในการวิจัยนี้ใช้รูปแบบของตัวหนังสือที่นิยมใช้กันบนระบบปฏิบัติการวินโดวส์ 98 โดยใช้ 6 รูปแบบคือแบบ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC และ FreesiaUPC แต่ละรูปแบบของตัวอักษรจะประกอบด้วยตัวอักษรขนาด 14, 16, 18, 20, 22, 24, 28 และ 36 จุด ซึ่งแต่ละรูปแบบและแต่ละขนาดจะมีจำนวนตัวอักษร 178 ตัว ดังนี้

   ก ข ฃ ค ฅ ฆ ง จ ฉ ช 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

   ะ า ิ ี ึ ื ุ ู ฺ เ แ โ ใ ไ ๆ ฤ ฤๅ ฦ ฦๅ

   A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

   a b c d e f g h i j k l m n o p q r s t u v w x y z

   ๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙ 0 1 2 3 4 5 6 7 8 9

   ! @ # $ % ^ & * ( ) _ - + = | \ { [}] : ; ' ' < , > . ? /

   และ เพิ่ม ตัวอักษร '!' โดยไม่มี '.' อยู่ด้านล่าง

   เพิ่ม ตัวอักษร '?' โดยไม่มี '.' อยู่ด้านล่าง

   เพิ่ม ตัวอักษร 'i' โดยไม่มี '.' อยู่ด้านบน

   เพิ่ม ตัวอักษร 'j' โดยไม่มี '.' อยู่ด้านบน

   เพิ่ม ตัวอักษร 'ญ' โดยไม่มี สัญลักษณ์พิเศษที่อยู่ด้านล่าง

   เพิ่ม ตัวอักษร 'ฐ' โดยไม่มี สัญลักษณ์พิเศษที่อยู่ด้านล่าง

   เพิ่ม ตัวอักษร สัญลักษณ์พิเศษที่อยู่ด้านล่างของ 'ญ'

   เพิ่ม ตัวอักษร สัญลักษณ์พิเศษที่อยู่ด้านล่างของ 'ฐ'

2. รูปภาพในชุดทดสอบ เป็นรูปภาพของเอกสารโดยเก็บไว้ในแฟ้มข้อมูล 1 หน้าเอกสารต่อหนึ่งแฟ้ม การจัดเก็บรูปภาพของตัวอักษรจัดเก็บโดยใช้รูปแบบ Windows Bitmap (BMP) โดยตัวอักษรที่ใช้ในการวิจัยนี้ใช้รูปแบบของตัวหนังสือที่นิยมใช้กันบนระบบปฏิบัติการวินโดวส์ 98 โดยใช้ 6 รูปแบบคือแบบ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC,

EucrosiaUPC และ FreesiaUPC แต่ละรูปแบบของตัวอักษรจะประกอบด้วยตัวอักษรขนาด 14, 16, 18, 20, 22, 24, 28 และ 36 จุด เป็นจำนวนทั้งหมด 48 หน้า และมีจำนวนตัวอักษรตามตารางที่ 4.1

รูปแบบ	14	16	18	20	22	24	28	36	รวม
Angsana	2634	1767	1834	1184	1428	1117	851	545	11360
Browallia	2773	2076	1639	1533	1425	1208	916	578	12148
Cordia	2534	1775	1715	1477	1367	911	877	532	11188
Dillenia	3106	2147	1831	1591	1637	1107	769	552	12740
Eucrosia	2983	2163	1578	1285	1379	1259	909	401	11957
Freesia	2785	2384	1490	1716	1435	1229	877	523	12439
									71832

ตารางที่ 4.1 แสดงจำนวนตัวอักษรในเอกสารต้นฉบับของข้อมูลในชุดทดสอบ

โดยที่เอกสารทั้งหมดนี้พิมพ์จากโปรแกรม Microsoft Word 97 ด้วยเครื่องพิมพ์แบบเลเซอร์ ที่มีความละเอียด 600 จุดต่อนิ้ว แล้วนำเอกสารที่ได้ไปทำการสแกนด้วยเครื่องสแกนเนอร์ โดยใช้ความละเอียด 300 จุดต่อนิ้ว แล้วทำการจัดเก็บแบบบิตแมพ

วิธีการทดสอบ

การทดสอบนั้นทำได้โดยการใช้รูปภาพในชุดทดสอบ ทั้ง 48 รูป มาทำการรู้จำด้วยโปรแกรมไทยโอซีอาร์ที่ได้จากวิทยานิพนธ์ฉบับนี้ เพื่อหาเปอร์เซ็นต์ตัวอักษรที่ผิดพลาดเมื่อเทียบกับเอกสารต้นฉบับ และเปรียบเทียบผลกับโปรแกรมไทยโอซีอาร์ที่มีใช้กันอยู่ในประเทศไทยอีก 2 โปรแกรมคือ

1. โปรแกรมอ่านไทย 1.0 (สร้างแผ่น Installation เมื่อวันที่ 10/29/97 เวลา 05:15:04 อ้างอิงจากแฟ้มข้อมูล Arnthai.asu ในแผ่น Installation) ของ ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) [20]
2. โปรแกรมไทยโอซีอาร์ 1.5b (Evaluation Copy) ของบริษัท เอเทรี่ยม เทคโนโลยี จำกัด (Atrium Technology Co.,Ltd.) [21]

ผลการทดสอบ

รูปแบบ	14	16	18	20	22	24	28	36	เฉลี่ย
Angsana	14.69%	7.41%	6.00%	6.08%	4.55%	2.86%	1.76%	4.04%	7.34%
Browallia	9.88%	8.86%	6.28%	7.24%	6.67%	3.97%	6.66%	5.02%	7.45%
Cordia	10.66%	8.00%	11.14%	9.68%	9.44%	12.62%	8.21%	9.40%	9.94%
Dillenia	17.48%	13.18%	10.05%	7.92%	8.31%	8.94%	9.10%	13.95%	11.92%
Eucrosia	5.53%	7.12%	5.39%	3.50%	5.22%	2.22%	3.08%	2.00%	4.89%
Freesia	7.86%	11.33%	7.52%	7.05%	7.60%	8.22%	5.25%	7.46%	8.18%
									8.31%

ตารางที่ 4.2 แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรม อ่านไทย 1.0

รูปแบบ	14	16	18	20	22	24	28	36	เฉลี่ย
Angsana	3.42%	1.08%	0.93%	5.41%	5.25%	4.57%	7.87%	10.46%	3.87%
Browallia	2.13%	1.49%	1.59%	3.52%	4.21%	4.55%	9.50%	6.57%	3.38%
Cordia	4.38%	2.59%	3.27%	5.89%	6.73%	8.01%	8.78%	9.21%	5.28%
Dillenia	2.00%	5.17%	2.46%	1.07%	xxx	5.69%	11.70%	12.50%	4.12%
Eucrosia	1.81%	3.93%	2.47%	3.11%	9.43%	9.85%	7.81%	9.73%	4.87%
Freesia	3.52%	5.62%	7.45%	7.58%	6.34%	9.68%	11.63%	9.94%	6.73%
									4.73%

ตารางที่ 4.3 แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรม เอเทรี่ยม ไทย-โอซีอาร์ 1.5 b

xxx เอกสารชุดที่ใช้ตัวอักษรแบบ Dillenia ขนาด 22 จุด ไม่สามารถทำการรู้จำได้โดยโปรแกรม เอเทรี่ยม ไทย-โอซีอาร์ 1.5 b ดังนั้นจึงหาเปอร์เซ็นต์ผิดพลาดเฉลี่ย ของเอกสารที่ใช้ตัวอักษรแบบ Dillenia โดยไม่คิดเปอร์เซ็นต์ผิดพลาดจาก เอกสารชุดที่ใช้ตัวอักษรแบบ Dillenia ขนาด 22 จุด

รูปแบบ	14	16	18	20	22	24	28	36	เฉลี่ย
Angsana	2.35%	0.96%	1.09%	1.10%	0.77%	0.45%	0.47%	0.92%	1.21%
Browallia	2.24%	1.64%	1.22%	1.17%	2.81%	1.90%	1.42%	1.04%	1.78%
Cordia	3.28%	2.31%	2.10%	2.44%	1.83%	2.31%	1.14%	1.50%	2.32%
Dillenia	3.44%	6.38%	1.80%	0.94%	1.71%	1.08%	1.69%	1.63%	2.78%
Eucrosia	2.18%	1.80%	1.52%	0.78%	1.23%	0.08%	1.43%	0.00%	1.41%
Freesia	2.05%	2.14%	1.68%	0.93%	0.91%	1.14%	1.14%	1.53%	1.56%
									1.85%

ตารางที่ 4.4 แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้
(ยังไม่ได้ทำการแก้ไขคำผิดด้วยไตรแกรมของประเภทของคำ)

รูปแบบ	14	16	18	20	22	24	28	36	เฉลี่ย
Angsana	1.63%	2.21%	1.20%	1.10%	0.21%	0.81%	0.47%	0.55%	1.20%
Browallia	1.01%	0.53%	1.04%	1.50%	0.77%	1.16%	0.44%	0.87%	0.93%
Cordia	1.89%	1.07%	1.52%	2.78%	0.59%	0.33%	1.03%	0.75%	1.41%
Dillenia	2.29%	6.24%	2.68%	1.13%	1.89%	1.08%	1.95%	0.91%	2.63%
Eucrosia	3.29%	0.83%	1.39%	0.86%	0.15%	0.08%	1.65%	0.00%	1.40%
Freesia	1.83%	1.51%	1.48%	0.52%	0.63%	0.73%	0.68%	0.76%	1.17%
									1.47%

ตารางที่ 4.5 แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้
(ทำการแก้ไขคำผิดที่ไม่เป็นคำ ด้วยไตรแกรมของประเภทของคำ)

รูปแบบ	14	16	18	20	22	24	28	36	เฉลี่ย
Angsana	1.86%	1.98%	1.20%	1.35%	0.14%	0.81%	0.47%	0.55%	1.23%
Browallia	1.01%	0.53%	1.04%	1.50%	0.77%	1.24%	0.44%	0.87%	0.94%
Cordia	1.89%	0.90%	1.40%	2.91%	0.51%	0.33%	1.14%	0.75%	1.39%
Dillenia	2.29%	6.43%	2.79%	1.19%	1.95%	1.17%	2.08%	0.91%	2.71%
Eucrosia	3.32%	0.92%	1.39%	0.86%	0.15%	0.08%	1.65%	0.25%	1.43%
Freesia	1.83%	1.51%	1.54%	0.52%	0.77%	0.90%	0.68%	0.76%	1.21%
									1.50%

ตารางที่ 4.6 แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้
(ทำการแก้ไขคำผิดที่ไม่เป็นคำและคำผิดที่เป็นคำ ด้วยไตรแกรมของประเภทของคำ)

ทำการทดสอบบนเครื่องไมโครคอมพิวเตอร์ ใช้หน่วยประมวลผลกลาง Pentium MMX ความเร็ว 233 MHz มีหน่วยความจำ 64 MB ที่ใช้ระบบปฏิบัติการ Microsoft Windows 98 ซึ่งผลการทดสอบจะคำนวณเปอร์เซ็นต์จำนวนตัวอักษรที่ผิดพลาดในแต่ละรูปแบบและแต่ละขนาด โดยช่องเฉลี่ยจะคำนวณเปอร์เซ็นต์เฉลี่ยจากจำนวนตัวอักษรที่ผิดพลาดในรูปแบบนั้น และช่องเฉลี่ยสุดท้ายจะคำนวณเปอร์เซ็นต์เฉลี่ยจากจำนวนตัวอักษรที่ผิดพลาดของโปรแกรมโอซีอาร์นั้น (การนับตัวอักษรที่ผิดนั้น ไม่รวมการเพิ่มหรือลดจำนวนเว้นวรรคระหว่างตัวอักษร และ ไม่รวมการเพิ่มหรือลดของจำนวนบรรทัดว่างเช่นกัน)

ผลการเปรียบเทียบ

ผลการทดสอบในตารางที่ 4.2 ถึง 4.6 นั้นสามารถนำมาสรุปเพื่อเปรียบเทียบเปอร์เซ็นต์อักษรที่ผิดพลาดจากผลลัพธ์ของโปรแกรมไทยโอซีอาร์แต่ละตัวได้ ในตารางที่ 4.7

รูปแบบ	อ่านไทย 1.0	เอเทรียม ไทย-โอซีอาร์	A	B	C
Angsana	7.34%	3.87%	1.21%	1.20%	1.23%
Browallia	7.45%	3.38%	1.78%	0.93%	0.94%
Cordia	9.94%	5.28%	2.32%	1.41%	1.39%
Dillenia	11.92%	4.12%	2.78%	2.63%	2.71%
Eucrosia	4.89%	4.87%	1.41%	1.40%	1.43%
Freesia	8.18%	6.73%	1.56%	1.17%	1.21%
เฉลี่ย	8.31%	4.73%	1.85%	1.47%	1.50%

ตารางที่ 4.7 เปรียบเทียบเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด
จากผลลัพธ์ของโปรแกรมโอซีอาร์ทั้ง 3 โปรแกรม

A คือวิทยานิพนธ์ฉบับนี้

B คือวิทยานิพนธ์ฉบับนี้ + แก้คำผิดที่ไม่เป็นคำ

C คือวิทยานิพนธ์ฉบับนี้ + แก้คำผิดที่ไม่เป็นคำและคำผิดที่เป็นคำ

วิเคราะห์ผลการวิจัย

จากผลการวิจัย เปอร์เซ็นต์ตัวอักษรที่ผิดพลาดจากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้ โดยที่ยังไม่ได้แก้ไขคำผิดด้วยวิธีการไตรแกรมของประเภทของคำ คือ 1.85% และหลังจากแก้ไขคำผิดที่ไม่เป็นคำ ด้วยไตรแกรมของประเภทของคำแล้ว จะมีเปอร์เซ็นต์ความผิดพลาดเท่ากับ 1.47% แต่พบว่าหลังจากแก้ไขคำผิดที่ไม่เป็นคำและคำผิดที่เป็นคำ ด้วยไตรแกรมของประเภทของคำแล้ว กลับมีความผิดพลาดเท่ากับ 1.50% ซึ่งสูงกว่าผลลัพธ์ของการแก้ไขคำผิดที่ไม่เป็นคำเพียงอย่างเดียว เนื่องจากเอกสารที่เป็นชุดทดสอบนั้นมีคำทับศัพท์ภาษาอังกฤษและชื่อเฉพาะอยู่มาก อีกทั้งยังมีคำในภาษาพูดอยู่ในเอกสารอีกด้วย ซึ่งคำเหล่านั้นไม่มีอยู่ในพจนานุกรม ซึ่งการแก้ไขคำผิดที่เป็นคำ จะแก้ไขคำเหล่านั้นให้กลายเป็นคำอื่นไป

เปอร์เซ็นต์ตัวอักษรที่ผิดพลาดจากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้ ของเอกสารที่ใช้รูปแบบตัวอักษรแบบ Cordia และ Dillenia นั้นสูงกว่าเอกสารที่ใช้รูปแบบตัวอักษรแบบอื่น เป็นเพราะตัวอักษรแบบ Cordia และ Dillenia นั้นมักจะติดกับตัวอักษรอื่นที่อยู่ข้างเคียง

โดยเฉพาะตัวอักษรแบบ Dillenia บางครั้งเกิดการติดกันของตัวอักษรมากกว่า 2 ตัว ซึ่งไม่สามารถแก้ไขคำผิดเหล่านั้นด้วยวิธีการแก้คำผิดในวิทยานิพนธ์ฉบับนี้

อย่างไรก็ตามเปอร์เซ็นต์ตัวอักษรที่ผิดพลาดจากผลลัพธ์ของโปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้ทั้ง 3 วิธีก็ยังมีค่าต่ำกว่า โปรแกรมอ่านไทย ซึ่งมีเปอร์เซ็นต์ผิดพลาด 8.31% และโปรแกรมเอเทรี่ยม ไทย-โอซีอาร์ ซึ่งมีเปอร์เซ็นต์ผิดพลาด 4.73%

ปัญหาและข้อจำกัด

1. ปัญหาเรื่องความเร็วในการรู้จำ ในการทดสอบรู้จำข้อความในเอกสารหนึ่งหน้ากระดาษขนาด A4 ที่มีตัวอักษรขนาด 16 จุด จำนวนประมาณ 1800 ตัวอักษร ใช้เวลาในการรู้จำ ประมาณ 4 นาที 30 วินาที (บนเครื่องที่ใช้หน่วยประมวลผลกลาง Pentium MMX ความเร็ว 233 MHz มีหน่วยความจำ 64 Mb) ซึ่งช้ากว่าโปรแกรมอ่านไทย และโปรแกรมเอเทรี่ยม ไทย-โอซีอาร์ ทั้ง 2 โปรแกรมใช้เวลาประมาณ 1 นาที
2. ปัญหาเรื่องการรู้จำเอกสารที่ไม่ชัด รูปเอกสารที่จะใช้ในการรู้จำจะต้องมีความชัดเจน โดยที่จะต้องไม่มี จุดภาพรบกวนขนาดใหญ่และรูปตัวอักษรจะต้องไม่บางจนเส้นของตัวอักษรขาดออกจากกัน ซึ่งทำให้ผลการรู้จำตัวอักษรผิดพลาดสูงขึ้น
3. ข้อจำกัดเรื่องการตัดแยกตัวอักษร โปรแกรมโอซีอาร์ในวิทยานิพนธ์ฉบับนี้ไม่สามารถตัดแยกตัวอักษรที่ติดกันให้แยกออกจากกันได้ ซึ่งทำให้ผลการรู้จำผิดพลาดสูง
4. ข้อจำกัดในการรู้จำเอกสารที่มีตัวอักษรหลายขนาดในบรรทัดเดียวกัน เนื่องจากโปรแกรมโอซีอาร์นี้จะต้องทำการคำนวณหาเส้นแบ่งระดับ ด้วยความสูงของตัวอักษรในบรรทัดนั้นๆ ซึ่งจะคำนวณผิดพลาดหากมีตัวอักษรหลายขนาดอยู่ในบรรทัดเดียวกัน และจะทำให้รู้จำตัวอักษรผิด
5. ข้อจำกัดในการกำหนดขอบเขตของเอกสารในการรู้จำ โปรแกรมโอซีอาร์นี้ไม่สามารถกำหนดขอบเขตของเอกสารในการรู้จำได้ เช่นโปรแกรมโอซีอาร์นี้ไม่สามารถรู้จำเอกสารที่มีรูปปรากฏอยู่ได้ ซึ่งทำให้ไม่สะดวกต่อการใช้งาน
6. ข้อจำกัดในการหมุนภาพของเอกสาร ในบางครั้งผู้ใช้งานไม่สามารถกำหนดค่าองศาความเอียงให้แก่เอกสารได้ถูกต้อง ซึ่งควรจะมีฟังก์ชันคำนวณหาองศาความเอียงด้วยเครื่อง เพื่อความสะดวกต่อการใช้งาน
7. ข้อจำกัดในการแก้ไขคำผิดที่เป็นภาษาอังกฤษ เนื่องจากการแก้ไขคำผิดในโปรแกรมโอซีอาร์นี้ไม่สามารถแก้ไขคำผิดที่เป็นภาษาอังกฤษได้
8. ข้อจำกัดของการแก้ไขคำผิด เนื่องจากพจนานุกรมมีคำน้อยเกินไป โดยเฉพาะคำทับศัพท์, ชื่อเฉพาะและคำในภาษาพูด