

การออกแบบและพัฒนาโปรแกรมไอซีอาร์ภาษาไทย

นายชาญฤทธิ์ สันตินานาเลิศ



สถาบันวิจัยบริการ

จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

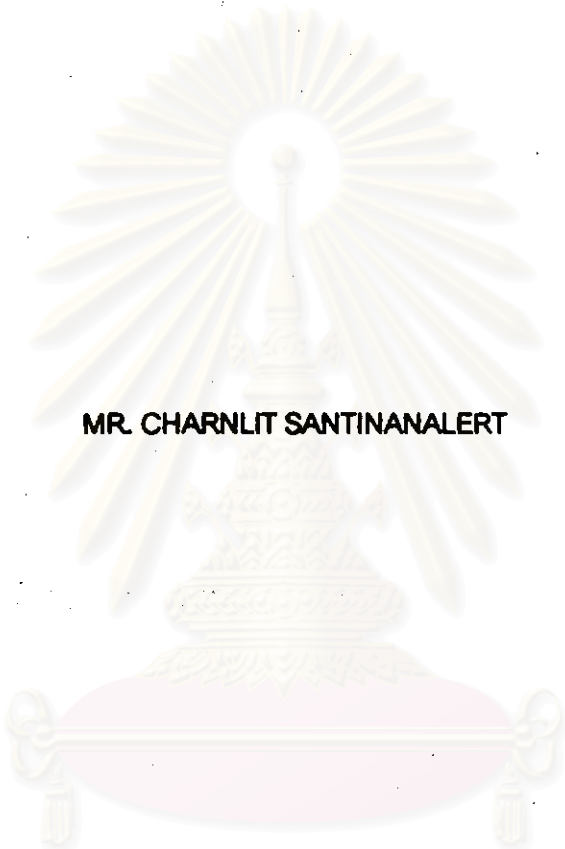
ปีการศึกษา 2542

ISBN 974-333-872-1

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

I19184323

DESIGN AND DEVELOPMENT OF A THAI-OCR PROGRAM



MR. CHARNLIT SANTINANALERT

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย
A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 1999

ISBN 974-333-872-1

ชาญฤทธิ์ สันตินานาเลิศ : การออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย. (DESIGN AND DEVELOPMENT OF A THAI-OCR PROGRAM) อ. ที่ปรึกษา : ดร.บุญเสริม กิจศิริกุล, 136 หน้า. ISBN 974-333-872-1.

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย เพื่อใช้ในการรู้จำตัวอักษรพิมพ์ในเอกสารภาษาไทยที่พิมพ์จากเครื่องคอมพิวเตอร์ด้วยแบบตัวอักษรมาตรฐาน วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการต่างๆ เพื่อใช้ในโปรแกรมโอซีอาร์ภาษาไทยคือ วิธีการประมวลผลภาพ, วิธีการตัดแยกตัวอักษร, วิธีการแยกลักษณะสำคัญของตัวอักษรแบบ เค-แอล ทรานส์ฟอร์ม, วิธีการแยกแยะตัวอักษรแบบแคพรอพาเกชันนิวโรลเน็ตเวิร์กและวิธีการแก้ไขค่าที่สะกดผิดแบบไตรแกรมของประเภทของคำ

ขั้นตอนในการทำงานของโปรแกรมโอซีอาร์ภาษาไทยที่พัฒนาขึ้นนี้ประกอบด้วย ขั้นตอนการนำเอกสารเข้าสู่โปรแกรม, ขั้นตอนการประมวลผลภาพ, ขั้นตอนการตัดแยกบรรทัด, ขั้นตอนการตัดแยกตัวอักษร, ขั้นตอนการรู้จำตัวอักษร, ขั้นตอนการแก้ไขผลลัพธ์ที่ได้จากขั้นตอนการรู้จำ, ขั้นตอนการสร้างบรรทัดและขั้นตอนการแก้ไขคำผิด

ในวิทยานิพนธ์ฉบับนี้ ได้นำภาพตัวอักษรและภาพของเอกสารที่ได้จากการพิมพ์ด้วยเครื่องพิมพ์เลเซอร์ที่ความละเอียด 600 จุดต่อนิ้ว นำเอกสารมาอ่านผ่านเครื่องสแกนเนอร์ที่ความละเอียด 300 จุดต่อนิ้ว ซึ่งประกอบด้วยตัวอักษรแบบ AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC และ FreesiaUPC แต่ละแบบประกอบด้วยตัวอักษรขนาด 14, 16, 18, 20, 22, 24, 28 และ 36 จุด โดยในการเรียนรู้นั้นใช้ภาพของตัวอักษรจำนวน 8544 ตัวอักษร และในการทดสอบการรู้จำใช้ภาพของเอกสารจำนวน 48 เอกสาร ซึ่งประกอบด้วยตัวอักษรจำนวน 71832 ตัวอักษร ได้ผลการรู้จำซึ่งยังไม่ได้แก้ไขคำผิดมีความผิดพลาดเฉลี่ยร้อยละ 1.85 ผลการรู้จำหลังจากแก้ไขคำผิดที่ไม่เป็นคำแล้วมีความผิดพลาดเฉลี่ยร้อยละ 1.47 และผลการรู้จำหลังจากแก้ไขคำผิดที่ไม่เป็นคำและคำผิดที่เป็นคำแล้วมีความผิดพลาดเฉลี่ยร้อยละ 1.50

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2542

ลายมือชื่อนิสิต
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

4070254721 : MAJOR COMPUTER SCIENCE

KEY WORD: CHARACTER RECOGNITION / CHARACTER SEGMENTATION /
BACKPROPAGATION / NEURAL NETWORKS / ERROR CORRECTION /
NON-WORD ERROR / REAL-WORD ERROR

CHARNLIT SANTINANALERT : DESIGN AND DEVELOPMENT OF A THAI-OCR
PROGRAM. THESIS ADVISOR : BOONSERM KIJSIRIKUL, Ph.D., 136 pp.
ISBN 974-333-872-1.

The objective of this thesis is to design and develop Thai-Optical Character Recognition (Thai-OCR) for recognizing printed characters in Thai documents, which are printed from a computer with standard fonts. The thesis employs several methods for Thai-OCR that are image pre-processing, character segmentation, K-L transform for feature extraction, backpropagation neural networks for character classification and part of speech trigram (pos trigram) for error correction.

The process of the developed Thai-OCR program is composed of image acquisition, image processing, line segmentation, character segmentation, character recognition, character correction, text line reconstruction and error correction.

In this thesis, character and document images are generated from a laser printer at 600 dots per inch and then are scanned with a scanner at 300 dots per inch. They compose of characters in 6 fonts: AngsanaUPC, BrowalliaUPC, CordiaUPC, DilleniaUPC, EucrosiaUPC and FreesiaUPC each font composed of size 14, 16, 18, 20, 22, 24, 28 and 36 points. In training process 8544 characters are used and in testing process 48 documents composed of 71832 characters are used. The error rate of recognition without error correction technique is 1.85%, the error rate of recognition with non-word error correction is 1.47% and the error rate of recognition with both non-word and real-word error correction is 1.50%.

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2542

ลายมือชื่อนิสิต *C.Santana*
ลายมือชื่ออาจารย์ที่ปรึกษา *Boonserm*
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จส่งไปได้ด้วยความช่วยเหลืออย่างดีของ อาจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ให้แนวทาง ข้อเสนอแนะ และคำปรึกษา ตลอดจน ขอรอบคุณ คุณธเนศ ศรีวิรุฬห์ชัย และ คุณอนันต์ลดา ไรติมงคล ที่ได้คำปรึกษา และให้ความช่วยเหลืออย่างดี ขอรอบคุณ คุณชวลิต ฉันทชัยสิริเวทย์ ที่ให้ความอนุเคราะห์ให้ใช้เครื่องสแกนเนอร์ในการทำวิทยานิพนธ์ ขอรอบคุณสมาชิกห้องปฏิบัติการ Machine Intelligence and Knowledge Discovery Laboratory ที่ให้ข้อเสนอแนะ ตลอดจนเพื่อนๆ พี่ๆ น้องๆ ที่คอยให้กำลังใจมาโดยตลอด

ท้ายที่สุดนี้ ผู้วิจัยใคร่กราบขอบพระคุณ บิดา มารดา ซึ่งให้การสนับสนุนและเป็นกำลังใจให้แก่ผู้วิจัยเสมอมา



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

บทที่	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญตาราง.....	ญ
สารบัญภาพ.....	ฉ
บทที่	
1. บทนำ.....	1
ความเป็นมาและความสำคัญของปัญหา.....	1
งานวิจัยและทฤษฎีที่เกี่ยวข้อง.....	2
วัตถุประสงค์.....	10
ขอบเขตของวิทยานิพนธ์.....	10
ขั้นตอนการทำวิทยานิพนธ์.....	11
ประโยชน์ที่คาดว่าจะได้รับ.....	11
2. แนวคิดและทฤษฎี.....	12
ค่าของจุดภาพในการประมวลผลรูปภาพ.....	12
การพิจารณาจุดภาพที่อยู่ติดกันในในการประมวลผลรูปภาพ.....	12
การแปลงระดับความเข้มสีของรูปภาพจากหลายระดับเป็นสองระดับ.....	12
การลดจุดภาพรบกวน.....	13
การหมุนภาพ.....	13
การกลับภาพ.....	14
การกลับสีของจุดภาพ.....	14
แผนภาพการทำงานของการรู้จำตัวอักษร.....	14
ข้อมูลภาพ.....	15
การเปลี่ยนขนาดภาพ.....	15
การหาค่าเมตริกซ์ค่าเฉลี่ยและเมตริกซ์ของไอเกนเวกเตอร์.....	16
การแปลงแบบเค-แอล.....	16
การใช้นิวรอลเน็ตเวิร์กแบบแบคพรอพาเกชัน.....	17

สารบัญ (ต่อ)

บทที่	หน้า
การทำงานของนิรอลเน็ตเวิร์กแบบแบคพรอพาทาเกชันในขั้นตอนการรู้จำ.....	18
การทำงานของนิรอลเน็ตเวิร์กแบบแบคพรอพาทาเกชันในขั้นตอนการเรียนรู้.....	19
การแก้ไขคำผิดที่เกิดจากโปรแกรมไฮซีอาร์ด้วยวิธีการไตรแกรมของประเภทของคำ.....	21
การหาค่าสถิติของความผิดพลาดที่เกิดขึ้นจากโปรแกรมไฮซีอาร์.....	22
ขั้นตอนในการแก้ไขคำผิดที่เกิดจากโปรแกรมไฮซีอาร์ภาษาไทย.....	24
ระดับของตัวอักษรภาษาไทย.....	25
การหาเส้นแบ่งระดับตัวอักษรภาษาไทย.....	26
3. การออกแบบและพัฒนา.....	28
การทำงานของโปรแกรมไทยไฮซีอาร์.....	28
โครงสร้างของข้อมูลในส่วนของกระบวนการประมวลผลรูปภาพ.....	30
การลดจุดภาพรบกวน.....	31
การหมุนภาพ.....	31
การกลับภาพ.....	31
อัลกอริทึมในกระบวนการตัดแยกบรรทัด.....	32
อัลกอริทึมในกระบวนการตัดแยกตัวอักษร.....	33
การรู้จำตัวอักษร.....	34
การแก้ไขผลลัพธ์ที่ได้จากขั้นตอนการรู้จำตัวอักษร.....	36
การสร้างบรรทัด.....	36
การแก้ไขคำผิดที่เกิดจากโปรแกรมไฮซีอาร์ภาษาไทย.....	37
4. ผลการวิจัย.....	38
แหล่งที่มาของรูปภาพเอกสาร.....	38
วิธีการทดสอบ.....	39
ผลการทดสอบ.....	40
ผลการเปรียบเทียบ.....	42
วิเคราะห์ผลการวิจัย.....	43
ปัญหาและข้อจำกัด.....	44
5. สรุปผลการวิจัยและข้อเสนอแนะ.....	45
สรุปผลการวิจัย.....	45
ข้อเสนอแนะ.....	45

สารบัญ (ต่อ)

บทที่	หน้า
รายการอ้างอิง.....	46
ภาคผนวก.....	48
ภาคผนวก ก. การใช้งานโปรแกรมไทยโอซีอาร์.....	48
ภาคผนวก ข. การใช้งานโปรแกรมต่างๆ ในขั้นตอนการเรียนรู้.....	57
ภาคผนวก ค. ตัวอย่างที่ใช้ในการเรียนรู้.....	62
ภาคผนวก ง. เอกสารที่ใช้ในการทดสอบการรู้จำ.....	87
ประวัติผู้เขียนวิทยานิพนธ์.....	136



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญตาราง

ตาราง	หน้า
2.1	แสดงกลุ่มของตัวอักษรโดยแบ่งตามตัวอักษรที่มีผลกระทบต่อความผิดพลาด..... 23
2.2	แสดงกลุ่มของตัวอักษรที่ปรากฏอยู่ในระดับต่างๆ..... 26
3.1	แสดงผลการทดลองของ network ขนาดต่างๆ และจำนวนความผิดพลาดที่เกิดขึ้น..... 35
4.1	แสดงจำนวนตัวอักษรในเอกสารต้นฉบับของข้อมูลในชุดทดสอบ..... 39
4.2	แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของโปรแกรม อ่านไทย 1.0..... 40
4.3	แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของ โปรแกรม เอเทรียม ไทย-ไอซีอาร์ 1.5 b..... 40
4.4	แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของ โปรแกรม ไอซีอาร์ในวิทยานิพนธ์ฉบับนี้ (ยังไม่ได้ทำการแก้ไขคำผิดด้วยโปรแกรมของประเภทของคำ) 41
4.5	แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของ โปรแกรม ไอซีอาร์ในวิทยานิพนธ์ฉบับนี้ (ทำการแก้ไขคำผิดที่ไม่เป็นคำ ด้วยโปรแกรมของประเภทของคำ) 41
4.6	แสดงเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของ โปรแกรม ไอซีอาร์ในวิทยานิพนธ์ฉบับนี้ (ทำการแก้ไขคำผิดที่ไม่เป็นคำและคำผิดที่เป็นคำ ด้วยโปรแกรมของประเภทของคำ)... 42
4.7	เปรียบเทียบเปอร์เซ็นต์ตัวอักษรที่ผิดพลาด จากผลลัพธ์ของโปรแกรมไอซีอาร์ทั้ง 3 โปรแกรม..... 43

สารบัญภาพ

ภาพประกอบ	หน้า
1.1 แสดงแบบจำลองของโปรแกรมโอซีอาร์.....	2
1.2 แสดงจุดภาพรวมกวนที่ปรากฏในภาพอาจทำให้เกิดความผิดพลาดขึ้นได้.....	3
1.3 แสดงการกลับภาพในแนวนอนและการกลับภาพในแนวตั้งตามลำดับ.....	4
1.4 แสดงการกลับสีของจุดภาพ.....	4
1.5 แสดงแบบจำลองของการตัดแยกตัวอักษร.....	5
1.6 แสดงรูปที่ได้จากการตัดแยกตัวอักษรแบบ Dissection.....	5
1.7 แสดงการแยกตัวอักษรโดยใช้ความรู้จากขั้นตอนการรู้จำตัวอักษร.....	5
1.8 แสดงแบบจำลองของการรู้จำตัวอักษร.....	7
2.1 แสดงจุดภาพ 9 จุด โดยจุดที่สนใจคือจุด P5.....	12
2.2 แสดงการกำหนดค่าระดับความเข้มของภาพ เมื่อต้องการรูปที่มี จำนวนระดับความเข้มเพียง 2 ระดับ.....	13
2.3 แสดงแผนภาพการทำงานของการรู้จำตัวอักษรพินท์ภาษาไทย โดยใช้เทคนิคด้านการวิเคราะห์ตัวประกอบสำคัญและนิเวศเน็ตเวิร์ก.....	15
2.4 ลักษณะการแทนจุดภาพด้วยเวกเตอร์.....	15
2.5 แสดงตัวอย่างของนิเวศเน็ตเวิร์ก.....	18
2.6 แสดงจำนวนจุดดำในแต่ละแถวตามแนวนอนของรูปภาพบรรทัด.....	26
3.1 แผนภาพแสดงการทำงานของโปรแกรมไทยโอซีอาร์.....	28
3.2 แสดงระนาบของการแสดงผล.....	30
3.3 แสดงระนาบของกราฟ.....	30
3.4 ก แสดงภาพและขนาดของภาพก่อนการหมุนภาพ.....	31
3.4 ข แสดงภาพหลังจากหมุนแล้ว แต่ยังใช้ขนาดภาพเท่าเดิม.....	31
3.4 ค แสดงภาพหลังจากหมุนแล้ว และเปลี่ยนขนาดภาพใหม่.....	31
3.5 ก แสดงวิธีหาแก่นซ้ายของเอกสาร.....	33
3.5 ข แสดงวิธีหาแก่นขวาของเอกสาร.....	33
3.5 ค แสดงวิธีหาบรรทัดในเอกสาร.....	33
3.6 แสดงวิธีหาตัวอักษรในบรรทัด.....	33
ก.1 แสดงหน้าจอหลักของโปรแกรมไทยโอซีอาร์.....	49
ก.2 แสดงคำสั่งย่อยภายใต้รายการคำสั่ง File.....	50

สารบัญภาพ (ต่อ)

ภาพประกอบ	หน้า
ก.3 แสดงคำสั่งย่อภายใต้รายการคำสั่ง Image.....	51
ก.4 แสดงคำสั่งย่อภายใต้รายการคำสั่ง OCR.....	51
ก.5 หน้าจอ Threshold.....	53
ก.6 หน้าจอ Rotate.....	54
ก.7 หน้าจอ Recognition Result โดยที่ ข้อความด้านบนแสดงผลของการรู้จำ และข้อความด้านล่างแสดงผลของการแก้ไขคำผิดจากผลลัพธ์ของการรู้จำ.....	55



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย