การออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย

นายชาญฤทธิ์ สันตินานาเลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2542

ISBN 974-333-872-1

DESIGN AND DEVELOPMENT OF A THAI-OCR PROGRAM

MR. CHARNLIT SANTINANALERT

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

หัวข้อวิทยานิพนธ์	การออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย
โดย	นายชาญฤทธิ์ สันตินานาเลิศ
ภาควิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	อาจารย์ ดร.บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร.สมศักดิ์ ปัญญาแก้ว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร.วันชัย ริ้วไพบูลย์)

..... อาจารย์ที่ปรึกษา

(อาจารย์ ดร.บุญเสริม กิจศิริกุล)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.สาธิต วงศ์ประทีป)

..... กรรมการ

(อาจารย์ ดร.สืบสกุล พิภพมงคล)

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

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาโปรแกรมโอซีอาร์ภาษาไทย เพื่อใช้ในการรู้จำตัวอักษรพิมพ์ในเอกสารภาษาไทยที่พิมพ์จากเครื่องคอมพิวเตอร์ด้วยแบบตัวอักษรมาตราฐาน วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการต่างๆ เพื่อใช้ในโปรแกรมโอซีอาร์ภาษาไทยคือ วิธีการประมวลผลภาพ, วิธีการตัดแยกตัวอักษร, วิธีการแยกลักษณะสำคัญของตัวอักษรแบบ เค-แอล ทรานส์ฟอร์ม, วิธีการแยกแยะตัวอักษรแบบแบคพรอพาเกชันนิวรอลเน็ตเวิร์กและวิธีการแก้ไขคำที่สะกดผิดแบบไตรแกรมของประเภทของคำ

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 ลายมือชื่อนิสิต .....
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2542 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 ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2542 ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ อาจารย์ ดร.บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ให้แนวทาง ข้อเสนอแนะ และคำปรึกษาตลอดมา ขอขอบคุณ คุณธเนศ ศรีวิรุฬห์ชัย และ คุณอนันต์ลดา โชติมงคล ที่ให้คำปรึกษา และให้ความช่วยเหลืออย่างดี ขอขอบคุณ คุณชวลิต อันชัยสิริเวทย์ ที่ให้ความอนุเคราะห์ให้ใช้เครื่องสแกนเนอร์ในการทำวิทยานิพนธ์ ขอขอบคุณสมาชิกห้องปฏิบัติการ Machine Intelligence and Knowledge Discovery Laboratory ที่ให้ข้อเสนอแนะ ตลอดจนเพื่อนๆ พี่ๆ น้องๆ ที่คอยให้กำลังใจมาโดยตลอด

ท้ายที่สุดนี้ ผู้วิจัยใคร่กราบขอบพระคุณ บิดา มารดา ซึ่งให้การสนับสนุนและเป็นกำลังใจให้แก่ผู้วิจัยเสมอมา

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

ภาพประกอบ หน้า