

การราย้คอมพิวเตอร์ตรวจรู้อักขระภาษาไทย



นายสมทิพ พรพนมชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2529

ISBN 974-567-193-2

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

012099

I 16/16A5

THAI CHARACTER RECOGNITION BY COMPUTER

Mr. Chomtip Pornpanomchai, 1915A-

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science

Department of Computer Engineering

Graduate School

1986

ISBN 974-567-193-2

หัวข้อวิทยานิพนธ์

การใช้คอมพิวเตอร์ตรวจรู้อักขระภาษาไทย

โดย

นายชมทิพ พรพนมชัย

ภาควิชา

วิศวกรรมคอมพิวเตอร์

อาจารย์ที่ปรึกษา

ดร.ศุภชัย ตั้งวงศ์ศานต์

ผู้ช่วยศาสตราจารย์ สุเมธ วัชรชัยสุรพล



บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้
เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

.....
(ศาสตราจารย์ ดร.ถาวร วัชรเกษม)

คณะกรรมการสอบวิทยานิพนธ์

.....

..... ประธานกรรมการ
(ดร.ยรรยง เต็งอำนวยการ)

..... กรรมการ
(ดร.ศุภชัย ตั้งวงศ์ศานต์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ สุเมธ วัชรชัยสุรพล)

..... กรรมการ
(ดร.จารุมาตร ปิ่นทอง)

หัวข้อวิทยานิพนธ์
ชื่อนิสิต
อาจารย์ที่ปรึกษา
ภาควิชา
ปีการศึกษา

การใช้คอมพิวเตอร์ตรวจรู้อักขระภาษาไทย
นายสมทิพ พรพนมชัย
ดร. ศุภชัย ตั้งวงศ์สานต์
ผู้ช่วยศาสตราจารย์ สุเมธ วัชรชัยสุรพล
วิศวกรรมคอมพิวเตอร์
2529

บทคัดย่อ



ในวิทยานิพนธ์นี้มีวัตถุประสงค์ที่จะศึกษาและพัฒนาวิธีการตรวจรู้อักขระภาษาไทย ซึ่งอยู่ในรูปแบบของตัวพิมพ์ดีด วิธีการและเทคนิคในการตรวจรู้อักขระภาษาไทยที่ใช้ในการวิจัยสามารถนำมาประยุกต์ใช้กับอุปกรณ์หรือเครื่องมือต่างๆ ได้เช่น เครื่องอ่านข้อมูล (optical character reader) เป็นต้น ซึ่งจะทำให้เราสามารถนำเครื่องมือชนิดนี้อ่านข้อความในเอกสารจากต้นฉบับได้โดยตรง อันเป็นการลดขั้นตอนและเวลาในการเตรียมข้อมูลเพื่อใช้ในการประมวลผลด้วยเครื่องคอมพิวเตอร์และยังเป็นการลดความผิดพลาดในการเตรียมข้อมูลได้อีกด้วย

อักขระที่ใช้ในการตรวจรู้จะสมมุติให้มีการจัดเตรียมอยู่ในรูปของภาพบิตเมทริกซ์ตัวอักษร (bit image matrix) ที่มีขนาด 20 X 20 หน่วย โดยอักขระเหล่านี้จะปราศจากสัญญาณรบกวน (noise free) ในเบื้องต้นได้สร้างอักขระภาษาไทยในรูปตัวพิมพ์จำนวน 5 ชุด ไว้ใช้เป็นต้นแบบในการทاملักษณ์ของการตรวจรู้อักขระ โดยที่วิธีการพัฒนาระบบการตรวจรู้อักขระนี้จะแบ่งเป็นขั้นตอนดังนี้

1. ขั้นตอนการเปลี่ยนภาพบิตแมทริกซ์ตัวอักษรให้เป็นโครงร่างของอักขระ (skeletal form) ขั้นตอนนี้อาจเรียกว่าเป็นขั้นตอนการลดความหนา (thinning process) ของภาพบิตแมทริกซ์ตัวอักษรที่เตรียมไว้ก่อนการนำเข้ามาใช้ในการตรวจรู้อักขระ
2. ขั้นตอนการเปลี่ยนโครงร่างของอักขระให้อยู่ในรูปของรหัส ซึ่งรหัสเหล่านี้จะหมายถึงลักษณะของตัวอักขระตามแนวแถวและแนวสดมภ์
3. ขั้นตอนการสร้างความสัมพันธ์ระหว่างรหัสในชั้นที่ 2 กับอักขระต้นแบบที่สร้างไว้เพื่อการตรวจรู้ โดยความสัมพันธ์นี้จะมีโครงสร้างเป็นรูปต้นไม้ที่เรียกว่า "recognizer tree"

จากการพัฒนาวิธีการตรวจรู้อักขระนี้ได้เขียนเป็นชุดคำสั่ง (program) และทดสอบชุดคำสั่งบนเครื่องไมโครคอมพิวเตอร์ โดยใช้ไมโครโพรเซสเซอร์ 8088 ภายใต้ชุดคำสั่งที่ใช้จัดระบบงาน MS-DOS 3.1 และยังใช้โพรเซสเซอร์ร่วม (co-processor) 8087 เข้าร่วมด้วย

ผลการวิจัยพบว่าเทคนิคการตรวจรู้อักขระที่พัฒนาขึ้นสามารถตรวจรู้อักขระต้นแบบได้ทุกรูป ด้วยอัตราความเร็วเฉลี่ย 72 ตัวอักษรต่ออนาที จากนั้นได้ทำการทดสอบอักขระเพิ่มเติมโดยการสร้างอักขระชุดใหม่เพิ่มอีกจำนวน 2 ชุด ผลปรากฏว่าสามารถตรวจรู้ได้ประมาณ 70% ของอักขระทั้งหมด

Thesis Title THAI CHARACTER RECOGNITION BY COMPUTER
 Name Mr. Chomtip Pornpanomchai
 Thesis Advisors Dr. Supachai Tangwongsan
 Assistant Professor Sumet Vacharachaisurapol
 Department Computer Engineering
 Academic Year 1986



ABSTRACT

This thesis is concerned with the development of a technique for recognizing Thai characters, particularly in printed form. The technique can be applied to devices, such as optical character reader, so that they are capable of reading printed source documents. Thus, time and error can be greatly reduced in the stage of data preparation for computer input.

It is assumed that input characters are clearly separated, noise free and pre-processed in the form of bit image matrix with the size of 20 X 20. In the proposed technique, we preliminarily use five sets of Thai printed characters as basic patterns to develop the scheme of recognizer. The technique essentially consists of the following procedures:

First, transform each bit image matrix which represents a base character into a skeletal form. This step is so-called thinning process. Next, convert the skeletal form into a string of codes, which represents the pattern characteristics of rows and columns. Finally, establish links between base characters and their code strings, a recognizer tree is then formed in this step.

The technique is programmed and implemented on 16-bit 8088 based microcomputer. The machine is running under MS-DOS 3.1 operating system and also with 8087 co-processor. It is found that the proposed technique is able to recognize all base characters with the speed of 72 characters per minute. Further test is performed by preparing two new sets of printed Thai characters, the result shows that only 70% of correct recognition is obtained.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ ดร.ศุภชัย ตั้งวงศ์ศานต์ และผู้ช่วยศาสตราจารย์ สุเมธ วัชรชัยสุรพล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านทั้งสองได้กรุณาสละเวลาอันมีค่าของท่านให้คำแนะนำและข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดี

การวิจัยครั้งนี้ได้ใช้อุปกรณ์บริภัณฑ์ของสำนักคอมพิวเตอร์ มหาวิทยาลัยมหิดล โดยมี คุณอรรรณพ สถิรปัญญา ผู้ช่วยในการจัดเตรียมอุปกรณ์บริภัณฑ์สำหรับงานวิจัย คุณปรีชา งามเสาวรส ผู้ให้คำปรึกษาด้านชุดคำสั่งภาษาแอสเซมบลี ตลอดจน คุณกอบกุล आयुวัฒน์ และคุณศรารุติ ทรงเจริญ ผู้ให้คำแนะนำด้านการใช้ภาษาสำหรับการเขียนวิทยานิพนธ์ ผู้เขียนจึงขอขอบพระคุณมา ณ ที่นี้

สมทิพ พรพนมชัย



สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	ฉ
กิตติกรรมประกาศ	ช
รายการตารางประกอบ	ฉ
รายการรูปประกอบ	ฉ
บทที่	
1. บทนำ	1
1.1 ความเป็นมาของปัญหา	1
1.2 วัตถุประสงค์ของการวิจัย	1
1.3 ขอบเขตของงานวิจัย	2
1.4 วิธีดำเนินการวิจัย	10
2. แนวความคิดเกี่ยวกับการตรวจรู้อักขระ	11
2.1 การรับภาพของมนุษย์	11
2.2 ระบบการตรวจรู้อักขระ	12
2.3 ประวัติการค้นคว้าเกี่ยวกับการตรวจรู้อักขระภาษาไทย	16
2.4 เทคนิคการตรวจรู้อักขระ	18
3. การตรวจรู้อักขระภาษาไทย	21
3.1 โครงสร้างของอักขระและคำภาษาไทย	21
3.2 สมาชิกของภาพบิตแมทริกซ์ตัวอักษร	22
3.3 การแปลงภาพบิตแมทริกซ์ตัวอักษรให้เป็นโครงร่าง ..	22
3.4 การกำหนดรหัสให้อักขระแต่ละรูป	27
3.5 การค้นหารหัสต้นแบบเพื่อการตรวจรู้	30
3.6 ผลลัพธ์ที่ได้จากการตรวจรู้อักขระ	36

บทที่	หน้า
4. การพัฒนาระบบการตรวจรู้อักขระ	38
4.1 ฝั่งงานของระบบการตรวจรู้อักขระ	38
4.2 ลักษณะของจอภาพที่แสดงรายการต่าง ๆ	38
4.3 การสับเปลี่ยนข้อมูลภายในหน่วยความจำหลัก	40
4.4 การตรวจรู้อักขระภาษาไทย	49
4.5 การแสดงผลของการตรวจรู้อักขระ	49
5. สรุปผลการวิจัยและ ข้อเสนอแนะ	51
5.1 ผลการวิจัย	51
5.2 สรุปผลการวิจัย	53
5.3 ข้อเสนอแนะ	54
เอกสารอ้างอิง	56
ภาคผนวก ก. ตารางแสดงรหัสต้นแบบของอักขระแต่ละรูป	58
ข. ฝั่งงานแสดงส่วนของชุดคำสั่งที่ใช้ในการตรวจรู้อักขระ ..	67
ค. แสดงอักขระทดสอบภาษาไทย	71
ประวัติผู้เขียน	74

รายการตารางประกอบ

ตารางที่		หน้า
3.1	ตารางกำหนดรหัสสำหรับการตรวจรู้อักขระภาษาไทย	28
3.2	ตารางแสดงขนาดของข้อมูลที่บรรจุในหน่วยความจำหลัก ..	31
4.1	ตารางแสดงตำแหน่งของข้อมูลที่บรรจุในหน่วยความจำหลัก	47
5.1	ตารางแสดงผลของการตรวจรู้อักขระทดสอบภาษาไทย ...	51
ก	ตารางแสดงรหัสต้นแบบของอักขระแต่ละรูป	58

รายการรูปประกอบ

รูปที่	หน้า
1.1 แสดงภาพบิตแมทริกซ์ตัวอักษร "ก"	2
1.2 แสดงอักขระภาษาไทย	3
1.3 แสดงอักขระต้นแบบภาษาไทยแบบพิมพ์ที่ 1	4
1.4 แสดงอักขระต้นแบบภาษาไทยแบบพิมพ์ที่ 2	5
1.5 แสดงอักขระต้นแบบภาษาไทยแบบพิมพ์ที่ 3	6
1.6 แสดงอักขระต้นแบบภาษาไทยแบบพิมพ์ที่ 4	7
1.7 แสดงอักขระต้นแบบภาษาไทยแบบพิมพ์ที่ 5	8
1.8 แสดงอักขระต้นแบบภาษาอังกฤษแบบพิมพ์ที่ 1	9
2.1 แสดงโครงสร้างของนัยน์ตามนุษย์	11
2.2 แสดงส่วนประกอบของเครื่อง OCR ทัวๆ ไป.....	13
2.3 แสดงส่วนรับภาพของเครื่อง OCR	13
2.4 แสดงการแบ่งอักขระ "ก" เพื่อการตรวจรู้จากข้อมูลทางสถิติ ของอักขระ	19
3.1 แสดงค่าในภาษาไทย	21
3.2 แสดงตำแหน่งของ N_0 ถึง N_8	23
3.3 แสดงแบบการคำนวณที่ใช้พิเคราะห์ค่า N_0 จากด้านขวาไป ด้านซ้าย	24
3.4 แสดงแบบการคำนวณที่ใช้พิเคราะห์ค่า N_0 จากด้านซ้ายไป ด้านขวา	24
3.5 (ก)-(จ) แสดงภาพบิตแมทริกซ์ตัวอักษร "ก" ก่อนและหลัง แปลงให้เหลือเพียงโครงร่าง	25
3.6 แสดงโครงร่างของอักขระ "ก"	29
3.7 แสดงโครงสร้างของการค้นหาแบบต้นไม้	32
3.8 แสดงการเก็บข้อมูลในแต่ละระเบียนของแฟ้มข้อมูลดัชนี	33

รูปที่	หน้า
3.9 แสดงการเก็บข้อมูลในแต่ละระดับของแฟ้มข้อมูลอักษร	33
3.10 แสดงการค้นหาที่สรวม "BDDGDAC" ของอักษร "ก"	34
3.11 แสดงลำดับการจัดเรียงอักษรภาษาไทยในแต่ละระดับให้อยู่ ในระดับเดียวกัน	37
3.12 แสดงการจัดเรียงอักษรทั้ง 4 ระดับของคำว่า "คนมุ่มมัน" ให้อยู่ในระดับเดียวกัน	37
4.1 แสดงผังงานของระบบตรวจรู้อักขระที่พัฒนาขึ้น	39
4.2 แสดงจอภาพของรายการหลัก	41
4.3 แสดงจอภาพของรายการจัดคำภาษาไทยทั้ง 4 ระดับ	42
4.4 แสดงจอภาพของรายการปรับปรุงแก้ไขรูปแบบของอักษร	43
4.5 แสดงจอภาพของรายการตารางแอสกีของอักษรภาษาไทย	44
4.6 แสดงจอภาพที่แสดงผลของอักษรที่สามารถตรวจรู้ได้	45
4.7 แสดงชุดคำสั่งที่ใช้ในการสับเปลี่ยนข้อมูลในหน่วยความจำหลัก	48
4.8 แสดงจอภาพที่แสดงผลของอักษรที่ไม่สามารถตรวจรู้ได้	50
ข.1 แสดงผังงานของส่วนของชุดคำสั่งที่ใช้ในการจัดเตรียมคำใน ภาษาไทย	68
ข.2 แสดงผังงานของส่วนของชุดคำสั่งที่ใช้ในการคัดเลือกอักษร เพื่อการปรับปรุงแก้ไขรูปแบบ	69
ข.3 แสดงผังงานของส่วนของชุดคำสั่งที่ใช้ในการเปลี่ยนรูปแบบของ อักษร	70
ค.1 แสดงอักษรทดสอบภาษาไทยแบบพิมพ์ที่ 1	72
ค.2 แสดงอักษรทดสอบภาษาไทยแบบพิมพ์ที่ 2	73