การใช้คอมพิวเตอร์ตรวจรู้อักขระภาษาไทย

นายชมทิพ พรพนมชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
ภาควิชาวิศวกรรมคอมพิวเตอร์
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
พ.ศ. 2529

ISBN 974-567-193-2

THAI CHARACTER RECOGNITION BY COMPUTER

Mr. Chomtip Pornpanomchai, 1959-

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 วัชระชัยสุรพล

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีบัณฑิตวิทยาลัย

(ศาสตราจารย์ ดร.ถาวร วัชระภัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ดร.ยรรยง เต็งอำนวย)

..... กรรมการ

(ดร.ศุภชัย ตั้งวงศ์ศานต์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ สุเมธ วัชระชัยสุรพล)

..... กรรมการ

(ดร.จารุมาตร ปิ่นทอง)

หัวข้อวิทยานิพนธ์ การใช้คอมพิวเตอร์ตรวจรู้อักขระภาษาไทย
ชื่อนิสิต นายชมทิพ พรพนมชัย
อาจารย์ที่ปรึกษา ดร. สุภชัย ตั้งวงศ์ศานต์
ผู้ช่วยศาสตราจารย์ สุเมธ วัชระชัยสุรพล
ภาควิชา วิศวกรรมคอมพิวเตอร์
ปีการศึกษา 2529

บทคัดย่อ

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

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

1. ขั้นตอนการเปลี่ยนภาพบิทเมทริกซ์ตัวอักษรให้เป็นโครงร่างของอักขระ (skeletal form) ขั้นตอนนี้อาจเรียกว่าเป็นขั้นตอนการลดความหนา (thinning process) ของภาพบิทเมทริกซ์ตัวอักษรที่เตรียมไว้ก่อนการนำเข้ามาใช้ในการตรวจรู้อักขระ

2. ขั้นตอนการเปลี่ยนโครงร่างของอักขระให้อยู่ในรูปของรหัส ซึ่งรหัสเหล่านี้จะหมายถึงลักษณะของตัวอักขระตามแนวแกวและแนวสดมภ์

3. ขั้นตอนการสร้างความสัมพันธ์ระหว่างรหัสในขั้นที่ 2 กับอักขระต้นแบบที่สร้างไว้เพื่อการตรวจรู้ โดยความสัมพันธ์นี้จะมีโครงสร้างเป็นรูปต้นไม้ที่เรียกว่า "recognizer tree"

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

ผลการวิจัยพบว่าเทคนิคการตรวจรู้อักขระที่พัฒนาขึ้นสามารถตรวจรู้อักขระต้นแบบได้ทุกรูป ด้วยอัตราความเร็วเฉลี่ย 72 ตัวอักษรต่อนาที 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 perposed 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 วัชระชัยสุรพล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านทั้งสองได้กรุณาสละเวลาอันมีค่าของท่านให้คำแนะนำและข้อคิดเห็นต่างๆ ของการวิจัยมาด้วยดี

การวิจัยครั้งนี้ได้ใช้อุปกรณ์บริภัณฑ์ของสำนักคอมพิวเตอร์ มหาวิทยาลัยมหิดล โดยมี คุณอรรณพ สถิรบัญชา ผู้ช่วยในการจัดเตรียมอุปกรณ์บริภัณฑ์สำหรับงานวิจัย คุณบรีชา งามเสาวรส ผู้ให้คำปรึกษาด้านชุดคำสั่งภาษาแอสเซมบลี้ ตลอดจน คุณกอบกุล อายุวัฒน์ และคุณศราวุฒิ ทรงเจริญ ผู้ให้คำแนะนำด้านการใช้ภาษาสำหรับการเขียนวิทยานิพนธ์ ผู้เขียนจึงขอขอบพระคุณมา ณ ที่นี้

ชมทิพ พรพนมชัย

สารบัญ

บทที่ หน้า

รายการตารางประกอบ

รายการรูปประกอบ

รูปที่ หน้า