

ระบบรู้จำอักษรภาษาไทย  
โดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย

นาย วิชา พานิช



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2539

ISBN 974-635-567-8

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

**A THAI CHARACTER RECOGNITION SYSTEM  
BASED ON DISTINCTIVE FEATURES OF THAI CHARACTERS**

**Mr. Wicha Panich**

**A Thesis Submitted in Partial Fulfillment of the Requirements**

**for the Degree of Master of Engineering**

**Department of Electrical Engineering**

**Graduate School**

**Chulalongkorn University**

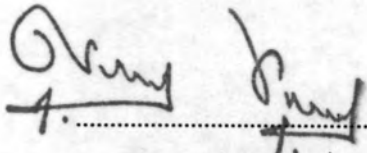
**Academic Year 1996**

**ISBN 974-635-567-8**

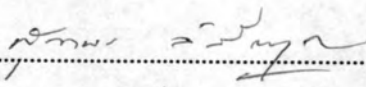
หัวข้อวิทยานิพนธ์      ระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย  
โดย                              นาย วิชา พานิช  
ภาควิชา                              วิศวกรรมไฟฟ้า  
อาจารย์ที่ปรึกษา              รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล

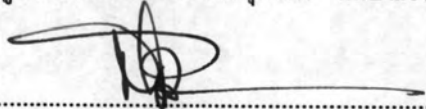
---

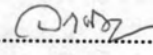
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

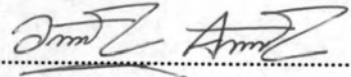
.....รักษาราชการแทนคณบดีบัณฑิตวิทยาลัย  
( ศาสตราจารย์ นายแพทย์ สุภวัฒน์ ชุตินวงศ์ )

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ  
( ผู้ช่วยศาสตราจารย์ ดร.สุดาพร ลักษณ์ินาวิน )

..... อาจารย์ที่ปรึกษา  
( รองศาสตราจารย์ ดร.สมชาย จิตะพันธ์กุล )

..... กรรมการ  
( อาจารย์ ดร.วาทิต เบญจพลกุล )

..... กรรมการ  
( ดร. จุฬารัตน์ ตันประเสริฐ )

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

วิชา พานิช : ระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของตัวอักษรไทย (A THAI CHARACTER RECOGNITION SYSTEM BASED ON DISTINCTIVE FEATURES OF THAI CHARACTERS ) อ.ที่ปรึกษา : รศ.ดร.สมชาย จิตะพันธ์กุล , 82 หน้า. ISBN 974-635-567-8

วิทยานิพนธ์ฉบับนี้มีจุดมุ่งหมายเพื่อสร้างระบบรู้จำอักษรภาษาไทยโดยใช้ลักษณะบ่งความต่างของอักษรไทย ซึ่งประกอบด้วยงาน 3 ส่วนหลักคือ ส่วนรู้จำอักษรเดี่ยว ส่วนแยกอักษรที่ติดกัน และ ส่วนวิเคราะห์เอกสาร

ในส่วนการรู้จำตัวอักษรภาษาไทยใช้การแบ่งกลุ่มโดยใช้ลักษณะของโครงสร้างหลักร่วมกับระดับของอักษร โดยแบ่งเป็นอักษรระดับบน 1 กลุ่ม ระดับล่าง 1 กลุ่ม และระดับกลางอีก 7 กลุ่ม แล้วจึงแยกแยะในกลุ่มย่อยโดยใช้ลักษณะบ่งความต่างของอักษรไทย ในส่วนการตัดแยกอักษรที่ติดกันนั้นใช้ลักษณะบ่งความต่างของอักษรไทยแบ่งประเภทของการติดกันโดยใช้ระดับของอักษรได้เป็น 10 กลุ่มแล้วใช้วิธีเฉพาะของแต่ละกลุ่มในการตัดแยก ในส่วนการวิเคราะห์เอกสารมีการแก้ความเอียงของเอกสาร การแยกคอลัมน์และแยกบรรทัดตัวอักษร

โดยทำการทดสอบบนเครื่องไมโครคอมพิวเตอร์ CPU 80486DX2-80 กับอักษรกว่า 50,000 ตัวอักษร ได้ผลการรู้จำร้อยละ 97.6 และใช้เวลาเฉลี่ยในการรู้จำ 36.4 อักษรต่อวินาที

ภาควิชา ..... วิศวกรรมศาสตร์  
สาขาวิชา ..... วิศวกรรมไฟฟ้า  
ปีการศึกษา ..... 2539

ลายมือชื่อนิสิต .....  
ลายมือชื่ออาจารย์ที่ปรึกษา .....  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม .....

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

## C815897 : MAJOR ELECTRICAL ENGINEERING  
KEY WORD: CHARACTER RECOGNITION / THAI DISTINCTIVE FEATURES

WICHA PANICH : A THAI CHARACTER RECOGNITION SYSTEM BASED ON DISTINCTIVE FEATURES OF THAI CHARACTERS. THESIS ADVISOR : ASSO. PROF. SOMCHAI JITAPUNKUL, Dr.Ing. 82 pp . ISBN 974-635-567-8

The objective of this thesis is to create a Thai character recognition system based on Thai distinctive features that consists of 3 main parts : a recognition module of single character , a module of segmentation of connected characters image , and a document analysis module.

In the recognition module of single character , the primary structures and the level of the character are used to classify thai characters into 9 groups . They composed of one upper level group , one lower level group and seven middle level groups . Distinctive features of Thai characters are used to classify the member in each groups. In the part of segmentation of connected characters , Thai distinctive features are used to identify the group of connected character . We have 10 groups and 10 methods to segment that connected characters to be single character. The document analysis module provides an algorithm to deskew , detect the columns , detect the lines of the scanned document to create a text file of the character strings in each column and each line.

By using microcomputer of CPU 80486DX2-80 to test the document that contains about 50,000 characters, recognition rate is 97.6% . The average processing time is 36.4 characters per second.

ภาควิชา.....วิศวกรรมศาสตร์  
สาขาวิชา.....วิศวกรรมไฟฟ้า  
ปีการศึกษา..... 253๗

ลายมือชื่อนิสิต..... วชิร พานิช  
ลายมือชื่ออาจารย์ที่ปรึกษา.....  
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....



### กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของ รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ของการวิจัยมาด้วยดีตลอด ขอขอบคุณทุกท่านที่ไม่ได้เอ่ยนามในที่นี้ ที่ช่วยให้ข้อคำแนะนำและให้กำลังใจเสมอมา

สุดท้ายนี้ ผู้วิจัยใคร่ขอกราบของพระคุณบิดา มารดา ซึ่งได้ให้การสนับสนุนและให้กำลังใจแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

## สารบัญ

	หน้า
บทคัดย่อภาษาไทย .....	ง
บทคัดย่อภาษาอังกฤษ .....	จ
กิตติกรรมประกาศ .....	ก
สารบัญตาราง .....	ฉ
สารบัญภาพ .....	ญ
บทที่ 1 บทนำ	
ความเป็นมาและความสำคัญของปัญหา .....	1
วัตถุประสงค์ .....	2
เป้าหมายและขอบเขตของวิทยานิพนธ์ .....	2
ขั้นตอนและวิธีการดำเนินงาน .....	3
ประโยชน์ที่คาดว่าจะได้รับ .....	3
บทที่ 2 แนวคิด ทฤษฎีและผลงานที่ผ่านมา .....	4
ทฤษฎีในการรู้จำ.....	6
ทฤษฎีที่เกี่ยวข้องกับลักษณะบ่งความต่าง .....	7
งานวิจัยด้านการรู้จำอักษรเดี่ยว .....	11
งานวิจัยด้านการแยกอักษรที่ติดกัน.....	13
งานวิจัยด้านกระบวนการวิเคราะห์เอกสาร .....	14
ตัวอักษรภาษาไทย .....	15
ทฤษฎีของการรู้จำแบบจีนแตกคิก.....	16
บทที่ 3 กระบวนการที่ใช้ในระบบรู้จำอักษรภาษาไทย .....	17
กล่าวนำ .....	17
3.1 ส่วนวิเคราะห์เอกสาร .....	19
3.1.1 การแก้ความเอียง .....	19
3.1.2 แยกคอลัมน์และแยกบรรทัด .....	21
3.1.3 การดึงเกาะของอักษร .....	23
3.1.4 การหาระดับของอักษร .....	24
3.1.5 การตรวจสอบระดับของอักษร .....	24
3.1.6 การตรวจสอบความถูกต้องทางภาษา .....	25
3.1.7 การสร้าง Text file .....	25

สารบัญ ( ต่อ )

	หน้า
3.2 กรรมวิธีในการแยกตัวอักษรที่ติดกัน .....	26
3.2.1 ประเภทของตัวอักษรที่ติดกัน .....	26
3.2.2 ขั้นตอนการตรวจสอบว่าเป็นอักษรที่ติดกันหรือไม่ .....	28
3.2.3 ส่วนตัดแยกของลักษณะที่ 4 .....	30
3.2.4 ส่วนตัดแยกของลักษณะที่ 3 .....	31
3.2.3 ส่วนตัดแยกของลักษณะที่ 2 .....	33
3.2.4 ส่วนตัดแยกของลักษณะที่ 1 .....	35
3.2.5 ส่วนตัดแยกพิเศษ .....	36
3.3 การแยกแยะอักษรเดี่ยว (Classification of Single Character).....	37
3.3.1 เครื่องมือที่ใช้ในการหาลักษณะบ่งความต่าง .....	39
3.3.2 การรู้จำอักษรระดับกลาง .....	42
3.3.3 การรู้จำอักษรระดับบน .....	55
3.3.4 การรู้จำอักษรระดับล่าง .....	58
บทที่ 4 ขั้นตอนการทดลอง ผลการทดลองและวิเคราะห์ผลการทดลอง .....	60
4.1 ขั้นตอนการทดลอง .....	60
4.2 ผลการทดลอง .....	61
4.3 การวิเคราะห์ผลการทดลอง .....	64
บทที่ 5. สรุปผลการทดลอง .....	69
เอกสารอ้างอิง .....	72
ภาคผนวก .....	74
ประวัติผู้เขียน .....	82



## สารบัญตาราง

หน้า

ตารางที่ 2.1	เปรียบเทียบแนวทาง StatPR SyntPR และ NeurPR .....	5
ตารางที่ 2.2	แสดงการแบ่งลักษณะตัวอักษร.....	15
ตารางที่ 3.1	แสดงจำนวนหน้าที่ทำเอกสารเรียงในมุมขนาดต่างๆ.....	19
ตารางที่ 3.2	แสดงแบบของตัวอักษรที่ติดกัน ตัวอย่าง และร้อยละที่พบ .....	27
ตารางที่ 3.3	ร้อยละที่พบของแบบของตัวอักษรที่ติดกัน .....	27
ตารางที่ 3.4	แสดงจำนวนอักษรที่ทำการทดสอบและที่ติดสำหรับอักษรแบบและขนาดต่าง ๆ .....	28
ตารางที่ 3.5	แสดงว่าลักษณะที่ได้สามารถเป็นอักษรปกติแบบใดและเป็นอักษรติดแบบใดได้ .....	28
ตารางที่ 3.6	แสดงการแบ่งกลุ่มของอักษรระดับกลาง ( ซึ่งแยกตัวเลขบางตัวออกไปแล้ว ) .....	42
ตารางที่ 3.7	แสดงเงื่อนไขการตรวจสอบว่าเป็นตัวเลขไทยหรือไม่ .....	44
ตารางที่ 4.1	แสดงผลการทดสอบของอักษรที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ .....	61
ตารางที่ 4.2	แสดงความผิดพลาดของอักษรลักษณะต่างๆ .....	62
ตารางที่ 4.3	แสดงจำนวนอักษรที่ทำการทดสอบและที่ติดสำหรับอักษรแบบและขนาดต่าง ๆ .....	62
ตารางที่ 4.4	แสดงผลการทดสอบจากเอกสารจริง .....	62
ตารางที่ 4.5	แสดงผลการทดสอบของอักษรที่พิมพ์ด้วยเครื่องพิมพ์เลเซอร์ .....	63
ตารางที่ 4.6	แสดงผลการรู้จำเอกสารที่เรียงด้วยมุมต่างๆ .....	63

## สารบัญภาพ

	หน้า
รูปที่ 1.1 กระบวนการหลักของระบบรู้จำเอกสารภาษาไทย .....	1
รูปที่ 2.1 แสดงลักษณะบ่งความต่างของอักษรไทย.....	9
รูปที่ 2.2 แสดงลำดับการตรวจสอบลักษณะบ่งความต่างของสระ.....	10
รูปที่ 2.3 แสดงตัวอย่างการแทนจุดภาพอักษรด้วยรหัสทิศทาง .....	11
รูปที่ 3.1 กระบวนการหลักของระบบรู้จำเอกสารภาษาไทย .....	17
รูปที่ 3.2 แสดง Flow Chart ของระบบรู้จำเอกสารภาษาไทย .....	18
รูปที่ 3.3 แสดงขั้นตอนการภาพอักษรก่อนและหลังการ Paste .....	20
รูปที่ 3.4 แสดงภาพก่อนและหลังการ Paste ทั้งแนวนอนและแนวตั้ง .....	21
รูปที่ 3.5 แสดง Modified Horizontal Projection และระดับที่ใช้ในการตัดบรรทัด .....	22
รูปที่ 3.6 แสดงตัวอย่างภาพบรรทัดที่ตัดมาได้ก่อนการดึงเกาะของอักษร .....	23
รูปที่ 3.7 แสดงเกาะของอักษรที่ดึงได้ทั้ง 3 เกาะ ก) ข) ช้ ก) ลี .....	23
รูปที่ 3.8 แสดงภาพบรรทัดที่ตัดได้ และ Horizontal projection ที่ได้ .....	23
รูปที่ 3.9 แสดงเส้นตัดสินระดับ .....	24
รูปที่ 3.10 แสดงตัวอย่างเกาะของอักษรที่ติดกันแบบต่างๆ .....	28
รูปที่ 3.11 แสดงขั้นตอนหลักๆ การทำงานของส่วนตัดแยก .....	29
รูปที่ 3.12 แสดงอักษรที่จะถูกตัดแยกของอักษรลักษณะที่ 4 .....	30
รูปที่ 3.13 แสดงการหาจุดตัด โดยอาศัยเส้นระดับล่างและช่วงการพิจารณา .....	30
รูปที่ 3.14 แสดงการตัดแยกระดับกลางกับล่างออกจากกัน .....	30
รูปที่ 3.15 แสดงปัญหาที่เกิดจากการตัดแยก และวิธีแก้ไข .....	31
รูปที่ 3.16 แสดง ( a ) Vertical Projection ของอักษรที่ติด ( b ) Averaged Vertical Projection.....	32
( c ) อนุพันธ์อันดับ 1 ( d ) อนุพันธ์อันดับ 2	
( f ) ผลการหารที่ได้จากสูตร เทียบกับ ( e ) อักษรที่ติด	
รูปที่ 3.17 แสดง flow chart การแบ่งกลุ่มของอักษรที่เข้ามาในส่วนแยกแยะของลักษณะที่ 2 .....	34
รูปที่ 3.18 แสดงจุดตัดแยกของอักษรเดี่ยวกลุ่มย่อยที่ 2 และ 3 .....	34
รูปที่ 3.19 แสดงการหาจุดตัดของอักษรติดแบบต่างๆ .....	34
รูปที่ 3.20 แสดงการตัดแยกอักษรที่มีหางกับสระหรือวรรณยุกต์ .....	35
รูปที่ 3.21 แสดงอักษรติดที่ถูกระบุว่าเป็นลักษณะที่ 1 .....	35
รูปที่ 3.22 แสดงการตัดแยกสระหรือวรรณยุกต์ติดกับ สระ ใ โ โ .....	36
รูปที่ 3.23 แสดงลักษณะ โครงสร้างหลักที่เหมือนกันระหว่างหลายๆ แบบอักษร .....	37
รูปที่ 3.24 แสดงขั้นตอนหลักในการรู้จำอักษรเดี่ยว .....	38
รูปที่ 3.25 แสดงอักษร ช ที่ 12 14 และ 16 point .....	39

สารบัญภาพ ( ต่อ )

	หน้า
รูปที่ 3.26 a) แสดงจุดภาพอักษร b) แสดง Vertical Projection c) แสดงค่าเฉลี่ยของ Vertical Projection	40
รูปที่ 3.27 (a) แสดงจุดภาพอักษร (b) top [ x ] ของภาพตัวอย่าง (c) bottom [ x ] ของภาพตัวอย่าง	40
รูปที่ 3.28 (a) แสดงจุดภาพอักษร (b) left [ x ] ของภาพตัวอย่าง (c) right [ x ] ของภาพตัวอย่าง	41
รูปที่ 3.29 แสดง Flow Chart การแบ่งกลุ่มอักษร การแยกตัวเลขและสระอะ ออกจากกระบวนการ	43
รูปที่ 3.30 แสดงตำแหน่งของลักษณะบ่งความต่างที่ปรากฏในตัวเลขไทย	44
รูปที่ 3.30 แสดงขั้นตอนการหาแบบของ vert_av	45
รูปที่ 3.31 แสดง Vertical Projection ของอักษรที่นับยอดเขาได้ 2 ยอด	45
รูปที่ 3.32 แสดง Vertical Projection ของอักษรแบบ 3 ยอดและนับได้ 3 ยอด	46
รูปที่ 3.33 แสดง Vertical Projection ของอักษรที่ไม่ใช่แบบ 3 ยอดแต่นับได้ 3 ยอด	46
รูปที่ 3.34 แสดง flow chart การรู้จำอักษรกลุ่มที่ 1	47
รูปที่ 3.35 แสดง flow chart การรู้จำอักษรกลุ่มที่ 2	48
รูปที่ 3.36 แสดงลักษณะบ่งความต่างระหว่างหัวเข้ากับหัวออก	49
รูปที่ 3.37 แสดงลักษณะบ่งความต่างของหัวข้างใน ถ.ดู	49
รูปที่ 3.38 แสดง flow chart การรู้จำอักษรกลุ่มที่ 3	50
รูปที่ 3.39 แสดง flow chart การรู้จำอักษรกลุ่มที่ 4	51
รูปที่ 3.40 แสดงลักษณะบ่งความต่างของ ข แบบอักษร a) CordiaUPC b) AngsanaUPC	51
รูปที่ 3.41 แสดงการตรวจสอบพยักภายในซึ่งเป็นลักษณะบ่งความต่างระหว่าง ข.ยัก กับ บ.ใบไม้	51
รูปที่ 3.42 แสดง flow chart การรู้จำอักษรกลุ่มที่ 2	51
รูปที่ 3.43 แสดง flow chart การรู้จำอักษรกลุ่มที่ 6	53
รูปที่ 3.44 แสดง flow chart การรู้จำอักษร ข ข ข ซ	53
รูปที่ 3.45 แสดง flow chart การรู้จำอักษรกลุ่มที่ 7	54
รูปที่ 3.46 แสดงแบบของหลังคา a) เรียบธรรมดา b) มีหยัก c) , d) มีหาง	54
รูปที่ 3.47 แสดงตัวอย่างการหาหัวของ ว.แหวนใช้ฟังก์ชัน Head ( left , H / 2 , H )	55
รูปที่ 3.48 แสดง flow chart แสดงการรู้จำอักษรในระดับบน	56
รูปที่ 3.49 แสดงลักษณะบ่งความต่างของไม้โท	56
รูปที่ 3.50 แสดงการหาลักษณะบ่งความต่างชนิดปากนก	57
รูปที่ 3.51 แสดงการตรวจไม้หันอากาศ	57
รูปที่ 3.52 แสดงลักษณะบ่งความต่างที่แยก อี อี อี อี ออกจากกัน	57
รูปที่ 3.53 แสดงสระ โ ใ ก่อนและหลังการตัดแยกในบทที่ 3	57
รูปที่ 3.54 แสดงลักษณะบ่งความต่างระหว่างสระ โ กับ ใ	58
รูปที่ 3.55 แสดงความกว้างและสูงของส่วนล่างของอักษร ฐ และ ฎ	58

สารบัญภาพ ( ต่อ )

หน้า

รูปที่ 4.1 ภาพอักษรตัวอย่างที่สแกนได้จากหนังสือพิมพ์ไทยรัฐ .....	65
รูปที่ 4.2 ภาพอักษรตัวอย่างที่สแกนได้จากนิตยสารผู้หญิงวันนี้ .....	65
รูปที่ 4.3 แสดงภาพอักษรตัวอย่างของแบบอักษร LilyUPC ที่สแกนได้ .....	65
รูปที่ 4.4 แสดงอักษรแบบ JasmineUPC , KodchiangUPC เทียบกับ AngsanaUPC .....	66
รูปที่ 4.5 แสดงสัญญาณรบกวนที่เกิดขึ้นเมื่อแก้เอกสารที่เอียง .....	66
รูปที่ ก.1 แสดงอักษรทั้งหมดของ AngsanaUPC และ CordiaUPC .....	74
รูปที่ ก.2 แสดงตัวอย่างอักษรของแบบอักษร BrowalliaUPC , DilleniaUPC , EucrosiaUPC , FreeciaUPC , IrisUPC , JasmineUPC , KodchiangUPC .....	75
รูปที่ ก.3 แสดงตัวอย่างอักษรที่สแกนได้จากเอกสารจริง .....	76
รูปที่ ก.4 แสดงภาพถ่ายเอกสารของเอกสารจริง ( หลายชีวิต ) .....	77
รูปที่ ก.5 แสดงภาพถ่ายเอกสารของเอกสารจริง ( ไทยรัฐ ) .....	78
รูปที่ ก.6 แสดงภาพถ่ายเอกสารของเอกสารจริง ( ไทยรัฐ ) .....	79
รูปที่ ก.7 แสดงภาพถ่ายเอกสารของเอกสารจริง ( Advance Thailand Geographic ) .....	80
รูปที่ ก.8 แสดงภาพถ่ายเอกสารของเอกสารจริง ( ผู้หญิงวันนี้ ) .....	81