

การรู้จำสายอักขระไทยตัวพิมพ์โดยวิธีซินแทกติก



นาย สมศักดิ์ คงถาวรวัฒนา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2539

ISBN 974-634-368-8

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

I 1๖8๙148x

RECOGNITION OF THAI PRINTED CHARACTER STRING
BY THE SYNTACTIC METHOD

Mr Somsak Kongthawornwattana

A Thesis Submitted in Partial Fulfillment of Requirements
for the Degree of Master of Engineering

Department of Electrical

Graduate School

Chulalongkorn University

1996

ISBN 974-634-368-8

หัวข้อวิทยานิพนธ์ การรู้จำสายอักขระไทยด้วยวิธีอินเทกติก
โดย นาย สมศักดิ์ คง วารวัฒนา
ภาควิชา วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษา รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล



บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยเป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร. สันติ ฤงสุวรรณ)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สุดาพร ลักษณะียนาวิน)

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

.....กรรมการ
(ดร. วิวัฒน์ วงศ์วารวิฑร์)

.....กรรมการ
(ดร. บวรกุล จิตต์ประเสริฐ)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว



สมศักดิ์ คงถาวรวัฒนา : การรู้จำสายอักขระไทยด้วยวิธีซินแทกติก (RECOGNITION of Thai-Printed character string by the Syntactic Method) อ. ที่ปรึกษา : รศ. ดร. สมชาย จิตะพันธ์กุล, 69 หน้า. ISBN 974-634-368-8

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อหาอัลกอริทึมต้นแบบที่สามารถนำไปใช้การรู้จำสายอักขระตัวพิมพ์ไทยเพื่อให้ผลลัพธ์ที่ได้ออกมาเป็นลักษณะของแฟ้มข้อมูลของตัวอักษร ตามมาตรฐานภาษาไทยของ สมอ. โดยตัวอักษรที่จะนำมาทดสอบจะเป็นตัวอักษรแบบ EucrosiaUPC ขนาด 18 points

สำหรับระบบการรู้จำสายอักขระตัวพิมพ์ไทยประกอบด้วยขั้นตอนต่าง ๆ ที่สำคัญคือ ขั้นตอนการแยกกลุ่มของข้อมูลภาพโดยใช้เทคนิคการหาขอบภาพ การจัดเรียงตัวอักษรโดยใช้วิธีการพิจารณาเส้นฐานและขนาดตัวอักษร ขั้นตอนการรู้จำตัวอักษรโดยใช้วิธีซินแทกติก (สนธยา, 2537)

ผลการทดลองซึ่งใช้สายตัวอักษร 150 สายรวม 1,974 ตัวอักษร พบว่าเมื่อนำภาพตัวอักษรที่ต้องการรู้จำมาทดสอบกับระบบการรู้จำที่พัฒนาขึ้นนั้น สามารถที่จะรู้จำสายอักขระเหล่านั้นได้ โดยผลของการรู้จำสำหรับการวิจัยนี้มีอัตราการรู้จำ 92.70% ไม่สามารถรู้จำได้ 2.90% และรู้จำผิด 4.40%

ภาควิชา
สาขาวิชา
ปีการศึกษา

ลายมือชื่อนิสิต
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C515692 : MAJOR ELECTRICAL ENGINEERING
KEY WORD: RECOGNITION/SYNTACTIC/THAI-PRINTED CHARACTER

SOMSAK KONGTHAWORNWATTANA : RECOGNITION OF THAI-PRINTED CHARACTER STRING
BY THE SYNTACTIC METHOD. THESIS ADVISOR : ASSO. PROF. SOMCHAI JITAPUNKUL.
Ph.D. 69pp. ISBN 974-634-368-8

This thesis proposed to use the syntactic method to recognize Thai-printed character string. The process composed of 3 steps, image segmentation based on edge detection technique, character sorting obtained from the determination of base line and the size of characters and lastly the character recognition process using the syntactic method (Sonthaya 1994). The input file was limited to EucrosiaUPC font of 18 points only.

From the experiment using 150 character strings with 1,974 characters in total, resulted in 92.70%, 4.4%, and 2.9% of correct, wrong, and undecided recognition rates respectively.

ภาควิชา.....
สาขาวิชา.....
ปีการศึกษา.....

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



กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างยิ่งของ รองศาสตราจารย์ ดร. สมชาย จิตะพันธ์กุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ของการวิจัย มาด้วยดีตลอด และ คุณชนิษฐา พรายงาม ที่มีส่วนช่วยเหลือในการให้คำปรึกษาการเขียนโปรแกรมภาษา C และในการจัดพิมพ์วิทยานิพนธ์นี้จนสำเร็จลุล่วงไปด้วยดี

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา-มารดา ซึ่งสนับสนุนในด้านการเงินและให้กำลังใจแก่ ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ



	หน้า
บทคัดย่อภาษาไทย-----	ง
บทคัดย่อภาษาอังกฤษ-----	ฉ
กิตติกรรมประกาศ-----	ฅ
สารบัญตาราง-----	ญ
สารบัญภาพ-----	ณ

บทที่

1. บทนำ-----	1
1.1. ความเป็นมาของปัญหา-----	1
1.2. แนวทางแก้ไขปัญหา-----	2
1.3. วัตถุประสงค์-----	3
1.4. ขอบเขตในการวิจัย-----	3
1.5. ขั้นตอนและวิธีการดำเนินการ-----	3
1.6. ประโยชน์ที่คาดว่าจะได้รับ-----	4
2. ระบบการรู้จำสายตัวอักษรไทย-----	5
2.1. ระบบการรู้จำสายตัวอักษรไทย-----	5
2.2. การปรับปรุงภาพ-----	6
2.3. การแยกข้อมูลภาพออกเป็นข้อมูลของตัวอักษรเดี่ยว ๆ-----	7
2.4. การจัดเรียงตัวอักษร-----	25
2.5. การรู้จำตัวอักษรไทย-----	35
2.6. การแก้ไขรหัสของตัวอักษร-----	36
3. การทดสอบระบบการรู้จำ-----	38
3.1. บทนำ-----	38
3.2. ตัวอักษรต้นแบบ-----	38
3.3. ข้อมูลที่ใช้ในการทดสอบ-----	39
3.4. วิธีการทดสอบ-----	41
3.5. ผลการทดสอบ-----	42
4. สรุปการวิจัยและข้อเสนอแนะ-----	46

4.1. สรุปผลการวิจัย -----	46
4.2. ข้อเสนอแนะ -----	47
รายการอ้างอิง -----	49
ภาคผนวก -----	50
ภาคผนวก ก. ตารางรหัสแอสกีภาษาไทยมาตรฐาน สมอ. -----	51
ภาคผนวก ข. การหาขอบของตัวอักษร -----	52
ภาคผนวก ค. ตัวอย่างการกำจัดจุดภาพข้างเคียง -----	56
ภาคผนวก ง. ตัวอย่างคำที่ใช้ทดสอบ -----	59
ภาคผนวก จ. ตัวอย่างของคำและผลลัพธ์ที่ได้ในแต่ละขั้นตอนของระบบ -----	64
ภาคผนวก ฉ. ตัวอย่างของการทำภาพให้เหลือความเข้มเพียง 2 ระดับ -----	67
ประวัติผู้เขียน -----	69

สารบัญตาราง

หน้า

ตารางที่ 3.1 ผลลัพธ์ของการทดสอบระบบรู้จำสายอักขระตัวพิมพ์ไทย----- 44



สารบัญภาพ

	หน้า
รูปที่ 2.1. ระบบการรู้จำสายตัวอักษรไทย-----	6
รูปที่ 2.2. ทิศทางในการตรวจสอบเพื่อปรับขนาดของภาพ -----	7
รูปที่ 2.3. แสดงโครงสร้างของการแยกข้อมูลภาพออกเป็นตัวอักษรเดี่ยว ๆ-----	7
รูปที่ 2.4. ขั้นตอนการหาขอบของภาพโดยการใช้ตารางหน้าต่าง -----	8
รูปที่ 2.5. แสดงภาพตัวอย่างและผลลัพธ์ที่ได้จากการหาขอบภาพของตัวอักษร-----	9
รูปที่ 2.6. แสดงทิศทางในการแยกข้อมูลภาพ -----	9
รูปที่ 2.7. การแยกข้อมูลภาพของกลุ่มคำออกทีละตัวอักษร-----	10
รูปที่ 2.8. แสดงการแยกข้อมูลที่ได้ตัวอักษรปกติ-----	10
รูปที่ 2.9. แสดงการแยกข้อมูลภาพที่ได้ตัวอักษรติดกัน-----	11
รูปที่ 2.10. แสดงการเตรียมข้อมูลสำหรับการรู้จำ -----	13
รูปที่ 2.11. โครงสร้างการเก็บข้อมูลของตำแหน่งของตัวอักษร -----	13
รูปที่ 2.12. แสดงกลุ่มข้อมูลของตัวอักษรที่ต่อเนื่องกัน-----	14
รูปที่ 2.13. แสดงการเปลี่ยนข้อมูลภาพของตัวอักษรเป็นกราฟของตัวอักษร-----	14
รูปที่ 2.14. แสดงรูปสัญลักษณ์ที่ใช้แทนทิศทางของปริมาณข้อมูล-----	15
รูปที่ 2.15. แสดงข้อมูลภาพและการเปลี่ยนกราฟเป็นรูปสัญลักษณ์-----	16
รูปที่ 2.16. แสดงการแยกตัวอักษรทางแนวตั้ง -----	20
รูปที่ 2.17. แสดงการแยกตัวอักษรทางแนวนอน -----	20
รูปที่ 2.18. แสดงจุดภาพที่เกิดขึ้นจากการเชื่อมลั้กันของข้อมูลภาพที่ใกล้เคียงกัน-----	21
รูปที่ 2.19. แสดงโครงสร้างของการเก็บข้อมูล -----	22
รูปที่ 2.20. แสดงข้อมูลกรณีไม่มีจุดภาพข้างเคียง-----	22
รูปที่ 2.21. ตัวอย่างกรณีมีจุดภาพข้างเคียง 2 กลุ่ม-----	22
รูปที่ 2.22. แสดงภาพตัวอย่างในการเปลี่ยนจุดภาพ -----	23
รูปที่ 2.23. ผลลัพธ์ที่ได้จากการเปลี่ยนจุดภาพบริเวณขอบ-----	24
รูปที่ 2.24. ผลลัพธ์ที่ได้จากขั้นตอนการเปลี่ยนจุดภาพ -----	24
รูปที่ 2.25. โครงสร้างของตัวอักษรที่ไม่มีจุดภาพของตัวอักษรข้างเคียง -----	25
รูปที่ 2.26. ตัวอักษรที่มีจุดภาพจากตัวอักษรข้างเคียง -----	25
รูปที่ 2.27. แสดงโครงสร้างของการจัดเรียงตัวอักษร-----	26
รูปที่ 2.28. แสดงรูปแบบมาตรฐานที่ใช้แบ่งลักษณะของตัวอักษร-----	26
รูปที่ 2.29. ตัวอักษรระดับกลาง-----	27
รูปที่ 2.30. ตัวอักษรระดับล่าง -----	27

รูปที่ 2.31.	ตัวอักษรระดับบน	27
รูปที่ 2.32.	ตัวอักษรระดับที่เกินกว่าเส้นขอบบน	28
รูปที่ 2.33.	ตัวอักษรที่ต่ำกว่าเส้นขอบล่าง	28
รูปที่ 2.34.	ตัวอักษรที่เป็นตัวเลข	28
รูปที่ 2.35.	แสดงโครงสร้างของตัวอักษรปกติ	33
รูปที่ 2.36.	ผลลัพธ์ของการจัดเรียงตัวอักษรของตัวอักษรธรรมดา	33
รูปที่ 2.37.	แสดงโครงสร้างของตัวอักษรที่มีจุดภาพของตัวอักษรข้างเคียง	34
รูปที่ 2.38.	ผลลัพธ์ของการจัดเรียงตัวอักษรของตัวอักษรที่มีจุดภาพจากตัวอักษรข้างเคียง	34
รูปที่ 2.39.	ผลลัพธ์ของตัวอักษรที่เป็นอักษรเว้นวรรค	34
รูปที่ 2.40.	โครงสร้างของระบบการรู้จำโดยวิธีซินแทกติก	35
รูปที่ 3.1.	แสดงผลลัพธ์ที่ได้จากการแยกตัวอักษร	42
รูปที่ 3.2.	การตัวอักษรทางแนวตั้ง	42
รูปที่ 3.3.	การแยกตัวอักษรทางแนวนอน	43
รูปที่ 3.4.	การแยกตัวอักษรติดกันที่ผิดพลาด	43
รูปที่ 3.5.	ตัวอักษรที่มีขนาดมากกว่าความกว้างของตัวอักษรปกติที่กำหนด	43