

การแยกตัวอักษรภาษาไทยที่ติดกัน เพื่อการรู้จำตัวอักษรพิมพ์ภาษาไทย



นายสมเกียรติ วัฒนาประเสริฐ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2543

ISBN 974-347-001-8

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

๓๐ เม.ย. 2546

I 200๕9.164

TOUCHING THAI CHARACTER SEGMENTATION FOR THAI CHARACTER RECOGNITION

Mr. Somkiat Wattanaprasobsuk

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 2000

ISBN 974-347-001-8



สมเกียรติ วัฒนาประเสริฐ : การแยกตัวอักษรภาษาไทยที่ติดกัน เพื่อการรู้จำตัวอักษรพิมพ์  
ภาษาไทย (TOUCHING THAI CHARACTER SEGMENTATION FOR THAI CHARACTER  
RECOGNITION)

อ. ที่ปรึกษา : ดร. บุญเสริม กิจศิริกุล, 50 หน้า. ISBN 974-347-001-8.

วิทยานิพนธ์ฉบับนี้มีจุดมุ่งหมายเพื่อสร้างระบบการตัดแยกตัวอักษรภาษาไทยที่ติดกัน โดย  
ลักษณะบ่งความต่างของอักษรไทย ซึ่งสามารถแบ่งออกเป็น 3 ขั้นตอนคือ ส่วนวิเคราะห์หาตัวอักษรที่ติด  
กัน โดยแบ่งกลุ่มของตัวอักษรตามระดับของตัวอักษรออกเป็น 5 กลุ่ม แล้วจัดกลุ่มที่สามารถติดกันได้  
โดยสามารถแบ่งออกเป็นกลุ่มตัวอักษรที่ติดกันได้ถึง 11 กลุ่ม ส่วนวิเคราะห์หาจุดตัดแยก ในส่วนนี้ใช้วิธี  
โปรเจกชันในแนวดิ่ง และแนวนอน วิธี นิวเมตริกซ์ ในแนวดิ่งและแนวนอน และวิธี หาอนุพันธ์อันดับที่สอง  
ของค่าโปรเจกชัน กับค่าโปรเจกชัน ส่วนวิเคราะห์ตัดแยกตัวอักษร ใช้วิธีการตัดแบบตรง การตัดโดยตัด  
ตามขอบของตัวอักษรที่ติดกัน และการตัดอักษรที่ไขว้กัน

ผลการวิจัยพบว่า การตัดแยกตัวอักษรพิมพ์ภาษาไทย มีความถูกต้องร้อยละ 63.14% โดย  
ประมาณ และใช้เวลาโดยเฉลี่ย 22.07 ตัวอักษรต่อวินาที โดยทำการทดสอบบนเครื่องไมโครคอมพิวเตอร์  
CPU Pentium II 400 MHz จำนวนตัวอักษรที่ทำการทดสอบประมาณ 13,563 ตัวอักษร

ภาควิชา วิศวกรรมคอมพิวเตอร์  
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์  
ปีการศึกษา 2543

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

## XXXXXXXX : MAJOR CCMPUTER SCIENCE

KEYWORD: XXX/XXX/XXX/XXX/XXX

SOMKIAT WATTANAPRASOBSUK : TOUCHING THAI CHARACTER SEGMENTATION  
FOR THAI CHARACTER RECOGNITION

THESIS ADVISOR : Dr. BOOMSERM KITSIRIKUL, 50 pp. ISBN 974-347-001-8.

The objective of this thesis is to propose a method for segmentation of connected Thai characters' based on Thai character distinctive features. The method consist of 3 main parts ; 1) a module of connected character analysis that classifies Thai characters into 5 groups by using levels of the characters and then grouping them into 11 groups of connected characters if possible, 2) a module of connected point analysis that includes vertical and horizontal projection techniques, new matrix and the 2<sup>nd</sup> derivative of projection techniques, and 3) a module of character segmentation analysis that uses vertical cut, contour cut and cross cut techniques.

The result of this thesis shows that the accuracy of character segmentation is 63.14% on average, processing time is 22.07 characters per second. The experiment is conducted on a microcomputer of CPU Pentium II 400 MHz by using documents consisting of 13,563 characters.

Department Computer Engineer  
Field of study Computer Science  
Academic year 2000

Student's signature .....  
Advisor's signature .....  
Co-advisor's signature .....

## กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างยิ่งของ อาจารย์ ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ และคณะกรรมการ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ ในการวิจัยด้วยดีมาโดยตลอด ขอขอบคุณ สำนักงานคณะกรรมการกำกับหลักทรัพย์ และตลาดหลักทรัพย์ ที่ได้สนับสนุนอุปกรณ์คอมพิวเตอร์ เพื่อใช้ในการทดลอง

ท้ายนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณ บิดา มารดา ที่ได้สนับสนุนในด้านการเรียน และให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

## สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ .....	ฉ
สารบัญตาราง .....	ฅ
สารบัญภาพ.....	ญ
บทที่ 1 บทนำ.....	1
ความเป็นมาและความสำคัญของปัญหา.....	1
วัตถุประสงค์ของการวิจัย.....	2
ขอบเขตของการวิจัย.....	2
ประโยชน์ที่คาดว่าจะได้รับ.....	3
วิธีดำเนินการวิจัย.....	3
บทที่ 2 เอกสารและงานวิจัยที่เกี่ยวข้อง.....	5
แนวคิดและทฤษฎี.....	5
เอกสารและงานวิจัยที่เกี่ยวข้อง.....	10
บทที่ 3 วิธีดำเนินการ และเทคนิคในการตัดแยกตัวอักษร.....	13
3.1. การจัดกลุ่มตัวอักษร.....	14
3.2. การวิเคราะห์ภาพเอกสาร.....	14
3.2.1. การตรวจสอบระดับตัวอักษร.....	15
3.2.2. การตรวจสอบลักษณะคอรัปชันตัวอักษร.....	16
3.2.3. ส่วนแยกแยะในลักษณะที่ 1.....	21
3.2.4. ส่วนแยกแยะในลักษณะที่ 2.....	22
3.2.5. ส่วนแยกแยะในลักษณะที่ 3.....	24
3.2.6. ส่วนแยกแยะในลักษณะที่ 4.....	25
3.3. การหาจุดตัดตัวอักษรที่ติดกัน.....	26
3.4. วิธีการตัดแยกตัวอักษรที่ติดกัน.....	27
บทที่ 4 วิธีการทดลอง ผลการทดลอง และวิเคราะห์ผลการทดลอง.....	31
วิธีการทดลอง.....	31
การวิเคราะห์ผลการทดลอง.....	35

บทที่ 5 สรุปผลการทดลอง ข้อเสนอแนะ.....	38
สรุปผลการทดลอง .....	38
ข้อเสนอแนะ .....	39
รายการอ้างอิง .....	40
ภาคผนวก .....	42
ประวัติผู้เขียน .....	50



## สารบัญตาราง

	หน้า
ตารางที่ 2-1 แสดงการแบ่งกลุ่มตัวอักษรไทย.....	10
ตารางที่ 3-1 แสดงกลุ่มตัวอักษรที่ติดกัน .....	14
ตารางที่ 3-2 แสดงตัวอักษรที่สามารถเป็นตัวอักษรเดี่ยว หรืออักษรติดกลุ่มต่างๆ .....	17
ตารางที่ 3-3 แสดงการแยกแยะลักษณะที่ 2 โดยใช้ลักษณะบ่งความต่าง.....	22
ตารางที่ 4-1 แสดงผลการทดสอบการตัดตัวอักษร.....	32

## สารบัญญภาพ

หน้า

รูปที่ 2-1 แสดงโปรเจกชันแนวนอน.....	5
รูปที่ 2-2 แสดงโปรเจกชันแนวตั้ง .....	6
รูปที่ 2-3 นิเวตริกซ์สำหรับตัดตัวอักษรที่ติดกัน.....	6
รูปที่ 2-4 แสดงลักษณะบ่งความต่างของตัวอักษร .....	8
รูปที่ 3-1 กระบวนการหลักของการตัดแยกตัวอักษรภาษาไทยที่ติดกัน.....	13
รูปที่ 3-2 แสดงภาพบรรทัดตัวอักษร.....	15
รูปที่ 3-3 แสดง Horizontal Projection ของภาพบรรทัดตัวอักษร .....	15
รูปที่ 3-4 แสดงเส้นตัดลีนกลุ่มตัวอักษร .....	16
รูปที่ 3-5 แสดงตัวอย่างกลุ่มที่ 3 ติดกับกลุ่มที่ 1 .....	17
รูปที่ 3-6 แสดงผังหลักการทำงานของส่วนแยกแยะลักษณะตัวอักษร .....	18
รูปที่ 3-7 แสดงจุดภาพตัวอักษรที่ซ้อนทับกันในแนวตั้ง .....	20
รูปที่ 3-8 แสดง Vertical Projection ของตัวอักษรในรูปที่ 3-7 .....	20
รูปที่ 3-9 แสดง Flow การทำงานของส่วนแยกแยะลักษณะที่ 1 .....	21
รูปที่ 3-10 แสดง Flow การทำงานของส่วนแยกแยะลักษณะที่ 2 .....	23
รูปที่ 3-11 แสดง Flow การทำงานของส่วนแยกแยะลักษณะที่ 3 .....	25
รูปที่ 3-12 แสดง Flow การทำงานของส่วนแยกแยะลักษณะที่ 4 .....	26
รูปที่ 3-13 แสดงตัวอย่างการตัดด้วย Vertical cut.....	27
รูปที่ 3-14 แสดงตัวอย่างการตัดแยกด้วยวิธี Contour cut.....	28
รูปที่ 3-15 แสดงตัวอักษรที่ไขว้กัน.....	29
รูปที่ 3-16 แสดงวิธีการตัดแบบ Cross cut .....	30
รูปที่ 4-1 แสดงแบบตัวอักษรต่างๆ .....	35
รูปที่ ก.1 แสดงตัวอักษรทั้งหมดของ AngsanaUPC .....	43
รูปที่ ก.2 แสดงตัวอักษรทั้งหมดของ BrowalliaUPC .....	43
รูปที่ ก.3 แสดงตัวอักษรทั้งหมดของ CordiaUPC.....	43
รูปที่ ก.4 แสดงตัวอักษรทั้งหมดของ DilleniaUPC.....	43
รูปที่ ก.5 แสดงตัวอักษรทั้งหมดของ EucrosiaUPC .....	44
รูปที่ ก.6 แสดงตัวอักษรทั้งหมดของ FreesiaUPC.....	44
รูปที่ ก.7 แสดงตัวอักษรทั้งหมดของ IrisUPC.....	44
รูปที่ ก.8 แสดงตัวอักษรทั้งหมดของ JasmineUPC.....	44

## สารบัญภาพ (ต่อ)

หน้า

รูปที่ ก.9 แสดงตัวอักษรทั้งหมดของ KodchiangUPC.....	44
รูปที่ ก.10 แสดงตัวอักษรทั้งหมดของ LilyUPC .....	45
รูปที่ ก.11 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร AngsanaUPC ขนาด14 Point.....	45
รูปที่ ก.12 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร BrowalliaUPC ขนาด14 Point.....	46
รูปที่ ก.13 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร CordiaUPC ขนาด14 Point.....	46
รูปที่ ก.14 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร DilleniaUPC ขนาด14 Point.....	47
รูปที่ ก.15 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร EucrosiaUPC ขนาด14 Point.....	47
รูปที่ ก.16 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร FreesiaUPC ขนาด14 Point.....	48
รูปที่ ก.17 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร IrisUPC ขนาด14 Point.....	48
รูปที่ ก.18 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร JasmineUPC ขนาด14 Point.....	49
รูปที่ ก.19 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร KodchiangUPC ขนาด14 Point.....	49
รูปที่ ก.20 แสดงภาพเอกสารที่ใช้ทดลอง แบบตัวอักษร LilyUPC ขนาด14 Point .....	49