THE RECOGNITION OF THAI HANDWRITTEN CHARACTERS USING FEATURE-BASED APPROACH

SURASIT KIWPRASOPSAK

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2001

ISBN 974-04-0582-7

3937013 SCCS/M : MAJOR : COMPUTER SCIENCE; M.Sc. (COMPUTER SCIENCE)
KEY WORDS : THAI CHARACTER RECOGNITION / THAI HANDWRITTEN CHARACTER RECOGNITION / FEATURE EXTRACTION / FEATURE EXTRACTION ALGORITHM / FEATURE-BASED APPROACH

SURASIT KIWPRASOPSAK : THE RECOGNITION OF THAI HANDWRITTEN CHARACTERS USING FEATURE-BASED APPROACH. THESIS ADVISORS: JARERNSRI L. MITRPANONT Ph.D., SUKANYA PHONGSUPHAP, Ph.D.

There are a very limited number of researches that are effective enough to extract most of the significant features of Thai handwritten characters. Many of them have concentrated on the extraction of only a few features mainly for Thai optical character recognition systems.

This research aims to develop a set of concrete feature extraction algorithms to be used in the recognition of off-line Thai handwritten characters by using the feature-based approach. These algorithms are used to exploit inherent characteristics or prominent features of Thai characters. Both the characteristics and the common human writing behaviors of Thai characters were studied and analyzed. The decision trees were used to classify Thai characters that share some common features into five classes.

A set of key features of Thai characters was identified. The major features covered were *an end-point (EP), a turning point (TP), a loop (LP), a zigzag (ZZ), a closed top (CT), a closed bottom (CB), and a number of legs.* These features were defined as standard feature or the "*Thai Character Feature Space.*" Then, we defined the 5x3 standard regions used in mapping the standard features of all Thai characters. The result of the mapping process was the "*Thai Character Solution Space,*" which can be used as a fundamental tool for recognition.

The twelve algorithms used to determine each feature of Thai characters were designed and developed along with the recognition system. This included the algorithms to determine an incomplete loop and a filled loop. Each algorithm has been tested thoroughly by using of more than 44,600 Thai characters handwritten by 22 individuals from 100 documents.

The experimental results of each algorithm are summarized as well as the performance. The feature extraction rate is as high as 98.66% with the average of 93.08% while the recognition rate is as high as 99.19% with the average of 91.42%. The results indicate that our proposed algorithms are well established and effective.

3937013 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์; วท. ม. (วิทยาการคอมพิวเตอร์)

สุรสิทธิ์ คิวประสพศักดิ์ : การรู้จำลายมือเขียนภาษาไทยโดยใช้วิธีการหาคุณลักษณะพื้นฐานในตัวอักษร (THE RECOGNITION OF THAI HANDWRITTEN CHARACTERS USING FEATURE-BASED APPROACH). คณะกรรมการควบคุมวิทยานิพนธ์ : เจริญศรี มิตรภานนท์, Ph.D., สุกัญญา พงศ์สุภาพ, Ph.D.

งานวิจัยนี้ได้นำเสนอวิธีการรู้จำตัวอักษรลายมือเขียนภาษาไทย โดยใช้วิธีหาคุณลักษณะพื้นฐานในตัวอักษร (Feature-based Approach) เพื่อใช้สำหรับการรู้จำตัวอักษรลายมือเขียนภาษาไทยแบบ Off-line โดยเรามุ่งเน้นที่การพัฒนา Algorithms พื้นฐานที่มีความถูกต้องสูง เพื่อใช้ในการรู้จำฟีเจอร์ (Feature) ของตัวอักษรภาษาไทย และส่งผลให้มีความถูกต้องในการรู้จำตัวอักษรภาษาไทยที่สูงตามมาด้วย

เราได้ทำการศึกษาคุณลักษณะของภาษาไทย และพฤติกรรมการเขียนอักษรไทยเพื่อใช้ออกแบบในขั้นต้น โดยได้นำเอา Decision Trees มาใช้เพื่อแยกแยะตัวอักษรออกมาเป็นกลุ่มที่มีฟีเจอร์ลักษณะเดียวกันเป็นจำนวนทั้งสิ้น 5 กลุ่ม ทำให้เราสามารถแยกแยะฟีเจอร์พื้นฐานของตัวอักษรภาษาไทยออกมาได้อย่างชัดเจน ตัวอย่างของคุณลักษณะที่สำคัญเหล่านี้ ได้แก่ *End-point (EP), Turning Point (TP), Loop (LP), Zigzag (ZZ), Closed Top (CT), Closed Bottom (CB)* และ *Number of Legs* ของตัวอักษร เป็นต้น ฟีเจอร์ต่าง ๆ เหล่านี้ได้ถูกกำหนดขึ้นมาเป็นฟีเจอร์มาตรฐาน และถูกเรียกว่า *"Thai Character Feature Space"* จากนั้นเราได้ออกแบบเมทริกซ์มาตรฐานขนาด 5x3 ซึ่งถูกเรียกว่า *"Thai Character Solution Space"* เพื่อใช้ในกระบวนการ Mapping ของคุณลักษณะมาตรฐานของตัวอักษรภาษาไทยทั้งหมด และใช้เป็นเครื่องมือพื้นฐานในขบวนการตรวจรู้ตัวอักษรภาษาไทย

หลังจากนั้น เราได้ออกแบบและพัฒนา Algorithms เพื่อใช้ในการตรวจรู้คุณลักษณะต่าง ๆ ทั้งสิ้น 12 Algorithms ซึ่ง Algorithms ต่าง ๆ ที่นำเสนอนี้นอกจากจะสามารถตรวจรู้ฟีเจอร์ที่สมบูรณ์ในตัวอักษรแล้ว ยังสามารถตรวจรู้ฟีเจอร์ที่ไม่สมบูรณ์ได้อีกด้วย เช่น Incomplete Loop และ Filled Loop เป็นต้น นอกเหนือจากนั้นแล้ว เรายังได้พัฒนาระบบเพื่อตรวจสอบความถูกต้องสมบูรณ์ของวิธีการและ Algorithms ที่นำเสนอ โดยได้ทำการทดลองกับตัวอักษรลายมือเขียนภาษาไทยมากกว่า 44,600 ตัวอักษร จากจำนวน 100 เอกสาร ที่ถูกเขียนขึ้นโดยคนไทยจำนวน 22 คน จากการวิเคราะห์ผลการทดลองเราพบว่า Algorithms ที่นำเสนอให้ผลที่น่าพอใจอย่างยิ่ง คือให้อัตราการตรวจรู้ฟีเจอร์ถึง 98.66% โดยอัตราการตรวจรู้ฟีเจอร์เฉลี่ย 93.08% ในขณะที่ได้อัตราการตรวจรู้ตัวอักษรภาษาไทยสูงสุด 99.19% ด้วยอัตรารู้จำเฉลี่ย 91.42% จากผลการทดลองที่ออกมาสามารถเป็นเครื่องบ่งชี้ได้ว่า แนวคิดที่นำเสนอนี้มีประสิทธิภาพสูงและสามารถนำมาประยุกต์ใช้งานได้เป็นอย่างดี