A HIGH PERFORMANCE SYSTEM FOR PRINTED THAI CHARACTER RECOGNITION WITH A STROKE STRUCTURE APPROACH

ORAWAN JUNGTHANAWONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2007

ระบบการรู้จำตัวพิมพ์อักษรภาษาไทยประสิทธิภาพสูงโดยใช้ลักษณะเส้นในการเขียนตัวอักษร
(A HIGH PERFORMANCE SYSTEM FOR PRINTED THAI CHARACTER RECOGNITION WITH A STROKE STRUCTURE APPROACH)

อรวรรณ จึงธนวงศ์ 4737226 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., สุกัญญา พงษ์สุภาพ, Ph.D., ชมทิพ พรพนมชัย, Ph.D., สุพัฒนา เอื้อทวีเกียรติ, Ph.D.

บทคัดย่อ

งานวิจัยนี้ มีวัตถุประสงค์ ในการพัฒนาระบบ การรู้จำตัวพิพม์ อักษรภาษาไทย ให้มีความแม่นยำสูง ในการรู้จำ การที่ต้องการระบบที่มีความแม่นยำสูงในการรู้จำ เพื่อสร้างความมั่นใจ ในการนำผลลัพธ์ไปต่อยอดในระบบงานอื่น เช่น Text to Speech หรือ Machine Translation เป็นต้น

เพื่อทำให้ระบบ มีความแม่นยำในการรู้จำสูง เราต้องหาสาเหตุ ที่ทำให้การรู้จำผิดพลาด เพื่อทำการแก้ไข ซึ่งพบว่า หนึ่งในสาเหตุหลัก คือ จำนวนตัวอักษรที่เป็นไปได้ ในการรู้จำแต่ละครั้ง มีมากเกินไป ทำให้ การค้นหาตัวอักษร ใช้เวลามากด้วย ดังนั้น เราได้เสนอ 2 ขั้นตอน สำหรับลดเวลา ในการค้นหาตัวอักษร ในขั้นตอนแรก ตัวอักษร ที่เป็นพยัญชนะ ตัวเลข และตัวอักขระพิเศษ แบ่งออกเป็น 12 กลุ่ม ตามลักษณะความคล้ายคลึงกัน เช่น กลุ่ม ข,ฆ,ช,ซ กลุ่ม บ,ป,ย เป็นต้น และตัวอักษร ที่เป็น วรรณยุกต์ และสระ ถูกแบ่งออกเป็น 8 กลุ่ม เช่น กลุ่ม โ, ไ, ใ กลุ่ม ิ-, ี-, ึ-, ื- เป็นต้น และในขั้นตอนที่สอง ตัวอักษรจะถูก แยกแยะ ตามแต่ละกลุ่ม ด้วยลักษณะสำคัญต่างๆ ที่ทำให้แต่ละ ตัวอักษรในกลุ่ม แตกต่างกัน ซึ่งทั้ง 2 ขั้นตอน จะทำให้ระบบ มีความแม่นยำในการรู้จำสูงขึ้น และในการตัดภาพอักษร เราได้ทำการรวมภาพของตัวอักษรที่เป็นสระบนและวรรณยุกต์เป็นภาพอักษร1ภาพ และนำไปรู้จำโดยถือว่าเป็นตัวอักษร 1 ตัว

การทดลองกับ แบบอักษรภาษาไทย Browallia New, Cordia New และFreesiaUPC ขนาด 14, 16 และ18 ในรูปแบบตัวหนาและธรรมดา จำนวนมากกว่า 100,000 ตัวอักษร ได้ Recognition rate เท่ากับ 100% ภายใต้การควบคุม ข้อจำกัดในการทดลอง ซึ่งใช้เวลาโดยเฉลี่ย ในการวิเคราะห์ตัวอักษร 1 ตัวอักษร และ ใช้ image and program storage โดยเฉลี่ย ในการวิเคราะห์ตัวอักษร 1 หน้า-A4 เท่ากับ 363 ms. และ5755.5 MB. ตามลำดับ

123 หน้า.

A HIGH PERFORMANCE SYSTEM FOR PRINTED THAI CHARACTER RECOGNITION WITH A STROKE STRUCTURE APPROACH

ORAWAN JUNGTHANAWONG 4737226 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORS: SUPACHAI TANGWONGSAN, Ph.D., SUKANYA PHONGSUPHAP, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D., SUPATANA AUETHAVEKIAT, Ph.D.

ABSTRACT

In this thesis, a highly accurate recognition system for Thai printed characters is proposed. High recognition accuracy leads to high efficiency in further processing, for example, text to speech or machine translation, etc.

To achieve a highly accurate recognition system, the causes of erroneous recognition have to be solved. From the investigation, it was found that one of major causes is that the number of characters to be recognized is too high. The higher the number is, the larger the search space will become. In order to reduce the search space, this thesis proposes a 2-stage process. In the first stage, the characters are grouped into 12 clusters according to the shape, which are consonants, special symbols and Thai digits such as clusters of {ข,ฃ,ช,ฺ}, {บ,ป,ษ}, etc. Similarly, 8 clusters of 21 characters which are vowels and tone marks can be categorized, for instance clusters of {โ, ไ, ใ}, {ิ-,ี-,ึ-,ื-}, etc. In the second stage, the character is recognized with the domain within each class. At this stage, features distinct within the class and indistinct across the class can be used. Thus, the accuracy is improved. In the segmentation process, characters of an upper-vowel level and a tonal level are combined into a single-character image.

The proposed system is tested with 100,000 Thai characters. The type tested fonts were Browallia New, Cordia New and FreesiaUPC .The sizes of the font were 14, 16 and 18. Both normal and bold styled fonts were tested. The proposed system achieved 100% recognition rate in the test under the controlled environment. The average analysis time and the average image and program storage in hard-disk were 363 ms per a character and 5755.5 MB per 1 A4 paper page, respectively.

KEYWORDS: PRINTED THAI CHARACTER RECOGNITION / CHARACTER RECOGNITION / STRUCTURE APPROACH

123 pp.