A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

BUNTIDA SUVACHARAKULTON 4836577 SCCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE : SUPACHAI TANGWONGSAN, Ph.D., SUKANYA PHONGSUPHAP, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D., PANJAI TANTASANAWONG, Ph.D.

ABSTRACT

This thesis proposes a model of optical character recognition with the technique of word prediction for bilingual documents in Thai and English. The model is hence named, BOCR-WP (Bilingual Optical Character Recognition with Word Prediction).

The BOCR-WP is an enhancement of conventional OCR with two additional and distinctive processes: language identification and word prediction. For language identification, the process attempts to distinguish which language mode those image strips should belong to, Thai or English, as a result of the identification. In word prediction, the process is actually followed by character verification after the processing of character recognition. The main idea is that instead of attempting to recognize each individual character via the conventional method, the new approach is trying to identify whole words, either in Thai or English, by using contextual analysis to predict those probable words. Then verify them to obtain the right one by template matching. Obviously, the longer the matched word is, the better the speed of recognition will be. Finally, the technique of dictionary look-up is used in order to improve the accuracy of the final answer for the whole recognition process.

A series of experiments showed that the BOCR-WP was able to classify the script modes, Thai or English, correctly with a high accuracy of 99.99% on average. This system also yielded a better performance compared to conventional OCR in terms of speed improvement with a best case of 28.85%, 22.20% on average, and a minimal improvement of 15.69% while still being able to maintain a quality of accuracy of 100% in Thai and 99% in English from a source of 141 bilingual documents or a total of 284,417 characters.

KEY WORDS: BILINGUAL OCR / N-GRAM / THAI CHARACTER RECOGNITION / WORD PREDICTION / WORD VERIFICATION

90 pages

ระบบประสิทธิภาพสูงสำหรับการรู้จำตัวพิมพ์ภาษาไทยและภาษาอังกฤษด้วยวิธีพยากรณ์คำศัพท์
A HIGHLY EFFECTIVE SYSTEM FOR THAI AND ENGLISH PRINTED CHARACTER RECOGNITION BY WORD PREDICTION METHOD

บุญธิดา สุวัชระกุลธร 4836577 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : ศุภชัย ตั้งวงศ์ศานต์, Ph.D., สุกัญญา พงษ์สุภาพ, Ph.D., ชมทิพ พรพนมชัย, Ph.D., ปานใจ ธารทัศนวงศ์, Ph.D.

บทคัดย่อ

งานวิจัยนี้ นำเสนอแบบจำลองสำหรับการรู้จำตัวพิมพ์อักษรไทยและอังกฤษ ด้วยวิธีการพยากรณ์คำ โดยแบบจำลองที่นำเสนอนี้ มีชื่อว่า ระบบ BOCR-WP

ระบบ BOCR-WP สามารถเพิ่มประสิทธิภาพของการรู้จำตัวอักษรของระบบรู้จำทั่วไป โดยเพิ่มเติมเทคนิคพิเศษ 2 วิธีการ กล่าวคือ การระบุภาษา และการพยากรณ์คำ สำหรับการระบุภาษา นำมาใช้ในการแยกโหมดของรูปภาพตัวอักษร ว่าเป็นภาษาไทยหรืออังกฤษ ส่วนวิธีพยากรณ์คำ นำมาใช้แทนการรู้จำตัวอักษรทีละตัวของระบบการรู้จำทั่วไป ในแนวทางใหม่ โดยพยายามทำนายเซตของคำที่น่าจะเป็น ของแถบรูปภาพตัวอักษร ด้วยการวิเคราะห์เชิงบริบท และตรวจสอบคำเหล่านี้ ด้วยวิธีการเข้าคู่รูปแบบ เพื่อหาคำที่เป็นคำตอบสำหรับการรู้จำ โดยคำที่เข้าคู่กับแถบรูปภาพตัวอักษร ยิ่งมีความยาวมาก ยิ่งทำให้การรู้จำรวดเร็วยิ่งขึ้น นอกจากนี้ ยังได้นำเทคนิคการค้นหาคำในพจนานุกรม มาช่วยปรับปรุงการรู้จำ ให้มีความถูกต้องมากขึ้น

ผลการทดลองกับเอกสารสองภาษา ไทยและอังกฤษ 141 หน้า จำนวนทั้งสิ้น 284,417 ตัวอักษร แสดงให้เห็นถึงความสามารถของระบบ BOCR-WP ในการระบุภาษา ไทยหรืออังกฤษ ด้วยความแม่นยำสูง โดยมีความถูกต้องโดยเฉลี่ย เท่ากับ 99.99% นอกจากนี้ ระบบสามารถรู้จำตัวอักษรไทยและอังกฤษ ได้เร็วขึ้น 28.85% สำหรับกรณีที่ดีที่สุด 22.20% สำหรับค่าเฉลี่ย และ 15.69% สำหรับกรณีที่แย่ที่สุด ในขณะเดียวกัน ระบบยังสามารถรักษาคุณภาพของการรู้จำ ที่ความถูกต้องเฉลี่ย 100% สำหรับตัวอักษรไทย และ 99% สำหรับตัวอักษรอังกฤษ

90 หน้า