บทคัดย่อ 188149

การเพิ่มความถูกต้องของระบบรู้จำอักษรภาษาไทย จำเป็นต้องมีระบบประมวลผลก่อนการรู้จำ ได้แก่กระบวนการปรับเอียงเอกสารรูปภาพ กระบวนการปรับความคมชัด กระบวนการลดสัญญาณรบกวน กระบวนการตัดแบ่งบรรทัดและตัวอักษร และกระบวนการวิเคราะห์ประเภทและตัดรูปภาพตัวอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน สำหรับการรู้จำอักษรภาษาไทย ภาพอักษรที่ติดกันมีผลกระทบต่อความถูกต้องของระบบรู้จำอักษรภาษาไทย ดังนั้น โครงงานวิจัยนี้จึงมุ่งเน้นในการออกแบบและพัฒนาโปรแกรมสำหรับตัดภาพอักษรที่ติดกันทั้งแบบสัมผัสและทับซ้อน กระบวนการตัดภาพอักษรที่ติดกันประกอบด้วย 3 ขั้นตอน ได้แก่ การสกัดภาพและจำแนกภาพอักษร การวิเคราะห์การติดกันของตัวอักษร และการหาตำแหน่งจุดตัดและปรับภาพอักษรให้สมบูรณ์ ในการดึงภาพอักษรจะประกอบด้วยการใช้ค่าสถิติของจุดดำในแนวนอนและการติดตามจุดดำเพื่อแยกตัวอักษรออกจากเอกสาร และใช้ระดับอักษรเพื่อจำแนกภาพอักษร ในการวิเคราะห์การติดกันของตัวอักษรจะใช้ข้อมูลภาพอักษร ได้แก่ ลักษณะเส้นอักษร อัตราส่วนความกว้างและความสูงของรูปภาพ สำหรับการหาตำแหน่งจุดตัดจะคำนวณหาจุดตัดวิกฤตหรือจุดตัดที่มีจุดร่วมน้อยสุดเพื่อตัดภาพอักษรติด โดยขึ้นกับประเภทอักษรติด ในกรณีที่การตัดภาพอักษรติด ทำให้ภาพอักษรไม่สมบูรณ์ จำเป็นต้องมีการปรับแต่งภาพอักษร ซึ่งงานวิจัยนี้ใช้ 2 เทคนิคคือ การคำนวณขอบเขตและชดเชยส่วนที่หายไปหลังการตัดจุดตัดวิกฤต หรือการเพิ่มพื้นที่อักษรเพื่อชดเชยอักษรหลังการตัด เพื่อเพิ่มความสมบูรณ์ให้กับตัวอักษร

ABSTRACT 188149

The accuracy of Thai OCR needs binarization, noise reduction, angle detection, document analysis, main line detection and segmentation of a cross-touch connected character image. In Thai language, a connected character image has a significant effect on the accuracy of Thai OCR. Problems of previous studies derive from the unrecognition of connected character in the image. Thus, this research aims at designing and developing the segmentation of cross-touch connected character image. The segmentation of a cross-touch connected character image consists of three steps, detecting and grouping character images, analyzing character connection patterns, and locating the connected character cutting lines and compensating character. Detecting and grouping character images requires horizontal projection and black point connected, to extract each character from the original document, and leveling each character to identify its type. The analysis of connected characters uses the characteristics of Thai characters and their Width-Height ratio. To locate the cutting point of each character, either peak to valley or break cost technique is used. If cutting causes a character defect due to overlapping of connected characters, the compensation technique will be used to restore and to refine them. Two compensation techniques used in this step are the identification of a connected character overlapping area and a compensation area, and area enlargement in order to increase the compensation area by using the peak to valley technique.