PRINTED TEXT AND HANDWRITING IDENTIFICATION IN THAI DOCUMENT IMAGES

TIPAWAN KHUNVIPAKORN

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE) FACULTY OF GRADUATE STUDIES MAHIDOL UNIVERSITY 2009

PRINTED TEXT AND HANDWRITING IDENTIFICATION IN THAI DOCUMENT IMAGES

TIPAWAN KHUNVIPAKORN 4737995 SCCS/M

M.Sc.(COMPUTER SCIENCE)

RESEARCH PROJECT ADVISORY COMMITTEE: SUKANYA PHONGSUPHAP, Ph.D., SUPATANA AUETHAVEKIAT, Ph.D.

ABSTRACT

General OCR systems are designed to recognize either the printed text or handwritten text. When they are applied in documents consisting of both printed text and handwritten text, their performance is greatly reduced. The objective of this research project was to propose a method for identifying printed text and handwritten text in Thai document images, so that each text block can be submitted to the appropriate OCR system. Document images studied were government documents: both internal and external ones. The technique consists of 4 stages: 1) the pre-processing stage consisting of 2 functions (converting original document images from gray-scale to binary images and correcting skew using Hough transformation technique); 2) the segmentation stage in which document images are segmented at the sentence level or text block; 3) the feature extraction stage in which 8 feature sets are considered for describing characteristics of text blocks; and 4) the classification stage in which minimum distance and Bayes classifiers are considered for identifying text blocks as printed text or handwritten text. The proposed technique was tested on 265 documents consisting of 200 known and 65 unknown images. The feature sets which gave the highest accuracy were the combination of Gabor features and bi-level 2x2 gram features. Bayes classifier yielded a better result than the minimum distance classifier. The proposed method gave 97.45% and 96.43% average accuracy rates for known text blocks and unknown text blocks, respectively.

KEY WORDS : THAI DOCUMENT IMAGE/ DOCUMENT IMAGE ANALYSIS / PRINTED TEXT IDENTIFICATION/ HANDWRITTEN TEXT IDENTIFICATION/ PATTERN CLASSIFICATION

243 pages.

การระบุตัวพิมพ์และลายมือเขียนในภาพเอกสารภาษาไทย

PRINTED TEXT AND HANDWRITING IDENTIFICATION IN THAI DOCUMENT IMAGES

ทิพวรรณ กุลวิภากร 4737995 SCCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาโครงการวิจัย: สุกัญญา พงษ์สุภาพ, Ph.D., สุพัฒนา เอื้อทวีเกียรติ, Ph.D.

บทคัดย่อ

ระบบ OCR โดยทั่วไปได้ออกแบบเพื่อรู้จำตัวอักษรประเภทตัวพิมพ์หรือลายมือเขียนอย่างใดอย่างหนึ่ง หากกลุ่มตัวอักษรแบบตัวพิมพ์และลายมือเขียนปรากฏอยู่ในเอกสารเดียวกันจะทำให้ประสิทธิภาพการรู้จำตัวอักษรลดลง วัตถุประสงค์ของโครงการวิจัยนี้ คือ เสนอวิธีการระบุตัวพิมพ์และลายมือเขียนในภาพเอกสารภาษาไทย ซึ่งจะเป็นประโยชน์ต่อระบบ OCR ภาพเอกสารที่ใช้ในการทดลองเป็นหนังสือราชการ ประกอบด้วย หนังสือราชการภายในและหนังสือราชการภายนอก ขั้นตอนการทำงานมี 4 ขั้นตอน ดังนี้ 1) การประมวลผลก่อน (Pre-processing) มี 2 ขั้นตอนย่อย (เปลี่ยนภาพเอกสารต้นฉบับจากภาพสเกลสีเทาเป็นภาพขาวดำ และปรับความเอียงของภาพให้ตั้งตรงโดยใช้วิธี Hough transformation) 2) การแบ่งส่วน (Segmentation) ภาพเอกสารจะถูกแบ่งส่วนให้เป็นบล็อกข้อความในระดับประโยค 3) การสกัดลักษณะสำคัญ (Feature extraction) ได้พิจารณากลุ่มลักษณะสำคัญ 8 กลุ่มเพื่ออธิบายคุณลักษณะของบล็อกข้อความ และ 4) การจำแนกประเภท (Classification) ได้พิจารณาตัวจำแนก 2 ชนิด คือ Minimum distance classifier และ Bayes classifier เพื่อระบุว่าบล็อกข้อความเป็นตัวพิมพ์หรือลายมือเขียน จากการทดลองใช้ภาพเอกสารจำนวน 265 ภาพ ประกอบด้วย กลุ่มภาพเอกสารที่เคยผ่านการเรียนรู้ จำนวน 200 ภาพ และกลุ่มภาพเอกสารที่ไม่เคยผ่านการเรียนรู้มาก่อน จำนวน 65 ภาพ ลักษณะสำคัญที่สามารถอธิบายลักษณะของบล็อกข้อความได้ดีที่สุด คือ Gabor features ร่วมกับ Bi-level 2x2 gram features และตัวจำแนกประเภท Bayes classifier ให้ผลลัพธ์ที่ดีกว่า Minimum distance classifier วิธีการที่เสนอสามารถจำแนกบล็อกข้อความจากภาพเอกสารที่เคยผ่านการเรียนรู้ได้ถูกต้องเฉลี่ย 97.45% ส่วนบล็อกข้อความจากภาพเอกสารที่ไม่เคยผ่านการเรียนรู้มาก่อน สามารถจำแนกได้ถูกต้องเฉลี่ย 96.43%

243 หน้า