THAI CHARACTER RECOGNITION ON TAX FORM

PHOORAT PAKAWATHANA

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

THAI CHARACTER RECOGNITION ON TAX FORM

PHOORAT PAKAWATHANA 4737985 ITCS/M

M.Sc. (COMPUTER SCIENCE)

RESEARCH PROJECT ADVISORY COMMITEE: SUKANYA PHONGSUPHAP, Ph.D., CHOMTIP PORNPANOMCHAI, Ph.D.

ABSTRACT

This research project develops a method to recognize the Thai characters on tax forms. There are five main steps in this method. 1) Preprocessing: The median filter is used to improve the quality of image by eradicating the noises and transform an image from 256 gray scale image to binary image. 2) Segmentation: The horizontal projection profile and heuristic rules are used to separate lines and then, the vertical and horizontal projection profiles are performed to segment characters. The result from this step is the single character images. 3) Feature Extraction: There are three different techniques used in this step consisting of finding the direction of the character's contour, the density of the character and the peripheral information of character. 4) Feature Selection: Sequential Forward Floating Selection (SFFS) is performed to select the proper features generated from the previous step. 5) Recognition: The Neural Networks are used to recognize the characters in each level.

In experiments, the test data having 4,000 characters, from the prepared documents obtained from a line printer and 8,391 characters from labels on tax forms are used to evaluate the system. The accuracy rates are approximately 98.88% and 93.77% respectively. The results indicate that the proposed method can work efficiently for recognition of printed Thai characters on labels on Tax form images.

KEY WORDS: CHARACTER RECOGNITION/ THAI CHARACTER RECOGNITION /NEURAL NETWORK/ CHARACTER SEGMENTATION/ FEATURE SELECTION/

148 pages.

การรู้จำตัวอักษรภาษาไทยบนแบบฟอร์มภาษี

THAI CHARACTER RECOGNITION ON TAX FORM

ภูรัฐ ภควัฒนะ 4737985 ITCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาโครงการวิจัย : สุกัญญา พงษ์สุภาพ, Ph.D., ชมทิพ พรพนมชัย, Ph.D.

บทคัดย่อ

งานวิจัยนี้ได้ศึกษาหาวิธีการรู้จำตัวอักษรภาษาไทยบนแบบฟอร์มภาษี โดยได้พิจารณาตัวอักษรภาษาไทยที่เป็นตัวพิมพ์ โดยแบ่งขั้นตอนการทำงานออกเป็น 5 ขั้นตอนหลักๆ ดังนี้ 1) การประมวลผลก่อน ได้ใช้ ตัวกรองมัธยฐาน (Median filter) ในการปรับปรุงคุณภาพและขจัดสัญญาณรบกวนของภาพตัวอักษร และทำการแปลงข้อมูลภาพจากความเทา 256 ระดับ (256 Gray Scale) ให้อยู่ในรูปแบบภาพสองระดับ (binary) 2) ขั้นตอนการตัดตัวอักษรภาพ โดยใช้วิธีการฉายเงาภาพในแนวตั้ง (Vertical Projection Profiles) การฉายเงาภาพในแนวนอน (Horizontal Projection Profiles) และใช้กฎในการตัดแยกตัวอักษร (Heuristic Rules) สำหรับการแบ่งแยกบรรทัด และ การตัดแยกตัวอักษรภาพ 3) การสกัดคุณลักษณะของตัวอักษรภาพได้นำเทคนิคต่างๆ มาใช้ 3 วิธีด้วยกันคือ การหาทิศทางของเส้นโครงร่างตัวอักษร การหาความหนาแน่นของตัวอักษร และการหาระยะห่างระหว่างขอบภาพและตัวอักษร 4) การเลือกคุณลักษณะของตัวอักษร ได้นำเอาวิธีการเลือกคุณลักษณะแบบ Sequential Forward Floating Selection (SFFS) มาใช้ในการเลือกคุณลักษณะของตัวอักษรภาพที่สกัดมาได้ 5) การรู้จำตัวอักษรภาพ ได้ใช้ Neural Network ในการรู้จำตัวอักษรโดยแบ่ง Neural Network ออกเป็น 3 กลุ่ม สำหรับการรู้จำตัวอักษรภาพในแต่ละระดับ จากการทดลองกับชุดทดสอบ 4,000 ตัวอักษรภาพที่พิมพ์มาจาก line printer สามารถรู้จำตัวอักษรได้ถูกต้องเฉลี่ย 98.88 เปอร์เซ็นต์ และเมื่อทดลองกับชุดข้อมูลภาพสลากบนแบบฟอร์มภาษี 8,391 ตัวอักษรสามารถรู้จำตัวอักษรได้ถูกต้องเฉลี่ย 93.77 เปอร์เซ็นต์ ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอทำงานได้อย่างมีประสิทธิภาพในการรู้จำตัวพิมพ์อักษรภาษาไทยจากสลากบนภาพแบบฟอร์มภาษี

148 หน้า