# RECOGNITION OF NUMERICAL DATA ON TAX FORMS

**BOONMA WANNABOOT**

**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**

RECOGNITION OF NUMERICAL DATA ON TAX FORMS

BOONMA WANNABOOT 4737984 ITCS/M

M.Sc.(COMPUTER SCIENCE)

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

ABSTRACT

This research project investigates a new approach to the recognition of numerical data appearing on tax documents. The proposed approach is able to recognize both handwritten and machine-printed numbers appearing on documents. The system uses an image processing technique that consists of five major steps: 1) Pre-processing using Gaussian filtering and Binarization; 2) Segmentation, which, in turn, consists of digit segmentation, broken digit restoration, and connected digit separation; 3) Feature extraction using Directional Distance Distribution, Concavities measurement, Crossing point, and Kirsch operators; 4) Feature selection using the Divergence value analysis; 5) Digit recognition using a Neural network classifier and the feature set from step 4.

To evaluate the recognition accuracy of the proposed approach, a total of 16,000 digits from actual tax documents were used in the experiments. The data can be divided as following: the training data set contained 14,400 digits (7,200 machine printed digits and 7,200 handwritten digits) and the testing data set contained 1,600 digits (800 machine-printed digits and 800 handwritten digits). The results showed that accuracy rates of recognition were 100% and 98.13% for machine-printed digits and handwritten digits respectively. The number of features used by the classifier was 235 out of the 442 available features. The results indicate that the proposed method work efficiently for the recognition of numerical data on tax forms, particularly for machine-printed digits.

KEY WORDS: HANDWRITTEN NUMERICAL RECOGNITION/ MACHINE-PRINTED NUMERICAL RECOGNITION/ FEATURE EXTRACTION/ FEATURE SELECTION/ NEURAL NETWORK

189 Pages

การรู้จำข้อมูลตัวเลขบนแบบฟอร์มภาษี

RECOGNITION OF NUMERICAL DATA ON TAX FORMS

บุญมา วรรณบุตร 4737984 ITCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

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

## บทคัดย่อ

สารนิพนธ์นี้เป็นการศึกษาวิธีการรู้จำและแยกแยะตัวเลขจากช่องจำนวนเงินบนแบบฟอร์มภาษี โดยพิจารณาทั้งตัวเลขที่เป็นตัวพิมพ์และลายมือเขียน โดยขั้นตอนที่ใช้ประกอบไปด้วย 5 ขั้นตอนดังนี้ 1) การเตรียมรูปภาพก่อนการประมวลผล ได้ใช้วิธีการปรับภาพด้วยเทคนิคการทำ Gaussian smoothing และการแปลงรูปภาพให้เป็นภาพขาวดำ 2) การตัดแยกรูปภาพ ได้ใช้ การตัดแยกรูปภาพตัวเลขออกเป็นตัวโดยวิธีการทำ Connected component labeling, การแก้ปัญหาการแตกของตัวเลข และ การแก้ปัญหาตัวเลขติดกันโดยใช้วิธีการทำ Drop fall algorithm 3) การสกัดคุณลักษณะสำคัญของตัวเลขได้หาลักษณะสำคัญต่อไปนี้ Directional Distance Distribution, Concavities measurement, Crossing point feature, และ Kirsch operators 4) การเลือกคุณลักษณะสำคัญ ใช้วิธีการหาค่า Divergence value 5) การรู้จำตัวเลขโดยใช้ Neural network เป็น Classifier และใช้ลักษณะสำคัญที่เลือกได้จากขั้นตอนที่ 4

จากการทดลองเพื่อทดสอบวิธีการที่เสนอกับข้อมูลตัวเลขจากแบบฟอร์มภาษีจำนวน 16,000 ตัว ซึ่งสามารถแบ่งออกเป็น ข้อมูลใช้ในการสอน 14,400 ตัว (เป็นตัวพิมพ์ 7,200 และลายมือเขียน 7,200 ตัว) และข้อมูลทดสอบ 1,600 ตัว (เป็นตัวพิมพ์ 800 ตัว และ ลายมือเขียน 800 ตัว) ได้อัตราความถูกต้องเฉลี่ย 100% และ 98.13% สำหรับการรู้ตัวเลขตัวพิมพ์และลายมือเขียนตามลำดับ โดยใช้ลักษณะสำคัญ 235 ลักษณะจาก ทั้งหมด 442 ลักษณะ ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอสามารถทำงานได้อย่างมีประสิทธิภาพในการรู้จำตัวเลขบนภาพแบบฟอร์มภาษีโดยเฉพาะอย่างยิ่งสำหรับการรู้จำตัวเลขที่เป็นตัวพิมพ์

189 หน้า