



การรู้จำตัวเลขอารบิกจากลายมือสำหรับระบุรหัสผู้สอบบนกระดาษคำตอบแบบปรนัย นุชนาฏ สัตยาภาวิ*, อีระเดช ราชไพบูลย์ และดวงเพ็ญ เจตนพิพัฒน์พงษ์

Handwritten Digits OCR for Identifying Examinee Number on Objective Test Answer Sheet

Nutchanat Sattayakawee*, Teeradetch Ratchapaiboon and Duangpen Jetpipattanapong

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ กำแพงแสน มหาวิทยาลัยเกษตรศาสตร์ อำเภอกำแพงแสน จังหวัดนครปฐม 73140
Department of Computer Engineering, Faculty of Engineering at Kamphaengsaen, Kasetsart University, Kamphaengsaen, Nakhonpathom, 73140

* Corresponding author. E-mail address: fengnncn@ku.ac.th

บทคัดย่อ

บทความนี้นำเสนอการรู้จำตัวเลขจากลายมือโดยการประมวลผลภาพ ซึ่งมีเป้าหมายเพื่อนำไปใช้ในการระบุเลขประจำตัวผู้สอบในกระดาษคำตอบของข้อสอบแบบปรนัย โดยใช้ภาพจากเครื่องสแกน ซึ่งรหัสประจำตัวผู้สอบจะต้องเขียนเป็นตัวเลขอารบิกด้วยปากกาบลูกลิ้งน้ำเงินไว้ที่มุมบนซ้ายของกระดาษคำตอบ หลักการที่ใช้ในการรู้จำตัวเลข คือ เทคนิคโปรเจกชันในแนวแกน 4 แกน ได้แก่ แนวนอน (Horizontal) แนวตั้ง (Vertical) แนวทแยงซ้าย (Left Diagonal) และแนวทแยงขวา (Right Diagonal) และนำวิธีทางสถิติอย่างง่ายมาช่วยวิเคราะห์คำตอบ จากการทดสอบความถูกต้องของการทำงานโดยตรวจสอบตัวเลขอารบิกจำนวน 600 ตัว โดยมีเลขอารบิก 0 - 9 อย่างละ 60 ตัว พบว่า ระบบสามารถแยกตัวเลขได้ถูกต้อง 465 ตัว คิดเป็นร้อยละ 77.50

คำสำคัญ: การรู้จำตัวเลข ลายมือ โปรเจกชัน ฮิสโตแกรม การรู้จำอักขระทางภาพ

Abstract

This paper presents the recognition of handwritten digits using image processing. It aims at identifying identification numbers of examinees on multiple choice answer sheets. Input images are from a scanner. In addition, an identification number must be written in Arabic numerals with a blue ballpoint pen on the upper left corner of an answer sheet. The proposed technique is based on projection profiles of 4 axes: horizontal, vertical, left diagonal and right diagonal axes. Simple statistical analysis is then applied in order to recognize the characters. With the test samples of 600 digits, each of which was represented by 60 samples, the test result showed that the algorithm recognized 465 digits correctly, giving the accuracy rate of 77.50 %.

Keywords: Digit Recognition, Handwritten, Projection, Histogram, Optical Character Recognition

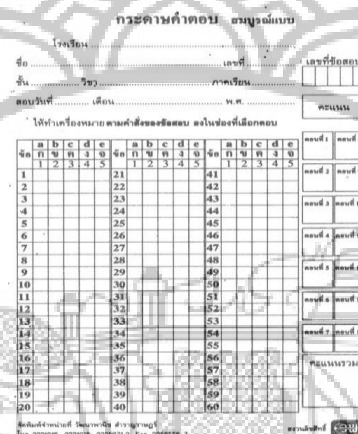
บทนำ

ข้อสอบแบบปรนัยถูกนำมาใช้อย่างแพร่หลายในการวัดผลหลายอย่าง เช่น การวัดระดับความรู้ทางการศึกษา การวัดความถนัด การทดสอบทางจิตวิทยา (Great Schools Partnership, 2013) ซึ่งข้อสอบมาตรฐานส่วนใหญ่ เช่น ข้อสอบ ก.พ. ข้อสอบ TOEIC เป็นต้น ก็มักจะมีส่วนของคำถามที่เป็นแบบปรนัยทั้งสิ้น (ชลสิทธิ์ ชำญนุชิต, 2557;

Educational Testing Service, 2012) ตามนิยามแล้วข้อสอบปรนัยคือข้อสอบที่มีการให้คะแนนที่ตายตัวโดยผู้ตรวจไม่สามารถใส่ความคิดเห็นลงไปได้ ทำให้รูปแบบของคำถามมักจะเป็นแบบที่มีตัวเลือกมากกว่า 1 ตัวเลือก แล้วให้ผู้ตอบตัดสินใจเลือกคำตอบที่พิจารณาว่าเหมาะสมที่สุด หรือการเติมคำตอบสั้นๆ (Murayama, 2009) โดยในที่นี้ จะสนใจเฉพาะคำถามแบบที่มีหลายตัวเลือก

ในสถาบันการศึกษาที่ต่ำกว่าระดับอุดมศึกษาของประเทศไทย มีการใช้กระดาษคำตอบของข้อสอบแบบปรนัยอยู่ 2 แบบ คือกระดาษคำตอบแบบแผ่นที่ตรวจด้วยเทคนิค OMR (Optical Mark Recognition) และกระดาษแบบกากบาท ดังรูปที่ 1 ซึ่งจากผลสำรวจการใช้กระดาษคำตอบในโรงเรียนระดับมัธยมศึกษาใน

(Sattayakawee, 2013, pp. 273-277) พบว่า 56 จาก 74 โรงเรียนใช้กระดาษคำตอบแบบกากบาท ซึ่งเป็นที่ทราบกันดีอยู่แล้วว่าราคาของกระดาษคำตอบแบบกากบาทมีราคาต่ำกว่ากระดาษคำตอบแบบแผ่นที่ใช้ตรวจกับเครื่องตรวจข้อสอบอัตโนมัติมาก



รูปที่ 1 กระดาษคำตอบแบบกากบาท

ในการตรวจข้อสอบใดๆ ผู้ประเมินจะต้องสามารถระบุผู้ที่ถูกประเมินหรือผู้ทำข้อสอบได้ การตรวจกระดาษคำตอบแบบ OMR ได้ถูกพัฒนาเป็นระบบที่มีการวางขายตามท้องตลาดและมีความแม่นยำสูง ในขณะที่กระดาษแบบกากบาทยังไม่มีการพัฒนาเพื่อการตรวจกระดาษคำตอบแบบกากบาทได้มีการพัฒนามาบ้างแล้วดังใน (Sattayakawee, 2013, pp. 273-277; พุทธินันท์ พัดกระจำ, 2555) แต่มีเพียงการตรวจเฉพาะส่วนของคำตอบเท่านั้นและยังไม่สามารถระบุผู้สอบได้เนื่องจากข้อมูลประจำตัวของผู้เข้าสอบเป็นการเขียนกรอกข้อมูลในส่วนหัวของกระดาษ ซึ่งจะต้องใช้เทคนิคการรู้จำตัวอักษรหรือ OCR (Optical Character Recognition) ซึ่งหมายถึง การแปลงข้อมูลตัวอักษรที่อยู่ในรูปแบบของภาพให้อยู่ในรูปแบบของข้อความในงานวิจัยที่ปรากฏใน (Theeramunkong, & Wongtapan, 2005, pp. 139-160; Pornpanomchai, Batanov, & Dimmitt, 2001, pp. 259-279; Chaowicharat, Cercone, & Naruedomkul, 2012, pp. 196-201). การรู้จำตัวอักษรภาษาไทยมีความถูกต้องร้อยละ

52.00-99.89 ในขณะที่การรู้จำตัวเลขอารบิกใน (Lauer, Suen, & Bloch, 2007, pp. 1816-1824; Choudhary, Rishi, & Ahlawat, 2012, pp. 119-126; Neves, Filho, Mello, & Band Zanchettin, 2011, pp. 501-515; Miri, Razavi, & Sadri, 2011, pp. 150-155) มีความถูกต้องร้อยละ 81.63-99.78 โดยจะเห็นได้ว่า การรู้จำตัวเลขอารบิกมีความถูกต้องอยู่ในช่วงที่แคบกว่าการรู้จำตัวอักษร จึงถือได้ว่าการรู้จำตัวเลขมีความน่าเชื่อถือกว่าอีกทั้งตัวเลขมีเพียง 10 ตัวซึ่งถือว่าน้อยกว่าตัวอักษรภาษาไทยมาก ทางผู้วิจัยจึงเห็นประโยชน์ในการพัฒนาการรู้จำตัวเลขอารบิกจากลายมือเพื่อนำมาใช้ในการระบุรหัสประจำตัวผู้สอบ ซึ่งถือเป็นส่วนเติมเต็มสำหรับการตรวจกระดาษคำตอบแบบกากบาทจาก (Sattayakawee, 2013; pp. 273-277; พุทธินันท์ พัดกระจำ, 2555) นอกจากนี้ เพื่อลดความผิดพลาดของการระบุรหัสผู้สอบที่อาจเกิดขึ้น รหัสผู้สอบจะถูกเขียนไว้ที่มุมบนซ้ายของกระดาษคำตอบในรูปที่ 1

เทคนิคที่ใช้ในการรู้จำตัวอักษรรวมถึงตัวเลขมีหลายเทคนิคด้วยกันซึ่งเทคนิคที่เห็นได้บ่อยอาจสรุปได้ดังนี้



1.1 การเข้าคู่รูปแบบ (Template Matching) (Du, Ibrahim, Shehata, & Badawy, 2013, pp. 311-325) ซึ่งจะมีรูปแบบของแต่ละตัวอักษร และการรู้จำ คือ กระบวนการจับคู่ว่าตัวอักษรนำเข้ามีความคล้ายคลึงกับรูปแบบของตัวอักษรใดมากที่สุดโดยส่วนใหญ่แล้วเทคนิคนี้มักนำมาใช้กับอักษรที่เป็นตัวพิมพ์

1.2 การวิเคราะห์โครงสร้าง (Structural Analysis) (Singh, 2013, pp. 545-550; Cheng, & Yan, 1998, pp. 235-255) ซึ่งหมายถึงการวิเคราะห์คุณสมบัติของตัวอักษรแต่ละตัวในเชิงโครงสร้างหรือรูปร่างของมัน เช่น การวิเคราะห์เส้นขอบ การวิเคราะห์การกระจายความเข้มของจุดสีของภาพ การวิเคราะห์จุดศูนย์กลางของภาพ เป็นต้น การรู้จำจะเกิดจากการสกัดลักษณะสำคัญ (feature extraction) ของแต่ละตัวอักษรจากโครงสร้างที่ใช้

1.3 วิธีทางสถิติ (Statistical Approach) (Theeramunkong, & Wongtapan, 2005, pp. 139-160) ซึ่งใช้ทฤษฎีความน่าจะเป็นในการรู้จำตัวอักษรจากชุดของลักษณะสำคัญที่ได้มา ซึ่งลักษณะสำคัญนั้นมักจะมาจากโครงสร้างของตัวอักษรหรือจุดสีของภาพ การรู้จำตัวอักษรด้วยวิธีทางสถิติ คือ การหาตัวอักษรที่มีค่าความน่าจะเป็นมากที่สุดจากชุดของภาพที่ผ่านการเรียนรู้ (train) มาแล้วเมื่อนำลักษณะสำคัญของตัวอักษรนำเข้าไปประมวลผลตัวอย่างการรู้จำด้วยวิธีทางสถิติ ได้แก่ โมเดล n-gram โมเดล Hidden Markov (HMM) เป็นต้น

1.4 โครงข่ายประสาทเทียม (Neural Network) (Singh, 2013, pp. 545-550; Lauer, Suen, & Bloch, 2007, pp. 1816-1824; Choudhary, Rishi, & Ahlawat, 2012, 119-126; Miri, Razavi, & Sadri, 2011, pp. 150-155) ซึ่งเป็นการเลียนแบบการทำงานของระบบประสาทของมนุษย์ โดยการรู้จำเกิดจากการนำจุดสีของภาพผ่านกระบวนการเรียนรู้ด้วยฟังก์ชันทางคณิตศาสตร์เพื่อให้ได้รูปแบบของตัวอักษร

1.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) (Lauer, Suen, & Bloch, 2007, pp. 1816-1824; Neves, Filho, Mello, Band Zanchettin, 2011, pp. 501-515) ซึ่งเป็นอัลกอริทึมสำหรับการจัดหมวดหมู่ (classification) โดยสามารถจัดได้เพียง 2 หมวดหมู่โดยจะมีฟังก์ชันในการแบ่งข้อมูลนำเข้า

ออกเป็น 2 กลุ่ม ดังนั้นการใช้ SVM ในการรู้จำจะต้องมีโครงสร้าง SVM หลายอันเพื่อให้เพียงพอต่อการรู้จำแต่ละตัวอักษรได้ วิธีการนี้เป็นวิธีการหนึ่งที่มีประสิทธิภาพในการรู้จำตัวเลข

นอกจากวิธีที่กล่าวไปแล้ว ยังมีอีกหลายวิธีที่ถูกนำมาใช้ในกระบวนการรู้จำดังปรากฏใน (Singh, 2013, pp. 545-550) และส่วนใหญ่มักจะเป็นการนำวิธีการมากกว่า 1 วิธีมาใช้ร่วมกันเพื่อประสิทธิภาพที่ดียิ่งขึ้น อย่างไรก็ตาม เทคนิคการรู้จำใน 1.3 - 1.5 มักจะมีความซับซ้อนมากและต้องใช้ข้อมูลจำนวนมากในการวิเคราะห์ ดังนั้น ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาเทคนิคการรู้จำอย่างง่าย โดยใช้หลักการวิเคราะห์โครงสร้างของตัวเลข ยิ่งไปกว่านั้น เมื่อมีข้อมูลเพิ่มขึ้น การสกัดลักษณะสำคัญของวิธีการที่นำเสนอไม่จำเป็นต้องเริ่มต้นขั้นตอนการเรียนรู้ใหม่ทั้งหมด

วิธีการ

ในงานวิจัยนี้ ผู้วิจัยได้นำเทคนิคโปรเจกชันมาใช้ควบคู่กับหลักสถิติอย่างง่ายในการสกัดลักษณะสำคัญและการจัดหมวดหมู่ โดยภาพจะถูกนำเข้าจากการสแกนด้วยเครื่องพิมพ์แบบมัลติฟังก์ชันรุ่น HP Laserjet M1536dnf MFP ด้วยความละเอียด 300 dpi และเป็นกระดาษคำตอบแบบกากบาทที่มีการเขียนตัวเลขอารบิกไว้ที่มุมบนซ้ายของกระดาษดังแสดงในรูปที่ 2 โดยตัวเลขที่เขียนไม่เขียนติดกันจนเกินไป ควรมีความกว้าง 0.3-0.5 ซม. และความสูง 0.4-0.6 ซม. เมื่อได้ภาพที่จะนำมาประมวลผลแล้ว ก็จะเข้าสู่ขั้นตอนในการรู้จำตัวเลขอารบิกจากลายมือ ซึ่งแบ่งเป็น 3 ขั้นตอน คือ การประมวลผลภาพเบื้องต้น การแยกองค์ประกอบของภาพ และการรู้จำภาพโดยอธิบายได้ ดังนี้

2.1 การประมวลผลภาพเบื้องต้น (Image Pre-processing)

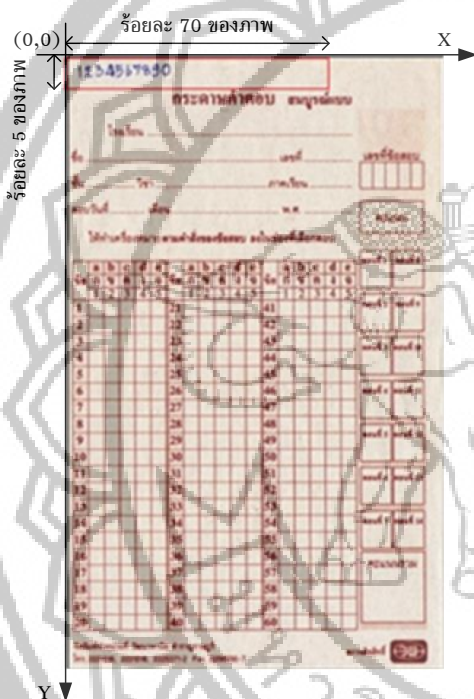
การประมวลผลภาพเบื้องต้นมีหลายขั้นตอน ได้แก่ การตัดภาพในส่วนที่ต้องการ โดยตัดภาพจากพิกัดที่ (0,0) ในแกน X และ Y โดยมีขนาดความกว้างเท่ากับร้อยละ 70 ของภาพและความสูงเท่ากับร้อยละ 5 ของภาพ การแปลงภาพสีให้เป็นภาพระดับเทาการแปลงภาพระดับเทาให้เป็นภาพขาวดำ และการกำจัดสัญญาณ

รบกวนด้วย Median smooth ซึ่งจะทำให้การแทนค่าจุดสีของภาพในแต่ละตำแหน่งของภาพด้วยค่ามัธยฐานของชุดของจุดสีในพื้นที่ของการกรอง (filter region) ขนาด $n \times n$ (Burger, & Burge, 2008, pp. 106 – 110,

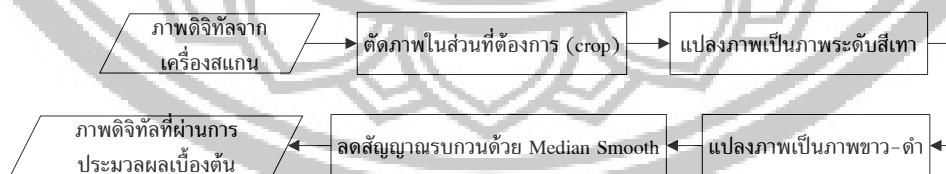
200 – 205) ซึ่งในที่นี้ ใช้ 5×5 โดยจุดสีของภาพใหม่ในแต่ละตำแหน่งเขียนได้ดังสมการที่ (1) โดย I' แทนภาพที่ผ่านการกำจัดสัญญาณรบกวนแล้ว และ I แทนภาพก่อนการกำจัดสัญญาณรบกวน

$$I'(x,y) = \text{median}\{I(u,v) | u \in \{i | x-2 \leq i \leq x+2\}, v \in \{j | y-2 \leq j \leq y+2\}\} \quad (1)$$

รูปที่ 3 แสดงขั้นตอนในการประมวลผลภาพเบื้องต้น ผลลัพธ์ที่ได้ คือ ภาพของชุดตัวเลขอารบิกที่พร้อมนำไปเข้าสู่กระบวนการแยกองค์ประกอบของรูปภาพดังแสดงในรูปที่ 4



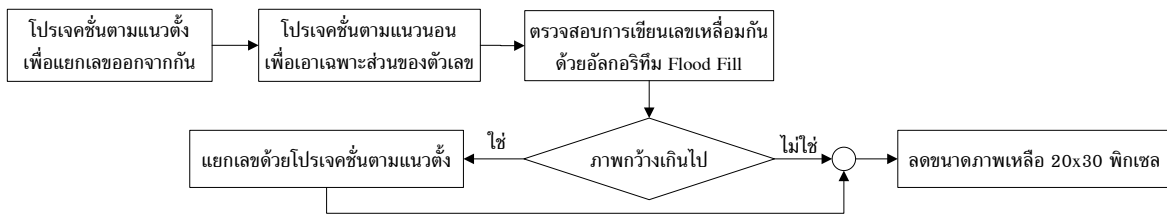
รูปที่ 2 ตัวอย่างรูปภาพกระดาษคำตอบแบบกากบาทที่มีการเขียนตัวเลขอารบิกไว้ที่มุมบนซ้ายของกระดาษ



รูปที่ 3 กระบวนการประมวลผลเบื้องต้น

1234567890

รูปที่ 4 ตัวอย่างรูปภาพที่ผ่านการประมวลผลภาพเบื้องต้น



รูปที่ 5 กระบวนการแยกองค์ประกอบของรูปภาพ

2.2 การแยกองค์ประกอบของรูปภาพ (Image Segmentation)

ขั้นตอนในการแยกองค์ประกอบของรูปภาพแสดงในรูปที่ 5 โดยมีเป้าหมายเพื่อแยกภาพของตัวเลขแต่ละตัวออกจากกัน วิธีการโปรเจกชันหรือฮิสโตแกรมได้ถูกนำมาใช้โดยเป็นการนับจำนวนจุดสีดำของภาพตามแนวแกนใดแกนหนึ่ง โดยในที่นี้ใช้การโปรเจกชันใน 2 แกน คือ

2.2.1 โปรเจกชันตามแนวตั้ง (Vertical Projection) ซึ่งใช้ในการแยกตัวเลขแต่ละตัวออกจากกัน โดยจะตัดตัวอักษรเมื่อมีจำนวนจุดสีดำเป็น 0 ในแนวแกน Y ซึ่งภาพที่ต้องการ คือ ภาพที่มีความกว้างพอดีกับเลขแต่ละตัว

2.2.2 โปรเจกชันตามแนวนอน (Horizontal Projection) ซึ่งนำมาใช้ในการหาแนวของบรรทัดบนกระดาษที่มีการเขียนตัวเลขหลังจากที่ผ่านโปรเจกชันตามแนวตั้งแล้ว โดยจะตัดส่วนของกระดาษที่มีจำนวนจุดสีดำเป็น 0 ในแนวแกน X ทำให้ภาพที่ได้มีความสูงพอดีกับความสูงของเลขแต่ละตัว

การแยกตัวเลขด้วยโปรเจกชันตามแนวตั้งอาจมีความผิดพลาดเกิดขึ้นได้ เนื่องจากตัวเลขอาจมีการเขียนเหลื่อมล้ำกันทำให้ไม่สามารถตัดตัวอักษรได้ ดังตัวอย่างในรูปที่ 6 ซึ่งเลข 3 มีการเขียนเหลื่อมกับเลข 2 อัลกอริทึม Flood Fill (Burger, & Burge, 2008, pp. 106 – 110, 200 – 205) จึงถูกนำมาประยุกต์ใช้เพื่อแยกเลขที่เขียนเหลื่อมกันออกจากกัน โดยอัลกอริทึมนี้สามารถใช้ในการหาพื้นที่ของภาพที่เราสนใจ โดยในที่นี้จะทำการสร้างภาพใหม่ที่มีขนาดเท่ากับภาพต้นฉบับและมีพื้นหลังเป็นสีขาว ซึ่งในตัวอย่างภาพต้นฉบับคือรูปที่ 6 (ก) แล้วตรวจหาพิกเซลสีดำจากตำแหน่งแรกและตำแหน่งที่อยู่รอบๆ ทั้ง 8 ด้านของตำแหน่งที่กำลัง

พิจารณาและคัดลอกไปยังภาพที่สร้างขึ้นใหม่เรื่อยๆ จนกว่าจะเจอพิกเซลสีขาวล้อมรอบทั้งหมด โดยจากตัวอย่างจะทำให้เราคัดลอกได้ภาพแรกเป็นเลข 3 ในรูปที่ 6 (ข) หลังจากนั้นจึงแทนที่ภาพต้นฉบับในตำแหน่งของเลข 3 ด้วยสีขาวทั้งหมด ตัดภาพ (crop) ใหม่ ทำให้เหลือภาพเฉพาะพื้นที่ของเลข 20 แล้วเริ่มการ Flood Fill ใหม่ ซึ่งในรอบต่อมา เราจะได้ภาพต้นฉบับและภาพหลังจากการ Flood Fill เป็นภาพเหมือนกัน จึงถือได้ว่าไม่มีภาพที่มีการเขียนเลขเหลื่อมกันแล้ว จึงถือเป็นอันสิ้นสุดการ Flood Fill

ในกรณีที่ตัวเลขเขียนติดกันทำให้ไม่สามารถตัดเลขได้อย่างถูกต้อง โดยจะดูจากขนาดของภาพที่ตัดได้ซึ่งมีความกว้างมากเกินไป การวิเคราะห์ฮิสโตแกรมตามแนวตั้งจะถูกนำมาใช้ โดยตรวจสอบหาตำแหน่งตามแนวแกน X ที่มีจำนวนของจุดสีดำน้อยที่สุด ซึ่งจะตรวจหาเฉพาะตำแหน่งในแกน X ในช่วง 25 – 75 % ของส่วนของรูปภาพ ตัวอย่างของตัวเลขที่เขียนติดกันและฮิสโตแกรมแสดงได้ดังรูปที่ 7 ซึ่งจะทำให้เราสามารถตัดเลข 2 และ 0 ออกจากกัน ณ ตำแหน่งที่ชี้โดยลูกศรเมื่อได้รูปภาพตัวอักษรที่ผ่านการแยกองค์ประกอบของรูปภาพแล้ว จึงทำการลดขนาดภาพเพื่อเพิ่มความเร็วในการประมวลผลต่อไป

2.3 การรู้จำภาพ (Image Recognition)

ก่อนอื่น เราจะต้องกำหนดชุดข้อมูลที่จะนำมาใช้ ซึ่งแบ่งออกเป็น 2 ชุด เรียกว่า ชุดเรียนรู้ (Training Set) และ ชุดทดสอบ (Test Set) โดยข้อมูลได้มาจากลายมือของนักเรียนโรงเรียนชลกันยานุกูล จ.ชลบุรี จำนวน 105 คน โดยเขียนรหัสนักเรียนซึ่งเป็นเลข 5 หลัก คนละ 2 ชุด หลังจากผ่านกระบวนการประมวลผลภาพเบื้องต้นแล้วมีตัวเลขที่นำมาใช้ได้ทั้งสิ้น 1008 ตัว เนื่องจากมีความผิดพลาดในการตัดแบ่งเลข

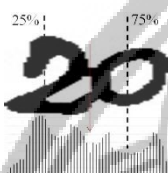


และการปรับสภาพแสงของเลขบางตัว เลข 1008 ตัวถูกนำมาใช้เป็นชุดทดสอบ 300 ตัวหรือประมาณร้อยละ 30 ซึ่งเป็นเลข 0 - 9 เลขละ 30 ตัว และใช้เป็นชุดเรียนรู้อีก

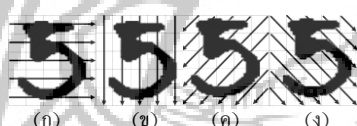
708 ตัว ซึ่งจำนวนตัวเลขแต่ละตัวแจกแจงในตารางที่ 1 กระบวนการรู้จำแบ่งออกเป็น 2 ขั้นตอน ดังนี้

320 3 20
(ก) (ข)

รูปที่ 6 ตัวอย่างภาพที่มีการเขียนเลขเหมือนกัน (ก) ภาพเริ่มต้น (ข) ภาพหลังจากการแยกเลขด้วย Flood Fill



รูปที่ 7 ตัวอย่างฮิสโตแกรมของรูปภาพตัวเลขติดกัน



รูปที่ 8 วิธีโปรเจกชันทั้ง 4 แกนได้แก่ (ก) แนวนอน (ข) แนวตั้ง (ค) แนวทแยงซ้าย และ (ง) แนวทแยงขวา

ตารางที่ 1 จำนวนตัวเลขชุดที่ใช้ในการทดลอง

ตัวเลข	0	1	2	3	4	5	6	7	8	9
จำนวนสำหรับชุดเรียนรู้ (ตัว)	64	67	71	80	79	73	71	72	61	70
จำนวนสำหรับชุดทดสอบ (ตัว)	30	30	30	30	30	30	30	30	30	30

2.3.1 การเรียนรู้ตัวเลข

ขั้นตอนนี้จะเกี่ยวข้องกับการนำข้อมูลทั้งหมดมาสกัดลักษณะสำคัญเพื่อจัดหมวดหมู่ของเลข 0-9 โดยใช้วิธีการโปรเจกชัน ซึ่งแจกแจงเป็น 2 ขั้นตอนย่อยและมีรายละเอียดดังนี้

1. โปรเจกชัน 4 แกน

ในเบื้องต้น ลักษณะของตัวเลขจะถูกสกัดจากโปรเจกชันใน 4 แกน คือ โปรเจกชันตามแนวนอน (Horizontal Projection) โปรเจกชันตามแนวตั้ง (Vertical Projection) โปรเจกชันตามแนวทแยงซ้าย (Left diagonal Projection) และโปรเจกชันตามแนวทแยงขวา (Right diagonal Projection) ดังรูปที่ 8

2. การสกัดลักษณะสำคัญของเลข 0 - 9

ลักษณะสำคัญของเลข 0 - 9 จะได้จากการวิเคราะห์ค่าจากการโปรเจกชันทั้ง 4 แกน ของเลขทั้งหมด 708 ตัว โดยแยกเป็นหมวดหมู่ตามเลขตั้งแต่ 0 ถึง 9 ค่าจากการโปรเจกชันทั้ง 4 แกนของตัวเลขที่อยู่ในแต่ละหมวดหมู่จะถูกนำมาหาค่ากลางทางสถิติ 3 แบบ คือ ค่าเฉลี่ย (Average) มัธยฐาน (Median) และฐานนิยม (Mode) ทำให้ลักษณะสำคัญของเลขแต่ละหมวดหมู่มีด้วยกัน 12 ลักษณะ โดยมาจากแกนของการโปรเจกชัน 4 แกนและแต่ละแกน นำมาคำนวณค่ากลางทางสถิติ 3 แบบ ลักษณะสำคัญเหล่านี้จะถูกใช้ในการอ้างอิงเพื่อตัดสินใจว่าภาพนำเข้าตรงกับลักษณะของเลขใดในขั้นตอนต่อไป



2.3.2 การหาคำตอบ

ในขั้นตอนนี้ ลักษณะสำคัญของเลขที่ได้จากการโปรเจกชันใน 2.3.1 จะถูกนำมาใช้เป็นค่ากลางทางสถิติในการหาคำตอบให้กับภาพนำเข้าโดยเราจะหาค่าจากการโปรเจกชันทั้ง 4 แกนของภาพนำเข้า

แล้วเทียบกับค่าลักษณะสำคัญที่ได้จาก 2. ของเลขทั้ง 10 ตัว โดยหลักการที่ใช้คือ การเลือกตัวเลขที่มีระยะห่างระหว่างลักษณะสำคัญของเลขกับค่าโปรเจกชันของภาพนำเข้าน้อยที่สุด (minimum distance) เป็นคำตอบ โดยสมการหาระยะห่างแสดงดังสมการที่ (2)

$$\sum_{i=0}^{n-1} |x_{(i)} - xm_{(i)}| = |x_{(0)} - xm_{(0)}| + |x_{(1)} - xm_{(1)}| + \dots + |x_{(n-1)} - xm_{(n-1)}| \quad (2)$$

โดยที่ $x_{(i)}$ คือ ค่าการโปรเจกชันของภาพนำเข้าตำแหน่งที่ i
 $xm_{(i)}$ คือ ค่ากลางทางสถิติของการโปรเจกชันตำแหน่งที่ i
 n คือ จำนวนข้อมูลของการโปรเจกชัน
 i คือ ตำแหน่งของข้อมูลในการโปรเจกชัน

การหาคำตอบสำหรับภาพนำเข้าภาพหนึ่งแบ่งออกเป็น 2 ขั้นตอนย่อยโดยมีรายละเอียด ดังนี้

1. วิเคราะห์คำตอบจากโปรเจกชัน 4 แกน

การวิเคราะห์คำตอบจากการโปรเจกชัน 4 แกนจะทำทั้งหมด 3 ครั้ง โดยใช้ค่ากลาง 3 แบบแตกต่างกันในการแทนค่า $xm_{(i)}$ ในสมการที่ (2) คำตอบที่ได้จากการใช้ค่ากลางแต่ละแบบจะได้ตัวเลข 4 ตัว ซึ่งเป็นตัวเลขที่มีระยะห่างระหว่างลักษณะสำคัญของเลขกับค่าโปรเจกชันของภาพนำเข้าที่น้อยที่สุดจากการโปรเจกชันทั้ง 4 แกนอัลกอริทึมนี้จะเลือกตอบตัวเลขที่มีความถี่มากที่สุด เช่น หากการเปรียบเทียบด้วยการหาระยะห่างที่น้อยที่สุดของการใช้ค่าเฉลี่ยเป็นค่ากลาง ได้เลข 2 2 4 1 ผลการตัดสินใจของอัลกอริทึมจะได้เลข 2

ในกรณีที่คำตอบที่ได้มีความถี่เท่ากันหมด ร้อยละความถูกต้องของคำตอบโปรเจกชัน 4 แกนจากตัวอย่างเรียนรู้ในตารางที่ 2 จะถูกนำมาใช้อ้างอิง โดยเลือกเลขที่มีร้อยละความถูกต้องสูงสุดในแกนโปรเจกชันนั้น ๆ ทั้งนี้ ข้อมูลจากตารางที่ 2 มาจากการสุ่มตัวเลขจากชุดเรียนรู้จำนวนเลขละ 30 ตัว แล้วบันทึกผลความถูกต้องของการตัดสินใจของเลขเหล่านั้นเมื่อเทียบกับค่าลักษณะสำคัญของตัวเลขในแต่ละแกนโดยใช้ค่ากลางทางสถิติแต่ละแบบ แต่หากร้อยละความถูกต้องของตัวเลขที่ได้ยังคงเท่ากัน ระบบจะตัดสินใจโดยเลือกแกนโปรเจกชันที่น่าเชื่อถือเรียงจากอันดับ 1 ไปอันดับ 4 ดังตารางที่ 3 ซึ่งเป็นการเรียงลำดับร้อยละความถูกต้องเฉลี่ยที่ได้มาจากตารางที่ 2 คำตอบที่ได้จากขั้นตอนนี้มี 3 คำตอบซึ่งได้จากการใช้ค่ากลางแต่ละแบบ

ตารางที่ 2 ร้อยละของความถูกต้องของคำตอบโปรเจกชัน 4 แกนจากตัวอย่างเรียนรู้ของค่ากลางทั้ง 3 แบบ

ตัวเลข	แนวนอน (%)			แนวตั้ง (%)			แนวทแยงซ้าย (%)			แนวทแยงขวา (%)		
	ค่าเฉลี่ย	มัธยฐาน	ฐานนิยม	ค่าเฉลี่ย	มัธยฐาน	ฐานนิยม	ค่าเฉลี่ย	มัธยฐาน	ฐานนิยม	ค่าเฉลี่ย	มัธยฐาน	ฐานนิยม
0	33.33	60.00	56.67	70.00	66.67	66.67	63.33	60.00	40.00	66.67	66.67	73.33
1	86.67	83.33	90.00	70.00	73.33	40.00	86.67	80.00	76.67	63.33	53.33	56.67
2	70.00	73.33	83.33	50.00	60.00	63.33	53.33	63.33	76.67	60.00	63.33	70.00
3	50.00	50.00	46.67	23.33	40.00	60.00	53.33	60.00	36.67	53.33	56.67	56.67
4	73.33	80.00	53.33	6.67	30.00	36.67	70.00	80.00	70.00	63.33	63.33	80.00
5	50.00	53.33	33.33	13.33	13.33	13.33	40.00	53.33	40.00	36.67	53.33	56.67
6	66.67	73.33	80.00	70.00	70.00	63.33	36.67	40.00	33.33	90.00	86.67	70.00
7	90.00	90.00	83.33	70.00	70.00	73.33	93.33	96.67	96.67	90.00	80.00	86.67
8	56.67	73.33	86.67	20.00	16.67	40.00	43.33	46.67	30.00	63.33	70.00	56.67
9	63.33	73.33	70.00	73.33	76.67	80.00	53.33	46.67	56.67	56.67	56.67	56.67
เฉลี่ย	64.00	71.00	68.33	46.67	51.67	53.67	59.33	62.67	55.67	64.33	65.00	66.33



ตารางที่ 3 ลำดับความสำคัญสำหรับการตัดสินใจเมื่อคำตอบที่ได้จากขั้นตอนวิเคราะห์จากคำตอบโปเจคชั่น 4 แกนมีน้ำหนักเท่ากัน

ชนิดของค่ากลาง	ลำดับ			
	1	2	3	4
ค่าเฉลี่ย	แนวทแยงขวา	แนวนอน	แนวทแยงซ้าย	แนวตั้ง
มัธยฐาน	แนวนอน	แนวทแยงขวา	แนวทแยงซ้าย	แนวตั้ง
ฐานนิยม	แนวนอน	แนวทแยงขวา	แนวทแยงซ้าย	แนวตั้ง

ตารางที่ 4 ร้อยละของความถูกต้องของคำตอบของตัวเลขอารบิกของชุดเรียนรู้จากค่ากลางทั้ง 3 แบบ

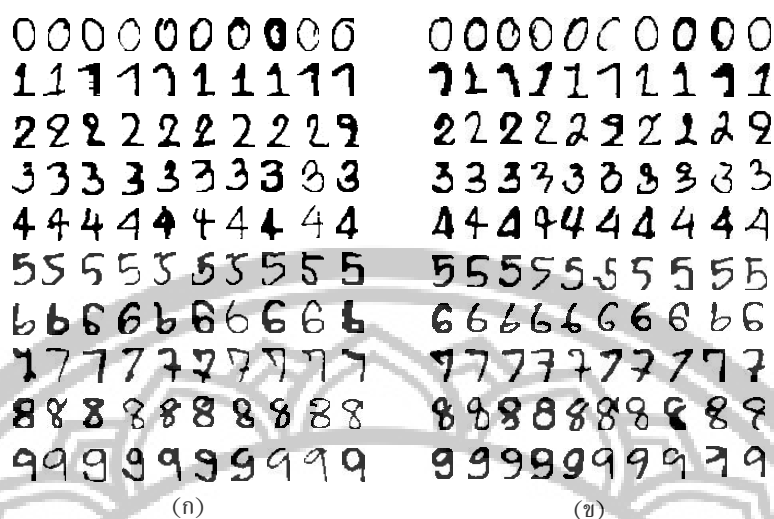
ตัวเลข	ค่าเฉลี่ย (%)	มัธยฐาน (%)	ฐานนิยม (%)
0	76.67	73.33	76.67
1	86.67	86.67	96.67
2	70.00	73.33	86.67
3	46.67	50.00	60.00
4	70.00	83.33	73.33
5	43.33	56.67	40.00
6	86.67	90.00	80.00
7	96.67	96.67	96.67
8	60.00	70.00	73.33
9	76.67	80.00	83.33
เฉลี่ย	71.33	76.00	76.67

ตารางที่ 4 สรุปร้อยละความถูกต้องของคำตอบของแต่ละเลขจากการใช้ค่ากลางทางสถิติแต่ละแบบในการตัดสินใจ โดยใช้ระยะห่างที่น้อยที่สุดร่วมกับค่าร้อยละของความถูกต้องของคำตอบในตารางที่ 2 หลังจากนั้นจึงเข้าสู่ขั้นตอนที่ 2 ของการวิเคราะห์

2. วิเคราะห์จากคำตอบของค่ากลางทาง

สถิติ 3 แบบ

ในขั้นตอนนี้ จะวิเคราะห์จากคำตอบ 3 คำตอบโดยตัดสินใจจากอย่างน้อย 2 ใน 3 ของคำตอบว่าเป็นตัวเลขเดียวกัน ในกรณีที่ไม่มีคำตอบเหมือนกันเลย จะตัดสินใจเลือกคำตอบของค่ากลางแบบฐานนิยม เนื่องจากตารางที่ 4 แสดงให้เห็นว่าค่าเฉลี่ยของความถูกต้องของเลขทุกตัวในฐานนิยมมีค่าสูงที่สุด



รูปที่ 9 ตัวอย่างภาพที่นำมาทดสอบสำหรับ (ก) ชุดเรียนรู้ (ข) ชุดทดสอบ



รูปที่ 10 ตัวอย่างภาพการเขียนเลขเจ็ด (ก) แบบมีขีดกลาง (ข) แบบมีหัว (ค) แบบไม่มีหัว

ผลและอภิปราย

ในการทดสอบจะใช้ตัวเลข 2 ชุดข้อมูล ชุดละ 300 ตัว โดยมีเลข 0 - 9 อย่างละ 30 ตัว โดย ชุดแรกเป็นชุดเรียนรู้ที่สุ่มมาแล้วใน 2.3.2 และ ชุดที่สองเป็นชุดทดสอบ ดังมีตัวอย่างแสดงในรูปที่ 9 ผลการทดสอบแสดงดังตารางที่ 5 และ 6 ซึ่งพบว่า ระบบสามารถตรวจสอบตัวเลขอารบิกชุดเรียนรู้ได้อย่างถูกต้องร้อยละ 78.00 และสามารถตรวจสอบตัวเลขอารบิกชุดทดสอบได้อย่างถูกต้องร้อยละ 77.00 และมีความถูกต้องเฉลี่ยร้อยละ 77.50 ซึ่งอยู่ในเกณฑ์ดี โดยตัวเลขที่สามารถรู้จำได้ดีที่สุดคือ เลข 7 เนื่องจากการสังเกตลักษณะสำคัญของเลข 7 ได้แบ่งการเขียนไว้สามประเภทดังรูปที่ 10 และตัวเลขที่สามารถรู้จำได้น้อยที่สุด คือ เลข 3 โดยมีร้อยละความถูกต้องเฉลี่ยเพียง 48.33 ซึ่งรู้จำผิดพลาดเป็นเลข 7 บ่อยที่สุด โดยสาเหตุมาจากลักษณะการเขียนเลข 3 ที่มีความใกล้เคียงกับการเขียนเลข 7 ที่มีขีดกลางดังรูปที่ 10 (ก)

เพื่อวัดประสิทธิภาพของวิธีการที่นำเสนอ ผู้เขียนได้ทดสอบตัวเลขในชุดทดสอบ 300 ตัวกับซอฟต์แวร์สำหรับการค้าของ (ABBY, 2014) ดังแสดงในตาราง

ที่ 6 ซึ่งปรากฏว่า ได้ร้อยละความถูกต้องเฉลี่ยเป็น 69.00 ซึ่งน้อยกว่าวิธีการที่นำเสนอร้อยละ 8 ถึงกระนั้นวิธีการนี้ยังมีประสิทธิภาพน้อยกว่าวิธีการอื่นที่นำเสนอใน (Lauer, Suen, & Bloch, 2007, pp. 1816-1824; Choudhary, Rishi, & Ahlawat, 2012, pp. 119-126; Cheng, & Yan, 1998, pp. 235-255; Neves, Filho, Mello, & Band Zanchettin, 2011, pp. 501-515; Miri, Razavi, & Sadri, 2011, pp. 150-155) ทั้งนี้ มีสาเหตุมาจากการเน้นอัลกอริทึมที่มีความซับซ้อนน้อยกว่าเพื่อให้พัฒนาได้ง่ายและสามารถประมวลผลได้อย่างรวดเร็ว และอีกปัจจัยหนึ่งที่น่าสนใจ คือ จำนวนข้อมูลที่ใช้ในการเรียนรู้ซึ่งนำมาจากข้อมูลจริงมีจำนวนน้อยมากเมื่อเทียบกับวิธีการใน (Lauer, Suen, & Bloch, 2007, pp. 1816-1824; Choudhary, Rishi, & Ahlawat, 2012, pp. 119-126; Cheng, & Yan, 1998, pp. 235-255; Neves, Filho, Mello, & Band Zanchettin, 2011, pp. 501-515; Miri, Razavi, & Sadri, 2011, pp. 150-155) ที่ใช้จำนวนข้อมูลสำหรับการเรียนรู้ประมาณ 5,000 - 600,000 ตัวอย่าง อย่างไรก็ตาม จะเห็นได้ว่าความถูกต้องในการรู้จำระหว่างชุดเรียนรู้และชุดทดสอบของวิธีการที่นำเสนอต่างกันเพียง 1.0% ซึ่ง

น้อยกว่าวิธีที่ปรากฏใน (Neves, Filho, Mello, & Band Zanchettin, 2011, pp. 501-515; Miri, Razavi, & Sadri, 2011, pp. 150-155) ที่ค่าความถูกต้องต่างกัน 1.5 – 4.0 %

สรุปผลและข้อเสนอแนะ

บทความนี้นำเสนอขั้นตอนวิธีสำหรับการรู้จำตัวเล
อการบิกจากลายมือสำหรับระบุรหัสผู้สอบบน
กระดาษคำตอบแบบปรนัยโดยใช้เทคนิคการโปรเจกชัน
ผสมกับวิธีทางสถิติ ซึ่งถือเป็นกระบวนการรู้จำที่ไม่
ซับซ้อนประสิทธิภาพในการตัดสินใจคำตอบมีความถูกต้อง
ร้อยละ 78.00 และ 77.00 สำหรับชุดเรียนรู้และชุด
ทดสอบ ตามลำดับ ด้วยการสกัดลักษณะสำคัญและการ
จัดหมวดหมู่ด้วยการหาค่ากลางของข้อมูล ทำให้เมื่อมีชุด
ข้อมูลสำหรับเรียนรู้มากขึ้นหรือมีการเปลี่ยนชุดข้อมูลก็
สามารถคำนวณใหม่ได้อย่างรวดเร็ว

ความผิดพลาดในการรู้จำเกิดได้ทั้งจากภาพหน้าเข้า
และกระบวนการรู้จำเอง เนื่องจากกระต่ายคำตอบเป็น

กระดาษที่ไม่มีคุณภาพมากมักก่อปรกกับการใช้ปากกาถูกลื่นที่มีการไหลของน้ำหมึกไม่สม่ำเสมอทำให้ภาพนำเข้าไม่คมชัดและเส้นปากกาไม่สม่ำเสมอ การเลือกชุดข้อมูลและจำนวนข้อมูลที่นำมาใช้ในการเรียนรู้และการทดสอบก็เป็นสิ่งหนึ่งมีผลกับประสิทธิภาพการเรียนรู้จำตัวเลข นอกจากนี้ จากการวิเคราะห์คำตอบจากโปรเจกต์ชั้น 4 แทน ซึ่งจากตารางที่ 4 จะสังเกตได้ว่าคำตอบที่ได้จากการใช้ค่าเฉลี่ยมีความน่าเชื่อถือน้อยที่สุด จึงอาจปรับปรุงการวิเคราะห์จากค่ากลางเพียง 2 แบบ คือฐานนิยมและมัธยฐานก่อนแต่ในกรณีที่คำตอบที่ได้มีความถี่เท่ากันหมด จะใช้ค่ากลางทั้ง 3 แบบ และนอกจากการปรับปรุงทั้งกระบวนการแล้ว อาจต้องมีส่วนตอนอื่นเฉพาะสำหรับแก้ปัญหาตัวเลขที่มีความน่าเชื่อถือต่ำ ถึงแม้ว่า ประสิทธิภาพในการรู้จำที่ปรากฏอาจยังมีความถูกต้องไม่มากนัก แต่หากมีการพัฒนาต่อยอดโดยการนำรหัสนักเรียนที่รู้จำได้ไปจับคู่กับรหัสนักเรียนในฐานข้อมูล โดยตรวจสอบความคล้ายคลึงกันของชุดของเลข จะมีความถูกต้องมากขึ้น

ตารางที่ 5 ผลการทดสอบความแม่นยำในการจำเลขจากตัวเลขชุดเรียนรู้ 300 ตัว ซึ่งมีตัวอย่างละ 30 ตัว

ตัวเลืข	ผลการร้จ้า										ร้ยลละความถุกต้ง
	0	1	2	3	4	5	6	7	8	9	
0	23	0	0	0	0	0	1	2	3	1	76.67
1	0	27	0	0	0	0	0	2	0	1	90.00
2	0	4	23	0	0	0	0	2	1	0	76.67
3	1	0	0	16	0	0	1	4	4	4	53.33
4	0	3	0	0	23	0	0	4	0	0	76.67
5	0	1	2	2	2	18	0	4	0	1	60.00
6	1	0	0	0	0	0	28	1	0	0	93.33
7	0	1	0	0	0	0	0	29	0	0	96.67
8	0	1	1	1	0	1	0	3	23	0	76.67
9	0	1	1	0	0	0	0	4	0	24	80.00
ร้ยลละความถุกต้งเฉลีย											78.00



ตารางที่ 6 ผลการทดสอบความแม่นยำในการรู้จำเลขจากตัวเลขชุดทดสอบ 300 ตัว ซึ่งมีตัวเลขอย่างละ 30 ตัว

ตัวเลข	ผลการรู้จำ										ร้อยละความถูกต้องใน	
	0	1	2	3	4	5	6	7	8	9	ร้อยละความถูกต้อง	การรู้จำของ ABBYY
0	29	0	0	0	0	0	0	1	0	0	96.67	100.00
1	0	28	0	0	0	0	0	2	0	0	93.33	96.67
2	1	7	19	0	0	0	0	2	1	0	63.33	70.00
3	0	0	4	13	0	0	0	8	2	3	43.33	73.33
4	0	0	0	0	25	0	1	2	2	0	83.33	86.67
5	0	1	1	0	0	19	1	4	0	4	63.33	86.67
6	3	0	2	0	0	0	24	1	0	0	80.00	20.00
7	0	1	0	0	0	0	0	29	0	0	96.67	50.00
8	0	1	0	0	1	0	0	4	22	2	73.33	23.33
9	0	0	0	1	0	0	0	3	3	23	76.67	83.33
ร้อยละความถูกต้องเฉลี่ย											77.00	69.00

กิตติกรรมประกาศ

ขอขอบพระคุณคณะวิศวกรรมศาสตร์ กำแพงแสน มหาวิทยาลัยเกษตรศาสตร์ที่เอื้อเฟื้อสถานที่ในการทำวิจัย ดร.ปาริฉัตร เสริมวุฒิสาร และดร.ศิวดล เสถียรพัฒนากุล สำหรับคำแนะนำที่ดีในการเขียนและทำวิจัย และอาจารย์ วินัย วิทยาลัยสำหรับตัวอย่างข้อมูลที่ใช้ในการทำวิจัย

เอกสารอ้างอิง

พุทธินันท์ พัดกระจำง. (2555). ระบบตรวจข้อสอบ แบบปรนัยบนระบบปฏิบัติการแอนดรอยด์. ปริญญาโท, สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ กำแพงแสน:มหาวิทยาลัยเกษตรศาสตร์.

ชลสิทธิ์ ชาญนุชิต. (2557). คู่มือเตรียมสอบ ภาค ก. ก.พ. ระดับ 3 - 4. กรุงเทพฯ: สถาบันวิชาการ Sure Group.

ABBYY. (2014). ABBYY Cloud OCR SDK Console. Retrieved September 8, 2014, from <http://cloud.ocrsdk.com/>

Burger, W., & Burge, M. J. (2008). *Digital Image Processing: An Algorithmic Introduction using Java*. New York: Springer.

Chaowicharat, E., Cerceone, N., & Naruedomkul, K. (2012). Novel Curve Signatures and a Combination Method for Thai On-Line Handwriting Character Recognition. In Proc. FLAIRS Conference. (196-201). California, USA: AAAI Press.

Cheng, D., & Yan, H. (1998). Recognition of handwritten digits based on contour information. *Pattern Recognition*, 31(3), 235-255.

Choudhary, A., Rishi, R., & Ahlawat, S. (2012). Unconstrained Handwritten Digit OCR Using Projection Profile and Neural Network Approach. *In Proc of the InConINDIA*, 132, 119-126.



- Du, S., Ibrahim, M., Shehata, M., & Badawy, W. (2013). Automatic License Plate Recognition (ALPR): A State-of-the-Art Review. *Circuits and Systems for Video Technology, IEEE Transactions on*, 23(2), 311-325.
- Educational Testing Service. (2012). Examinee Handbook Listening & Reading. Retrieved September 8, 2014, from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf
- Great Schools Partnership. (2013). STANDARDIZED TEST. Retrieved September, 8, 2014, from <http://edglossary.org/standardized-test/>
- Lauer, F., Suen, C. Y., & Bloch G. (2007). A trainable feature extractor for handwritten digit recognition. *Pattern Recognition*, 40, 1816-1824.
- Miri, E., Razavi, S., M., & Sadri, J. (2011). Performance Optimization of Neural Networks in Handwritten Digit Recognition Using Intelligent Fuzzy C-Means Clustering. In Proc. of International eConference on Computer and Knowledge Engineering (ICCKE), (150-155). NJ, USA: IEEE.
- Murayama, K. (2009). Objective Test Items. Retrieved September 8, 2014, from <http://www.education.com/reference/article/objective-test-items/>
- Neves, R., F., P., Filho, A., N., G., L., Mello, C., A., & Band Zanchettin, C. (2011). A SVM Based Off-Line Handwritten Digit Recognizer. In Proc. of IEEE international conf. on Systems, Man, and Cybernetics (501-515). Anchorage, AK: USA.
- Pornpanomchai, C., Batanov, D., N., & Dimmitt, N. (2001). Recognizing Thai handwritten characters and words for human computer interaction. *Int. J. Human-Computer Studies*, 55, 259-279.
- Sattayakawee, N. (2013). Test Scoring for Non-Optical Grid Answer Sheet Based on Projection Profile Method. *International Journal of Information and Education Technology*, 3(2), 273-277.
- Singh, S. (2013). Optical Character Recognition Techniques: A Survey. *Journal of Emerging Trends in Computing and Information Sciences*, 4(6), 545-550.
- Theeramunkong, T., & Wongtapan, C. (2005). Offline Isolated Handwritten Thai OCR Using Island-Based Projection with N-Gram Models and Hidden Markov Models. *Information Processing & Management*, 41(1), 139-160.