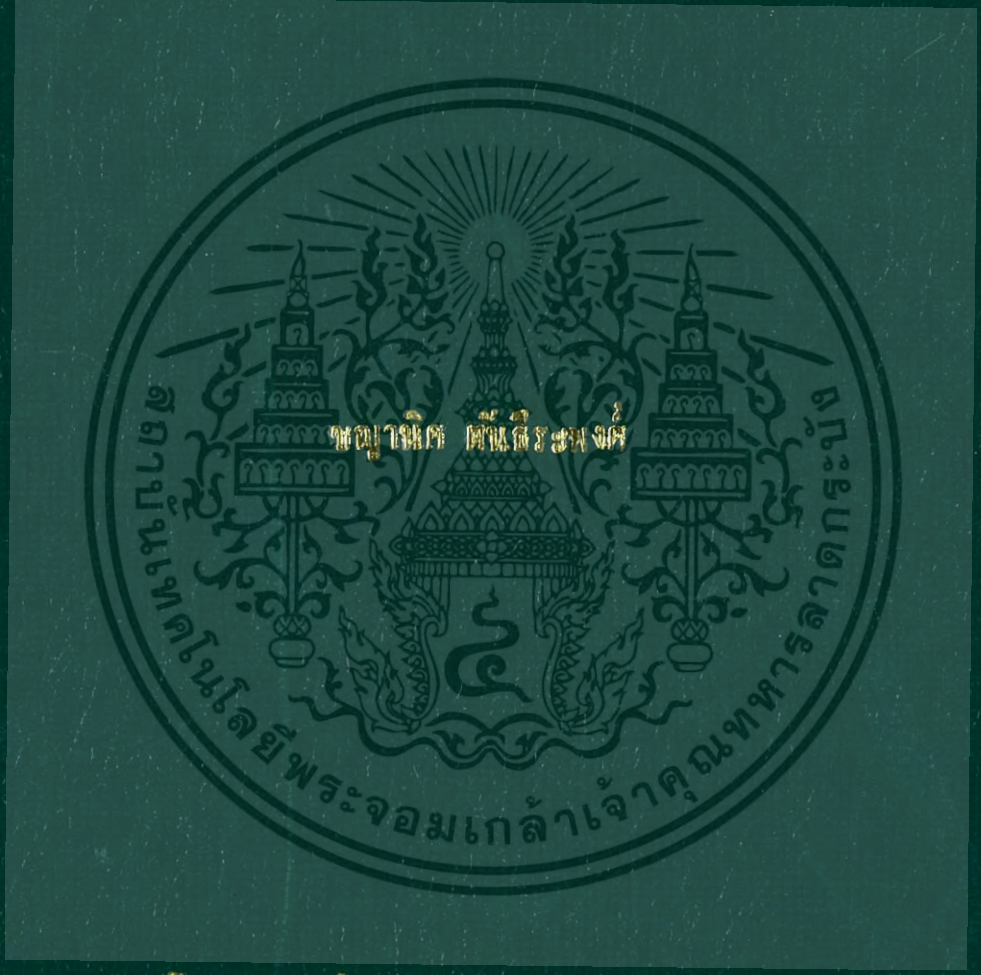
การตรวจจับข้อความเคลื่อนที่แบบเสถียรสูงสุดโดยใช้ MAXIMALLY STABLE EXTREMAL REGIONS กับ SUPPORT VECTOR MACHINE

TEXT DETECTION USING MAXIMALLY STABLE EXTREMAL REGIONS AND SUPPORT VECTOR MACHINE FOR TEXT-TO-SPEECH SYNTHESIS



ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2558

การตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดโดยใช้ MAXIMALLY STABLE EXTREMAL REGIONS และ SUPPORT VECTOR MACHINE

TEXT DETECTION USING MAXIMALLY STABLE EXTREMAL REGIONS AND SUPPORT VECTOR MACHINE FOR TEXT-TO-SPEECH SYNTHESIS

โดย

ชญานิศ ตันธีระพงศ์

CHAYANIS TUNTEARAPONG

อาจารย์ที่ปรึกษา

ผู้ช่วยศาสตราจารย์ ดร. กันต์พงษ์ วรรัตน์ปัญญา

อาจารย์ที่ปรึกษาร่วม

ผู้ช่วยศาสตราจารย์ ดร. กิติ์สุชาต พสุภา

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2558

การตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดโดยใช้ MAXIMALLY STABLE EXTREMAL REGIONS และ SUPPORT VECTOR MACHINE

TEXT DETECTION USING MAXIMALLY STABLE EXTREMAL REGIONS AND SUPPORT VECTOR MACHINE FOR TEXT-TO-SPEECH SYNTHESIS

โดย

ชญานิศ ตันธีระพงศ์

CHAYANIS TUNTEARAPONG

อาจารย์ที่ปรึกษา

ผู้ช่วยศาสตราจารย์ ดร. กันต์พงษ์ วรรัตน์ปัญญา

อาจารย์ที่ปรึกษาร่วม

ผู้ช่วยศาสตราจารย์ ดร. กิติ์สุชาต พสุภา

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2558

TEXT DETECTION USING MAXIMALLY STABLE EXTREMAL REGIONS AND SUPPORT VECTOR MACHINE FOR TEXT-TO-SPEECH SYNTHESIS

CHAYANIS TUNTEARAPONG

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

BACHELOR OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/2015

ใบรับรองปริญญานิพนธ์ ประจำปีการศึกษา 2558

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดโดยใช้ MAXIMALLY STABLE EXTREMAL REGIONS และ SUPPORT VECTOR MACHINE

TEXT DETECTION USING MAXIMALLY STABLE EXTREMAL REGIONS AND SUPPORT VECTOR MACHINE FOR TEXT-TO-SPEECH SYNTHESIS

ผู้จัดทำ นางสาวชญานิศ ตันธีระพงศ์ รหัสนักศึกษา 55070020

.....อาจารย์ที่ปรึกษา

(ผู้ช่วยศาสตราจารย์ ดร. กันต์พงษ์ วรรัตน์ปัญญา)

.....อาจารย์ที่ปรึกษาร่วม

(ผู้ช่วยศาสตราจารย์ ดร. กิติ์สุชาต พสุภา)

หัวข้อโครงงาน	การตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดโดยใช้ Maximally Stable Extremal Regions และ Support Vector Machine
นักศึกษา	นางสาวชญานิศ ตันธีระพงศ์ รหัสนักศึกษา 55070020
ปริญญา	วิทยาศาสตรบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
ปีการศึกษา	2558
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. กันต์พงษ์ วรรัตน์ปัญญา
อาจารย์ที่ปรึกษาร่วม	ผู้ช่วยศาสตราจารย์ ดร. กิติ์สุชาต พสุภา

บทคัดย่อ

การตรวจจับข้อความจากฉากธรรมชาติ มีหลายปัจจัยที่มีผลกระทบต่อการตรวจจับและรู้จำ เช่น มุมมองของภาพและความหลากหลายของตัวอักษร เป็นต้น จากปัจจัยที่กล่าวมาข้างต้นนั้นมีความท้าทายเป็นอย่างยิ่งสำหรับการตรวจจับข้อความเพื่อแปลงเป็นเสียงพูด นอกจากนี้การตรวจจับข้อความในภาษาไทยนั้นยังไม่มีวิธีที่หลากหลาย เพราะว่าลักษณะโครงสร้างของภาษาไทยมีความแตกต่างจากภาษาอื่น ๆ โดยรายงานนี้จะนำเสนอการพัฒนาระบบตรวจจับข้อความเพื่อแปลงเป็นเสียงพูด โดยมี 4 ขั้นตอนหลัก คือ 1).แบ่งข้อมูลของภาพแยกออกจากกันตามช่องสี 2).ตรวจจับคุณลักษณะ 3).แยกกลุ่มของตัวอักษรและกลุ่มที่ไม่ใช่ตัวอักษรโดยใช้หลักการเรียนรู้ของเครื่อง 4). การรวมกลุ่มของข้อความ จากการทดลองพบว่าวิธีการที่พัฒนาสามารถตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดได้ โดยมีความแม่นยำและครบถ้วนของภาษาไทยอยู่ที่ 0.63 และ 0.67 และค่าแม่นยำและค่าครบถ้วนสำหรับภาษาอังกฤษอยู่ที่ 0.58 และ 0.69 ตามลำดับ

Project	Text Detection Using Maximally Stable Extremal Regions and Support Vector Machine for Text-To-Speech Synthesis
Student	Miss Chayanis Tuntearapong Student ID 55070020
Degree	Bachelor of Science
Program	Information Technology
Academic Year	2015
Advisor	Asst. Prof. Dr. Kuntpong Woraratpanya
Co-Advisor	Asst. Prof. Dr. Kitsuchart Pasupa

ABSTRACT

Text detection in natural scenes has many factors that take effect, such as perspective and a variety of characters, etc. These factors become a grand challenge for text detection of text-to-speech synthesis system. In addition, the text detection for Thai language has a few methods to support, because the structure of Thai characters differs from other languages. This project proposes an approach to detect Thai text for text-to-speech synthesis system. This approach is composed of four main procedures: (i) color channel decomposition, (ii) feature detection, (iii) text and non-text classification using machine learning and (iv) text area grouping. The experimental results demonstrate that the proposed approach can achieve the average precision and recall for Thai text at 0.63 and 0.67. Furthermore, the average precision and recall for English text are 0.58 and 0.69, respectively.

กิตติกรรมประกาศ

ขอขอบคุณผศ.ดร.กันต์พงษ์ วรรัตน์ปัญญา และ ผศ.ดร.กิติ์สุชาต พสุภา อาจารย์ที่ปรึกษา ที่ได้กรุณาให้คำปรึกษาและชี้แนะแนวทาง ตลอดจนแก้ไขปริญญานิพนธ์นี้ให้ผ่านพ้นไปด้วยดี

ขอขอบคุณ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง สำหรับความรู้และได้มอบโอกาสให้ข้าพเจ้าได้ไปแลกเปลี่ยนที่ญี่ปุ่นเป็นเวลา 1 อาทิตย์ รวมไปถึงประสบการณ์ต่าง ๆ ที่มอบให้ ตลอดเวลา 4 ปีที่ศึกษา อยู่ในสถานที่แห่งนี้

ขอขอบคุณ คณาจารย์คณะเทคโนโลยีสารสนเทศ และพี่ ๆ สมาชิกห้องปฏิบัติการวิจัยด้านการรู้จำแบบและการประมวลผลภาพ (PRIP) ที่ได้ให้คำปรึกษา และถ่ายทอดวิชาความรู้ต่าง ๆ ให้แก่ข้าพเจ้า

ขอขอบคุณน้อง ๆ จากคณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สำหรับความรู้ทางด้านการเรียนรู้ของเครื่อง และ แนะนำแหล่งความรู้ต่าง ๆ ที่สามารถใช้ในการจบปริญญานิพนธ์นี้ได้

ท้ายที่สุด ขอขอบคุณบิดา มารดา และครอบครัวที่สนับสนุนข้าพเจ้าตลอดเวลา ไม่ว่าข้าพเจ้าจะตัดสินใจอย่างไร ครอบครัวก็ยังคอยเคียงข้างข้าพเจ้าเสมอจนกระทั่งสำเร็จการศึกษา

ชญานิศ ตันธีระพงศ์

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันการใช้ชีวิตประจำวันเรามักพบกับข้อมูลต่าง ๆ ที่มีที่มาหลากหลาย เช่น ป้ายบอกทาง ป้ายโฆษณา เป็นต้น สิ่งเหล่านี้เป็นข้อมูลและสารสนเทศที่จำเป็นสำหรับการเรียนรู้และใช้ในชีวิตประจำวัน ดังนั้น การพัฒนาระบบตรวจจับข้อความเพื่อแปลงเป็นเสียงพูด จึงเป็นขั้นตอนที่สำคัญที่ใช้ผลลัพธ์ที่ได้นั้น มาประยุกต์เข้ากับงานประเภทต่าง ๆ เช่น เครื่องมือสำหรับช่วยเหลือผู้พิการทางสายตา การพัฒนาระบบตรวจจับป้ายทะเบียนรถ เป็นต้น

ความท้าทายของการพัฒนาระบบตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดนั้นมีหลากหลายประการ เช่น การเรียงตัวของข้อความ, สีของข้อความ, ความหลากหลายของรูปแบบตัวอักษร, มุมมองของภาพและความคมชัดของภาพ สิ่งเหล่านี้ล้วนเป็นปัจจัยที่ส่งผลให้การตรวจจับข้อความที่มาจากภาพถ่ายนั้นมีความผิดพลาดได้ ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำ จึงมีการควบคุมปัจจัยเหล่านั้น เพื่อรักษาคุณภาพของข้อความ ให้สามารถตรวจจับและรู้จำได้มากที่สุด

ในการตรวจจับข้อความจากภาพถ่ายนั้นมีหลากหลายวิธี ซึ่งแต่ละวิธีมีข้อดีและข้อเสียแตกต่างกันไป สิ่งสำคัญคือการสร้างกระบวนการที่มีประสิทธิภาพที่ให้ผลลัพธ์ที่มีความแม่นยำได้มากที่สุด โดยวิธีการตรวจจับข้อความจากภาพถ่ายนั้น มีหลากหลายประเภท [1] ทั้ง การตรวจจับข้อความโดยพิจารณาจากพื้นที่ (Region-based) การตรวจจับโดยพิจารณาจากส่วนประกอบที่เชื่อมต่อกัน (Connected Component-based: CC-based) และ การตรวจจับแบบผสม (hybrid methods) เมื่อเปรียบเทียบวิธีการจะค้นพบว่า วิธีการตรวจจับข้อความโดยพิจารณาจากพื้นที่สามารถตรวจจับข้อความที่มีสัญญาณรบกวนสูงได้ แต่มีข้อเสียในเรื่องของเวลาที่ใช้เวลานานในการประมวลผล และมีความจำเป็นที่ต้องใช้การเรียนรู้ของเครื่อง (Machine Learning) เพื่อให้พิจารณาข้อมูลเหล่านั้นได้ วิธีการที่สองคือการพิจารณาจากส่วนประกอบที่เชื่อมต่อกัน โดยใช้ปัจจัยต่าง ๆ ในการค้นหาส่วนประกอบที่เชื่อมต่อกัน เช่น ขอบของวัตถุในภาพ, สี, ความเป็นปึกแผ่นของส่วนประกอบ หรือตรวจสอบจากการเปลี่ยนแปลงของค่าความเข้มสีของวัตถุ จากการศึกษาวิธีตรวจจับโดยพิจารณาจากส่วนประกอบที่เชื่อมต่อกันนั้น มีข้อดีคือ สามารถนำเข้าระบบรู้จำได้ง่ายและใช้เวลาประมวลผลที่น้อยกว่าและแบบสุดท้ายคือการนำข้อดีของการตรวจจับแต่ละประเภทมาใช้งานร่วมกัน ซึ่งในปัจจุบันวิธีการนี้เป็นที่นิยมใช้ เพราะนอกจากสามารถกำจัดสัญญาณรบกวนได้อย่างแม่นยำและ

ประสิทธิภาพแล้ว ยังสามารถประยุกต์ใช้กับการเรียนรู้ของเครื่องเพื่อเพิ่มประสิทธิภาพของการตรวจจับได้อีกด้วย

Ross G., Jeff D., Trevor D. และ Jitendra M. นำเสนอวิธี R-CNN (Regions - Convolutional neural network) ที่ใช้การตรวจจับข้อมูลโดยนำภาพเข้ามาสกัดคุณลักษณะผ่านคอนโวลูชันนอลนิวรอลเน็ตเวิร์ค (Convolutional neural network) [2] โดยข้อดีของวิธีการนี้คือ ได้ผลลัพธ์ที่แม่นยำ และ ใช้ระยะเวลาในการประมวลผลสั้น และสามารถนำผลลัพธ์ที่ได้ไปใช้งานได้ทันที แต่ข้อเสียที่ได้คือ วิธีการนี้ยังไม่ได้ถูกใช้ในการตรวจจับข้อความและนอกจากนี้ ยังมีความจำเป็นที่ต้องใช้ฮาร์ดแวร์ที่มีคุณสมบัติสูงและมีความซับซ้อนในการเตรียมสร้างแบบโครงข่ายประสาทเทียมซึ่งใช้เวลานาน

Xiaobing W., Yonghong S., Yuanlin Z. และ Jingmin X. นำเสนอวิธี multi-layer CC segmentation และ Higher order conditional random field based analysis [3] ที่ใช้ตรวจจับข้อความโดยแบ่งส่วนประกอบของภาพเป็นหลายๆชั้น แล้วนำผลลัพธ์ของการแบ่งชั้นนำมาตรวจจับข้อความบนรูปภาพ โดยการตรวจจับจะจัดเป็นลำดับชั้น เช่น ส่วนประกอบที่เป็นตัวอักษร ส่วนประกอบที่เป็นคำ และส่วนประกอบที่เป็นวลีหรือข้อความ ซึ่งวิธีการนี้มีผลดีคือ มีความแม่นยำสูง แม้ว่าจะมีสัญญาณรบกวน แต่ข้อเสียคือ ใช้เวลาประมวลผลนานเนื่องจากการวิเคราะห์รายละเอียดของภาพ

Weilin H., Yu Q. และ Xiaoon T. นำเสนอวิธี Convolutional neural network Induced MSER Trees ซึ่งใช้วิธีการตรวจจับโดยพิจารณาพื้นที่ โดยใช้ MSER และ Convolutional neural network [4] สร้างแบบจำลองเพื่อคัดแยกตัวอักษรและสิ่งที่ไม่ใช่ตัวอักษรออกจากกัน ซึ่งข้อดีคือมีความแม่นยำ เนื่องจาก การทำงานควบคู่กันของ MSER และ Convolutional neural network นั้นเสริมประสิทธิภาพซึ่งกันและกัน แต่ก็มีข้อเสียกับรูปภาพบางกรณี เช่น ภาพที่มีความคมชัดต่ำ หรือภาพที่มีตัวอักษรและพื้นหลังที่ซับซ้อน เช่น ตัวอักษรที่อยู่บนกำแพงอิฐ ตัวอักษรที่มีสีเดียวกับพื้นหลัง หรือมีสีที่ใกล้เคียงกับพื้นหลังและภาพตัวอักษรที่มองเห็นไม่ชัด จะไม่สามารถตรวจจับได้โดยวิธีการนี้ เพราะว่าการตรวจจับของ MSER นั้นบางครั้งมักจะให้ส่วนประกอบของภาพพื้นหลังติดมาด้วย และลักษณะภาพของอิฐนั้นส่วนมากจะมีค่าสีที่ไม่ห่างกันมาก จึงทำให้ MSER สามารถตรวจจับได้ นอกจากนี้ การเสริมสร้างประสิทธิภาพประมวลผลของงานวิจัยนี้ให้รวดเร็วได้นั้นควรการใช้ฮาร์ดแวร์ที่มีคุณสมบัติสูง และมีความซับซ้อนในการเตรียมสร้างแบบโครงข่ายประสาทเทียมซึ่งใช้เวลานาน

พิมพ์ลักษ์ บุญชูกุศล [5] ได้นำเสนอวิธีการวิเคราะห์ส่วนประกอบที่เชื่อมต่อกันของวัตถุ (Modified Connected Component Analysis: MCCA) ในการค้นหาตำแหน่งของข้อความบนภาพถ่าย โดยทำการตรวจจับตัวอักษรภาษาไทยในส่วนลำตัวหลัก (Modified Body of Text Detection) การรวมกลุ่มข้อความที่อยู่ในแนวเดียวกัน(Text Grouping) และทำการขยายกรอบล้อมรอบข้อความ (Text Boundary Padding) เพื่อตรวจจับสระบน และล่างของข้อความ โดยใช้ร่วมกับกฎที่สร้างขึ้นมาใหม่จากสัดส่วนมาตรฐานของตัวอักษรไทย ซึ่งวิธีการนี้เหมาะสมสำหรับใช้ในการตรวจจับตัวอักษรภาษาไทย แต่สำหรับรูปภาพที่มีปัญหาเรื่องมุมมองเช่น ลักษณะการวางตัวอักษร และ ความหลากหลายของตัวอักษรของบางรูปไม่สามารถจับตัวอักษรได้ครบถ้วน เนื่องจากมีสัดส่วนที่แตกต่างจากมาตรฐานและสูตรที่สร้างไว้

Datong C., Herve B., Jean-P. T [6] ได้นำเสนอวิธียืนยันพื้นที่ตัวอักษรบนรูปภาพโดยใช้การตรวจสอบคุณสมบัติของกลุ่มตัวอักษร โดยวิเคราะห์จากการวางตัวอักษรในช่วงแนวตั้งและแนวนอน และหลังจากนั้นนำพื้นที่ที่มีแนวโน้มว่าจะเป็นกลุ่มของตัวอักษรนั้นมาสกัดคุณลักษณะภายในกลุ่ม และใช้หลักการ SVM(Support Vector Machine) ในการตรวจสอบว่ากลุ่มของพื้นที่นั้นใช้กลุ่มของตัวอักษรหรือไม่ โดยผลลัพธ์สุดท้ายที่ได้คือพื้นที่ตัวอักษรที่ได้รับการยืนยันเรียบร้อยแล้ว โดยวิธีการนี้มีข้อดีคือ ประมวลผลไว และแม่นยำ แต่มีข้อเสียคือสามารถรองรับได้แค่บางภาษาเท่านั้น

Wei, Yi Cheng, และ Chang Hong Lin.[7] ได้นำเสนอวิธีตรวจจับตัวอักษรโดยใช้ หลักการประมวลผลภาพแบบ Pyramidal Method ตามแนวตั้งและแนวนอนของภาพเพื่อตรวจสอบลักษณะพื้นที่และพื้นผิวทั้งหมดของภาพ และใช้ SVM ในการแยกประเภทของพื้นที่ที่ถูกประมวลผลภาพว่าบริเวณนั้นเป็นพื้นที่ของตัวอักษรหรือไม่ ซึ่งวิธีนี้สามารถสอนตัวอย่างได้และประมวลผลได้อย่างมีประสิทธิภาพและมีความแม่นยำสูงสำหรับภาษาอังกฤษ

1.2 ความมุ่งหมายและวัตถุประสงค์

1.2.1 เพื่อพัฒนาระบบตรวจจับข้อความและระบุตำแหน่งข้อความให้มีความแม่นยำ

1.2.2 เพื่อพัฒนาระบบตรวจจับข้อความและระบุตำแหน่งข้อความให้มีประสิทธิภาพ

1.3 กรอบแนวคิดของโครงงาน

ขั้นตอนภายในระบบนั้นจะแบ่งเป็น 3 ขั้นตอนหลัก คือ 1. การตรวจจับข้อความ 2.การรู้จำข้อความ 3. และขั้นตอนการแปลงข้อความเป็นเสียง โดยในโครงงานนี้ จะมีจุดมุ่งหมายไปที่ขั้นตอนหลักที่ 1 โดยขั้นตอนทั้งหมดจะแสดง ดังรูปที่ 1.1

รูปที่ 1.1 แสดงกระบวนการทำงานของระบบตรวจจับและรู้จำข้อความเพื่อแปลงเป็นเสียงพูด

เนื่องจากจุดมุ่งหมายของโครงงานเน้นไปที่การพัฒนาขั้นตอนการตรวจจับข้อความ ดังนั้นจึงมีการแสดงการทำงานของขั้นตอนการตรวจจับข้อความ ดังรูปที่ 1.2

รูปที่ 1.2 แสดงกระบวนการทำงานของการตรวจจับข้อความ

โดยกระบวนการนั้นจะเริ่มโดยทำกระบวนการ Pre-Processing โดยนำภาพสีระบบ RGB ทำการแบ่งเป็นภาพสองมิติ โดยที่แบ่งตาม Channel ของสี RGB โดยผลลัพธ์จะได้เป็นภาพระดับสีเทา (Grayscale Image) หลังจากนั้นจึงตรวจจับคุณลักษณะตามพื้นที่ของภาพโดยใช้หลักการ MSER(Maximally stable extremal regions) จากนั้นนำลักษณะของพื้นที่ในแต่ละช่องสี นั้นมาคัดแยกคุณลักษณะ โดยแบ่งออกเป็น 2 กลุ่มคือกลุ่มของข้อมูลที่เป็นตัวอักษรและไม่เป็นตัวอักษร โดยใช้หลักการ SVM(Support Vector Machine) คัดแยกข้อมูลออกจากกันและนำผลลัพธ์ของการทำนาย ซึ่งเป็นพื้นที่ที่เป็นตัวอักษรนั้นมาวางตามตำแหน่งของพื้นที่บนภาพ โดยผลลัพธ์สุดท้ายที่ได้ในขั้นตอนนี้คือตำแหน่งของตัวอักษร ที่ถูกตัดออกมาและนำมาทำการปรับปรุงข้อมูลให้เหมาะสมก่อนจะเข้าสู่ขั้นตอนที่ 2 ซึ่งแสดงผลกระบวนการการทำงานดังรูปที่ 1.3

รูปที่ 1.3 แสดงกระบวนการการปรับปรุงข้อมูลเพื่อเข้าสู่กระบวนการรู้จำข้อความ

สำหรับขั้นตอนที่ 2 นั้นคือการรู้จำของข้อความนั้นทางผู้จัดทำใช้ Tesseract [8] OCR โดยได้ปรับปรุงโดยการสอนแบบตัวอักษรและคำในภาษาไทยเพิ่มเติมเพื่อให้สามารถรองรับข้อความภาษาไทยและแบบตัวอักษรที่มากขึ้นทั้งในภาษาไทยและอังกฤษ เพื่อให้ผลลัพธ์ที่ได้มีความแม่นยำสูงพอที่จะเข้าขั้นตอนการที่ 3

โดยในขั้นตอนที่ 3 นั้น จะนำผลลัพธ์ที่ได้ ทำการแปลงข้อความผ่าน Microsoft Speech API [9]และเนื่องจาก Microsoft Speech API นั้นไม่รองรับภาษาไทย ดังนั้นผลลัพธ์ของภาษาไทยที่ออกมา จะมีผลลัพธ์เฉพาะแค่ตัวอักษรที่แก้ไขได้เท่านั้น

1.4 ขอบเขตของโครงงาน

1.4.1 ภาพที่รับมาในระบบคือภาพสีระบบ RGB และมีขนาดมากกว่า 300×300 pixels และสูงสุดที่ขนาด 3840×2400 pixels

1.4.2 สามารถรับไฟล์ภาพได้แค่นามสกุล .jpg และ .png เท่านั้น

1.4.3 ระบบการแปลงข้อความเป็นเสียงพูดสามารถรองรับได้เฉพาะภาษาอังกฤษเท่านั้น

1.4.4 ระบบการรู้จำข้อความนั้นพัฒนาจาก Tesseract OCR โดยการสอนข้อมูลเพิ่มเติม เพื่อสร้างความแม่นยำในการรู้จำข้อความ

1.4.5 ระบบสามารถตรวจจับและรู้จำข้อความได้ โดยสามารถรองรับได้สองภาษา คือ ภาษาไทยและภาษาอังกฤษ

1.4.6 ชุดข้อมูลจำแนกกลุ่มข้อความและไม่ใช่กลุ่มข้อความ ผู้จัดทำใช้ชุดข้อมูลทั้งหมด 5 ชุด ได้แก่ Chars 74K English Dataset, ชุดข้อมูล BEST 2012 – 2014, ชุดข้อมูล ICDAR 2003 และชุดข้อมูลที่ผู้จัดทำนั้นทำการเก็บภาพเอง

1.4.7 ชุดข้อมูลที่ใช้ทดสอบการตรวจจับข้อความคือ ชุดข้อมูลจากงานวิจัย Thai Text Detection From Medium Shot Of Natural Scenes by using Fast Boundary Clustering and Modified Connected Component Analysis[5] , ชุดข้อมูล The Street View Text Dataset และ ชุดข้อมูล ICDAR 2003 Robust Reading

1.4.8 ผลลัพธ์ที่ได้จะมี 2 ลักษณะ คือ ตัวอักษรที่แก้ไขได้ และเสียงพูด

1.4.9 โครงงานนี้จะไม่มีการครอบคลุมกรณีต่างๆ ดังนี้

1. ตัวอักษรที่ไม่สมบูรณ์ อันเนื่องมาจากปัจจัยต่างๆ เช่น แสง เงา ความคมชัด และความไม่สมบูรณ์ของตัวอักษรจากรูปภาพที่มีมาตั้งแต่แรก
2. รูปภาพที่มีลักษณะของตัวอักษรที่เป็นลายมือ (Hand Written)
3. ระบบการแปลงข้อความเป็นเสียงพูด
4. ผลลัพธ์ที่ได้สำหรับภาษาไทย จะมี 1 ลักษณะ คือ ตัวอักษรที่แก้ไขได้ เท่านั้น

1.5 ประโยชน์ที่ได้รับ

1.5.1 สามารถนำไปประยุกต์ใช้เพื่อช่วยเหลือผู้พิการทางสายตา เช่น อ่านฉลากข้อความ ป้ายบอกทางและแปลงเป็นเสียงให้ผู้พิการทางสายตาเข้าใจมากขึ้น

1.5.2 สามารถนำไปประยุกต์ใช้เพื่อช่วยเหลือทางด้านจราจรและการเดินทาง เช่น ระบบอ่านป้ายเพื่อนำทาง การอ่านป้ายสัญญาณจราจร หรือระบบตรวจจับและรู้จำข้อความจากป้ายทะเบียนรถ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง โดยทฤษฎีที่มีการนำมาใช้ในโครงงาน ได้แก่ ทฤษฎีการเรียนรู้ของเครื่อง (Machine Learning) เช่น การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ SVM (Support Vector Machine) ทฤษฎีคอมพิวเตอร์วิทัศน์ (Computer Vision) เช่น Maximally Stable Extremal Regions (MSER) และทฤษฎีการประมวลผลภาพดิจิตอล (Digital Image Processing) เช่น ภาพสี (RGB)และภาพระดับสีเทา (Intensity Image)

2.1 ทฤษฎีการเรียนรู้ของเครื่อง (Machine Learning)

2.1.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบมีผู้สอน คือ การเรียนรู้ของเครื่องประเภทหนึ่ง ซึ่งใช้ชุดข้อมูลที่ผู้สอนนั้นระบุอย่างชัดเจนแล้วว่าสิ่งนั้นคือประเภทอะไร และต้องมีคุณสมบัติอะไรถึงจะสามารถแยกประเภทออกจากกันได้ เช่น สัดส่วน หรือหลักเกณฑ์ เป็นต้น โดยสิ่งที่นำมาทดสอบการเรียนรู้ของเครื่องนั้นจะต้องให้ผลลัพธ์เป็นไปตามที่ชุดข้อมูลที่ใช้ในการเรียนรู้นั้นสอนไว้ โดยการเรียนรู้ของเครื่องแบบมีผู้สอนนั้น สิ่งที่สามารถวัดผลได้คือ ชุดข้อมูลที่ใช้สอนและทดสอบต้องมีคำตอบไปในแนวโน้มหรือทิศทางเดียวกัน การเรียนรู้โดยมีผู้สอนนั้นสามารถแบ่งออกได้เป็น 2 ประเภทหลัก ๆ คือ การวิเคราะห์ความถดถอย (Regression) และการแบ่งประเภทของข้อมูล (Classification) [10][11]

2.1.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้แบบไม่มีผู้สอนคือการเรียนรู้ของเครื่องที่ชุดข้อมูลที่ระบุแน่ชัดว่าสิ่งนั้นคืออะไรแต่ไม่ได้สอนให้เครื่องนั้นรู้จักว่าสิ่งที่ใช้สอนไปนั้น ควรจะมีคุณสมบัติอย่างไรที่สามารถระบุได้อย่างแน่ชัดว่าเป็นสิ่งนั้น โดยการเรียนรู้แบบไม่มีผู้สอนนั้นจะสร้างแบบจำลองขึ้นมา เพื่อให้สามารถแยกกลุ่มของสิ่งที่ใช้เรียนรู้ได้ โดยมีโครงสร้างข้อมูลที่สามารถ สามารถแบ่งข้อมูล (Clustering) ออกเป็นกลุ่ม ๆ เพื่อลดความซ้ำซ้อนของข้อมูลได้ และทำให้เข้าใจได้ว่ากลุ่มของข้อมูลนั้นคือข้อมูลอะไรและควรทำนายให้อยู่ในประเภทไหน [10][11]

2.1.3 SVM (Support Vector Machine)

SVM คือประเภทหนึ่งของการเรียนรู้แบบมีผู้สอน SVM เกิดขึ้นในปี 1992 โดย Vapnik และคณะ [12] SVM นั้นถูกพัฒนามาจาก Logistic Regression เนื่องจากบ่อยครั้งข้อมูลที่มีการสอนนั้นมี

การกระจายตัวแยกออกจากกันและสร้างความสับสนของข้อมูลจนเกิดปัญหาในการทำนายบน Logistic Regression [11]

โดยแนวคิดของ SVM นั้นจะใช้การคัดแยกกลุ่มของวัตถุที่มี 2 ชนิดขึ้นไปโดยแนวคิดของ SVM นั้นจะมีความแตกต่างจาก Logistic Regression ตรงที่ มีการกำหนด Large Margin Classifiers ซึ่งแสดงดังสมการต่อไปนี้

$$\text{if } y=1,\ \Theta^T x \geq 1 \ (\text{Not just} \geq 0) \tag{2.1}$$

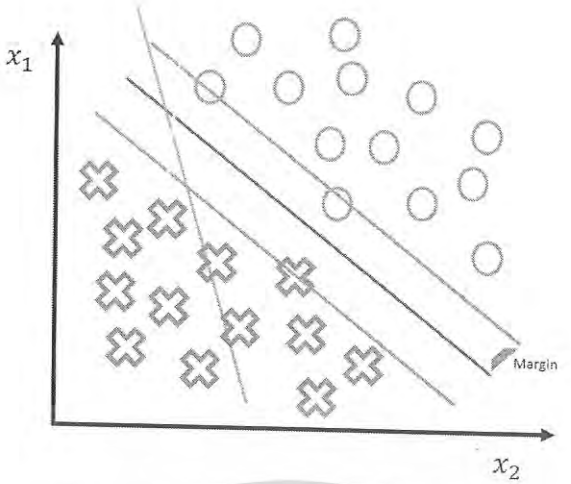
$$\text{if } y=0,\ \Theta^T x \leq 1 \ (\text{Not just} < 0) \tag{2.2}$$

โดยการคำนวณเพื่อหาสมการที่ดีที่สุดสำหรับข้อมูลนั้น สามารถทำได้โดยการสอนข้อมูล พร้อมคุณสมบัติที่ต้องการแยกประเภทเข้าไปในการเรียนรู้ โดยที่การสร้างสมการ SVM ที่นี่ได้นั้นจำเป็นต้องมีค่า Cost Function ที่น้อยที่สุด โดยสามารถคำนวณ Cost Function เบื้องต้นได้จากสมการดังนี้

$$J(\theta) = C \cdot 0 + \frac{1}{2}\sum_{j=1}^{n} \Theta_j^2 \tag{2.3}$$

$$J(\theta) = \frac{1}{2}\sum_{j=1}^{n} \Theta_j^2 \tag{2.4}$$

โดยจากสมการข้างต้น สามารถสร้างสมการที่ดีที่สุดสำหรับข้อมูลนั้น ๆ ได้ จากรูปที่ 2.1นั้นแสดงข้อมูลที่ถูกวางบนกราฟ 2 มิติที่แสดงการกระจายของข้อมูล โดยการคำนวณของ SVM นั้นสร้างสมการไปเรื่อย ๆ จนกว่าจะได้สมการเส้นตรงที่ดีสุดที่จะใช้แบ่งข้อมูล



รูปที่ 2.1 แสดงการเลือกสมการเส้นตรงที่ดีที่สุด

โดยระยะห่างระหว่างสมการเส้นตรงและเส้นตรงที่อยู่ขนานกับสมการเส้นตรงหลักนั้นเรียกว่า **Margin** นอกจากนี้สำหรับการคำนวณข้อมูลแบบ Non-linear โดยวิธีการ SVM นั้นจะมีการกำหนดค่า Kernel ซึ่งเป็นฟังก์ชั่นที่สามารถนำข้อมูลที่มีมิติ (dimension) น้อยกว่า มาทำการปรับปรุงทำให้มีมิติที่สูงขึ้นเพื่อใช้ในการการแบ่งข้อมูลแบบ Linear Model

2.2 ทฤษฎีคอมพิวเตอร์วิทัศน์ (Computer Vision)

2.2.1 Maximally Stable Extremal Regions(MSER)

MSER เป็นแนวคิดในการตรวจจับพื้นที่ของข้อมูลที่เป็นรูปภาพ โดย MSER นั้นจะใช้การคำนวณการหาคุณลักษณะจากภาพระดับสีเทา (Grayscale Image) และวัดผลของคุณลักษณะจากการเติบโตของพื้นที่ภายในภาพ โดยที่พื้นที่นั้นจะต้องคงทนต่อการเปลี่ยนแปลงของค่าความเข้มของสี [13] [14] โดยแนวคิดของ MSER นั้นสามารถแบ่งออกเป็น 4 ขั้นตอน [15] ดังนี้

2.2.1.1 การเตรียมข้อมูล (Pre-processing)

ข้อมูลที่เข้ามาครั้งแรกของ MSER คือข้อมูล Matrix ที่มีขนาด 2 มิติและเป็นจำนวนจริง ที่มีค่าของความเข้มของสีอยู่ที่ 0 - 255 โดยการเตรียมข้อมูลนั้น จะนำข้อมูลมาเรียงลำดับตามค่าความเข้มของสี โดยในแต่ละความเข้มจะจัดเก็บตำแหน่งพิกเซลและค่าความเข้มของภาพ โดยตัวอย่างการเรียงลำดับของค่าสี จะแสดงในรูปที่ 2.2

32	31	152	151	150
31	152	151	150	136
152	151	150	136	135
151	150	136	135	131
150	136	135	131	131

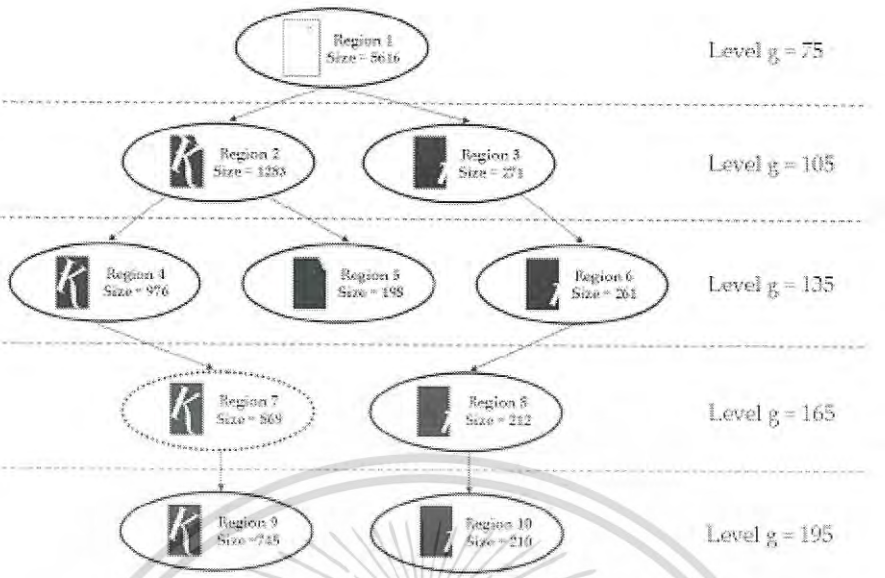
A part of Image

Intensity	Number of pixels	Position of pixels
31	2	(1,2),(2,1)
32	1	(1,1)
131	3	(4,5),(5,4),(5,5)
135	3	(3,5),(4,4),(5,3)
136	4	(2,5),(3,4),(4,3),(5,2)
150	5	(1,5),(2,4),(3,3),(4,2),(5,1)
151	4	(1,4),(2,3),(3,2),(4,1)
152	3	(1,3),(2,2),(3,1)

รูปที่ 2.2 แสดงการเรียงลำดับของพิกเซลโดยเรียงจากค่าความเข้มของสีในภาพระดับสีเทา

2.2.1.2 การแบ่งกลุ่มข้อมูล (Clustering)

หลังจากที่เรียงข้อมูลตามความเข้มของสีแล้ว จึงนำข้อมูลเหล่านี้มาทำการค้นหาว่าอยู่ในพื้นที่เดียวกันหรือไม่ โดยหลักการ MSER นั้นจะไม่ตรวจสอบกับจุดพิกเซลที่มีการเชื่อมต่อน้อยกว่า 4 จุด โดยในขั้นตอนนี้จะนำ Set ข้อมูลนั้นมาทำการ Union กันแล้วทำการค้นหาว่าอยู่ในพื้นที่เดียวกันหรือไม่ ถ้าใช่ ในพื้นที่นั้นก็จะมีเซตข้อมูลของพิกเซลภาพและความเข้มของสีอยู่ ถ้าไม่ก็จะค้นหาต่อไปหาพิกเซลเหล่านี้อยู่ในพื้นที่อื่นหรือไม่ โดยการค้นหาจะค้นหาจนกว่าได้เซตของพื้นที่ทั้งหมด โดยเรียงต้วกันตามการเปลี่ยนแปลงของค่าความเข้มของสี โดยผลลัพธ์ที่ได้จะเป็นต้นไม้ของพื้นที่ที่เรียงลำดับตามค่าความเข้มของสี โดยผลลัพธ์จะแสดงดังรูปที่ 2.3



รูปที่ 2.3 แสดงผลลัพธ์ของต้นไม้ของพื้นที่โดยเรียงจากค่าความเข้มของสี

(ภาพ : M. Donoser and H. Bischof. 2006. Efficient Maximally Stable Extremal Region (MSER) Tracking.)

2.2.1.3 การตรวจจับ MSER (MSER Detection)

การตรวจจับของ MSER นั้นจะนำผลลัพธ์จากขั้นตอนก่อนหน้านี้มาทำการตรวจสอบทีละระดับค่าสี โดยแบ่งเป็นการไล่ระดับค่าสีสองรูปแบบคือ จากสว่างไปมืด (255 - 0) และจากมืดไปสว่าง (0 - 255) โดยในการตรวจจับ MSER นั้นจะมีการเปรียบเทียบค่าสีทั้งหมดเปรียบเสมือนเซตเซตหนึ่ง ซึ่งมีการกำหนดคุณสมบัติของ MSER ผ่านหลักทางคณิตศาสตร์ ดังนี้ [13]

Image I คือแผนผังของภาพที่ถูกกำหนดค่า $I \subset Z^2 \rightarrow S$. โดยที่ Extremal Region บนภาพจะถูกกำหนดจากเงื่อนไขดังนี้

1. S คือที่ถูกคำนวณทั้งหมดและจะต้องเป็น set ที่มีคุณสมบัติมีความสะท้อน, ไม่สมมาตรและเป็นความสัมพันธ์แบบถ่ายทอด โดยที่มีค่าน้อยกว่า Extremal Region ที่มีอยู่ซึ่งถูกกำหนดตามค่าต่าง ๆ เช่น ค่าจริงของภาพ (S = R)

2. ค่า adjacency (neighborhood) relation ที่ถูกกำหนดค่า $A \subset D \times D$ นั้นคือ $p, q \in D$ ที่อยู่ติดกัน (ใน MSERs จะใช้การติดกันของจุดขั้นต่ำที่ 4 จุด) โดยที่ (pAq) if $\sum_{i=1}^{d} |p_i - q_i| \leq 1.f$

Region Q คือเซตย่อยต่อเนื่องของ D นั้นคือแต่ละ $p, q \in Q$ จะเป็นลำดับ $p, a_1, a_2, \ldots, a_n, q$ และ $pAa_1, a_iAa_{i+1}, a_nAq$

(Outer) Region Boundary $\partial Q = \{q = D \setminus Q: \exists p \in Q: qAp\}$. นั้นคือเส้นขอบ ∂Q ของ Q ซึ่งเป็นเซตของพิกเซลที่อยู่ติดกันอย่างน้อย 1 พิกเซลของ Q แต่ค่าค่านั้นไม่อยู่ใน Q

Extremal Region $Q \subset D$ คือพื้นที่สำหรับทุกค่าของ $p \in Q, q \in \partial Q: I(p) > I(q)$ (ค่าความเข้มของแสงที่สูงสุดในพื้นที่หรือ $I(p) < I(q)$ (ค่าความเข้มของแสงที่ต่ำสุดในพื้นที่)

Maximally Stable Extremal region (MSER). ให้ $Q_1, \ldots, Q_{i-1}, Q_i, \ldots$ เรียงตามลำดับของการเชื่องโยงของ Extremal Regions นั้นคือ $Q_1 \subset Q_{i+1}$ โดยที่ Q_{i*} จะเป็น Maximally Stable ก็ต่อเมื่อค่า $q(i) = |Q_{i+\Delta} \setminus Q_{i-\Delta}| / |Q_i|$ นั้นเกิดค่า local minimum ที่ i^* (. |หมายถึงการนับจำนวนสมาชิก). $\Delta \in S$ นั้นคือ พารามิเตอร์ของวิธีการนี้

ตามหลักคณิตศาสตร์นั้นพื้นที่ที่เป็น MSER จะต้องมีจุด Local Minimum เสมอและจะต้องไม่มีการเปลี่ยนแปลงของขนาดพื้นที่เมื่อผ่านช่วงการเปลี่ยนแปลงความเข้มในระยะหนึ่ง โดยการหาว่าพื้นที่นั้นใช่ MSER หรือไม่ สามารถค้นหาได้จากสมการนี้

$$\frac{|Q_{i+\Delta} \setminus Q_{i-\Delta}|}{|Q_i|} \tag{2.5}$$

โดยสมการนี้สามารถกำหนดค่าการค้นหา MSER ได้จากค่า Δ โดยที่ค่าสีของภาพทุกภาพต้องอยู่ในช่วง 0-255

2.2.1.4 แสดงผล(Result)

หลังจากตรวจสอบแล้วว่าพื้นที่นั้นใช้ MSER หรือไม่แล้ว ก้จะนำ Set ของพิกเซลที่เป็น MSER มาทำการแสดงผล โดยการแสดงผลของ MSER นั้นสามารถแสดงได้ทั้งรูปแบบเซตของพิกเซล

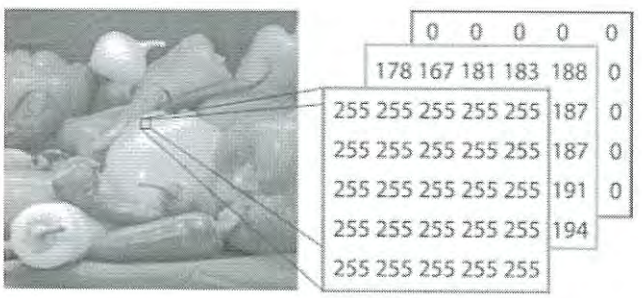
ภาพ Binary สีขาวดำ หรือการวางจุดบนรูปแบบ โดยผลลัพธ์สุดท้ายที่ได้จาก MSER จะเป็น Set ของตำแหน่งของพื้นที่ที่เป็น MSER โดยมีตัวอย่างผลลัพธ์ของ MSER ดังรูปที่ 2.4

รูปที่ 2.4 ตัวอย่างการแสดงผลลัพธ์ของ MSER

2.3 ทฤษฎีการประมวลผลภาพดิจิตอล (Digital Image Processing)

2.3.1 ภาพสี (RGB)

กระบวนการจำแนกกลุ่มเส้นขอบจะพิจารณาจากค่าสีของพิกเซล โดยแต่ละพิกเซลของ ภาพสีจะประกอบด้วยระดับของความเข้มของแสง 3 ค่า คือ สีแดง สีเขียว และสีน้ำเงิน [16], [17], [18] ดังรูปที่ 2.5



รูปที่ 2.5 ตัวอย่างภาพสีและค่าความเข้มแสงของแต่ละพิกเซล

2.3.2 ภาพระดับสีเทา (Intensity Image)

ภาพระดับสีเทาโดยทั่วไปนั้นมักมีขนาด 8 บิต โดยสามารถแสดงระดับค่าสีเท่าได้ 256 ระดับ โดยมีค่าที่เป็นไปได้คือ[0-255] โดยค่าความเข้มแสงที่น้อยที่สุดจะมีค่าเท่ากับ 0 จะทำให้พิกเซลนั้นเป็นสีดำ แต่ในทางตรงกันข้าม หากค่าความเข้มแสงมีระดับสูงสุด คือมีค่าเท่ากับ 255 นั้นจะทำให้พิกเซลนั้นๆ เป็นสีขาว แต่อย่างไรก็ตาม ระดับภาพสีเทา สามารถมีค่าน้อยหรือมากกว่าขนาด 8 บิต ได้เสมอ เนื่องจากทุกอย่างขึ้นอยู่กันความละเอียดของภาพ[16], [17], [18] โดยรูปที่ 2.6 จะเป็นการแสดงตัวอย่างของภาพระดับสีเทา และค่าความเข้มของแสงในแต่ละพิกเซล

รูปที่ 2.6 ตัวอย่างของภาพระดับสีเทาที่มีขนาด 8 บิต และค่าความเข้มของแสงในแต่ละพิกเซล

2.4 งานวิจัยที่เกี่ยวข้อง

2.4.1 การตรวจจับข้อความโดยพิจารณาจากพื้นที่

Ross G., Jeff D., Trevor D. และ Jitendra M. นำเสนอวิธี R-CNN (Regions - Convolutional neural network) ที่ใช้การตรวจจับข้อมูลโดยนำภาพเข้ามาสกัดคุณลักษณะผ่านคอนโวลูชันนอลนิวรอลเน็ตเวิร์ค [2] โดยข้อดีของวิธีการนี้คือ ได้ผลลัพธ์ที่แม่นยำ และ ใช้ระยะเวลาในการประมวลผลสั้น และสามารถนำผลลัพธ์ที่ได้ไปใช้งานได้ทันที แต่ข้อเสียที่ได้คือ การใช้ฮาร์ดแวร์ที่มีคุณสมบัติสูง และมีความซับซ้อนในการเตรียมสร้างแบบโครงข่ายประสาทเทียมซึ่งใช้เวลานาน

Weilin H., Yu Q. และ Xiaoon T. นำเสนอวิธี Convolutional neural network Induced MSER Trees ซึ่งใช้วิธีการตรวจจับโดยพิจารณาพื้นที่ โดยใช้ MSER และใช้คอนโวลูชันนอล นิวรอลเน็ตเวิร์ค [4] สร้างแบบจำลองเพื่อคัดแยกตัวอักษรและสิ่งที่ไม่ใช่ตัวอักษรออกจากกัน ซึ่งข้อดีคือมีความแม่นยำ เนื่องจาก การทำงานควบคู่กันของ MSERs และ Convolutional neural network นั้น

เสริมประสิทธิภาพซึ่งกันและกัน แต่ก็มีข้อเสียกับรูปภาพบางกรณี เช่น ภาพที่มีความคมชัดต่ำ หรือภาพที่มีตัวอักษรและพื้นหลังที่ซับซ้อน เช่น ตัวอักษรที่อยู่บนกำแพงอิฐ ตัวอักษรที่มีสีเดียวกับพื้นหลังหรือมีสีที่ใกล้เคียงกับพื้นหลังและภาพตัวอักษรที่มองเห็นไม่ชัด จะไม่สามารถตรวจจับได้โดยวิธีการนี้ เพราะว่าการตรวจจับของ MSER นั้นบางครั้งมักจะให้ส่วนประกอบของภาพพื้นหลังติดมาด้วย และลักษณะภาพของอิฐนั้นส่วนมากจะมีค่าสีที่ไม่ห่างกันมาก จึงทำให้ MSER สามารถตรวจจับได้ นอกจากนี้ การเสริมสร้างประสิทธิภาพประมวลผลของงานวิจัยนี้ให้รวดเร็วได้นั้นควรการใช้ฮาร์ดแวร์ที่มีคุณสมบัติสูง และมีความซับซ้อนในการเตรียมสร้างแบบโครงข่ายประสาทเทียมซึ่งใช้เวลานาน

Datong C., Herve B. และ Jean-P. T [6] ได้นำเสนอวิธียืนยันพื้นที่ตัวอักษรบนรูปภาพโดยใช้การตรวจสอบคุณสมบัติของกลุ่มตัวอักษร โดยวิเคราะห์จากการวางตัวอักษรในช่วงแนวตั้งและแนวนอน และหลังจากนั้นนำพื้นที่ที่มีแนวโน้มว่าจะเป็นกลุ่มของตัวอักษรนั้นมาสกัดคุณลักษณะภายในกลุ่ม และใช้หลักการ SVM ในการตรวจสอบว่ากลุ่มของพื้นที่นั้นใช่กลุ่มของตัวอักษรหรือไม่ โดยผลลัพธ์สุดท้ายที่ได้คือพื้นที่ตัวอักษรที่ได้รับการยืนยันเรียบร้อยแล้ว โดยวิธีการนี้มีข้อดีคือ ประมวลผลไว และแม่นยำ แต่มีข้อเสียคือสามารถรองรับได้แค่บางภาษาเท่านั้น

Wei, Yi Cheng และ Chang Hong Lin.[7] ได้นำเสนอวิธีตรวจจับตัวอักษรโดยใช้ หลักการประมวลผลภาพแบบ Pyramidal Method ตามแนวตั้งและแนวนอนของภาพเพื่อตรวจสอบลักษณะพื้นที่และพื้นผิวทั้งหมดของภาพ และใช้ SVM ในการแยกประเภทของพื้นที่ที่ถูกประมวลผลภาพว่าบริเวณนั้นเป็นพื้นที่ของตัวอักษรหรือไม่ ซึ่งวิธีนี้สามารถสอนและประมวลผลได้อย่างมีประสิทธิภาพและมีความแม่นยำสำหรับภาษาอังกฤษ

2.4.2 การตรวจจับข้อความโดยพิจารณาส่วนประกอบที่เชื่อมต่อกัน

Xiaobing W., Yonghong S., Yuanlin Z. และ Jingmin X. นำเสนอวิธี multi-layer CC segmentation [3] ที่ใช้ตรวจจับข้อความโดยแบ่งส่วนประกอบของภาพเป็นหลายๆชั้น แล้วนำผลลัพธ์ของการแบ่งชั้นนำมาตรวจจับข้อความบนรูปภาพ โดยการตรวจจับจะจัดเป็นลำดับชั้น เช่น ส่วนประกอบที่เป็นตัวอักษร ส่วนประกอบที่เป็นคำ และส่วนประกอบที่เป็นวลีหรือข้อความ ซึ่งวิธีการนี้มีผลดีคือ มีความแม่นยำสูง แม้ว่าจะมีสัญญาณรบกวน แต่ข้อเสียคือ ใช้เวลาประมวลผลนาน เนื่องจากการวิเคราะห์รายละเอียดของภาพ

พิมพ์ลักษ์ บุญชูกุศล [5] ได้นำเสนอวิธีการวิเคราะห์ส่วนประกอบที่เชื่อมต่อกันของวัตถุ (Modified Connected Component Analysis: MCCA) ในการค้นหาตำแหน่งของข้อความบน

ภาพถ่าย โดยทำการตรวจจับตัวอักษรภาษาไทยในส่วนลำตัวหลัก (Modified Body of Text Detection) การรวมกลุ่มข้อความที่อยู่ในแนวเดียวกัน(Text Grouping) และทำการขยายกรอบล้อมรอบข้อความ (Text Boundary Padding) เพื่อตรวจจับสระบน และล่างของข้อความ โดยใช้ร่วมกับกฎที่สร้างขึ้นมาใหม่จากสัดส่วนมาตรฐานของตัวอักษรไทย

บทที่ 3

การดำเนินงานวิจัย

จากแนวคิดของโครงการตามที่ระบุไว้ในบทที่ 1 โดยในบทที่ 3 นั้น จะกล่าวถึงแนวความคิดเพิ่มเติมโดยลงไปในรายละเอียดของขั้นตอนหลักต่างๆ โดยในบทนี้ จะขยายขั้นตอนการตรวจจับข้อความ (Text Detection) การปรับปรุงข้อมูลเพื่อเข้าสู่ระบบรู้จำข้อความ ขั้นตอนการรู้จำข้อความ (Text Recognition) และแปลงข้อความเป็นเสียงพูด (Text - To - Speech) ซึ่งมีการทำงานในแต่ละขั้นตอน ดังนี้

3.1 การตรวจจับข้อความ

3.1.1 กระบวนการตระเตรียมภาพ (Pre-processing)

นำภาพสีระบบ RGB เข้ามาทำการแบ่งส่วนประกอบของภาพตาม Channel ของสี โดยระบบสี RGB นั้นจะนำเข้ามาในรูปของ Matrix 3 มิติ เมื่อนำมาแบ่งแล้วจะได้ Matrix 2 มิติและเมื่อแสดงผลออกมา จะค้นพบว่าภาพที่ได้จะเป็นภาพระดับสีเทา ซึ่งแตกต่างกัน ในกรณีที่ค่าสีของแต่ละช่องนั้นไม่เท่ากัน โดยแสดงดังรูปที่ 3.1

รูปที่ 3.1 แสดงการแยกส่วนประกอบของภาพตาม Color Channel ของภาพที่รับเข้ามา

3.1.2 ตรวจจับคุณลักษณะของภาพโดยใช้หลักการ MSER

เมื่อได้ภาพที่ทำการแยกส่วนประกอบของสีแล้ว เนื่องจากภาพสี RGB เมื่อทำการตรวจจับคุณลักษณะ โดยใช้หลักการMaximally stable extremal regions โดยในโครงงงานนี้ จะใช้การปรับ

ค่า Delta เพื่อยกเว้นการประมวลผลทุก ๆ 2 ค่าโดยการแสดงผลนั้น จะแสดงผลเป็นภาพ binary และมีความแตกต่างตามรูปที่ 3.2

รูปที่ 3.2 แสดงความแตกต่างของภาพในแต่ละช่องสีที่ผ่านการตรวจจับคุณลักษณะโดยใช้ MSER

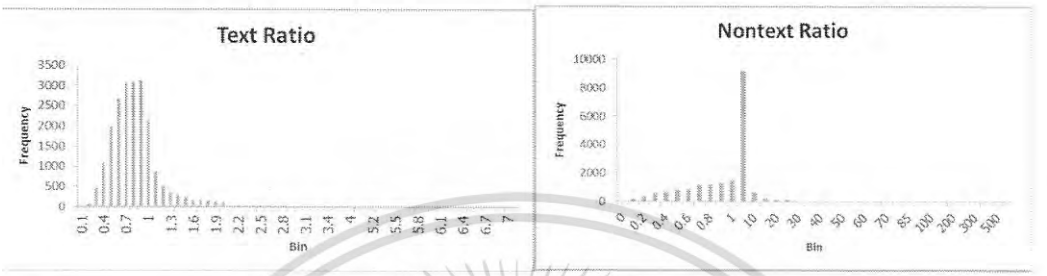
3.1.3 แยกกลุ่มของตัวอักษรโดยใช้หลักการ Support Vector Machine

ในการแยกกลุ่มตัวอักษรนั้น ทางผู้จัดทำได้ทำการสอนข้อมูลโดยใช้พื้นที่ของตัวอักษรเป็นคุณลักษณะที่ใช้ในการสอน โดยผู้จัดทำได้แบ่งข้อมูลเป็น 2 ส่วนและใช้คุณสมบัติในการสร้างสมการ ซึ่งแสดงในตารางที่ 3.1 ดังนี้

ตารางที่ 3.1 คุณสมบัติของพื้นที่ที่ใช้ในการสร้างสมการและแยกกลุ่ม[19]

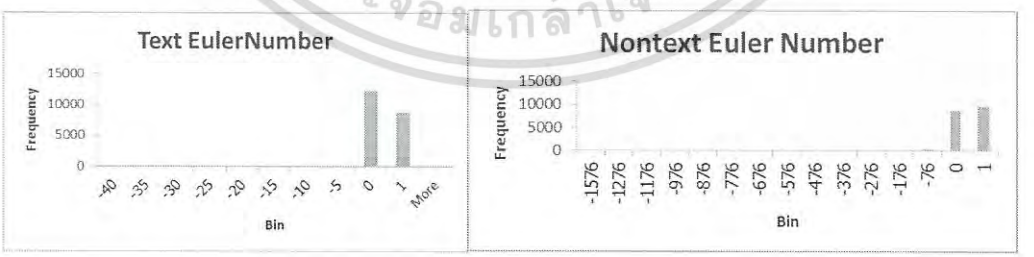
คุณสมบัติ	สมการ/คุณสมบัติ
Ratio	Width/Height
Eccentricity	The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1.
Euler number	objects in the region - the number of holes in objects.
Extent	Area/bounding box
Solidity	area/convex area
Perimeter	distance between each adjoining pair of pixels around the border of the region

โดยจากการใช้ชุดข้อมูลทั้งหมด 40000 ชุด โดยแบ่งเป็นชุดตัวอักษร 21,000 พื้นที่ และไม่ใช่ตัวอักษร 19000 พื้นที่ สามารถแจกแจงความถี่เพื่อแสดงให้เห็นความแตกต่างได้ ดังรูปที่ 3.3 – 3.8

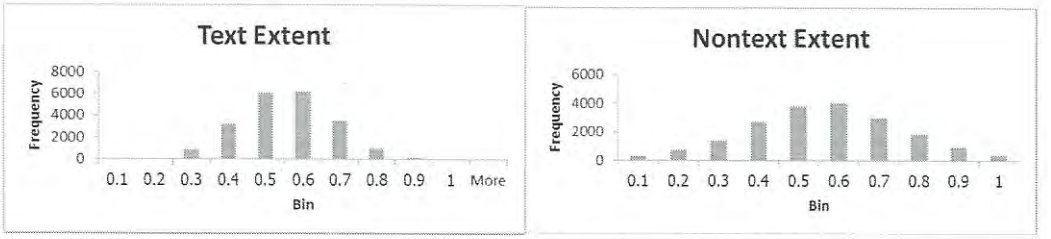


รูปที่ 3.3 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Ratio

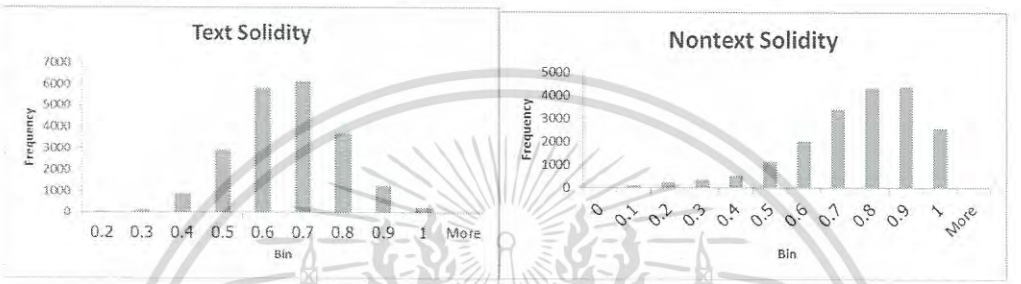
รูปที่ 3.4 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Eccentricity



รูปที่ 3.5 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Euler number



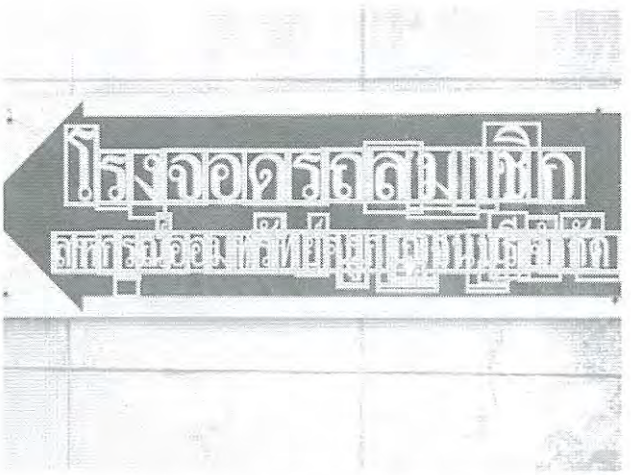
รูปที่ 3.6 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Extent



รูปที่ 3.7 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Solidity

รูปที่ 3.8 การแจกแจงความถี่ระหว่างชุดข้อมูลตัวอักษรและไม่ใช่ตัวอักษร โดยอ้างอิงจากค่า Perimeter

เมื่อทำการสอนข้อมูลทั้งหมดแล้ว ระบบจะสร้างแบบจำลองพร้อมทำนาย โดยการทำนายนั้นจะทำนายจากพื้นที่ MSER ที่ถูกตรวจจับได้ในขั้นตอนที่ 3.1.2 แล้วเก็บเฉพาะสิ่งที่ถูกทำนายได้ว่าเป็นตัวอักษร ซึ่งแสดงการวางของกรอบภาพตามรูปที่ 3.9



รูปที่ 3.9 แสดงการวางกรอบสีเหลี่ยมของพื้นที่ที่ถูกทำนายว่าเป็นตัวอักษร

3.2 การปรับปรุงข้อมูลก่อนเข้าสู่ระบบรู้จำข้อความ

3.2.1 การรวมกลุ่มของพื้นที่ที่ถูกทำนายว่าเป็นตัวอักษร

ก่อนจะทำพื้นที่ตัวอักษรเข้าสู้กระบวนการรู้จำข้อความนั้น มีความจำเป็นที่จะต้องรวมกลุ่มของตำแหน่งให้เป็นกลุ่มของข้อความเสียก่อน โดยการรวมกลุ่มตัวอักษรนั้น ใช้วิธีการตรวจสอบการคาบเกี่ยว (Overlap) ของแต่ละพื้นที่ตัวอักษรแล้วนำมารวมเข้าด้วยกัน โดยผลลัพธ์ที่ได้จะแสดงดังรูปที่ 3.10

รูปที่ 3.10 แสดงผลลัพธ์ของการรวมกลุ่มของพื้นที่ที่ถูกทำนายว่าเป็นอักษร

3.2.2 การเลือกใช้กล่องเพื่อเข้าสู่ระบบรู้จำข้อความ

เมื่อได้กลุ่มข้อความแล้ว ระบบจะสั่งให้เก็บกลุ่มข้อความเป็นตำแหน่งบนภาพ เมื่อนำมาเข้ากระบวนการรู้จำข้อความ ระบบจะส่งให้ทำการรู้จำข้อความทีละกล่อง และทำให้ได้ผลลัพธ์ที่ได้คือตัวอักษรที่แก้ไขได้ เพื่อใช้ในการแปลงเป็นเสียงพูดต่อไป

3.3 การรู้จำข้อความ

ขั้นตอนนี้ทางผู้จำทำได้ใช้ Tesseract OCR ในระบบการรู้จำข้อความ ซึ่งสามารถรองรับได้ทั้งภาษาไทยและภาษาอังกฤษ ซึ่งการรู้จำข้อความโดยใช้ Tesseract นั้นมีประสิทธิภาพที่ดี แต่จะมีข้อจำกัดกับภาพที่มีฟ้อนต์ที่ไม่คุ้นเคยหรือภาพที่มีพื้นหลังซับซ้อน นอกจากนี้สำหรับภาษาไทยยังมีข้อจำกัดในเรื่องของภาพที่มาจากฉากธรรมชาติ จึงทำให้ผลลัพธ์ของภาษาไทยที่ได้บางครั้งจะถูกเปลี่ยนไปเป็นภาษาอื่น ๆ แทน

3.4 การแปลงข้อความเป็นเสียงพูด

ขั้นตอนนี้ผู้จัดทำใช้ Microsoft Speech API (SAPI 5.4) ซึ่งสามารถรองรับได้เฉพาะภาษาอังกฤษ แต่สามารถใช้งานง่าย สะดวกเพราะไม่ต้องลงโปรแกรมเสริมใด ๆ เนื่องจากจะมีมาให้เมื่อใช้กับ Windows อยู่แล้ว แต่ SAPI เองก็มีข้อจำกัดในเรื่องของการอ่านเครื่องหมายและถ้าหากผลลัพธ์จากขั้นตอนที่ 3.3 ออกมาผิดพลาด ระบบก็จะอ่านตามตัวอักษรที่ถูกตรวจจับได้เท่านั้น

สำหรับการเรียกใช้งาน Microsoft Speech API นั้นสามารถเรียกใช้ได้โดยการเรียกใช้งานผ่าน Function TTS ของ Matlab ซึ่งสามารถดาวน์โหลดได้ผ่านทาง File Exchange ของทาง Matlab

บทที่ 4

ผลการทดลอง

4.1 การวัดประสิทธิภาพ

ในการวัดผลประสิทธิภาพนั้นทางผู้จัดทำจะวัดค่าที่ได้จากการตรวจจับข้อความบนรูปภาพ โดยมีการคำนวณค่าความแม่นยำ(Precision) ค่าความครบถ้วน(Recall) และค่าวัดประสิทธิภาพ (F-Measure) ตามตารางที่ 4.1 โดยสมการนี้อ้างอิงมาจากการแข่งขัน BEST 2010 [21]

ตารางที่ 4.1 แสดงสมการคำนวณ ค่า Precision, Recall และ F-Measure

ค่าทดสอบความแม่นยำ	สมการ
Precision	Cor / Output
Recall	Cor / Ref
F-Measure	F-Measure = 2 x Precision x Recall / (Precision + Recall)

*Cor = จำนวนข้อมูลที่ประมวลผลได้อย่างถูกต้อง, Output = จำนวนข้อมูลที่ประมวลผลทั้งหมด, Ref = จำนวนข้อมูลที่เฉลยมาจาก Dataset

4.2 ผลการทดลอง

จากการทดลองมีการแบ่งผลการทดลองความแม่นยำ ความครบถ้วนและประสิทธิภาพ ของภาษาไทยและภาษาอังกฤษแยกออกจากกัน โดยทางผู้จัดทำใช้ภาพในการทดลองทั้งหมด 400 ภาพ โดยแบ่งเป็นภาพจากฉากธรรมชาติที่มีข้อความภาษาไทยทั้งหมด 200 ภาพ และ ภาพจากฉากธรรมชาติที่มีข้อความภาษาอังกฤษทั้งหมด 200 ภาพ โดยภาพที่มีข้อความแต่ละภาษาจะถูกแบ่งกลุ่มออกเป็น 2 กลุ่ม กลุ่มละ 100 ภาพ โดยแยกกลุ่มจากความยากง่ายของความซับซ้อนของพื้นหลังและการวางรูปแบบข้อความที่หลากหลาย จากการทดลองทั้งหมดนั้นสามารถแสดงผลการทดลองตามตารางที่ 4.2 และ ตารางที่ 4.3

ตารางที่ 4.2 แสดงผลการทดลองภาษาไทย

กลุ่มที่	จำนวนตัวอักษรทั้งหมด	จำนวนของวัตถุที่ตรวจจับได้		เวลาที่ใช้ในการประมวล(นาที/ภาพ)	ประสิทธิภาพ		
		ตัวอักษร	ไม่ใช่ตัวอักษร		P	R	F
1	1,987	1,349	942	1.3757	0.58	0.67	0.62
2	2,675	1,791	899	1.7045	0.67	0.66	0.66

ตารางที่ 4.3 แสดงผลการทดลองภาษาอังกฤษ

กลุ่มที่	จำนวนตัวอักษรทั้งหมด	จำนวนของวัตถุที่ตรวจจับได้		เวลาที่ใช้ในการประมวล(นาที/ภาพ)	ประสิทธิภาพ		
		ตัวอักษร	ไม่ใช่ตัวอักษร		P	R	F
1	1,254	849	478	0.5187	0.63	0.67	0.65
2	984	692	635	1.9562	0.52	0.70	0.59

4.3 ผลการทดลองเมื่อเทียบกับวิธีการอื่น ๆ

จากการเปรียบเทียบการทดลองกับวิธีการต่าง ๆ เพื่อวัดประสิทธิภาพของวิธีการที่ใช้ในรายงาน โดยเปรียบเทียบประสิทธิภาพระหว่างการใช้ MSER ร่วมกับกฎพื้นฐานของโครงสร้างตัวอักษร และการใช้วิธีการเดียวกับรายงานแต่ลดขั้นตอนการแยกข้อมูลภาพตามค่าแสงของช่องสี (RGB Cluster) ซึ่งผู้จัดทำใช้ภาพชุดเดียวกันทั้งหมด แต่ไม่แบ่งกลุ่มตามมุมมองและความหลากหลายของตัวอักษร จากการทดลองภาพจากฉากธรรมชาติที่มีข้อความภาษาไทยทั้งหมด 200 ภาพ และภาพจากฉากธรรมชาติที่มีข้อความภาษาอังกฤษ 200 ภาพ สามารถเปรียบเทียบประสิทธิภาพและสามารถแสดงผลการทดลองตามตารางที่ 4.4 และ 4.5

ตารางที่ 4.4 แสดงผลการทดลองเมื่อเทียบกับวิธีการอื่น ๆ สำหรับภาษาไทย

วิธีการ	จำนวนตัวอักษรทั้งหมด	จำนวนของวัตถุที่ตรวจจับได้		เวลาที่ใช้ในการประมวล(นาที/ภาพ	ประสิทธิภาพ		
		ตัวอักษร	ไม่ใช่ตัวอักษร		P	R	F
Rules base	4662	3556	2756	1.0447	0.56	0.76	0.64
SVM (without RGB clustering)	4662	3150	1720	1.2658	0.65	0.68	0.66
SVM	4662	3140	1841	1.7321	0.63	0.67	0.65

ตารางที่ 4.5 แสดงผลการทดลองเมื่อเทียบกับวิธีการอื่น ๆ สำหรับภาษาอังกฤษ

วิธีการ	จำนวนตัวอักษรทั้งหมด	จำนวนของวัตถุที่ตรวจจับได้		เวลาที่ใช้ในการประมวล(นาที/ภาพ	ประสิทธิภาพ		
		ตัวอักษร	ไม่ใช่ตัวอักษร		P	R	F
Rules base	2238	1854	1698	0.7845	0.48	0.83	0.61
SVM (without RGB clustering)	2238	1547	1042	1.0245	0.60	0.69	0.64
SVM	2238	1541	1113	1.2374	0.58	0.68	0.63

บทที่ 5
สรุปผลการทดลองและข้อเสนอแนะ

5.1 สรุปผลการทดลอง

จากการทดลองค้นพบว่าการใช้วิธีนี้ สามารถทำได้ในกรณีที่จำกัด แต่ว่าถ้าหากมีการปรับปรุงกระบวนการเรียนรู้ของเครื่องให้มีประสิทธิภาพในการแยกกลุ่มให้ได้มากกว่านี้ อาจจะทำให้ ค่าครบถ้วนและค่าความแม่นยำนั้นมีค่าสูงขึ้นได้ เพราะว่าจากการเปรียบเทียบการให้คะแนนความมั่นใจของระบบการเรียนรู้ของเครื่อง จะค้นพบว่ามีการให้คะแนนระหว่างสองวัตถุที่ใกล้เคียงกัน สำหรับบางกรณีการเรียนรู้ของเครื่องไม่สามารถทำนายข้อมูลได้ถูกต้องเลย ถึงแม้จะปรับการตรวจจับของ MSER ให้ไม่มีความคงทนต่อการเปลี่ยนแปลงของค่าความเข้มของแสงและไม่ได้ปรับช่วงของพื้นที่ให้มีความจำกัด ซึ่งจากผลการทดลองนี้ จึงทำให้ผู้จัดทำเล็งเห็นว่าควรจะแก้ไขที่การเรียนรู้ของเครื่อง เพื่อเพิ่มประสิทธิภาพในการตรวจจับข้อความ

5.2 สรุปผลการทดลองเมื่อเทียบกับวิธีอื่น ๆ

จากการทดลองเมื่อเทียบกับวิธีอื่น ๆ จะพบว่าประสิทธิภาพของการใช้วิธี MSER และ SVM มีค่าที่สูงที่สุดสำหรับกรณีที่ไม่มีการแบ่งแยกช่องสีก่อนตรวจจับคุณลักษณะ และค้นพบว่าการใช้กฎในการกรองพื้นที่นั้น สามารถให้ความครบถ้วนที่มากกว่า แต่เมื่อเปรียบเทียบความแม่นยำและค่าประสิทธิภาพของการตรวจจับข้อความ จะให้ค่าที่น้อยกว่าการใช้การเรียนรู้ของเครื่องในการคัดแยกข้อมูลพื้นที่นั้นออกจากกัน สำหรับกรณีที่ภาพนั้นไม่มีสัญญาณรบกวนมาก การทำงานของทั้งสองวิธีนั้นสามารถทำได้ดีในระดับที่ใกล้เคียงกัน

5.3 ข้อเสนอแนะ

5.3.1 ปรับปรุงการสอนการเรียนรู้ของเครื่องให้ดีขึ้น เพื่อลดความสับสนในการทำนายของพื้นที่บางพื้นที่

5.3.2 เพิ่มวงจำกัดของการวิจัย ให้มีความคงทนในการตรวจจับข้อความมากขึ้น

5.3.3 ปรับปรุงข้อมูลที่ใช้ในการสอนให้เหมาะสมกว่านี้

บรรณานุกรม

[1] Zhenyu Z. , Cong F. , Zhouchen L. , Yi W. 2015. **A robust hybrid method for text detection in natural scenes by learning-based partial differential equations**. ScienceDirect Neurocomputing 168 (2015). pp. 23–34

[2] Ross G., Jeff D., Trevor D, Jitendra M. 2012. **feature hierarchies for accurate object detection and semantic segmentation** .Computer Science Division. UC Berkeley

[3] Xiaobing W., Yonghong S., Yuanlin Z. , Jingmin X. 2015. **Natural scene text detection with multi-layer segmentation and higher order conditional random field based analysis .** ScienceDirect Pattern Recognition Letters 60–61 (2015). pp. 41–47

[4] Weilin H., Yu Q., Xiaoon T. 2014.**Robust Scene Text Detection with Convolutional Neural Network Induced MSER Tree**. ECCV 2014 Part IV LNCS 8692. pp 497-511

[5] PIMLAK BOONCHUKUSOL. 2014 **.Thai Text Detection From Medium Shot Of Natural Scenes By Using Fast Boundary Clustering And Modified Connected Component Analysis.**วิทยานิพนธ์ สาขาวิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

[6] Datong C.,Herve B.,Jean-P. T. 2001. **Text identification in Complex Background Using SVM**. Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Vol. 2. IEEE, 2001.

[7] Wei, Yi Cheng, and Chang Hong Lin.2012. **A robust video text detection approach using SVM**. "Expert Systems with Applications 39.12 (2012): 10832-10840.

[8] Smith, R. 2007. **An overview of the Tesseract OCR** engine. In icdar (pp. 629-633). IEEE.

[9] Microsoft. **Microsoft Speech API Overview (SAPI 5.4).** [Online]https://msdn.microsoft.com/en-us/library/ee125077(v=vs.85).aspx

[10] Kevin P. M. 2012. **Machine learning: a probabilistic perspective**. The MIT Press Cambridge, Massachusetts London, England

บรรณานุกรม (ต่อ)

[11] Andrew N. 2015. **Machine Learning**. Coursea. [Online] https://www.coursera.org/learn/machine-learning

[12] Amnon S. 2008. **Support Vector Machines and Kernel Functions.** Introduction to Machine Learning67577 - Fall, 2008 School of Computer Science and Engineering. The Hebrew University of Jerusalem Jerusalem, Israel

[13] Matas, J., O. Chum, M. Urba, and T. Pajdla. **"Robust wide baseline stereo from maximally stable extremal regions."** Proceedings of British Machine Vision Conference, pages 384-396, 2002.

[14] Nister, D., and H. Stewenius, **"Linear Time Maximally Stable Extremal Regions"**. Lecture Notes in Computer Science. 10th European Conference on Computer Vision, Marseille, France: 2008, no. 5303, pp. 183–196.

[15] Fredrik Kristensen and W. James MacLean. 2007.Real-**Time Extraction of Maximally Stable Extremal Regions on an FPGA.** Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on (pp. 165-168). IEEE

[16] สมเกียรติ อุดมหรรษากุล. 2554. **การประมวลผลภาพดิจิตอลเบื้องต้น.** กรุงเทพฯ : สำนักพิมพ์ท็อป

[17] อรฉัตร จิตต์โสภักตร์ 2552. **ทฤษฎีการประมวลผลภาพดิจิตอล.** กรุงเทพฯ : สงวนกิจพรินท์ แอนด์ มีเดีย.

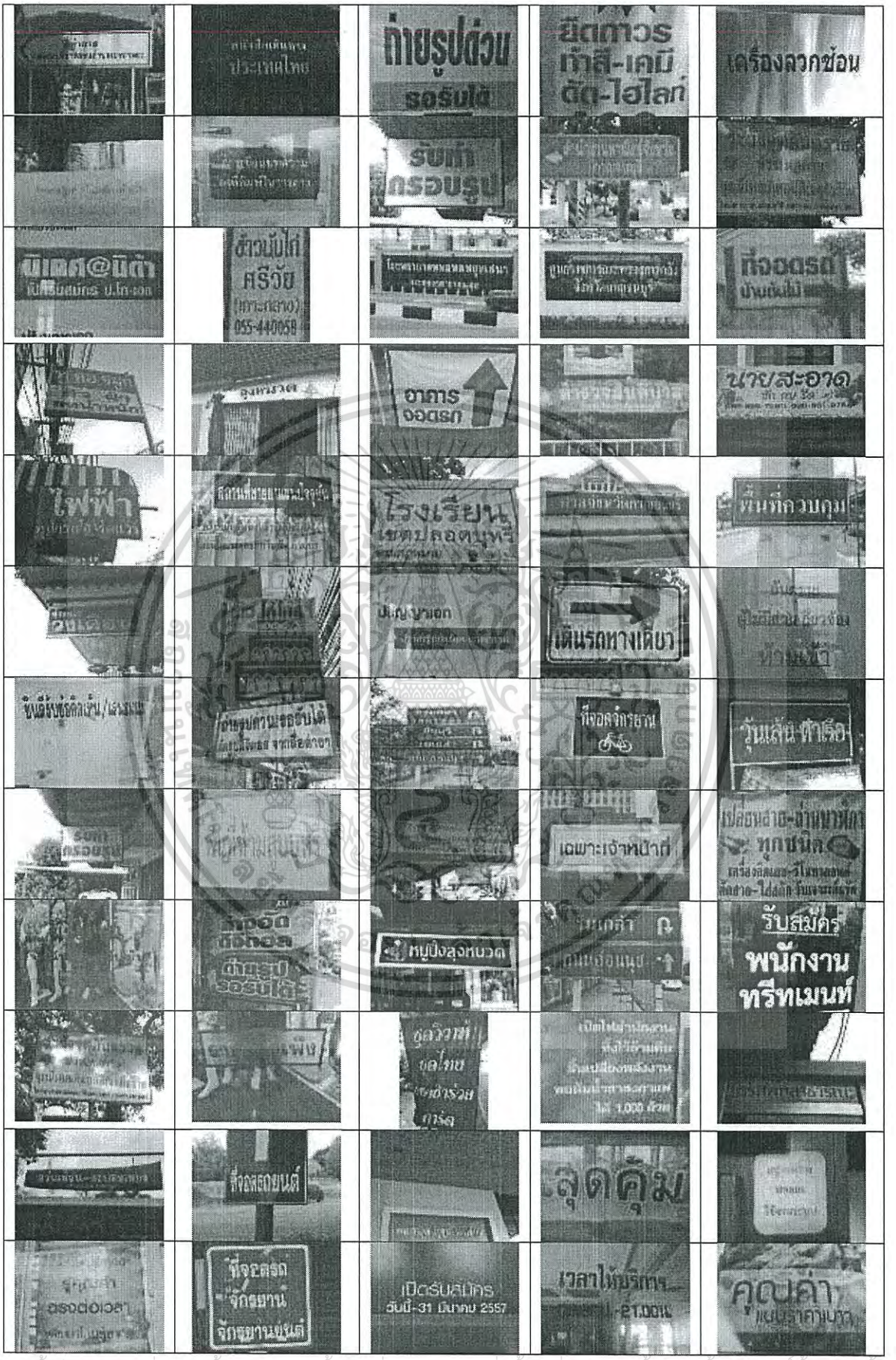
[18] Gonzalez, R. C. and Woods, R. E. 2002. **Digital Image Processing (2nd Edition)**. New Jersey : Prentice Hall.

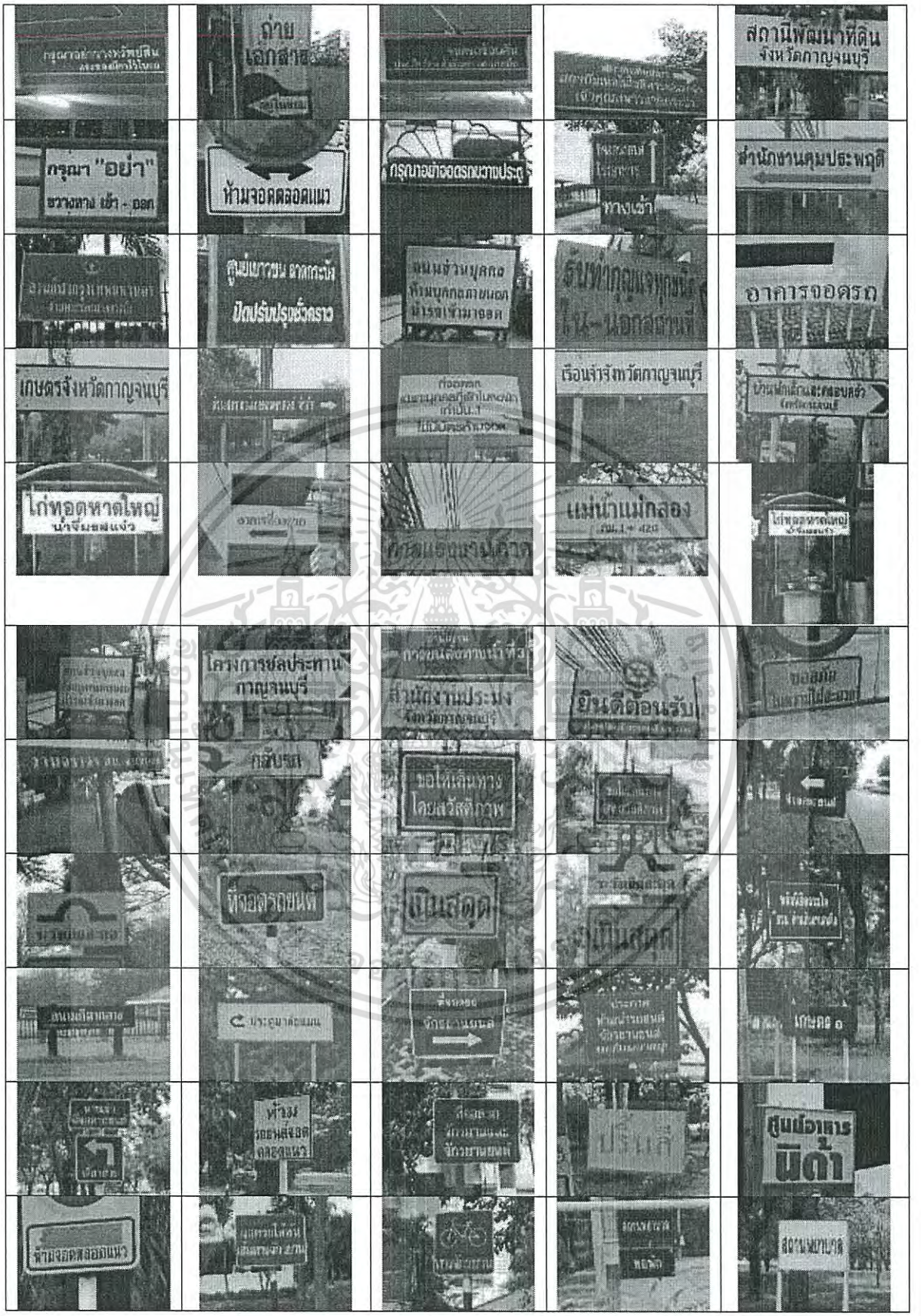
[19] regionprops function. **Matlab2016a Documentation** [Online]. http://www.mathworks.com/help/images/ref/regionprops.html

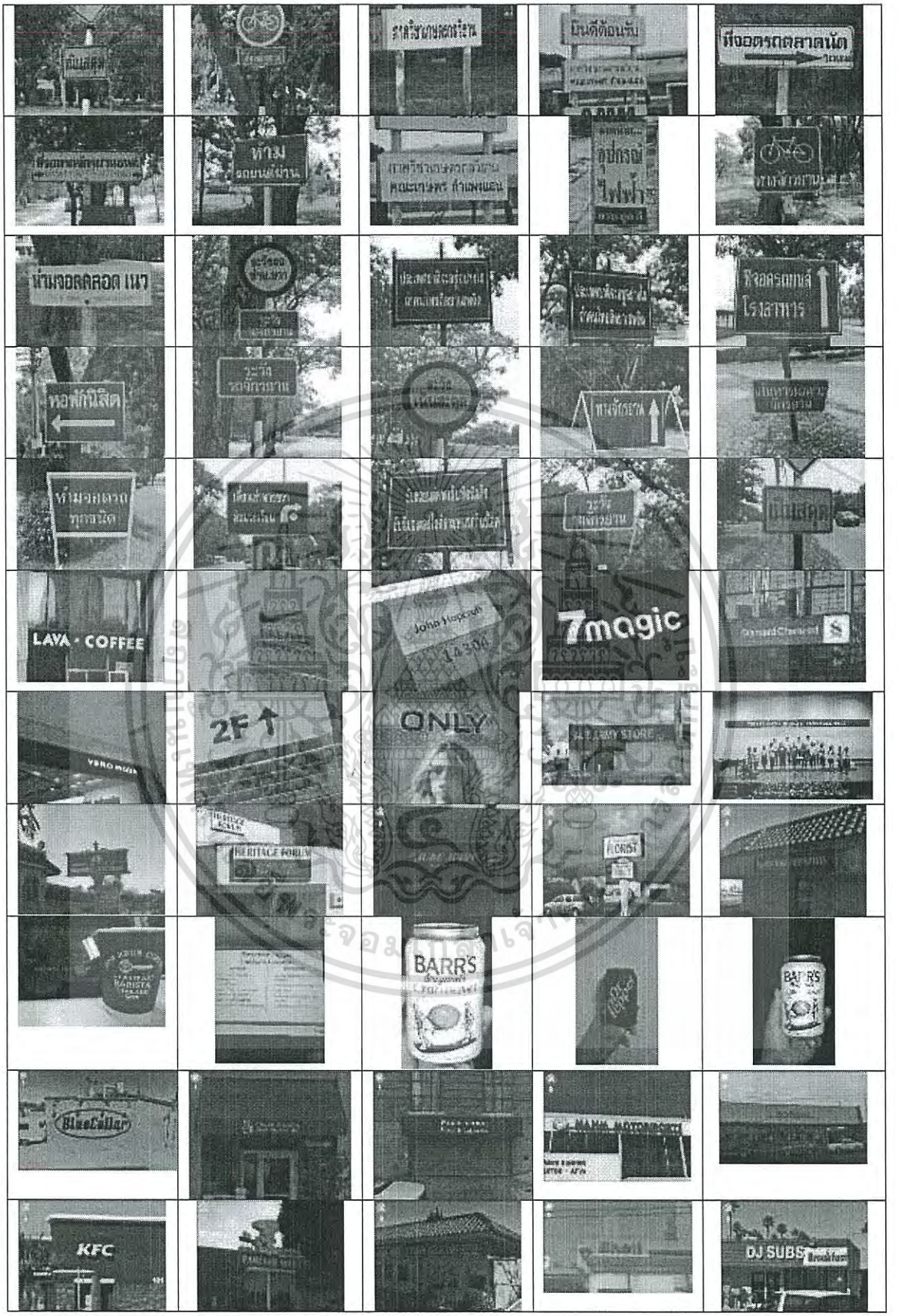
[20] หลักเกณฑ์ BEST 2010. **Human Language Technology Laboratory,National Electronics and Computer Technology Center** [Online].http://thailang.nectec.or.th/best/?q=node/13

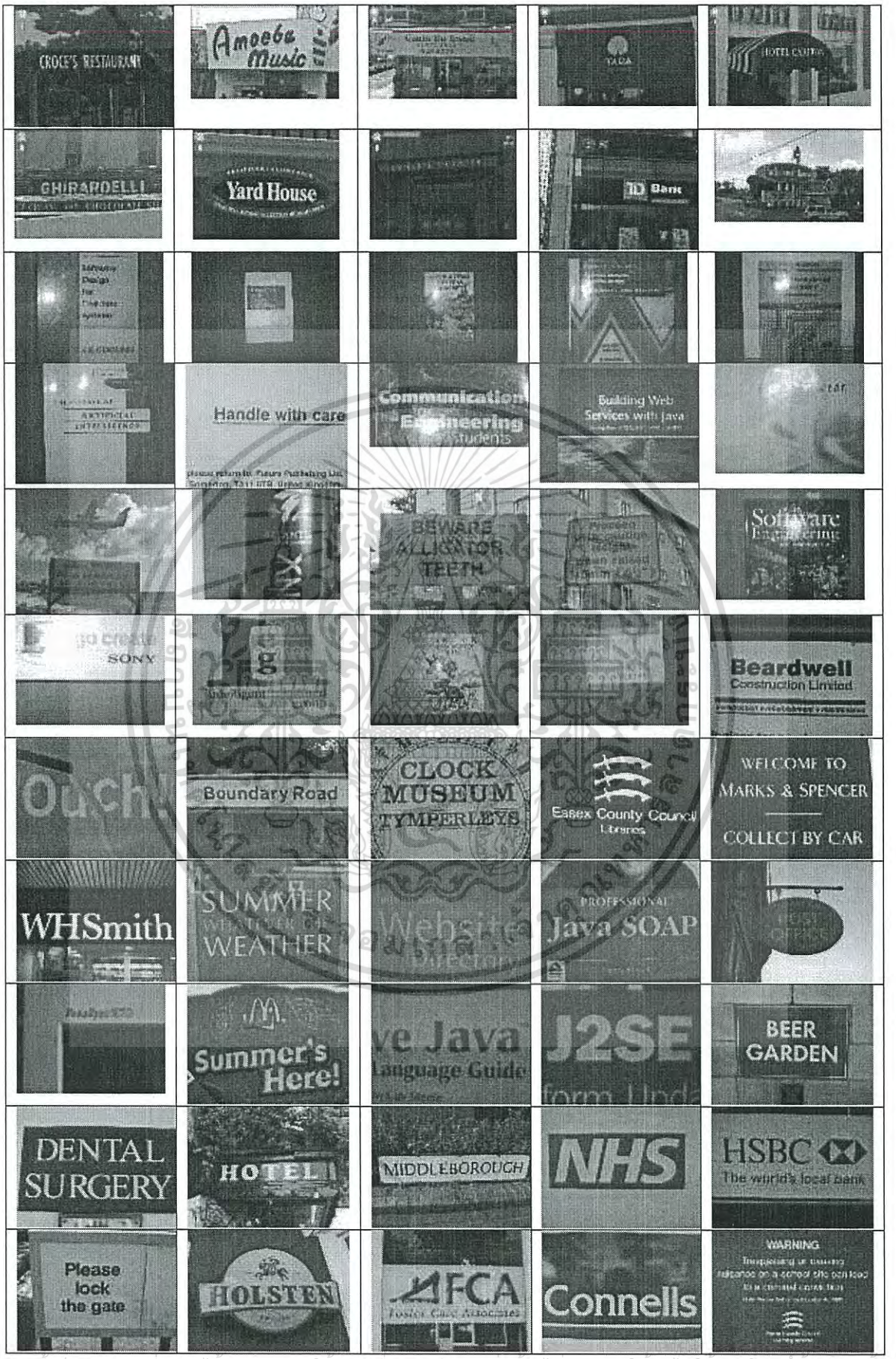
ภาคผนวก

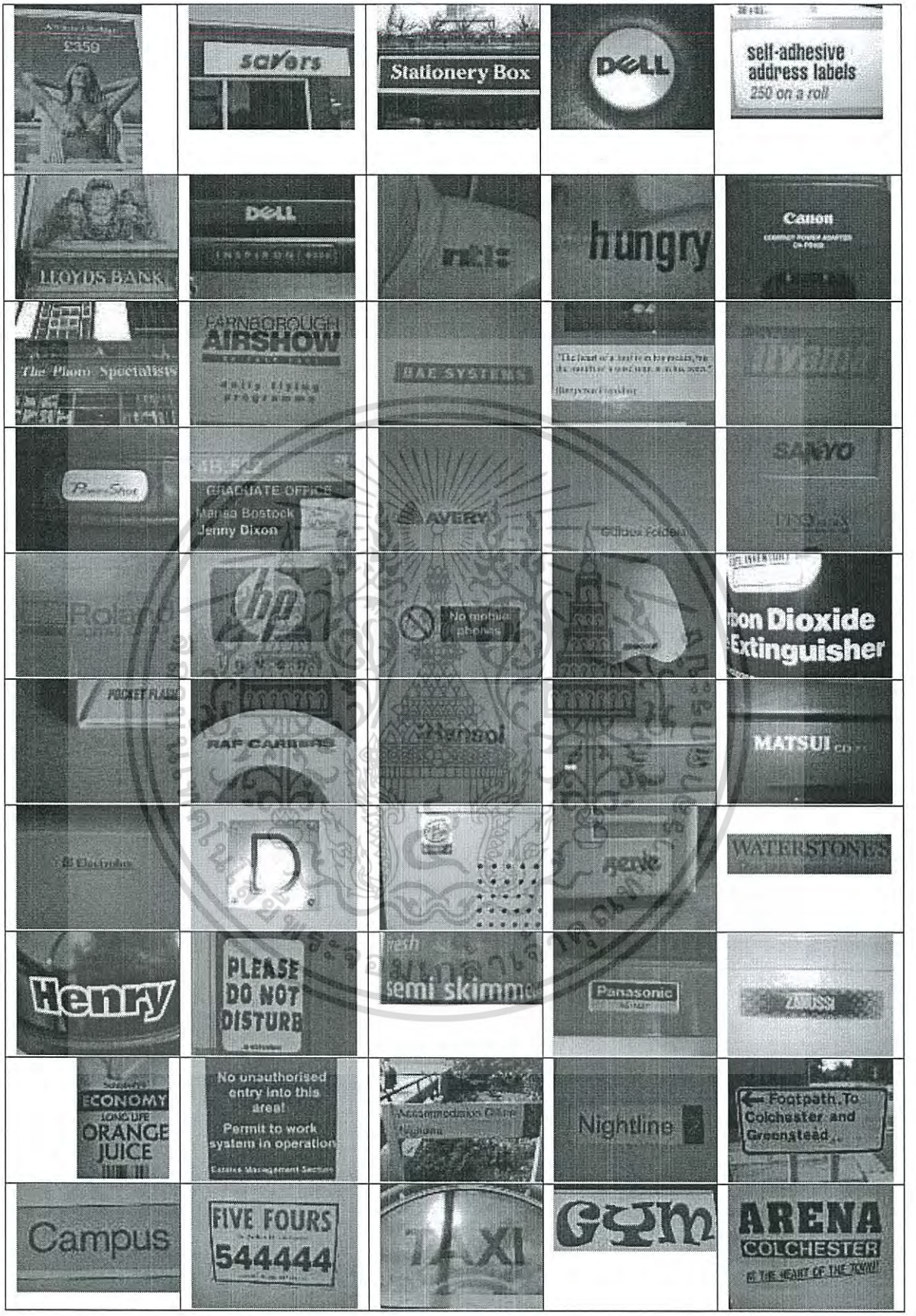
ภาคผนวก ก
ภาพที่ใช้ในการทดลอง



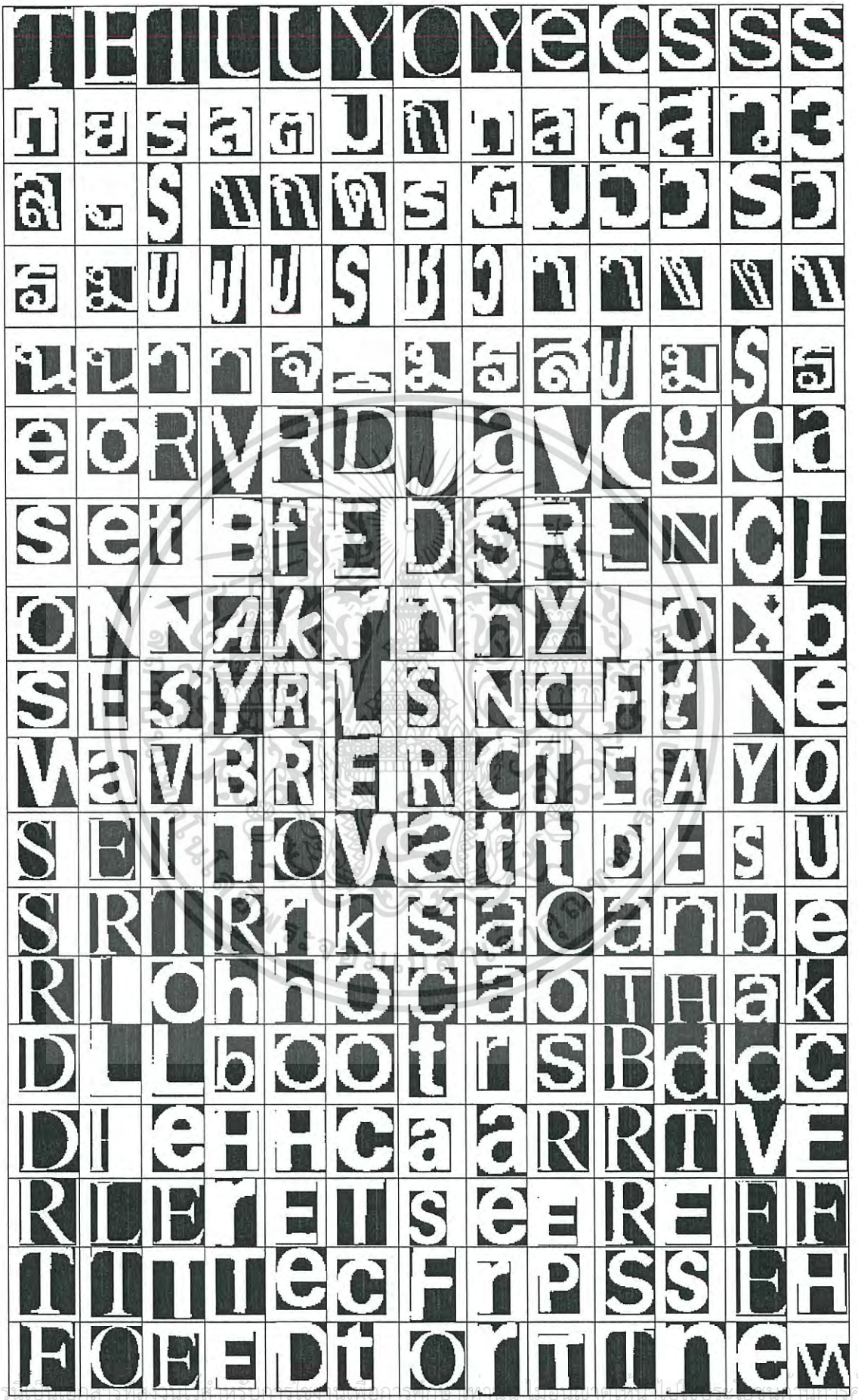








ภาคผนวก ข
ภาพตัวอย่างที่ใช้ในการสอนแบบจำลอง



ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวชญานิศ ตันธีระพงศ์
วัน-เดือน-ปีเกิด	11 กุมภาพันธ์ 2537
สถานศึกษา	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
สถานที่ติดต่อ	99/52 ม.ศิริชัย ซอยรามอินทรา 14 แยก 22 ถนนรามอินทรา ท่าแร้ง บางเขน กทม. 1023

การตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดโดยใช้ Maximally Stable Extremal Regions และ Support Vector Machine

ชญานิศ ตันธีระพงศ์[1]

[1]*คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ*

Emails: Chayais.t@outlook.com

บทคัดย่อ

การตรวจจับข้อความจากฉากธรรมชาติ มีหลายปัจจัยที่มีผลกระทบต่อการตรวจจับและรู้จำ เช่น มุมมองของภาพและความหลากหลายของตัวอักษร เป็นต้น จากปัจจัยที่กล่าวมาข้างต้นนั้นมีความท้าทายเป็นอย่างยิ่งสำหรับการตรวจจับข้อความเพื่อแปลงเป็นเสียงพูด นอกจากนี้การตรวจจับข้อความในภาษาไทยนั้นยังไม่มีวิธีที่หลากหลาย เพราะว่าลักษณะโครงสร้างของภาษาไทยมีความแตกต่างจากภาษาอื่น ๆ โดยรายงานนี้จะนำเสนอการพัฒนาระบบตรวจจับข้อความเพื่อแปลงเป็นเสียงพูด โดยมี 4 ขั้นตอนหลัก คือ 1).แบ่งข้อมูลของภาพแยกออกจากกันตามช่องสี 2).ตรวจจับคุณลักษณะ 3).แยกกลุ่มของตัวอักษรและกลุ่มที่ไม่ใช่ตัวอักษรโดยใช้หลักการเรียนรู้ของเครื่อง 4).การรวมกลุ่มของข้อความ จากการทดลองพบว่าวิธีการที่พัฒนาสามารถตรวจจับข้อความเพื่อแปลงเป็นเสียงพูดได้ โดยมีความแม่นยำและครบถ้วนของภาษาไทยอยู่ที่ 0.63 และ 0.67 และค่าแม่นยำและค่าครบถ้วนสำหรับภาษาอังกฤษอยู่ที่ 0.58 และ 0.69 ตามลำดับ

คำสำคัญ – ภาษาไทย,การตรวจจับตัวอักษร,MSER,SVM

1. บทนำ

ในปัจจุบันการใช้ชีวิตประจำวันเรามักพบกับข้อมูลต่างๆรอบตัว ที่มีที่มาหลากหลาย เช่น ป้ายบอกทาง ป้ายโฆษณา โปสเตอร์ เป็นต้น สิ่งเหล่านี้เป็นข้อมูลและสารสนเทศที่จำเป็นสำหรับการเรียนรู้และใช้ในชีวิตประจำวัน ดังนั้น การพัฒนาระบบตรวจจับและรู้จำข้อความเพื่อแปลงเป็นเสียงพูด จึงเป็นขั้นตอนที่สำคัญที่ใช้ผลลัพธ์ที่ได้มาประยุกต์ใช้งานในด้านต่างๆ เช่น เครื่องมือสำหรับช่วยเหลือผู้พิการทางสายตา เป็นต้น

ความท้าทายของการพัฒนาระบบตรวจจับและรู้จำข้อความเพื่อแปลงเป็นเสียงพูดนั้นมีหลากหลายประการ อาทิเช่น การเรียงตัวของข้อความ ความหลากหลายของรูปแบบตัวอักษร,มุมมองของภาพ และความคมชัดของภาพ สิ่งเหล่านี้ล้วนเป็นปัจจัยที่ส่งผลให้การตรวจจับและรู้จำข้อความที่มาจากภาพถ่ายนั้นมีความผิดพลาดสูง ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำ จึงมีการควบคุมปัจจัยต่างๆเหล่านี้ เพื่อรักษาคุณภาพของข้อความ ให้สามารถตรวจจับและรู้จำได้มากที่สุด

ในการตรวจจับข้อความจากภาพถ่ายนั้นมีหลากหลายวิธี ซึ่งแต่ละวิธีมีข้อดีและข้อเสียแตกต่างกันไป สิ่งสำคัญคือการสร้างกระบวนการที่มีประสิทธิภาพที่ให้ผลลัพธ์ที่มีความแม่นยำได้มากที่สุด โดยในรายงานนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้อง วิธีการ ผลการทดลองและสรุปผลการทดลอง ซึ่งจะกล่าวในหัวข้อต่อไป

2. งานวิจัยที่เกี่ยวข้อง

Weilin H., Yu Q. และ Xiaoon T. นำเสนอวิธี Convolutional neural network Induced MSER Trees ซึ่งใช้วิธีการตรวจจับโดยพิจารณาพื้นที่ (Region-based) โดยใช้ MSERs (Maximally Stable Extrenal Regions) และใช้คอนโวลูชันนอล นิวรอลเน็ตเวิร์ค (Convolutional neural network)[1] สร้างแบบจำลอง

เพื่อคัดแยกตัวอักษรและสิ่งที่ไม่ใช่ตัวอักษรออกจากกัน ซึ่งข้อดีคือมีความแม่นยำ เนื่องจาก การทำงานควบคู่กันของ MSERs และ Convolutional neural network นั้นเสริมประสิทธิภาพซึ่งกันและกัน แต่ก็มีข้อเสียกับรูปภาพบางกรณี เช่น ภาพที่มีความคมชัดต่ำ หรือภาพที่มีตัวอักษรและพื้นหลังที่ซับซ้อน เช่น ตัวอักษรที่อยู่บนกำแพงอิฐ ตัวอักษรที่มีสีเดียวกับพื้นหลังหรือมีสีที่ใกล้เคียงกับพื้นหลังและภาพตัวอักษรที่มองเห็นไม่ชัด จะไม่สามารถตรวจจับได้โดยวิธีการนี้ เพราะว่าการตรวจจับของ MSER นั้นบางครั้งมักจะให้ส่วนประกอบของภาพพื้นหลังติดมาด้วย และลักษณะภาพของอิฐนั้นส่วนมากจะมีค่าสีที่ไม่ห่างกันมาก จึงทำให้ MSER สามารถตรวจจับได้ นอกจากนี้ การเสริมสร้างประสิทธิภาพประมวลผลของงานวิจัยนี้ให้รวดเร็วได้นั้นควรการใช้ฮาร์ดแวร์ที่มีคุณสมบัติสูง และมีความซับซ้อนในการเตรียมสร้างแบบโครงข่ายประสาทเทียมซึ่งใช้เวลานาน

Datong C.,Herve B.,Jean-P. T [2] ได้นำเสนอวิธียืนยันพื้นที่ตัวอักษรบนรูปภาพโดยใช้การตรวจสอบคุณสมบัติของกลุ่มตัวอักษร โดยวิเคราะห์จากการวางตัวอักษรในช่วงแนวตั้งและแนวนอน และหลังจากนั้นนำพื้นที่ที่มีแนวโน้มว่าจะเป็นกลุ่มของตัวอักษรนั้นมาสกัดคุณลักษณะภายในกลุ่ม และใช้หลักการ SVM(Support Vector Machine) ในการตรวจสอบว่ากลุ่มของพื้นที่นั้นใช้กลุ่มของตัวอักษรหรือไม่ โดยผลลัพธ์สุดท้ายที่ได้คือพื้นที่ตัวอักษรที่ได้รับการยืนยันเรียบร้อยแล้ว โดยวิธีการนี้มีข้อดีคือ ประมวลผลไว และแม่นยำ แต่มีข้อเสียคือสามารถรองรับได้แค่บางภาษาเท่านั้น

3. วิธีการ

ภายในระบบนั้นมีขั้นตอนการทำงานหลัก 3 ขั้นตอน แต่ในระบบนั้นจะเน้นไปที่ขั้นตอนหลักที่ 1 ซึ่งการทำงานของระบบนั้นแสดงดังรูปที่ 1

รูปที่ 1. แสดงกระบวนการทำงานของระบบตรวจจับและรู้จำข้อความเพื่อแปลงเป็นเสียงพูด

โดยในระบบนั้น ผู้จัดทำได้มุ่งเน้นไปที่การพัฒนาขั้นตอนหลักที่ 1 ซึ่งสามารถแสดงการทำงาน ได้ดังรูปที่ 2

รูปที่ 2. แสดงการทำงานของขั้นตอนการตรวจจับข้อความ

โดยขั้นตอนของการทำงานของระบบตรวจจับจะเริ่มที่การนำภาพสีระบบ RGB นั้นเข้ามาแยกค่าสีในภาพออกจากกัน ก่อนจะทำการตรวจจับคุณลักษณะของภาพทั้งหมดโดยใช้หลักการ MSER โดยที่หลักการ MSER นั้นเน้นที่การตรวจจับคุณลักษณะจากความคงทนต่อการเปลี่ยนแปลงของค่าความเข้มของแสง โดยที่ระบบนี้เลือกใช้ระยะห่างในการคำนวณระหว่าง การคงทนต่อความเข้มของแสงอยู่ที่ 4 หลังจากนั้นจึงนำคุณลักษณะที่ถูกตรวจจับได้มาตรวจสอบผ่านการคัดแยกพื้นที่ตัวอักษรและไม่ใช่ตัวอักษรโดยใช้การสกัดคุณลักษณะตามตารางที่ 1

ตารางที่ 1. คุณสมบัติของพื้นที่ที่ใช้ในการสร้างสมการและแยกกลุ่ม

คุณสมบัติ	สมการ/คุณสมบัติ
Ratio	Width/Height
Eccentricity	The eccentricity is the ratio of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1.
Euler number	objects in the region - the number of holes in objects.

Extent	Area/bounding box
Solidity	area/convex area
Perimeter	distance between each adjoining pair of pixels around the border of the region

จากการใช้ชุดข้อมูลทั้งหมด 40000 ชุด โดยแบ่งเป็นชุดตัวอักษร 21,000 พื้นที่ และไม่ใช่ตัวอักษร 19000 พื้นที่และใช้เคอร์แนลแบบการ์เซียน

เมื่อทำการสอนข้อมูลทั้งหมดแล้ว ระบบจะสร้างแบบจำลองพร้อมทำนาย โดยการทำนายนั้นจะทำนายจากพื้นที่ MSER ที่ถูกตรวจจับได้ในขั้นตอนที่ 3.1.2 แล้วเก็บเฉพาะพื้นที่ที่ถูกทำนายได้ว่าเป็นตัวอักษร ซึ่งสามารถแสดงได้ ดังภาพที่ 3

ภาพที่ 3. แสดงผลลัพธ์ของการรวมกลุ่มของพื้นที่ที่ถูกทำนายว่าเป็นอักษร

หลังจากนั้นจึงนำพื้นที่ที่ถูกทำนายนั้นมารวมตัวกันเป็นกลุ่มข้อความ และนำกลุ่มข้อความทีละกลุ่มนั้นเข้าสู่ระบบรู้จำข้อความต่อไป

4. ผลการทดลอง

เนื่องจากการทดลองนั้น ในการทดลองแต่ละครั้งจะแบ่งผลการทดลองความแม่นยำของภาษาไทยและภาษาอังกฤษแยกออกจากกัน ในการทดลองนั้น ทางผู้จัดทำใช้ภาพในการทดลอง ทั้งหมด 400 ภาพ โดยแบ่งเป็นภาพภาษาไทยทั้งหมด 200 ภาพ และ ภาพภาษาอังกฤษทั้งหมด 200 โดยแต่ละภาพจะถูกแบ่งกลุ่มตามความซับซ้อนของพื้นหลัง และ การวางรูปแบบข้อความที่หลากหลาย จากผลการทดลองค้นพบว่าวิธีการนี้สามารถให้ค่าความแม่นยำและครบถ้วนของภาษาไทยอยู่ที่ 0.63 และ 0.67 และค่าแม่นยำและค่าครบถ้วนสำหรับภาษาอังกฤษอยู่ที่ 0.58 และ 0.69 ตามลำดับ

แต่เมื่อเปรียบเทียบกับการใช้กฎในการกรองพื้นที่ตัวอักษรและวิธีการเดียวกันแต่ไม่มีการแยกข้อมูลตาช่องสี ค้นพบว่าการใช้วิธีเดียวกันแต่ไม่แยกช่องสีกลับให้ผลลัพธ์ที่ดีกว่าในกรณีที่ภาพมีความซับซ้อนของพื้นหลัง และการใช้กฎจะให้ผลลัพธ์ที่ดีกว่าในกรณีที่ข้อมูลที่ใช้ทดสอบนั้น ไม่เคยถูกมองเห็นมาก่อน (Unseen Data)

5. สรุปผล

จากการทดลองค้นพบว่าการใช้วิธีนี้ สามารถทำได้ในกรณีที่จำกัด แต่ว่าถ้าหากมีการปรับปรุงกระบวนการเรียนรู้ของเครื่องให้มีประสิทธิภาพในการแยกกลุ่มให้ได้มากกว่านี้ อาจจะทำให้ ค่าครบถ้วนและค่าความแม่นยำนั้นมีค่าสูงขึ้นได้ เพราะว่าจากการเปรียบเทียบการให้คะแนนความมั่นใจของระบบการเรียนรู้ของเครื่อง จะค้นพบว่ามีการให้คะแนนระหว่างสองวัตถุที่ใกล้เคียงกัน สำหรับบางกรณีการเรียนรู้ของเครื่องไม่สามารถทำนายข้อมูลได้ถูกต้องเลย ถึงแม้จะปรับการตรวจจับของ MSER ให้ไม่มีความคงทนต่อการเปลี่ยนแปลงของค่าความเข้มของแสงและไม่ได้ปรับช่วงของพื้นที่ให้มีความจำกัด ซึ่งจากผลการทดลองนี้ จึงทำให้ผู้จัดทำเล็งเห็นว่าควรจะแก้ไขที่การเรียนรู้ของเครื่อง เพื่อเพิ่มประสิทธิภาพในการตรวจจับข้อความ

เอกสารอ้างอิง

[1] Weilin H., Yu Q., Xiaoon T. 2014.Robust Scene Text Detection with ConvolutionalNeural Network Induced MSER Tree. ECCV 2014 Part IV LNCS 8692. pp 497-511

[2] PIMLAK BOONCHUKUSOL. 2014 .Thai Text Detection From Medium Shot Of Natural Scenes By Using Fast Boundary Clustering And Modified Connected Component Analysis.วิทยานิพนธ์ สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

[3] Smith, R. 2007. An overview of the Tesseract OCR engine. In icdar (pp. 629-633). IEEE.

[4] Microsoft. Microsoft Speech API Overview (SAPI 5.4). [Online]https://msdn.microsoft.com/en-us/library/ee125077(v=vs.85).asp