



Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคสนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562



Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาชีวกรรมทุนยนต์และระบบอัตโนมัติ

สถาบันวิทยาการทุนยนต์ภาครสนา

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

Google แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุ์งาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรณาการ

วิทยานิพนธ์เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาชีวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคสนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

คณะกรรมการสอบวิทยานิพนธ์

ประธานกรรมการสอบวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

()

กรรมการสอบวิทยานิพนธ์

(อ.บวรศักดิ์ ศกุลเกื้อภูลสุข)

กรรมการสอบวิทยานิพนธ์

(ดร.บุญทริกา เกษมสันติธรรม)

ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ชื่อวิทยานิพนธ์	Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบบุคลากรกระทำของมนุษย์
หน่วยกิต	6
ผู้เขียน	นายปัจมพงศ์ สินธุจัน นายศุภกร เบญจวิกรัย นายอุตุษฐ์ เลิศวรรณการ
อาจารย์ที่ปรึกษา	ทีปรึกษาวิทยานิพนธ์หลัก ดร.วราสินี ฉายแสงมงคล
หลักสูตร	วิศวกรรมศาสตรบัณฑิต
สาขาวิชา	วิศวกรรมหุ่นยนต์และระบบอัตโนมัติ
คณะ	สถาบันวิทยาการหุ่นยนต์ภาคสนาม
ปีการศึกษา	2562

บทคัดย่อ

งานวิทยานิพนธ์นี้เป็นงานที่เกี่ยวกับการออกแบบและสร้างเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ รวมถึงระบบบุคลากรกระทำของมนุษย์ โดยใช้ชื่อว่า Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบบุคลากรกระทำของมนุษย์ ซึ่งมีจุดประสงค์เพื่อให้ผู้พัฒนาสามารถใช้งานเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ในการสร้างชุดข้อมูลสำหรับสร้างปัญญาประดิษฐ์ได้ง่ายและสะดวกขึ้น ภาพรวมของวิทยานิพนธ์นี้จะแบ่งออกเป็นหัวข้อหลักสองส่วน โดยที่ส่วนแรกเป็นการออกแบบและสร้างแอพพลิเคชันที่ใช้ในการสร้างชุดข้อมูลจากวิดีโอ สำหรับพัฒนาโมเดลปัญญาประดิษฐ์ และส่วนที่สองเป็นการออกแบบและสร้างระบบบุคลากรกระทำของมนุษย์ภายใต้ในสำนักงาน

คำสำคัญ : ระบบบุคลากรกระทำของมนุษย์ / เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ / Goggle

กิตติกรรมประกาศ

ขอขอบพระคุณ ดร.วราสินี ฉายแสงมงคล อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก ที่ได้สละเวลามาให้คำปรึกษา ชี้แนะแนวทาง ให้ความรู้ในด้านต่างๆ ที่จำเป็นต่องานวิจัย รวมถึงการให้การสนับสนุนในเรื่องอุปกรณ์ในการทำวิจัย ช่วยตรวจสอบและแก้ไขวิทยานิพนธ์ให้เป็นไปอย่างสมบูรณ์ ตลอดจนกรุณาให้เกียรติเป็นประธานกรรมการสอบวิทยานิพนธ์

ขอขอบพระคุณอาจารย์อาจารย์ บวรศักดิ์ สกุลเกื้อกูลสุข ที่กรุณาให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ ให้คำแนะนำที่เป็นประโยชน์ต่อการวิจัย และการแก้ไขปรับปรุงงานวิจัย ตลอดจนตรวจสอบแก้วิทยานิพนธ์ให้ดำเนินไปอย่างสมบูรณ์

ขอขอบพระคุณอาจารย์ ดร.บุญทริกา เกษมสันติธรรม ที่กรุณาให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ ให้คำแนะนำที่เป็นประโยชน์ต่อการวิจัย และการแก้ไขปรับปรุงงานวิจัย ตลอดจนตรวจสอบแก้วิทยานิพนธ์ให้ดำเนินไปอย่างสมบูรณ์

ขอขอบพระคุณคณาจารย์ และบุคลากรในสถาบันวิทยาการหุ่นยนต์ภาคนามทุกท่าน ที่ได้ให้คำปรึกษา และช่วยเหลือด้านสถานที่พร้อมทั้งส่งอำนวยความสะดวกต่างๆ ในระหว่างการทำวิทยานิพนธ์

ขอขอบคุณนักศึกษาปริญญาตรี สถาบันวิทยาการหุ่นยนต์ภาคนามทุกท่าน ที่ได้ให้คำแนะนำ ถามไถ่ และเป็นกำลังใจมาโดยตลอด

และสุดท้ายนี้ ขอน้อมรำลึกถึงพระคุณบิดา มารดา และครอบครัว ที่ส่งเสริมให้กำลังใจ และให้การสนับสนุนในเรื่องต่างๆ จนกระทั้งข้าพเจ้าประสบความสำเร็จในการศึกษา

นายปฐมพงศ์ สินธุรงาม
นายศุภกร เบญจวิกรัย
นายอุกฤษฎ์ เลิศวรรณาการ

สารบัญ

เรื่อง	หน้า
บทคัดย่อ	ค
กิตติกรรมประกาศ	๔
สารบัญ	๕
รายการรูปภาพ	๗
รายการตาราง.....	๘
รายการสัญลักษณ์.....	๙
ประมวลศัพท์และตัวย่อ.....	๙
บทที่ 1 บทนำ.....	๑
1.1 ที่มาและความสำคัญ.....	1
1.2 วัตถุประสงค์.....	1
1.3 ประโยชน์ที่คาดว่าจะได้รับ	1
1.4 ขอบเขตการดำเนินงาน.....	2
1.5 ขั้นตอนการดำเนินงาน	2
บทที่ 2 ทฤษฎี/การวิจัยที่เกี่ยวข้อง.....	4
2.1 การวิเคราะห์ผลวิดีโอ.....	4
2.1.1 การตรวจจับวัตถุ.....	5
2.1.2 ระบบท่านายตำแหน่งถัดไปของวัตถุ.....	7
2.1.3 ระบบระบุตัวตนของบุคคล	8
2.1.4 ระบบจำแนกการกระทำ	9
2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ	19
2.2.1 โมเดลปัญญาประดิษฐ์	19
2.2.2 เครื่องมือกำกับคุณลักษณะ	29
2.3 ทฤษฎีที่เกี่ยวข้อง	31
2.3.1 Optical flow.....	31
บทที่ 3 ระเบียบวิธีวิจัย	33
3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	34

สารบัญ (ต่อ)

เรื่อง	หน้า
3.2 ความต้องการของระบบ.....	34
3.2.1 ความต้องการเชิงการใช้งาน (functional requirements).....	34
3.2.2 ความต้องการเชิงวิศวกรรม (non-functional requirements).....	35
3.3 หน้าที่ความรับผิดชอบ.....	36
3.4 เครื่องมือที่ใช้ในงานวิจัย	36
3.5 ภาษาที่ใช้ในการพัฒนาระบบ	37
3.6 Program library ที่ใช้ในการพัฒนาระบบและแอพพลิเคชัน.....	37
3.7 แผนการดำเนินงาน	37
3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	38
3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	38
3.9 การออกแบบการทดสอบการตรวจจับวัตถุ.....	48
3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำงาน.....	48
3.10 การออกแบบการทดสอบการทำงานตามมาตรฐานต่อไปของมนุษย์.....	49
3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำงานตามมาตรฐานต่อไปของวัตถุในวิดีโอ	49
3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์.....	50
3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ	50
3.12 การออกแบบการทดสอบการจัดการกระทำของมนุษย์	51
3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกแทนที่ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในกราฟทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง.....	51
3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง	52
3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง	53
บทที่ 4 ผลการดำเนินงาน	54
4.1 เครื่องมือกำกับคุณลักษณะ	54
4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน	54
4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอพพลิเคชัน	58

สารบัญ (ต่อ)

เรื่อง	หน้า
4.2 ผลการทดลองการตรวจจับวัตถุ	61
4.2.1 ข้อมูลรายละเอียดประกอบการทดสอบ.....	61
4.2.2 ผลทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล	61
4.3 ผลการทดสอบการติดตามทำนายตำแหน่งของมนุษย์.....	62
4.3.1 ข้อมูลรายละเอียดประกอบการทดสอบ.....	62
4.3.2 ผลทดสอบประสิทธิภาพ และความเร็วในการประมวลผล.....	62
4.4 ผลการทดสอบระบบระบุตัวตนของมนุษย์	63
4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของบุคคล	63
4.5 ผลการทดสอบการจัดลำดับการทำของมนุษย์.....	66
4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้.....	66
4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง.....	66
4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่านชุด ข้อมูลสำหรับการเทรน์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบ และเทียบผลลัพธ์การทดสอบก่อนหน้า	66
เอกสารอ้างอิง.....	66
ภาคผนวก ก ตัวอย่างชุดข้อมูลที่ผู้วิจัยสร้างขึ้น	68

รายการรูปภาพ

รูป	หน้า
รูปที่ 2.1 กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO	6
รูปที่ 2.2 แนวคิดของระบบท่านายทำแห่งตัวไปของวัตถุ	7
รูปที่ 2.3 การแบ่งรูปออกเป็น 8 ส่วนของระบบระบุตัวตนของบุคคล	9
รูปที่ 2.4 ตัวอย่างโครงสร้างของ CNN	19
รูปที่ 2.5 ตัวอย่างเครอร์เนล	19
รูปที่ 2.6 ตัวอย่างการหาผังคุณลักษณะ	20
รูปที่ 2.7 ตัวอย่างการทำ max pooling	20
รูปที่ 2.8 ตัวอย่างการทำ average pooling	20
รูปที่ 2.9 โครงสร้างของ fully-connected	21
รูปที่ 2.10 ตัวอย่างการเคลื่อนที่ของลูกบอล	21
รูปที่ 2.11 หลักการของ Residual block ของ ResNet	22
รูปที่ 2.12 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[?]	24
รูปที่ 2.13 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[?]	25
รูปที่ 2.14 กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO	26
รูปที่ 2.15 โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ YOLO	26
รูปที่ 2.16 โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO	27
รูปที่ 2.17 โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ Faster RCNN	28
รูปที่ 2.18 UI ของโปรแกรม DarkLabel	29
รูปที่ 2.19 UI ของโปรแกรม OpenLabeling	30
รูปที่ 2.20 ตัวอย่างการเคลื่อนที่ของลูกบอล	31
รูปที่ 3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	34
รูปที่ 3.2 กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	38
รูปที่ 3.3 หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	39
รูปที่ 3.4 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select	40
รูปที่ 3.5 หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	41
รูปที่ 3.6 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect	42
รูปที่ 3.7 หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	43

รายการรูปภาพ (ต่อ)

รูป	หน้า
รูปที่ 3.8 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track.....	44
รูปที่ 3.9 หน้าต่าง Label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	45
รูปที่ 3.10 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Label.....	46
รูปที่ 3.11 ตัวอย่างข้อมูลภายในไฟล์ xml	47
รูปที่ 4.1 รูปหน้าต่างแสดงผลของหน้าต่าง Select	54
รูปที่ 4.2 รูปหน้าต่างแสดงผลของหน้าต่าง Detect	55
รูปที่ 4.3 รูปหน้าต่างแสดงผลของหน้าต่าง Track.....	56
รูปที่ 4.4 รูปหน้าต่างแสดงผลของหน้าต่าง Label	57
รูปที่ 4.5 รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่และไม่มีมนุษย์อยู่ภายในเฟรม	58
รูปที่ 4.6 รูปคิริ์เฟรมที่ถูกตีกรอบสีเหลี่ยมในส่วนที่มีมนุษย์อยู่	58
รูปที่ 4.7 รูปผลลัพธ์การทำงานของหน้าต่าง Track.....	59
รูปที่ 4.8 รูปผลลัพธ์การทำงานของหน้าต่าง Label	60
รูปที่ 4.9 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 1	63
รูปที่ 4.10 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 2.....	64
รูปที่ 4.11 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 3.....	64
รูปที่ 4.12 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 4.....	65
รูปที่ 4.13 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 5.....	65
รูปที่ 4.14 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 6.....	66
รูปที่ ก.1 รูปผลลัพธ์การทำงานของหน้าต่าง Track.....	68

รายการตาราง

ตาราง	หน้า
ตารางที่ 1.1 แผนการดำเนินงาน	3
ตารางที่ 2.1 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลสำหรับทดสอบของ YouTube-8M	11
ตารางที่ 2.2 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลสำหรับทดสอบของ Sports-1M	12
ตารางที่ 2.3 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลสำหรับทดสอบของ ActivityNet	12
ตารางที่ 2.4 ผลการทดลองของวิธีต่างๆ บนคุณลักษณะระดับเฟรม	15
ตารางที่ 2.5 ประสิทธิภาพของโมเดล Resnet50 3D ที่ใช้ชุดข้อมูล Kinetics และ Moments in Time ..	17
ตารางที่ 2.6 อัตราเร้อยล์ของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet	22
ตารางที่ 2.7 ค่าความผิดพลาดที่ได้จากการทดลองจำนวนขั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10	23
ตารางที่ 2.8 ประสิทธิภาพของโมเดล 3D แบบ two-stream เมื่อใช้ข้อมูลจาก UCF-101, HMDB-51 และ Kinetics ในการสร้างและทดสอบด้วยเครื่องมือวัดผลแบบความแม่นยำจากการทำนายอันดับแรกสุด	25
ตารางที่ 4.1 ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการทำตรวจสอบจับภาพบุคคล	61
ตารางที่ 4.2 ผลการทดสอบประสิทธิภาพของการตรวจสอบสีเหลี่ยมภายในวิดีโอ	62
ตารางที่ 4.3 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์	63
ตารางที่ 4.4 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 1	63
ตารางที่ 4.5 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 2	64
ตารางที่ 4.6 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 3	64
ตารางที่ 4.7 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 4	65
ตารางที่ 4.8 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 5	65
ตารางที่ 4.9 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 6	66
ตารางที่ 4.10 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์เทียบผลลัพธ์กับแหล่งอ้างอิง	66
ตารางที่ 4.11 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ เมื่อใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น	67
ตารางที่ 4.12 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น ใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น	67

รายการสัญลักษณ์

θ	เชิงตัว
d	distance
kg	Kilogram
m^2	Square Metre

ประมวลศัพท์และตัวย่อ

AVA	Atomic Visual Actions
Artificial intelligence	ปัญญาประดิษฐ์
Machine learning model	โมเดลปัญญาประดิษฐ์
Label	คำกำกับที่บ่งบอกถึงคุณลักษณะของสิ่งที่สนใจ
Labeling	การสร้างคำกำกับคุณลักษณะ
Human action classification	การจำแนกการกระทำของมนุษย์
Video labeling	การสร้างคำกำกับคุณลักษณะภายในวิดีโอ
Video analytics	การวิเคราะห์ผลวิดีโอ
Uniform label distribution	การที่มีจำนวนตัวอย่างภายใต้คำกำกับเท่ากันทุกประเภท
KMUTT	King Mongkut's University of Technology Thonburi

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

บริษัท เพอเช็ปตรา ดำเนินธุรกิจเกี่ยวกับการให้บริการและคำปรึกษาเกี่ยวกับปัญญาประดิษฐ์ (artificial intelligence) เนื่องจากปัจจุบันนี้ความสามารถและประสิทธิภาพของปัญญาประดิษฐ์มีความก้าวหน้าขึ้นจนสามารถก้าวข้ามความสามารถของมนุษย์ในงานหลายประเภท ทำให้ลูกค้าต้องการที่จะให้ทางบริษัทสร้างปัญญาประดิษฐ์เพื่อนำไปใช้งานหรือแก้ปัญหาที่ต่างกันออกไป เช่น ใช้ปัญญาประดิษฐ์มาช่วยประมวลผลภาพจากกล้องวงจรปิด เพื่อหาบุคคลที่มีท่าทางน่าสงสัย เป็นต้น ซึ่งการจะสร้างปัญญาประดิษฐ์ที่เหมาะสมกับการแก้ปัญหาเหล่านี้ จำเป็นต้องมีชุดข้อมูล (dataset) ที่เหมาะสม บางครั้งอาจต้องใช้มนุษย์ในการสร้างขึ้นมาโดยการเก็บข้อมูลวิดีโอ และลงมือสร้างชุดข้อมูลจากวิดีโอที่ได้ด้วยตัวเอง หนึ่งในปัจจัยสำคัญในการพัฒนาโมเดลปัญญาประดิษฐ์ให้มีประสิทธิภาพสูงคือจำนวนข้อมูล ซึ่งหากมีจำนวนวิดีโอเป็นจำนวนมาก การใช้มนุษย์ในการสร้างชุดข้อมูลนั้นอาจจะต้องใช้มนุษย์เป็นจำนวนมาก และใช้เวลานาน

ทางคณบัญชีจึงมีความต้องการที่จะออกแบบและสร้างต้นแบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ (AI-assisted video labeling tool) สำหรับสร้างชุดข้อมูลจากวิดีโอ เพื่อช่วยแบ่งเบาภาระของผู้พัฒนาในการสร้างชุดข้อมูลสำหรับการพัฒนาโมเดลปัญญาประดิษฐ์ในการแก้ปัญหาที่ลูกค้าต้องการ โดยโครงการสหกิจนี้เน้นศึกษาเกี่ยวกับการวิเคราะห์และจำแนกการกระทำการของมนุษย์ (human action classification) ภายในสำนักงานจากภาพเคลื่อนไหวเป็นหลัก

1.2 วัตถุประสงค์

- เพื่อสร้างต้นแบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ ที่ทำให้มนุษย์และปัญญาประดิษฐ์ ทำงานร่วมกันเพื่อสร้างชุดข้อมูลในการนำมาพัฒนาปัญญาประดิษฐ์อื่นๆ ที่เหมาะสมกับปัญหาที่ต้องการ
- เพื่อออกแบบและสร้างต้นแบบของระบบบวิเคราะห์วิดีโอที่สามารถตรวจจับมนุษย์และจำแนกการกระทำพื้นฐานของมนุษย์ภายในสำนักงาน ประกอบด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว พูดคุย นอน โดยใช้ปัญญาประดิษฐ์
- เพื่อสร้างเครื่องมือที่สามารถสร้างชุดข้อมูลสำหรับการจำแนกการกระทำการของมนุษย์ให้สามารถใช้งานได้ง่าย สะดวกสบายมากขึ้น และมีประสิทธิภาพที่สูงกว่าเครื่องมือตัวอื่นในปัจจุบัน

1.3 ประโยชน์ที่คาดว่าจะได้รับ

- เพิ่มความสะดวกในการสร้างชุดข้อมูลสำหรับพัฒนาโมเดลปัญญาประดิษฐ์จากวิดีโอ
- ต้นแบบระบบบวิเคราะห์วิดีโอที่สามารถจำแนกการกระทำการของมนุษย์ภายในสำนักงานได้

1.4 ขอบเขตการดำเนินงาน

1. สร้างต้นแบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ โดยระบบจะประกอบไปด้วยสี่ส่วนดังนี้
 - (a) หน้าต่างของแอพพลิเคชัน (user interface)
 - (b) ระบบตรวจจับมนุษย์ในภาพ (person detection)
 - (c) ระบบท่านายตำแหน่งต่อไปของมนุษย์ในภาพเคลื่อนไหว (person tracker)
 - (d) ระบบจำแนกการกระทำการของมนุษย์ ซึ่งประกอบไปด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว พูดคุย นอน
2. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการของมนุษย์กับชุดข้อมูลที่ได้จากเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ เพื่อที่จะทดสอบว่าชุดข้อมูลที่ได้สามารถใช้งานจริงได้หรือไม่
3. พัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการของมนุษย์ภายในสำนักงานอย่างน้อย 2 โมเดล

1.5 ขั้นตอนการดำเนินงาน

การดำเนินงานวิจัยถูกแบ่งออกเป็นสามส่วน โดยส่วนแรกคือการศึกษาเทคโนโลยีในปัจจุบันเพื่อหาความเป็นไปได้และกำหนดขอบเขตของงาน ส่วนที่สองคือเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ เป็นส่วนที่ออกแบบและสร้างเครื่องมือสำหรับช่วยผู้พัฒนาในการสร้างชุดข้อมูล และส่วนที่สุดท้ายคือการนำชุดข้อมูลที่ได้จากการใช้เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ไปพัฒนาโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง

1. ศึกษาเกี่ยวกับการวิเคราะห์วิดีโอ (video analytics)
2. ศึกษาเกี่ยวกับชุดข้อมูลสำหรับการวิเคราะห์วิดีโอ
3. ศึกษาเกี่ยวกับโมเดลปัญญาประดิษฐ์ที่ใช้ในการวิเคราะห์วิดีโอ
4. ศึกษาเครื่องมือที่ใช้ในการช่วยสร้างชุดข้อมูลจากวิดีโอ

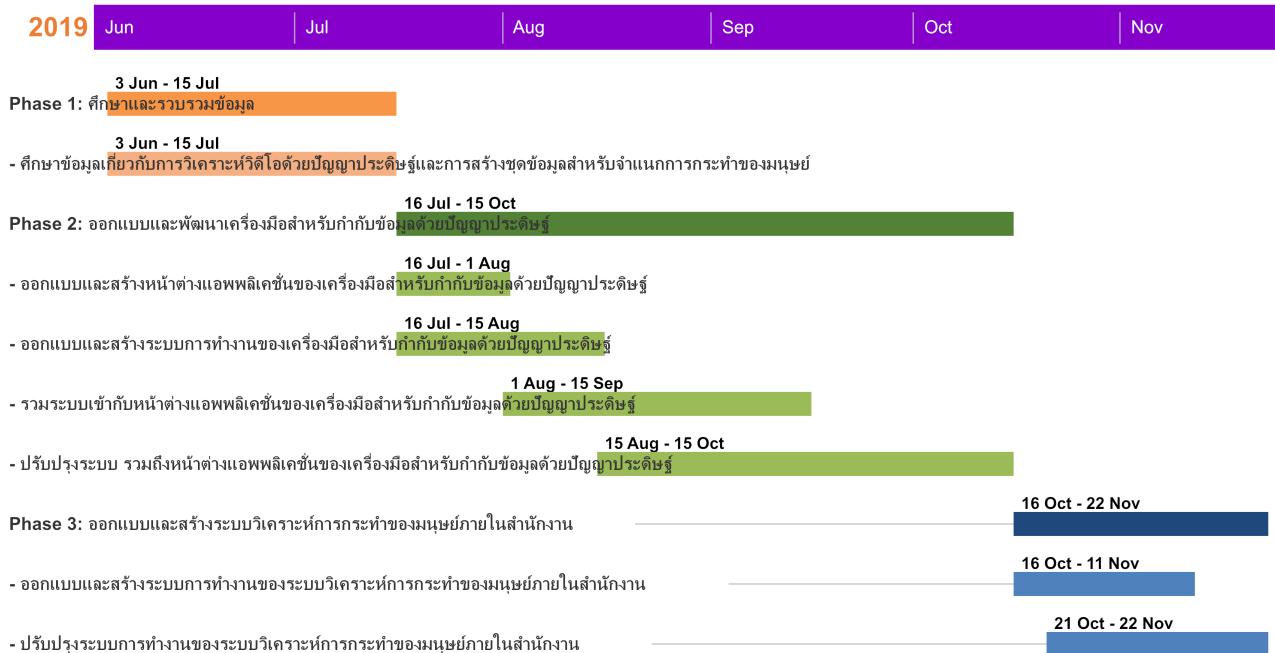
เครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์

1. ออกแบบและสร้างหน้าต่างแอพพลิเคชันของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. ออกแบบและสร้างระบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
3. ทดสอบและปรับปรุงการทำงานของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์

โมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

1. สร้างชุดข้อมูลสำหรับสร้างโมเดลปัญญาประดิษฐ์จากเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. สร้างโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน
3. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

แผนการดำเนินงาน



ตารางที่ 1.1: แผนการดำเนินงาน

บทที่ 2

ทฤษฎี/การวิจัยที่เกี่ยวข้อง

การวิเคราะห์วิดีโອในปัจจุบันนั้นมีวิธีและเทคนิคมากมาย ผู้วิจัยจึงต้องศึกษาองค์ความรู้และงานวิจัยที่เกี่ยวข้องกับวัตถุประสงค์ของงาน เพื่อศึกษาและใช้เป็นแนวทางในการประยุกต์สำหรับสร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ และโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ ซึ่งหัวข้อที่ผู้วิจัยได้ไปศึกษามา มีดังต่อไปนี้

1. การวิเคราะห์ผลวิดีโอ

- (a) การตรวจจับวัตถุ (object detection)
- (b) การนำทางตำแหน่งถัดไปของวัตถุ (object tracker)
- (c) การระบุตัวตนของบุคคล (person re-identification)
- (d) การจำแนกการกระทำ

2. เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

- (a) โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมนุษย์
- (b) เครื่องมือกำกับคุณลักษณะ (labeling tool)

3. ทฤษฎีที่เกี่ยวข้อง

- (a) Optical flow
- (b) IoU

2.1 การวิเคราะห์ผลวิดีโอ

ในส่วนของงานวิจัยสิ่งที่สนใจ คือ ข้อมูลการกระทำการของมนุษย์แต่ละคนภายในวิดีโอ เพื่อที่จะได้ผลลัพธ์ที่มีประสิทธิภาพอย่างมากเป็นข้อมูลของสิ่งที่สนใจ เช่น จำนวนคนที่เดินผ่านกล้อง หรือทิศทางการเดินของคนในวิดีโอ จึงจำเป็นต้องใช้การวิเคราะห์ผลวิดีโอเพื่อที่จะสกัดสิ่งที่สนใจออกมาจากวิดีโอ ซึ่งการวิเคราะห์ผลวิดีโอมีหลากหลายกระบวนการ โดยในแต่ละกระบวนการจะมีจุดประสงค์ของการทำและผลลัพธ์หลังการประมวลผลที่แตกต่างกัน ในหัวข้อนี้จะมาอธิบายถึงกระบวนการในการวิเคราะห์ผลของวิดีโอและผลลัพธ์ของกระบวนการนั้น

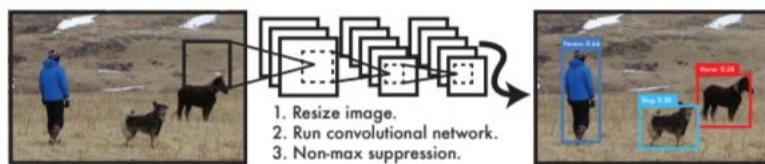
2.1.1 การตรวจจับวัตถุ

การตรวจจับวัตถุนั้นเป็นหนึ่งในกระบวนการวิเคราะห์ผลของวิดีโอ กล่าวคือกระบวนการที่ผู้วิจัยจะต้องทำการระบุสิ่งที่สนใจว่า คืออะไร อยู่ที่ตำแหน่งใด การตรวจจับวัตถุกูกคันพบเมื่อานมาแล้ว และในปัจจุบันนั้นสามารถทำได้หลากหลายวิธี โดยภายในบทความนี้จะสรุปใจความสำคัญของวิธีการต่างในการตรวจจับวัตถุ เช่น Sliding Window , Brute Force Search , R-CNN , Fast-RCNN , Faster-RCNN , YOLO , SSD

1. Sliding Window วิธีการที่เปรียบเสมือนมีหน้าต่าง (kernel) ค่อยๆเลื่อนไปยังแต่ละพิกเซลบนรูป ซึ่งก่อนการเลื่อนของหน้าต่างแต่ละครั้ง จะนำส่วนของรูปภาพที่ถูกหน้าต่างทับอยู่ไปคำนวณว่าใช้วัตถุที่เราต้องการหรือไม่ จากนั้นจึงค่อยเลื่อนตัวไป โดยจะทำการวนการแบบนี้จนครบทั้งรูปภาพ
2. Brute Force Search ถูกสร้างขึ้นมาเพื่อแก้ปัญหาขนาดของหน้าต่างไม่ตรงกับขนาดของวัตถุที่อยู่ในภาพ ทำให้มีโอกาสที่จะไม่พบวัตถุ โดยหลักการของวิธีการนี้ คือ การย่อ-ขยาย รูปภาพและนำเข้าในหลายอัตราส่วน ตั้งแต่ 0.1 เท่า จนถึง 2 เท่า แต่ข้อเสียของวิธีการนี้คือ มีการคำนวณพื้นที่ช้าๆ และใช้เวลานาน
3. R-CNN ใช้อัลกอริทึม Selective search เข้ามาช่วยในการเสนอพื้นที่ที่น่าจะมีวัตถุอยู่ทุกด้านการค้นหา ทุกๆตัวແเน่ง จากนั้นก็นำรูปภาพในส่วนพื้นที่นั้นไปคำนวณว่าวัตถุนั้นคืออะไร กรณีที่มีพื้นที่ที่อยู่ใกล้ๆ วัตถุกูกเสนอเข้ามาเป็นจำนวนมากด้วย เราจะใช้ Non-Maximum Suppression (NMS) หรือการเลือกพื้นที่ที่ถูกทับซ้อนมากที่สุดในบริเวณนั้น
4. Fast-RCNN จากวิธีการ R-CNN แต่ละพื้นที่จะถูกนำไปสกัดคุณลักษณะ และ คำนวณผลทีละพื้นที่ ทำให้เสียเวลา โดย Faster-RCNN จะมีส่วนที่คล้ายกับ R-CNN ในส่วนการทำ Selective search หาพื้นที่ที่น่าจะมีวัตถุเหมือนเดิม แต่ Faster-RCNN จะนำรูปภาพทั้งรูปภาพไปสกัดคุณลักษณะ หลังจากที่ได้คุณลักษณะแล้ว นำพิกัดของพื้นที่ที่น่าจะมีวัตถุ บนรูปภาพที่ถูกสกัดคุณลักษณะแล้วของ ไปผ่าน ROI Pooling (การลดขนาดข้อมูลให้มีขนาดคงที่เพื่อเป็นอินพุตให้กับโมเดลในการคำนวณผล)
5. Faster-RCNN พัฒนาจาก Fast-RCNN โดยวิธีของ Faster-RCNN จะรวมในส่วนของ Selective search และ การทำงานอื่นๆให้อยู่ในโครงข่ายเดียวกัน สรุปคือการทำงานของโครงข่ายของ Faster-RCNN จะมีการทำงาน 3 อย่างหลักคือ 1) สกัดคุณลักษณะ 2) การเสนอส่วนที่น่าจะมีวัตถุอยู่ในรูปภาพ 3) หลังจากได้รูปภาพจากการสกัดคุณลักษณะ นำพิกัดของพื้นที่ที่น่าจะมีวัตถุ บนรูปภาพที่ถูกสกัดคุณลักษณะแล้ว ของ ไปผ่าน ROI Pooling
6. YOLO เป็นวิธีการที่ใช้โครงข่ายประสาทแบบconvolutoinal ขึ้นเพียงตัวเดียวทำนายรูปภาพทั้งรูป โดยโครงข่ายจะแบ่งรูปภาพออกเป็นพื้นที่ และใช้ Fully-connected (เป็นโครงข่ายประสาทเทียมที่นำเอาคุณลักษณะมาคำนวณผล) คำนวณตำแหน่งของกรอบสีเหลี่ยมและหมวดหมู่ของกรอบสีเหลี่ยมในแต่ละพื้นที่ไปพร้อมกัน
7. SSD ใช้โครงข่ายประสาทเทียมตัวเดียวเหมือนกับ YOLO แต่การออกแบบโครงสร้างแตกต่างกัน SSD จะใช้ VGG-16(เป็นโมเดล CNN ชนิดหนึ่ง) ในการสกัดคุณลักษณะ และ ใช้ Convolution layer ต่อ กันหลายขั้นเพื่อลดมิติและความละเอียดทำให้ตรวจจับวัตถุในหลายขนาด ซึ่งในแต่ละขั้นจะได้ผลลัพธ์เป็น Convolution filter งานนั้นจะนำ Convolution filter ไปคำนวณผลต่อ

YOLO

โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เป็นโครงสร้างที่มีความเร็วมาก มีความเร็วในการประมวลผลถึง 45 เฟรมต่อวิ ทำให้สามารถประมวลผลแบบเรียลไทม์ได้ นอกจากนั้นยังมีความแม่นยำ mAP มากกว่า



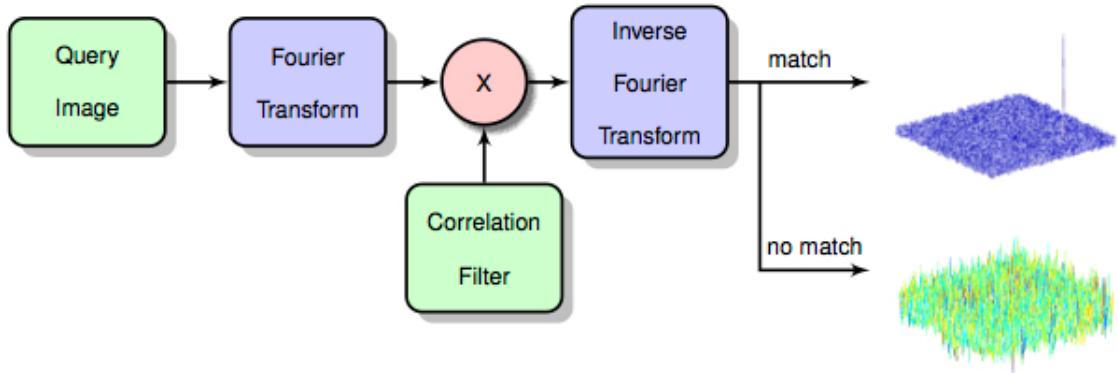
รูปที่ 2.1: กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO

โมเดลสำหรับตรวจจับวัตถุอื่นๆ ถึง 2 เท่า ซึ่งเหตุผลที่โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เร็วกว่าโมเดลปัญญาประดิษฐ์ตัวอื่นๆ เนื่องจาก มีแนวคิดที่ต่างออกไป คือ สำหรับการตรวจจับวัตถุในวิธีการก่อนหน้าจะใช้ วิธีทำนายกรอบสี่เหลี่ยมก่อน แล้วจึงค่อยนำกรอบสี่เหลี่ยมไปทำนายว่าเป็นหมวดหมู่อะไร ซึ่ง YOLO มีวิธีการที่ ต่างออกไป คือ ทำนายตำแหน่งของกรอบสี่เหลี่ยมและทำนายว่ากรอบสี่เหลี่ยมนั้นเป็นหมวดหมู่อะไรพร้อมกัน โดยใช้โครงข่ายประสาทแบบคอนโวลูชัน ด้วยแนวคิดนี้ จึงเป็นที่มาของชื่อ YOLO หรือ you only look once การมองแค่เพียงครั้งเดียว ซึ่งโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO ที่ถูกใช้ในงานวิจัยนี้ประกอบไปด้วย 1) YOLOv3-tiny 2) YOLOv3 3) YOLOv3-spp ซึ่งทั้ง 3 โครงสร้างจะมีความแตกต่างของโครงสร้างดังนี้

1. YOLOv3-tiny ใช้ Max-Pooling layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
2. YOLOv3 ใช้ Convolutional layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
3. YOLOv3-spp ใช้ Convolutional layers+ฟีเจอร์ที่ดีที่สุดของ Max-Pooling layers ในขั้นตอนของการ ลดจำนวนข้อมูลตัวอย่าง

2.1.2 ระบบที่นำมายำต์แหน่งถัดไปของวัตถุ

การนำมายำต์แหน่งถัดไปของวัตถุ^[2] คือระบบที่ใช้สำหรับการติดตามการเคลื่อนไหวของวัตถุที่สนใจที่อยู่ในรูปภาพ โดยใช้การคำนวณทางคณิตศาสตร์ และการประมวลผลภาพ (image processing) ทำให้การประมวลผลนั้นเร็วกว่าการใช้โมเดลปัญญาประดิษฐ์



รูปที่ 2.2: แนวคิดของระบบที่นำมายำต์แหน่งถัดไปของวัตถุ

จากรูปที่ 2.2 เป็นหลักการในการนำมายำต์แหน่งต่อไป [feihong this way] โดยการนำรูปมาผ่านกระบวนการแปลงฟูรีเยอร์ (fourier transform) และนำมาคูณกับ correlation filter ซึ่งเป็นตัวกรองที่ใช้สำหรับการหาความสัมพันธ์กับวัตถุในภาพ จากนั้นทำการแปลงฟูรีเยอร์ผกผัน (inverse fourier transform) เพื่อตรวจสอบว่าวัตถุในภาพนั้นอยู่ที่ตำแหน่งใด โดยมีการคำนวณเริ่มจากการหา correlation filter ที่ดีที่สุดโดยใช้วิธีลดผลรวมของข้อผิดพลาดกำลังสองให้น้อยที่สุดดังนี้

$$\epsilon = \left\| \sum_{l=1}^d h^l * f^l - g \right\|^2 + \lambda \sum_{l=1}^d \|h^l\|^2 \quad (2.1)$$

โดยที่

ϵ = ค่าความคลาดเคลื่อน

d = จำนวนมิติของผังคุณลักษณะ (feature map) ของภาพ

h = correlation filter

\star = circular correlation

f = พื้นที่สีเหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

g = ผลลัพธ์ correlation ที่ต้องการของ f

λ = regularization term

เมื่อพิจารณาจากรูปภาพเดี่ยวในกรณีที่เวลา (t) เท่ากับ 1 จะสามารถจัดรูปสมการด้านบนได้ดังนี้

$$H^l = \frac{\bar{G}F^l}{\sum_{k=1}^d \bar{F}^k F^k + \lambda} \quad (2.2)$$

$$H_t^l = \frac{A_t^l}{B_t} \quad (2.3)$$

$$A_t^l = (1 - \eta) A_{t-1}^l + \eta \bar{G}_t F_t^l \quad (2.4)$$

$$B_t = (1 - \eta) B_{t-1} + \eta \sum_{k=1}^d \bar{F}_t^k F_t^k \quad (2.5)$$

โดยที่

H = correlation filter

η = อัตราการเรียนรู้

\bar{G} = g ที่ผ่านการทำ complex conjugation

F = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

\bar{F} = f ที่ผ่านการทำ complex conjugation

t = เวลา

จากสมการที่ได้มาจะสามารถทำให้หาตำแหน่งต่อไปของวัตถุที่สนใจได้ด้วยสมการต่อไปนี้

$$y = F^{-1} \left\{ \frac{\sum_{l=1}^d \bar{A}^l Z^l}{B + \lambda} \right\} \quad (2.6)$$

โดยที่

y = correlation score

F^{-1} = การแปลงฟูร์เรียร์ผกผันแบบไม่ต่อเนื่อง (inverse discrete fourier transform)

\bar{A}^l = A^l ที่ผ่านการทำ complex conjugation

Z = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการหาผังคุณลักษณะของภาพใหม่

โดยค่าของ y ที่ได้ออกมาจะทำให้รู้ถึงตำแหน่งของวัตถุที่สนใจได้ ณ ตำแหน่งที่ y มีค่าสูงสุด

2.1.3 ระบบบุตัวตนของบุคคล

ระบบระบุตัวตนของบุคคล^[7] คือการระบุตัวตนของบุคคลภายในวิดีโอหรือระหว่างรูป 2 รูป สามารถนำมาประยุกต์ใช้ในด้านของการรักษาความปลอดภัย การตามหาบุคคล หรือการตรวจสอบการกระทำการของบุคคลนั้นในวิดีโอด้วย การระบุตัวตนของบุคคลนั้นเป็นปัญหาที่ท้าทาย เนื่องจากคุณลักษณะทั่วไปของบุคคลในรูปไม่เพียงพอต่อการระบุบุคคลภายในรูปว่าเป็นบุคคลคนเดียวกันได้ ซึ่งวิธีการที่ใช้ในการระบุตัวตนของบุคคลเรียกว่า Dynamically Matching Local Information (DMLI) ที่สามารถจัดแนวนรายละเอียดข้อมูลของรูป และเพิ่มประสิทธิภาพให้สูงขึ้น ถึงแม้ว่าวิธีการทำ Dynamically Matching Local Information นั้นจะไม่ใช้วิธีการที่มีประสิทธิภาพสูงสุดแต่มีประสิทธิภาพใกล้เคียงกัน และด้วยทางที่คนผู้วิจัยสามารถนำวิธีนี้มาประยุกต์เข้ากับงานวิจัยครั้งนี้ได้สะดวกที่สุด จึงนำวิธี Dynamically Matching Local Information มาใช้สำหรับงานวิจัยครั้งนี้

การทำงานของระบบบุตัวตนของบุคคลจะเริ่มจากการแบ่งรูปออกเป็น 8 ส่วนและนำคุณลักษณะของรูปมาผ่านกระบวนการ normalization เพื่อลดความซ้ำซ้อนของข้อมูล แล้วนำมาเปรียบเทียบความแตกต่างของคุณลักษณะของรูป หลังจากนั้นหากค่าเฉลี่ยของความแตกต่างของค่าที่ได้ออกมา โดยค่าที่ได้ออกมาจะเรียกว่า original distance ถ้าค่าที่ออกมากลับค่าเฉลี่ยของค่าที่ออกมากัน 0 จะหมายถึงบุคคลในรูปทั้งสองเป็นบุคคลเดียวกัน และมีการตั้งค่าเกณฑ์สำหรับ original distance เพื่อใช้สำหรับในการระบุบุคคลในรูปเป็นบุคคลเดียวกัน

โดยชุดข้อมูลที่นำมาใช้สำหรับการทำโน้มเดลปัญญาประดิษฐ์ได้แก่



รูปที่ 2.3: การแบ่งรูปออกเป็น 8 ส่วนของระบบระบุตัวตนของบุคคล

1. Market1501 เป็นชุดข้อมูลที่เก็บข้อมูลภาพของบุคคลโดยใช้กล้องจำนวน 6 ตัว ถ่ายภาพบุคคลที่ด้านหน้าของซุปเปอร์มาร์เก็ตในมหาวิทยาลัย Tsinghua
2. DukeMTMCReID เป็นชุดข้อมูลที่เก็บข้อมูลภาพของบุคคลโดยใช้กล้องจำนวน 8 ตัว ถ่ายภาพบุคคลที่วิทยาเขตของมหาวิทยาลัย Duke ซึ่งมีการเก็บภาพมากถึง 2 ล้านภาพของนักศึกษา 2 พันคน
3. CUHK-03 เป็นชุดข้อมูลที่เก็บภาพของบุคคลที่มหาวิทยาลัยจีนที่ห่องกง
4. MSMT17 เป็นชุดข้อมูลที่เก็บข้อมูลภาพของบุคคลโดยใช้กล้องจำนวน 15 ตัว โดยที่กล้องแต่ละตัวจะไม่ได้ตั้งอยู่สถานที่เดียวกัน และเก็บข้อมูลที่ในวันที่มีสภาพอากาศต่างกัน

โดยทุกชุดข้อมูลจะใช้ ResNet50 ในการสร้างโมเดลปัญญาประดิษฐ์ สำหรับการต่อมาการทดสอบโมเดลปัญญาประดิษฐ์ด้วยวิธีการ Global+DMIL คือวิธีการนำคุณลักษณะทั่วไปของภาพนำไปจัดแนวรายละเอียดของข้อมูล และนำไปเข้าโมเดลปัญญาประดิษฐ์เพื่อคำนวนหาค่า rank1 และ mAP โดยที่ค่า rank1 หมายถึงค่าเปอร์เซ็นต์ความมั่นใจสูงสุดของโมเดลปัญญาประดิษฐ์ที่ทำนายออกมากถูกต้อง และค่า mAP คือการหาค่าเฉลี่ยความแม่นยำในแต่ละหมวดหมู่ ซึ่งสามารถดูค่า rank1 และ mAP ของโมเดลปัญญาประดิษฐ์สำหรับการทำระบุตัวตนของบุคคลได้ที่การทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของมนุษย์

2.1.4 ระบบจำแนกการกระทำ

ระบบจำแนกการกระทำเป็นกระบวนการสำหรับทำนายการกระทำการของมนุษย์หรือสิ่งที่สนใจที่เกิดการกระทำขึ้นภายในวิดีโอ โดยในทั้งข้อนี้จะกล่าวถึงตั้งแต่ขั้นตอนการได้มาซึ่งชุดข้อมูลมีกระบวนการอย่างไร การนำโมเดลปัญญาประดิษฐ์มาใช้ในการจำแนกการกระทำ และการวัดผลของโมเดลปัญญาประดิษฐ์ โดยชุดข้อมูลที่ผู้วิจัยได้เลือกนำมาศึกษาจากชุดข้อมูลที่ถูกเป็นที่กล่าวถึงในปัจจุบัน และมีขนาดของชุดข้อมูลที่ใหญ่

จากข้อความข้างต้นชุดข้อมูลที่ผู้วิจัยได้เลือกนำมาใช้ได้แก่ YouTube-8M, AVA, Moment in Time โดยแต่ละชุดข้อมูลจะมีความแตกต่างกันในหลายๆ ด้าน แต่จะมีสิ่งที่เหมือนกัน คือ เป็นชุดข้อมูลสำหรับการวิเคราะห์วิดีโอที่สนใจจากการกระทำการของมนุษย์ โดยในบทความนี้จะกล่าวถึงความแตกต่างในด้านต่างๆ เช่น เป้าหมายของแต่ละชุดข้อมูล วิธีการเก็บข้อมูลสำหรับชุดข้อมูล วิธีการสร้างคำจำกัดคุณลักษณะ และรายละเอียดของชุดข้อมูลจากนั้นจะสรุปข้อมูลของแต่ละชุดข้อมูล

ชุดข้อมูล YouTube-8M

1. รายละเอียดของชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : เพื่อจำแนกสาระสำคัญของวิดีโอ (video theme)
- (b) จำนวนของวิดีโอ : 8,264,650 วิดีโอ

- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 229.6 วินาที
- (d) จำนวนของหมวดหมู่ของคำกำกับคุณลักษณะ : 4,800 หมวดหมู่
- (e) กฎในการรับรวมวิดีโอดังนี้

- i. ทุกๆหัวข้อต้องเป็นรูปธรรม
- ii. ในแต่ละหัวข้อต้องมีจำนวนวิดีโอยield ไม่น้อยกว่า 200 วิดีโอ
- iii. ความยาวของวิดีโอดังนี้อยู่ระหว่าง 120 - 500 วินาที

หลังจากได้กฎในการรับรวมวิดีโอล้วน ขั้นตอนต่อไปคือการสร้างคำศัพท์ที่ใช้ในการค้นหาข้อมูลวิดีโอดังนี้ใน YouTube

(f) ขั้นตอนในการสร้างคำศัพท์มีดังนี้

- i. กำหนดบัญชีขาว (whitelist) ของหัวข้อที่เป็นรูปธรรมมา 25 ชนิด เช่น กีฬา เป็นต้น
- ii. กำหนดบัญชีดำ (blacklist) ของหัวข้อที่คิดว่าไม่เป็นรูปธรรมไว้ เช่น software เป็นต้น
- iii. รวบรวมหัวข้อที่มีอยู่ในรายการที่อนุญาตอย่างน้อย 1 หัวข้อ และต้องไม่มีอยู่ในบัญชีดำซึ่งจะทำให้ได้หัวข้อที่ต้องการมาประมาณ 50,000 หัวข้อ
- iv. จักนั้นใช้ผู้ประเมินจำนวน 3 คน ในการคัดหัวข้อที่คิดว่าเป็นรูปธรรม และสามารถจัดจำหรือเข้าใจได้ง่ายโดยไม่ต้องใช้ข้อมูลในด้านนั้นๆ ซึ่งผู้ประเมิน ก็จะมีความว่า “มันยากขนาดไหนถึงจะระบุได้ว่ามีหัวข้อดังกล่าวอยู่ในรูปหรือวิดีโอด้วยเพียงแค่การมองเห่านั้น?” โดยแบ่งเป็นระดับดังนี้
 - A. บุคคลทั่วไปสามารถเข้าใจได้
 - B. บุคคลทั่วไปที่ผ่านการอ่านบทความที่เกี่ยวข้องมาแล้วสามารถเข้าใจได้
 - C. ต้องใช้ญี่ปุ่นในด้านใดซึ่งจะเข้าใจได้
 - D. เป็นไปไม่ได้ ถ้าไม่มีความรู้ที่ไม่ได้เป็นรูปธรรม
 - E. ไม่เป็นรูปธรรม
- v. หลังจากคำนึงข้างบนและการให้คะแนน จะทำการเก็บไว้เฉพาะหัวข้อที่มีคะแนนเฉลี่ยมากที่สุดอยู่ที่ประมาณ 2.5 คะแนนหรือต่ำกว่าเท่านั้น
- vi. ทำให้สุดท้ายเหลือเพียงประมาณ 10,000 หัวข้อที่สามารถใช้ได้
- vii. หลังจากได้หัวข้อที่คิดว่าเป็นรูปธรรมแล้วก็นำไปค้นหาและรวบรวมด้วย YouTube annotation system โดยมีขั้นตอนดังนี้
 - A. สุมเลือกวิดีโอมาก 10 ล้านวิดีโอ พร้อมกับหัวข้อของวิดีโอ โดยใช้กฎที่กำหนดไว้ เอาหัวข้อที่มีจำนวนวิดีโอน้อยกว่า 200 วิดีโอออก
 - B. ทำให้เหลือจำนวนวิดีโอยู่ 8,264,650 วิดีโอ
 - C. แยกออกเป็น 3 ส่วนคือ ชุดข้อมูลสำหรับสร้างโมเดล (train set) ชุดข้อมูลสำหรับตรวจสอบ (validate set) และชุดข้อมูลสำหรับทดสอบ (test set) ในอัตราส่วน 70:20:10 ตามลำดับ

2. โมเดลปัญญาประดิษฐ์

(a) การเตรียมข้อมูล

- i. คุณลักษณะระดับเฟรม : การลดขนาดของข้อมูล เนื่องจากมีข้อมูลที่มีขนาดใหญ่ทำให้ใช้เวลาในการเปิดนาน ซึ่งกระบวนการนี้จะมีการลดความเร็วเฟรมต่อวินาที เวกเตอร์ของคุณลักษณะ (feature vector) และแปลงข้อมูลจาก 32 บิต ให้เป็น 8 บิต
- ii. คุณลักษณะระดับวิดีโอ : การแยกเวกเตอร์คุณลักษณะระดับวิดีโອอกจากคุณลักษณะระดับเฟรมซึ่งการทำแบบนี้ทำให้ได้ประโยชน์ 3 ข้อ คือโมเดลทั่วไปที่ไม่ใช่โครงข่ายประสาทเทียบสามารถนำไปใช้งานได้ ขนาดข้อมูลเล็กลง และเหมาะสมกับการนำไปสร้างโมเดลในขอบเขตอื่นมากขึ้น

- (b) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับเฟรม
- i. one vs all logistic regression classifier + average pooling
 - ii. Deep bag of frames
 - iii. Long short-term memory (LSTM)
- (c) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับวิดีโอ
- i. Logistic regression
 - ii. Support vector machine (SVM)
 - iii. Mixture of Expert (MoE)
- (d) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือ
- i. Mean Average Precision (mAP)
 - ii. Hit@k (Top@k)
 - iii. Precision at equal recall rate (PERR)
- (e) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ YouTube-8M

Input features	Modeling approach	mAP	Hit@1	PERR
Frame-level	Logistic + average	11.0	50.8	42.2
	Deep bag of frames	26.9	62.7	55.1
	LSTM	26.6	64.5	57.3
Video-level	SVM	17.0	56.3	47.9
	Logistic regression	28.1	60.5	53.0
	Mixture-of-2-experts	30.0	63.3	55.8

ตารางที่ 2.1: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ YouTube-8M

- (f) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ Sports-1M

Approach	mAP	Hit@1	PERR
Logistic regression	58.0	60.1	79.6
Mixture-of-2-experts	61.3	63.2	82.6
LSTM	67.6	65.7	86.2
Hierarchical 3D convolutions ^[?]	-	61.0	80.0
Stacked 3D convolutions ^[?]	-	61.0	85.0
LSTM with optical flow and pixels ^[?]	-	73.0	91.0

ตารางที่ 2.2: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ Sports-1M

- (g) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ ActivityNet

Approach	mAP	Hit@1	Hit@5
Mixture-of-2-experts	77.6	74.9	91.6
LSTM	57.9	63.4	81.0
Ma, Bargal et al. ^[?]	53.8	-	-
Heilbron et al. ^[?]	43.0	-	-

ตารางที่ 2.3: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ ActivityNet

- (h) ปัญหาที่พบ

เนื่องจากว่า YouTube-8M นั้นมีจำนวนข้อมูลที่เยอะมาก ทำให้ไม่สามารถตรวจสอบความถูกต้องของชุดข้อมูลได้ทั้งหมดว่ามีความถูกต้องมากน้อยขนาดไหน ทำให้อาจเกิดข้อผิดพลาดได้ (ปัจจุบันปี 2019 YouTube-8M ได้มีการตรวจสอบข้อมูลอีกครั้ง เพื่อเพิ่มประสิทธิภาพของชุดข้อมูลซึ่งทำให้ปัจจุบันจำนวนข้อมูล และจำนวนหัวข้อลดน้อยลงจากข้อมูลที่ใช้อ้างอิงในบทความข้างต้นที่ได้กล่าวมา)

ชุดข้อมูล Atomic visual action (AVA)

AVA^[2] คือ ชุดข้อมูลที่รวบรวมวิดีโอที่มีความยาว 15 นาที ถูกแบ่งด้วยความถี่ 1 hz (900 keyframes) จากในหนังโดยยึดการกระทำของมนุษย์เป็นศูนย์กลาง เพื่อใช้สำหรับสร้างโมเดลที่เข้าใจกิจกรรมของมนุษย์ในวิดีโอด้วยคำอธิบาย (label) เป็นแบบ multiple label (ในหนึ่งกรอบสี่เหลี่ยม (bounding box) สามารถมีคำอธิบายได้หลายคำอธิบาย) และคำอธิบายของ AVA (label) มีจำนวน 80 class สามารถแบ่งได้เป็น 3 หมวดหมู่คือ ท่าทาง (Pose) , ปฏิสัมพันธ์กับวัตถุ (Interaction with object) และ ปฏิสัมพันธ์กับบุคคล (Interaction with people) และสามารถมีคำอธิบายได้มากสูงสุดถึง 7 คำอธิบาย

1. รายละเอียดชุดข้อมูล

(a) ขั้นตอนการเก็บข้อมูลสำหรับการทำชุดข้อมูลมีขั้นตอนการทำ 5 ขั้นดังนี้

- i. การสร้างคำศัพท์การกระทำจะมีหลักการ 3 ข้อในการรวบรวมคำศัพท์ดังนี้
 - A. เก็บรวบรวมคำศัพท์ที่นำไปที่เกิดขึ้นในชีวิตประจำวัน
 - B. จะต้องมีเอกสารยืนยันสามารถเห็นได้ชัดเจน เช่น การถือของ
 - C. กำหนดรูปแบบของคำศัพท์ขึ้นมา และใช้ความรู้จากชุดข้อมูลอื่นในการทำให้ได้หมวดหมู่การกระทำของมนุษย์ที่ครอบคลุม
- ii. ภาพยนต์และส่วนที่เลือกมาใช้ทำชุดข้อมูล AVA ทั้งหมดจะถูกนำมาจาก YouTube โดยเริ่มจากการรวบรวมรายการซื้อของนักแสดงที่มีชื่อเสียง ซึ่งจะมีความหลากหลายของเชื้อชาติ รวมกันอยู่ วิดีโอที่ถูกคัดเลือกจะมีเกณฑ์ดังนี้
 - A. วิดีโอต้องอยู่ในหมวด ภาพยนต์ และละครโทรทัศน์
 - B. วิดีโอจะต้องมีความยาวมากกว่า 30 นาที
 - C. เผยแพร่มาแล้วเป็นระยะเวลาอย่างน้อย 1 ปี
 - D. มีจำนวนยอดคนดูมากกว่า 1,000 คน
 - E. ลงทะเบียนวิดีโอของประเภท เป็นภาพขาว-ดำ มีความละเอียดต่ำ การ์ตูน หรือวิดีโอเกม
- iii. การสร้างกรอบสี่เหลี่ยมครอบมนุษย์ที่อยู่ภายในภาพประกอบด้วย 2 ขั้นตอน
 - A. สร้างกรอบสี่เหลี่ยมโดยใช้โมเดลปัญญาประดิษฐ์ faster RCNN สำหรับการตรวจจับมนุษย์
 - B. ใช้มนุษย์ในการตรวจสอบและแก้ไขกรอบสี่เหลี่ยมที่ผิดพลาด
- iv. การติดตามตำแหน่งของบุคคล

ทำการติดตามตำแหน่งของบุคคลที่อยู่ในช่วงเวลาเดียวกันด้วยใช้วิธีการแทร็คโดยยึดมนุษย์เป็นศูนย์กลาง โดยการคำนวณค่าความใกล้เคียงกันระหว่างบุคคล โดยใช้ person embedding (ใช้โครงข่ายประสาทเทียมในการหาพื้นที่เรอร์ขั้นสูงและใช้เมทริกซ์ในการหาความสัมพันธ์ของแต่ละคน) จากนั้นจะใช้อัลกอริทึม Hungarian distance (อัลกอริทึมสำหรับการหาข้อเสนอที่ดีที่สุด) ในการหาตัวเลือกคู่ของกรอบสี่เหลี่ยมที่ดีที่สุด
- v. การสร้างคำจำกัดความลักษณะ

การสร้างคำจำกัดของ การกระทำจะถูกสร้างขึ้นโดยมนุษย์ ซึ่งผู้วิจัยจะใช้โปรแกรมสำหรับช่วยเหลือในการสร้างคำจำกัดความลักษณะ โดยสามารถกำหนดคำจำกัดของ การกระทำได้สูงสุดถึง 7 คำต่อ 1 กรอบสี่เหลี่ยม นอกจากนั้นสามารถตั้งสถานะเนื้อหาที่ไม่เหมาะสม หรือ

กรอบสีเหลี่ยมที่ผิดพลาดได้อีกด้วย ซึ่งในทางปฏิบัติเพื่อลดโอกาสที่จะเกิดข้อผิดพลาด จึงแบ่งขั้นตอนในการสร้างคำกำกับออกเป็น 2 ขั้นตอนดังนี้

- A. สร้างข้อเสนอสำหรับคำกำกับของการกระทำ
- B. ข้อเสนอจะถูกตรวจสอบข้อเสนอที่ได้จากขั้นตอนแรก ซึ่งจะใช้มุขย์ในการตรวจสอบ 3 คน โดยคำกำกับจะต้องถูกตรวจสอบด้วยผู้ตรวจสอบอย่างน้อย 2 คน จึงจะถูกยืนยันว่าเป็นคำกำกับหลัก

2. โมเดลปัญญาประดิษฐ์

- (a) โมเดลปัญญาประดิษฐ์ที่งานวิจัยนี้ใช้ คือ two stream variant ซึ่งจะทำการประมวลผลทั้ง RGB flow และ optical flow โดยเป็นโครงสร้างของ faster RCNN ที่นำ Inception network เข้ามาใช้
- (b) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือค่า IoU และ 3D IoUs
 - i. ค่า IoU คือค่าที่ใช้วัดความสอดคล้องระหว่างสองกรอบสีเหลี่ยม(กรอบสีเหลี่ยมจริงของเฟรม และ กรอบสีเหลี่ยมที่ทำนายขึ้นมา) ซึ่งใช้สำหรับการวัดผลระดับเฟรม
 - ii. ค่า 3D IoUs คือค่าที่ใช้วัดความสอดคล้องระหว่างกรอบสีเหลี่ยมภายใน 2 วิดีโอ ซึ่งใช้สำหรับการวัดผลระดับวิดีโอ โดยเทียบกันระหว่างกรอบสีเหลี่ยมจริงในช่วงของเฟรมที่ต่อกัน (ground-truth tubes) และ กรอบสีเหลี่ยมที่ทำนายขึ้นมาในช่วงของเฟรมที่ต่อกัน (linked detection tubes)
- (c) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ในปัจจุบัน
ข้อมูลโมเดลปัญญาประดิษฐ์ที่นำมาทดสอบ
 - i. Actionness เป็นการหาความน่าจะเป็นของการกระทำ โดยใช้โครงสร้างของ hybrid fully convolutional network (HFCN) hybrid fully เป็นโครงสร้างที่ประกอบด้วยโครงข่ายประสานภาพและโครงข่ายconvolutional network ที่ทำงานร่วมกัน 2 ชนิด คือ
 - A. Appearance-FCN (A-FCN) คือ โครงข่ายประสานภาพที่นำมาระบบที่แสดงลักษณะของวัตถุ(ตำแหน่งวัตถุ, ความตื้นลึกวัตถุ) ที่ปรากฏบนรูป RGB1
 - B. MotionFCN (M-FCN) คือ โครงข่ายประสานภาพที่แยกการเคลื่อนไหว จากข้อมูลของ optical flow
 - ii. Peng without MR, Peng with MR (Multi-region two-stream R-CNN) เป็นโมเดลปัญญาประดิษฐ์ที่ใช้สำหรับตรวจจับวิดีโອในชีวิตจริง ซึ่งพื้นฐานของโมเดลนี้เป็น Faster R-CNN โดยโมเดลนี้มีกระบวนการ 3 กระบวนการคือ
 - A. สร้างข้อเสนอพื้นที่ที่มีการเคลื่อนไหว
 - B. สะสม Optical flow จากเฟรมหลายเฟรม เพื่อนำไปปรับปรุงการตรวจจับการกระทำ
 - C. นำพื้นที่หลายๆ ส่วนมาวิเคราะห์ผ่านโมเดล Faster R-CNN
 - iii. ACT Action Tubelet Detector เป็นการระบุตำแหน่งของการกระทำที่มีระยะเวลาสั้นๆ ซึ่งใช้วิธีการตรวจจับระดับเฟรม และ ใช้การติดตามตำแหน่งในการเชื่อมระหว่างเฟรมปัจจุบันไปยังเฟรมถัดไป. ACT ถูกสร้างต่อจาก SSD framework และ ใช้ค่อนโภคุณชั้นในการสกัดคุณลักษณะในแต่ละเฟรม ซึ่ง การคิด คะแนน และ ความน่าจะเป็นของหมวดหมู่จะคิดจาก การนำคุณลักษณะเรียงต่อกัน และ หาข้อมูลจากลำดับข้อมูลนั้น

จากการทดสอบการเทียบโมเดลปัญญาประดิษฐ์ของงานวิจัยนี้และวิธีการอื่นๆ โดยนำไปทดสอบกับชุดข้อมูลวิดีโอ JHMDB และ UCF101-24 ได้ผลลัพธ์ออกมาดังนี้

Frame-mAP	JHMDB (mAP)	UCF101-24 (mAP)
Actionness	39.9	-
Peng w/o MR	56.9	64.8
Peng w/ MR	58.5	65.7
ACT	65.7	69.5
2 stream(Our approach)	73.3	76.3

ตารางที่ 2.4: ผลการทดลองของวิธีต่างๆบนคุณลักษณะระดับเฟรม

- (d) ปัญหาที่พบ ในปัจจุบันยังไม่มีโมเดลปัญญาประดิษฐ์ที่ทดสอบด้วยชุดข้อมูล AVA และได้ผลการทำงานที่ดี เนื่องจากชุดข้อมูลนี้สนใจการกระทำของมนุษย์ที่มีรายละเอียดเล็กๆน้อยๆ ทำให้ยากต่อการทำนายสำหรับโมเดลปัญญาประดิษฐ์

ชุดข้อมูล Moments in Time

Moments in time^[7] คือชุดข้อมูลที่ใช้มนุษย์ในการกำกับข้อมูล ทั้งหมดให้กับวิดีโอด้วยสิ่งที่ 1 ล้านวิดีโอ และมีจำนวนกิจกรรมหรือกระทำต่างกัน 339 หมวดหมู่ โดยแต่ละวิดีโอมีความยาวอยู่ที่ 3 วินาที เนื่องจากเป็นเวลาเฉลี่ยที่มนุษย์ใช้ในการเข้าใจกับเหตุการณ์ที่เกิดขึ้น (human working memory) รูปแบบของชุดข้อมูลจะมีอยู่ทั้งหมดอยู่ 3 รูปแบบ ได้แก่ ภายนอก (spatial) เสียง (auditory) และการเคลื่อนไหว (temporal) นอกจากนี้ชุดข้อมูลนี้นั้นไม่รวมเพียงแค่การกระทำการของมนุษย์เท่านั้น ยังรวมไปถึง สัตว์ สิ่งของ และ ปรากฏการณ์ธรรมชาติ ทำให้ ชุดข้อมูลนี้เป็นการท้าทายรูปแบบใหม่ เพราะด้วยข้อมูลที่มีความซับซ้อนมากขึ้น เช่น การสร้างโมเดลที่สามารถอภิปรายการกระทำ ได้ถึงแม้ว่าสิ่งที่เราสนใจ (มนุษย์ สัตว์ สิ่งของ หรือปรากฏการณ์ธรรมชาติ) จะแตกต่างกัน เป็นต้น

1. รายละเอียดชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนับสนุนให้การเรียนรู้ในวิดีโอด้วยการกระทำการที่เกิดขึ้นในวิดีโอ เช่น การกระทำการของคนหรือสัตว์ เหตุการณ์ และปรากฏการณ์ธรรมชาติ
- (b) จำนวนของวิดีโอ : มากกว่า 1,000,000 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 3 วินาที
- (d) จำนวนของหมวดหมู่ : 339 หมวดหมู่
- (e) วิธีการเก็บรวบรวมข้อมูล
 - i. เริ่มจากการรวบรวมคำที่ใช้อยู่ทั่วไปในชีวิตประจำวันมา 4,500 คำ จาก VerbNet^[7] เว็บไซต์ที่เก็บรวบรวมคำกริยาภาษาอังกฤษขนาดใหญ่ จากนั้นนำมาแบ่งกลุ่มคำที่มีความหมายใกล้เคียงกันโดยใช้คุณลักษณะจาก Propbank^[7] และ FrameNet^[7] โดยเก็บข้อมูลเป็นแบบเวกเตอร์คุณลักษณะฐานสอง (binary feature vector) ซึ่งถ้าคำใดมีความเกี่ยวข้องกับคุณลักษณะใดก็จะให้ค่าเป็น 1 ถ้าไม่เกี่ยวข้องกันจะให้ค่าเป็น 0 จากนั้นจึงใช้วิธี k-means clustering ในการแบ่งกลุ่ม เมื่อแบ่งกลุ่มแล้วจากนั้นจะเลือกคำจากในแต่ละกลุ่มนั้น โดยคำที่เลือกมาจะเป็นคำที่ใช้บ่อยที่สุดในกลุ่มนั้น และลบคำน้ออกจากกลุ่มอื่นๆ ทั้งหมด (คำๆ หนึ่งสามารถอยู่ได้หลายกลุ่ม) จากนั้นจะทำการวนการนี้ไปเรื่อยๆ แต่คำที่เลือกมาจะต้องไม่มีความหมายคลุมเครือ หรือเป็นสิ่งที่ไม่สามารถมองเห็นหรือได้ยินได้ และต้องไม่มีความหมายเหมือนกับคำที่เคยเลือกมาก่อน จนสุดท้ายแล้วได้ออกมาที่ 339 หมวดหมู่
 - ii. ต่อมาทำการหาชุดข้อมูลวิดีโอด้วยทัศนคติที่ตัดออกมากเพียง 3 วินาทีที่เกี่ยวข้องกับคำใน 339 หมวดหมู่ที่เลือกมาจากวิดีโอแหล่งต่างกัน 10 แหล่ง การตัดวิดีโอนั้นจะไม่ใช้พวก Video2Gif (โมเดลที่ระบุตำแหน่งของสิ่งที่น่าสนใจในวิดีโอ) เพราะจะทำให้เกิดอคติขึ้นจะเกิดขึ้นตอนสร้างโมเดล ดังนั้นจึงใช้มนุษย์ในการตัดวิดีโอ จากนั้นจะทำการส่งข้อมูลของคำ และวิดีโอที่ตัดไปยัง Amazon Mechanical Turk (AMT หรือตลาดแรงงาน) เพื่อทำการสร้างคำกำกับโดยพนักงานของ AMT ทำให้ได้ 64 วิดีโอที่เกี่ยวข้องกับคำนั้น และอีก 10 วิดีโอที่มีคำกำกับอยู่แล้ว โดยวิดีโอด้วยวิดีโอด้วยที่มีคำกำกับอยู่แล้วนั้นถ้าพนักงานของ AMT ตอบเหมือนกันเกิน 90% ถึงจะนำเข้าไปรวมกับชุดข้อมูลส่วนอีก 64 วิดีโอ ถ้าเป็นชุดข้อมูลสำหรับสร้างโมเดลจะต้องผ่านพนักงานของ AMT อย่างน้อย 3 ครั้ง และต้องมีคำกำกับเหมือนกัน 75% ขึ้นไปถึงจะถือว่าเป็นคำกำกับที่ถูกต้อง ถ้าเป็นชุดข้อมูลสำหรับตรวจสอบ และชุดข้อมูลสำหรับทดสอบ จะต้องผ่านพนักงานของ AMT อย่างน้อย 4 ครั้ง และต้องมีคำกำกับเหมือนกัน 85% ขึ้นไป เหตุผลที่ไม่ต้องเกณฑ์ไว้ที่ 100% เพราะจะทำให้วิดีโอนั้นยากเกินไปที่จะทำให้สามารถจำการกระทำได้

2. การเตรียมข้อมูล

- (a) ชุดข้อมูลสำหรับสร้างโมเดลจะมี 802,264 วิดีโอ และมีวิดีโອนในแต่ละหมวดหมู่อยู่ที่ 500 ถึง 5,000 วิดีโอ
- (b) ชุดข้อมูลสำหรับตรวจสอบความต่อเนื่องจะมี 33,900 วิดีโอ และมีวิดีโອนในแต่ละหมวดหมู่อยู่ที่ 100 วิดีโอ
- (c) แยกเฟรม RGB ออกมาจากวิดีโอ และทำการเปลี่ยนขนาดให้เป็น 340×256 pixel
- (d) ใช้อัลกอริทึม TVL1 optical flow จาก OpenCV เพื่อลดข้อมูลรูบกวนที่จะเกิดขึ้น
- (e) ทำการแปลงค่าที่อยู่ใน optical flow ให้เป็นเลขจำนวนเต็มเพื่อทำให้การคำนวณนั้นเร็วขึ้น
- (f) ปรับค่า displacement ใน optical flow ให้ค่าสูงสุดเป็น 15 ต่ำสุดเป็น 0 และทำการปรับขนาดให้เป็นช่วง 0 - 255
- (g) เก็บข้อมูลออกแบบในรูปแบบของภาพขาวดำเพื่อลดพื้นที่ในการเก็บข้อมูล
- (h) แก้ปัญหาเรื่องการเคลื่อนไหวของกล้องด้วยการนำค่าเฉลี่ยของเวกเตอร์ไปลบกับ displacement
- (i) สุดท้ายจะเป็นสุ่มตัดภาพออกแบบเพื่อเพิ่มจำนวนข้อมูล

3. โมเดลปัญญาประดิษฐ์

- (a) ในงานวิจัยนี้มีการทดสอบโมเดลปัญญาประดิษฐ์หลายรูปแบบ โดยโมเดลปัญญาประดิษฐ์ที่มีประสิทธิภาพการทำงานที่ดีที่สุด 5 ลำดับแรกมีดังนี้
 - i. SVM มีรูปแบบข้อมูลที่ป้อนเข้า คือ เฟรมที่ต่อเนื่อง (spatial) + เฟรมเดี่ยว (temporal) + ข้อมูลเสียง (auditory)
 - ii. I3D มีรูปแบบข้อมูลที่ป้อนเข้า คือ เฟรมที่ต่อเนื่อง + เฟรมเดี่ยว
 - iii. TRN-Multiscale มีรูปแบบข้อมูลป้อนเข้า คือ เฟรมที่ต่อเนื่อง + เฟรมเดี่ยว
 - iv. TSN-2stream มีรูปแบบข้อมูลป้อนเข้า คือ เฟรมที่ต่อเนื่อง + เฟรมเดี่ยว
 - v. ResNet50-ImageNet มีรูปแบบข้อมูลป้อนเข้า คือ เฟรมที่ต่อเนื่อง
- (b) เครื่องมือที่ใช้วัดผลงานวิจัยนี้
 - i. Classification accuracy Top-1, Top-5
- (c) ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ในปัจจุบัน
 - i. ทำการทดสอบด้วยวิธี cross dataset transfer โดยการนำโมเดล ResNet50 I3D ที่สร้างด้วยชุดข้อมูล Kinetics และ Moments in Time แล้วนำหัว 2 โมเดลไปทดสอบกับชุดข้อมูลอื่น โดยจะปรับอัตราความถี่ของเฟรม (frame rate) ของวิดีโอให้เป็น 5 fps

Pretrained	Fine-Tuned		
	UCF-101	HMDB-51	Something Something
Kinetics	Top-1 : 92.6 Top-5 : 99.2	Top-1 : 62.0 Top-5 : 88.2	Top-1 : 48.6 Top-5 : 77.9
Moments	Top-1 : 91.9 Top-5 : 98.6	Top-1 : 65.9 Top-5 : 89.3	Top-1 : 50.0 Top-5 : 78.8

ตารางที่ 2.5: ประสิทธิภาพของโมเดล Resnet50 I3D ที่ใช้ชุดข้อมูล Kinetics และ Moments in Time

- ii. จะเห็นได้ว่า Kinetics ให้ผลลัพธ์ที่ดีกว่าใน UCF-101 เพราะว่ามีหมวดหมู่ที่ตรงกันอยู่หลายอย่าง ในขณะที่ HMDB-51 นั้นมีการรวมข้อมูลจากหลายแหล่ง และมีจำนวนหมวดหมู่ที่หลากหลายจึงทำให้มีความไม่คล้ายกับตัวข้อมูลของ Moments in Time ดังนั้นจึงเทียบผลลัพธ์จาก Something Something ซึ่งจะทำให้เห็นว่า Moments in Time มีประสิทธิภาพที่ดีกว่าและวิดีโอที่มีความยาวมากกว่า 3 วินาทีจะไม่ส่งผลกระทบกับประสิทธิภาพของ Moments in Time

4. ปัญหาที่พบ

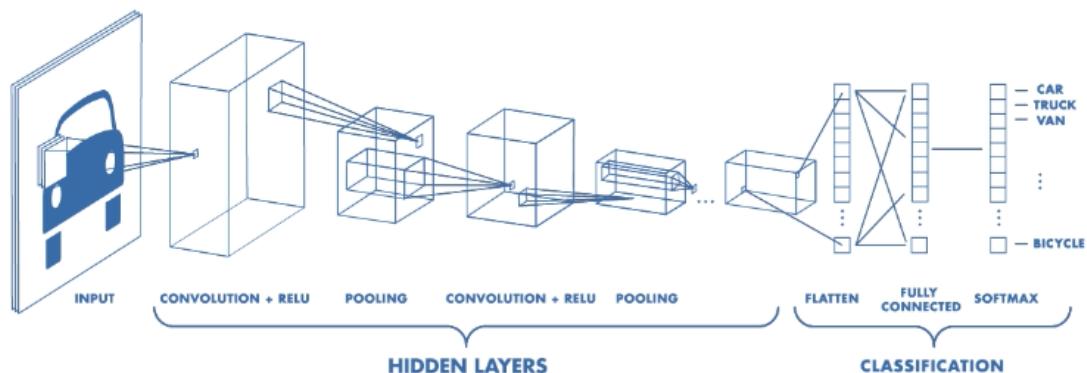
ผลลัพธ์จากการทำนายด้วยโมเดลถ้าผ่านรูปภาพที่มีรายละเอียดเยอะจะทำให้การ ทำนายโอกาสผิดนั้นค่อนข้างสูง ซึ่งปัญหานี้สามารถทำให้เกิดน้อยลงด้วยการนำวิธี class activation mapping (CAM) จะเป็นการเน้นรูปภาพในส่วนที่มีข้อมูลมากที่สุดและทำนายผลออกมา แต่ก็ยังมีจุดที่เป็นปัญหาอยู่ เช่น การกระทำที่เกิดขึ้นเรื่องมากจะทำให้การทำนายนั้นมีโอกาสผิดสูงขึ้น

2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

2.2.1 โมเดลปัญญาประดิษฐ์

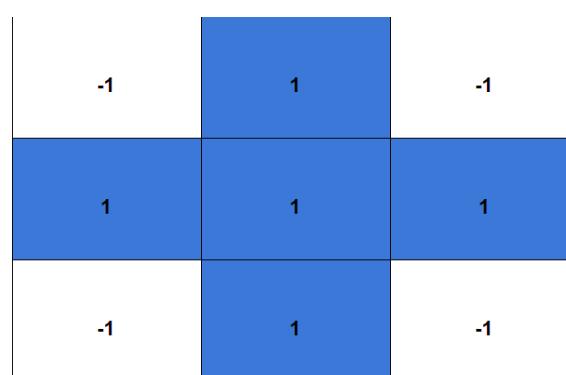
CNN

Convolution Neural Network (CNN) คือโมเดลปัญญาประดิษฐ์ประเภทหนึ่ง ซึ่งมักจะนำมาใช้กับงานที่เกี่ยวข้องกับรูป โดยการดึงจุดเด่นของภาพออกมามา เพื่อใช้สำหรับการจำแนกประเภทของสิ่งต่าง ๆ



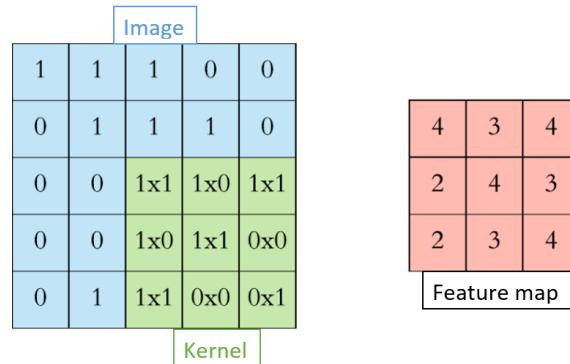
รูปที่ 2.4: ตัวอย่างโครงสร้างของ CNN

ซึ่งการคำนวณ CNN นั้นจะเริ่มจากการแยกคุณลักษณะออกจากรูป โดยการใช้เคอร์เนล (kernel) ซึ่งการที่เคอร์เนลมีลักษณะไม่มีเหมือนกันนั้น จะทำให้คุณลักษณะที่ต้องออกแบบจากรูปแตกต่างกัน โดยปกติแล้ว เคอร์เนลจะมีหลายแบบ เพราะ ใช้สำหรับการหาคุณลักษณะที่มีรูปแบบต่าง ๆ รูปแบบของเคอร์เนลจะเป็นตารางสองมิติที่มีขนาดขึ้นอยู่กับผู้สร้างที่จะออกแบบ รูปด้านล่างจะเป็นรูปตัวอย่างของเคอร์เนล



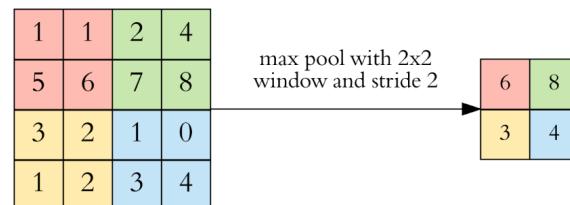
รูปที่ 2.5: ตัวอย่างเคอร์เนล

เมื่อนำเครื่องเรนเลิปทาบกับรูปจะทำให้สามารถถึงคุณลักษณะของมาได้ และเลื่อนตัวเครื่องเรนเลิปยังพิกเซลต่อไปจนครบทั้งรูป ซึ่งการเลื่อนนั้นขึ้นอยู่กับผู้สร้างว่าต้องการจะให้เลื่อนเท่าไหร แต่ระยะการเลื่อนที่มากขึ้นจะทำให้ความเกี่ยวข้องของคุณลักษณะที่ได้ออกมาน้อยลงด้วย โดยการวางแผนเครื่องเรนเลิปที่แบบรูปนี้จะวางแผนเครื่องเรนเลิปไม่ให้เกินกรอบรูป แต่ถ้าต้องการทاบทาเครื่องเรนเลิปทุกพิกเซลในรูป สามารถทำได้โดยการพื้นที่เกินขอบรูปเท่ากับ 0 ได้ เป็นต้น คุณลักษณะที่ได้ออกมาทั้งหมดจะเรียกว่าผังคุณลักษณะ ตามรูปด้านล่างดังนี้



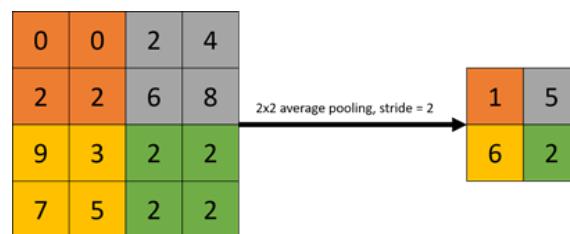
รูปที่ 2.6: ตัวอย่างการหาผังคุณลักษณะ

นอกจากนี้การทำเครื่องเรนเลิปมีการทำอีกแบบนึงซึ่งเรียกว่าการทำ pooling มีความสามารถในการลดขนาดของรูปภาพแบบนึง ซึ่งนิยมใช้กันอยู่สองประเภทได้แก่ max pooling และ average pooling โดยที่ max pooling เมื่อนำไปทาบกับรูป จะหาค่าที่มากที่สุดของมา ตัวอย่างตามรูปด้านล่างดังนี้



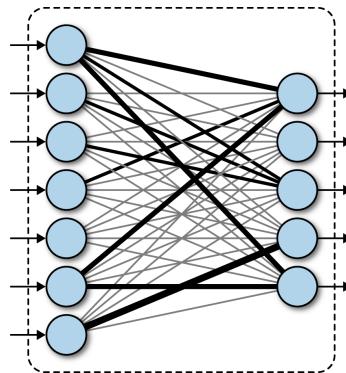
รูปที่ 2.7: ตัวอย่างการทำ max pooling

ในขณะที่ average pooling เมื่อนำไปเทียบกับรูป จะหาค่าเฉลี่ยของบริเวณที่เทียบของมาตามรูปด้านล่างดังนี้



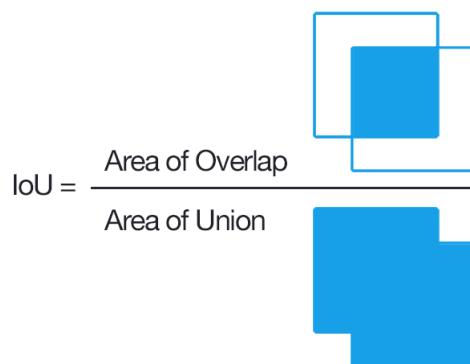
รูปที่ 2.8: ตัวอย่างการทำ average pooling

ในขั้นตอนการทำนาย ขั้นที่ชื่อว่า Fully connected ในขั้นนี้จะเป็นขั้นที่นิวรอนทุกตัวเชื่อมต่อกัน ซึ่งในงานประเกทตรวจจับวัตถุมักใช้ขั้นนี้ในการทำนายผลความน่าจะเป็นของแต่ละหมวดหมู่



รูปที่ 2.9: โครงสร้างของ fully-connected

การประเมินผล Intersection Over Union (IoU)



รูปที่ 2.10: ตัวอย่างการเคลื่อนที่ของลูกบอล

เป็นหนึ่งวิธีในการประเมินผลการทดลองสำหรับการตรวจจับวัตถุ โดยหลักการของการคำนวณ IoU สำหรับการประเมินผลการตรวจจับวัตถุ คือ การนำกรอบสีเหลี่ยมจริงของเฟรม และ กรอบสีเหลี่ยมที่ทำนายขึ้นมา มาหาอัตราส่วนระหว่าง พื้นที่ที่กรอบสีเหลี่ยมทั้งสองทับซ้อนกัน และ พื้นที่ทั้งหมดของกรอบสีเหลี่ยมทั้งสองรวมกัน ผลลัพธ์จะได้เป็นค่า IoU ซึ่งจะมีสมการดังนี้

$$IoU(P, G) = \frac{|P \cap G|}{|P \cup G|} \quad (2.7)$$

โดยที่

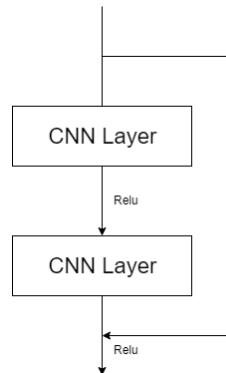
IoU = ค่าที่ใช้สำหรับวัดผลความใกล้เคียงระหว่างสองกรอบสีเหลี่ยม

P = พื้นที่ของกรอบสีเหลี่ยมที่ทำนายได้

G = พื้นที่ของกรอบสีเหลี่ยมจริงของรูปภาพ

ResNet

ในการสร้างโมเดลปัญญาประดิษฐ์นั้นการใช้จำนวนชั้น (layer) เยอะนั้นจะทำให้ได้คุณลักษณะของข้อมูลที่ออกมากเยอะตามไปด้วย แต่การที่คุณลักษณะของข้อมูล曳ะไม่ได้หมายความว่าโมเดลปัญญาประดิษฐ์จะให้ประสิทธิภาพที่ดีเสมอไป ซึ่งสามารถแก้ปัญหานี้ได้โดยใช้ residual network (ResNet) ซึ่งเป็น convolution neuron network (CNN) ประเภทหนึ่ง ที่ส่วนใหญ่จะนำมายังกับข้อมูลที่เป็นรูปภาพ เช่น การจดจำวัตถุ เป็นต้น โดย ResNet นี้จะสามารถทำการข้ามชั้นที่ไม่จำเป็นได้ การข้ามชั้นที่ไม่จำเป็นจะช่วยลดเวลาที่ใช้ในการสร้างโมเดลปัญญาประดิษฐ์ และทำให้ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ดีขึ้น



รูปที่ 2.11: หลักการของ Residual block ของ ResNet

การทดลองโมเดลปัญญาประดิษฐ์ ResNet ด้วยการทำจำแนกรูปภาพโดยใช้ชุดข้อมูลทดสอบ ImageNet ที่มีหมวดหมู่มากกว่า 1,000 หมวดหมู่ มาเทียบกับโมเดลปัญญาประดิษฐ์ทั่วไป (plain model) ที่จำนวนชั้น 18 ชั้น และ 34 ชั้น โดยโครงสร้างพื้นฐานของโมเดลปัญญาประดิษฐ์ ResNet และโมเดลปัญญาประดิษฐ์ทั่วไปเหมือนกัน ซึ่งผลลัพธ์อัตราเรื่อยละเอียดของความผิดพลาดจะได้ออกมาตามตารางที่ 2.6

จำนวนชั้นของโมเดลปัญญาประดิษฐ์	Training error	
	plain	ResNet
18	27.94	27.88
34	28.54	25.03

ตารางที่ 2.6: อัตราเรื่อยละเอียดของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet

จากตาราง 2.6 จะเห็นได้ว่าโมเดลปัญญาประดิษฐ์ทั่วไป 34 ขั้นมีค่าอัตราเร้อยละของความผิดพลาดสูงกว่า โมเดลปัญญาประดิษฐ์ ResNet ได้อย่างชัดเจน ในขณะที่โมเดลปัญญาประดิษฐ์ทั่วไปจะมีอัตราเร้อยละของความผิดพลาดสูงขึ้นเมื่อเทียบกันระหว่าง 18 ขั้นและ 34 ขั้น

ต่อมาจะนำโมเดลปัญญาประดิษฐ์ ResNet มาทดสอบกับชุดข้อมูล CIFAR-10 ซึ่งเป็นชุดข้อมูลที่มีรูปสำหรับใช้สร้างโมเดลปัญญาประดิษฐ์ 50,000 รูป รูปสำหรับทดสอบ 10,000 รูป และมีจำนวนหมวดหมู่ทั้งหมด 10 หมวดหมู่ โดยจะมีการออกแบบของจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet ตามจำนวนของชั้น convolution ที่มีผังคุณลักษณะเท่ากัน 6 ชั้นติดกันและการข้ามชั้นทีละ 2 ชั้น จึงทำให้ได้รูปแบบการคิดชั้นดังนี้ $6n + 2$ สำหรับการทดสอบจะให้ค่า $n = [3, 5, 7, 9, 200]$ ดังตารางต่อไปนี้

โมเดลปัญญาประดิษฐ์	จำนวนชั้น	Training error
ResNet	20	8.75
ResNet	32	7.51
ResNet	44	7.17
ResNet	56	6.97
ResNet	110	6.43
ResNet	1202	7.93

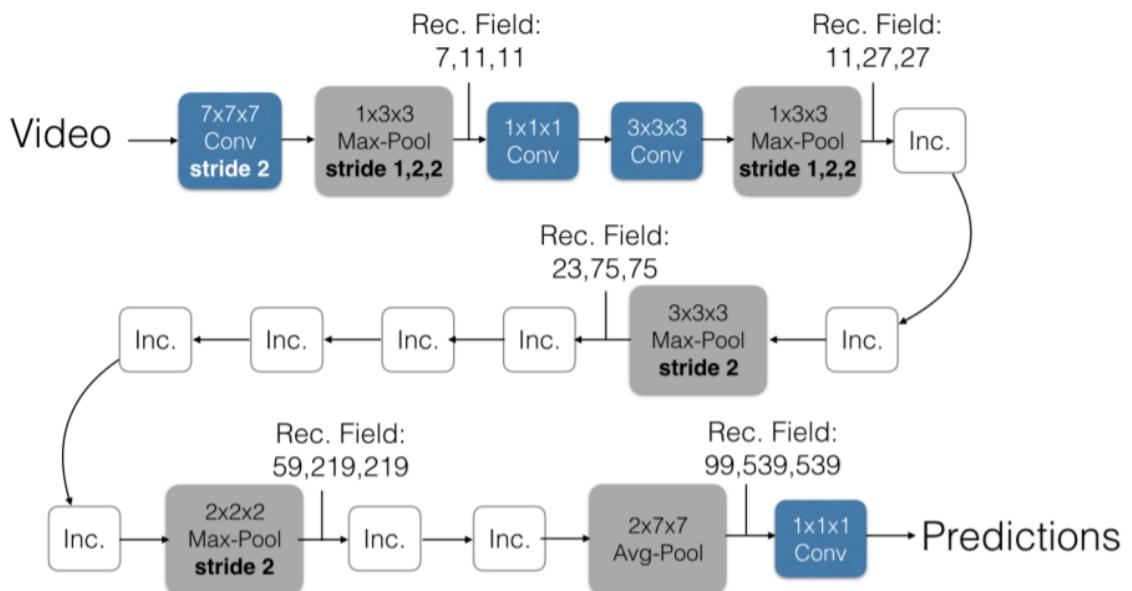
ตารางที่ 2.7: ค่าความผิดพลาดที่ได้จากการทดลองจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10

จากตาราง 2.7 จะเห็นได้ว่าที่โมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1,202 นั้นมีค่าความผิดพลาดเกิดขึ้นมากกว่าจำนวนชั้น 110 ซึ่งอาจจะเป็นไปได้ว่าขนาดของโมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1,202 นั้นมากเกินไปสำหรับชุดข้อมูลขนาดเล็กนี้

Inflated 3D convolutional network

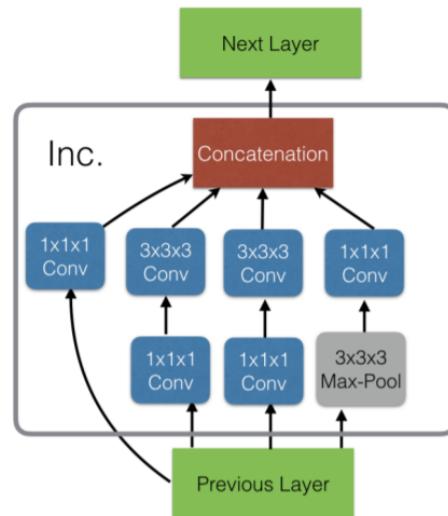
ในการพัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์นั้นมีพื้นฐานจากการจำแนกวัตถุ (object classification) หมายถึงการใช้รูปภาพหนึ่งรูปในการประมวลผลและคำนวณอุกมาดาว่าภายในรูปนั้นมีบริบทการกระทำอย่างไร โดยไม่ได้คำนึงถึงข้อมูลเชิงต่อเนื่อง (spatio-temporal information) จากบทความ ”Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”^[7] นั้นได้พัฒนาโครงสร้างของโมเดลปัญญาประดิษฐ์ (architecture) ที่มีประสิทธิภาพในการประมวลผลภาพเคลื่อนไหวได้ชื่อว่า I3D หรือ inflated 3D-convolution network โดยโครงสร้างพื้นฐานของ I3D นั้นมาจาก Inception-v1^[7] ที่ถูกพัฒนาโดย Google ซึ่งเป็นโครงสร้างที่มีประสิทธิภาพสูงในการจำแนกวัตถุในรูปภาพ แล้ว I3D นั้นได้ทำการขยายมิติของโครงสร้างจาก 2 มิติ เป็น 3 มิติ เพื่อให้โมเดลปัญญาประดิษฐ์สามารถเรียนรู้ข้อมูลเชิงต่อเนื่องได้

Inflated Inception-V1



รูปที่ 2.12: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[7]

Inception Module (Inc.)



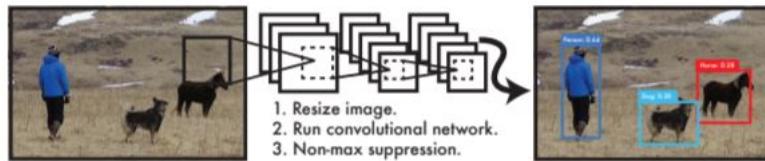
รูปที่ 2.13: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[?]

ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อเทียบกับ long-short term memory (LSTM), 3D-convolution network, two-stream และ 3D-fused โดยใช้เครื่องมือในการวัดผลคือ Top@1 accuracy ตามตารางที่ 2.8

Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
LSTM	81.0	–	–	36.0	–	–	63.3	–	–
3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
Two-Stream I3D	84.5	90.6	93.4	49.8	61.9	66.4	71.1	63.4	74.2

ตารางที่ 2.8: ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อใช้ข้อมูลจาก UCF-101, HMDB-51 และ Kinetics ในการสร้างและทดสอบด้วยเครื่องมือวัดผลแบบความแม่นยำจากการทำนายอันดับแรกสุด

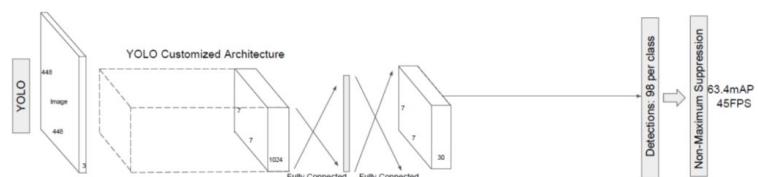
YOLO



รูปที่ 2.14: กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO

โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เป็นโครงสร้างที่มีความเร็วมาก มีความเร็วในการประมวลผลถึง 45 เฟรมต่อวินาที ทำให้สามารถประมวลผลแบบเรียลไทม์ได้ นอกจากนั้นยังมีความแม่นยำ mAP มากกว่าโมเดลสำหรับตรวจจับวัตถุอื่นๆ ถึง 2 เท่า ซึ่งเหตุผลที่โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เร็วกว่าโมเดลปัญญาประดิษฐ์ตัวอื่นๆ เนื่องจาก มีแนวคิดที่ต่างออกแบบไป คือ สำหรับการตรวจจับวัตถุในวิธีการก่อนหน้าจะใช้วิธีทำนายกรอบสี่เหลี่ยมก่อน และจึงค่อยนำกรอบสี่เหลี่ยมไปทำนายว่าเป็นหมวดหมู่อะไร ซึ่ง YOLO มีวิธีการที่ต่างออกแบบไป คือ ทำนายตำแหน่งของกรอบสี่เหลี่ยมและทำนายว่ากรอบสี่เหลี่ยมนั้นเป็นหมวดหมู่อะไรพร้อมกัน โดยใช้โครงข่ายประสาทแบบคอนโวลูชัน ด้วยแนวคิดนี้จึงเป็นที่มาของชื่อ YOLO หรือ you only look once การมองแค่เพียงครั้งเดียว

โครงสร้างของโมเดลปัญญาประดิษฐ์ของ YOLO

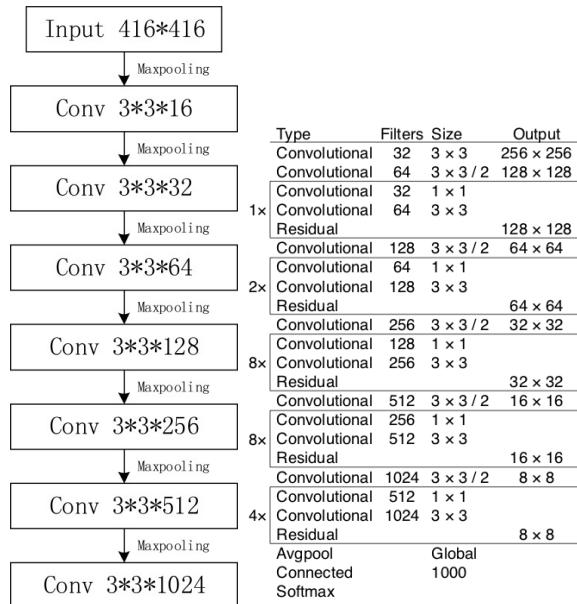


รูปที่ 2.15: โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ YOLO

จากรูปภาพที่ 2.15 จะเห็นได้ว่า YOLO ใช้โครงข่ายประสาทเทียมเพียงตัวเดียวซึ่งภายในโครงข่ายจะมีกระบวนการหลักๆ 3 อย่าง กระบวนการแรกคือการสกัดคุณลักษณะกระบวนการนี้จะมีจำนวนขั้นของเลเยอร์ที่แตกต่างกันไปตามความลึกของการสกัดแล้วแต่โมเดล ซึ่งจะมีตัวอย่างอยู่ในบทความด้านล่าง และ ขั้นตอนถัดมา คือ การทำนายผล หลังจากที่ได้คุณลักษณะมาแล้วจะนำไปทำนายผลผ่าน Fully connected ซึ่งจะได้ผลลัพธ์ออกเป็นหมวดหมู่และตำแหน่งของกรอบสี่เหลี่ยม ขั้นตอนสุดท้ายคือ การทำ NMS เพื่อให้ได้ผลลัพธ์ที่ดีที่สุดออกมานะ

ซึ่งโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO ที่ถูกใช้ในงานวิจัยนี้ประกอบไปด้วย 1) YOLOv3-tiny 2) YOLOv3 3) YOLOv3-spp ซึ่งทั้ง 3 โครงสร้างจะมีความแตกต่างของโครงสร้างดังนี้

1. YOLOv3-tiny ใช้ Max-Pooling layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
2. YOLOv3 ใช้ Convolutional layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
3. YOLOv3-spp ใช้ Convolutional layers+ฟีเจอร์ที่ดีที่สุดของ Max-Pooling layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง



(ก) โครงสร้างโมเดล

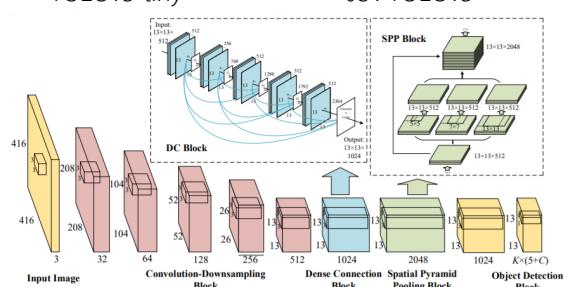
ปัญญาประดิษฐ์ของ

YOLOv3-tiny

Table 1. Darknet-53.

(ข) โครงสร้างโมเดลปัญญาประดิษฐ์

ของ YOLOv3

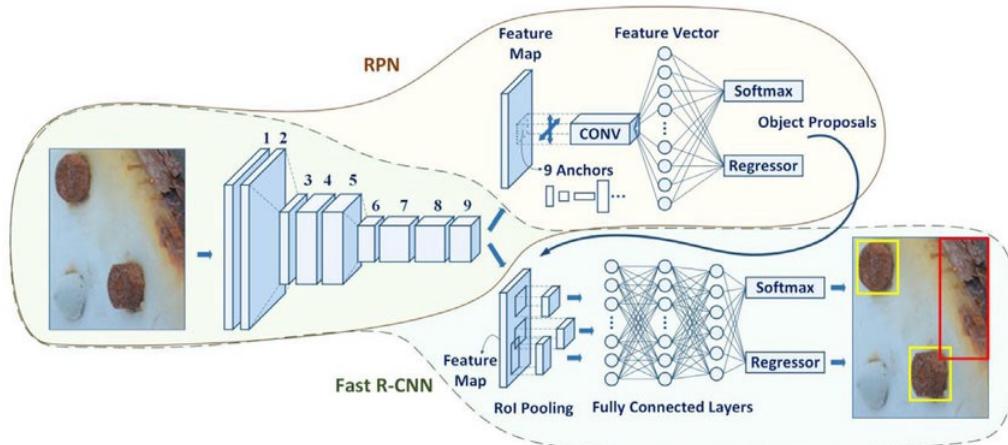


11

(ค) โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLOv3-spp

รูปที่ 2.16: โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO

Faster-RCNN



รูปที่ 2.17: โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ Faster RCNN

faster-rcnn มีการพัฒนาในการหาพื้นที่ที่สนใจ (ROI) โดยการเปลี่ยนจากใช้โครงข่ายหาพื้นที่ที่สนใจแยก เฉพาะ (selective search) นำมารวมในโครงข่ายเดียวกัน ดังนั้น faster-rcnn จึงมีโครงข่ายประสาทเทียมเดียว ในการทำงาน ซึ่งภายในโครงข่ายจะประกอบไปด้วยการทำงานหลัก 3 อย่าง คือ

1. การสกัดคุณลักษณะ
นำรูปภาพทั้งรูปภาพเข้าโครงข่ายคอนโวลูชันเพื่อการสกัดคุณลักษณะของรูปภาพ
2. การเสนอพื้นที่ที่คาดว่าจะมีวัตถุอยู่
หลังจากที่รูปภาพผ่านการสกัดคุณลักษณะแล้ว จะถูกนำเข้าไปใน region proposal network เพื่อสร้างข้อเสนอพื้นที่ที่คาดว่าจะมีวัตถุอยู่
3. การทำนายผล
ทำการ pooling คุณลักษณะของรูปภาพและพื้นที่ที่คาดว่าจะมีวัตถุอยู่ และ นำเข้าไปในขั้นการทำนายผล (full connected layer) สุดท้ายจะได้ผลลัพธ์เป็นหมวดหมู่ของกรอบสีเหลี่ยม และ ตำแหน่งของกรอบสีเหลี่ยม

region proposal network (RPN) คือ โครงข่ายที่เสนอพื้นที่ที่คาดว่าจะมีวัตถุอยู่ จะถูกใช้หลังรูปภาพผ่านการสกัดคุณลักษณะ RPN มีโครงสร้างที่เป็นเอกลักษณ์เฉพาะตัว คือมีการบอกว่าบริเวณนั้นมีวัตถุอยู่หรือไม่ (classification layer) และ สำหรับการระบุพิกัดของกรอบสีเหลี่ยมที่คาดว่าจะมีวัตถุอยู่ (regression layer) ซึ่งผลลัพธ์จะได้ ROI (พื้นที่บริเวณที่เราสนใจ)

2.2.2 เครื่องมือกำกับคุณลักษณะ

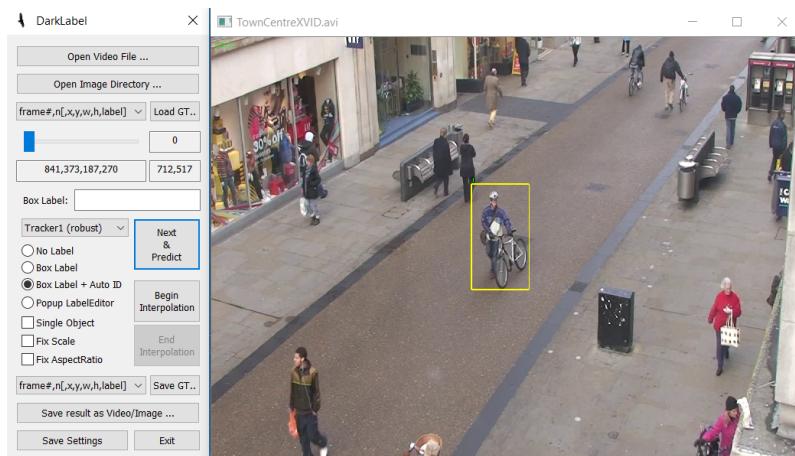
จากการค้นคว้าหาเครื่องมือในการสร้างคำกำกับข้อมูลเพื่อใช้เป็นแนวทางในการออกแบบเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ พบรเครื่องมือที่เปิดให้ใช้งานสาธารณะ (open source) 2 เครื่องมือ คือ DarkLabel และ OpenLabeling โดยสรุปข้อสำคัญได้ดังนี้

โปรแกรม DarkLabel

เป็นโปรแกรมที่ช่วยในการทำนายคำกำกับและบันทึกในรูปแบบต่างๆ รองรับข้อมูลป้อนเข้าในรูปแบบไฟล์วิดีโอ avi, mp4 หรือกลุ่มรูปภาพ มีขั้นตอนการสร้างคำกำกับดังนี้

- สร้างกรอบสี่เหลี่ยมครอบบริเวณวัตถุที่สนใจโดยใช้มนุษย์เป็นคนสร้าง
- กดปุ่ม Next และ Predict อย่างต่อเนื่อง เพื่อทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดๆไป จนกระทั่งการเกิดข้อผิดพลาด
- ลบกรอบสี่เหลี่ยมที่พลาด และเริ่มทำขั้นตอนที่ 1 ใหม่ อีกรังสีจักรบทุกเฟรมในวิดีโอ

หลังจากที่ผู้ใช้จัดตั้งค่าและเริ่มใช้โปรแกรม DarkLabel พบรฯ เป็นโปรแกรมที่การทำงานส่วนใหญ่เป็นการสร้างคำกำกับแบบใช้มนุษย์เป็นคนทำด้วยตัวเอง ซึ่งทำให้ใช้เวลาในการทำงาน



รูปที่ 2.18: UI ของโปรแกรม DarkLabel

โปรแกรม OpenLabeling

เป็นโปรแกรมที่ช่วยในการทำนายคำจำกัดนิยม โดยโปรแกรมจะมีการทำงานอยู่ 2 รูปแบบการทำงาน คือแบบทำด้วยตัวเอง (Mode Manual) และแบบอัตโนมัติ (Mode Auto) ซึ่งมีการทำงานแยกกันอย่างชัดเจน

1. การทำงานแบบอัตโนมัติ

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดดังนี้

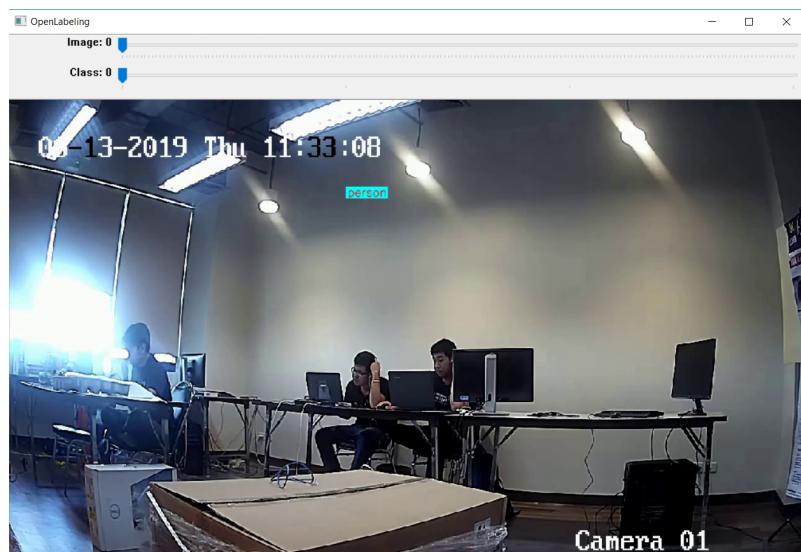
- (a) โปรแกรมจะทำงานอัตโนมัติโดยใช้โมเดลปัญญาประดิษฐ์ในการทำนายคีย์เฟรม (keyframe) และทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดไปด้วยอัลกอริทึมที่ใช้การคำนวนคณิตศาสตร์และการประมวลผลภาพในภาพที่เหลือ ผลลัพธ์ที่ได้คือรูปภาพและไฟล์คำจำกัดกับคุณลักษณะ

2. การทำงานแบบทำด้วยตัวเอง

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดดังนี้

- (a) สร้างกรอบสี่เหลี่ยมขึ้นมาโดยใช้มนุษย์เป็นคนสร้าง
- (b) กดปุ่มเพื่อทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดไป จนกว่าทั้งเกิดข้อผิดพลาด
- (c) ลบกรอบสี่เหลี่ยมที่พลาด และเริ่มทำขั้นตอนที่ 1 อีกครั้งจนครบทุกเฟรมในวิดีโอ

หลังจากที่ได้ทดลองใช้โปรแกรม OpenLabeling ทั้ง 2 รูปแบบการทำงานแล้วพบว่า การทำงานแบบอัตโนมัติ ไม่สามารถปรับแก้ไขสิ่งใดในระหว่างกระบวนการนั้น ทำให้หากเกิดกรณีที่โมเดลทำนายกรอบสี่เหลี่ยมพลาดหรือเกินมา จะไม่สามารถแก้ไขได้ และการทำงานแบบทำด้วยตัวเองไม่มีระบบตรวจสอบกรอบสี่เหลี่ยม ทำให้ผู้ใช้งานจะต้องสร้างกรอบสี่เหลี่ยมขึ้นมาเอง

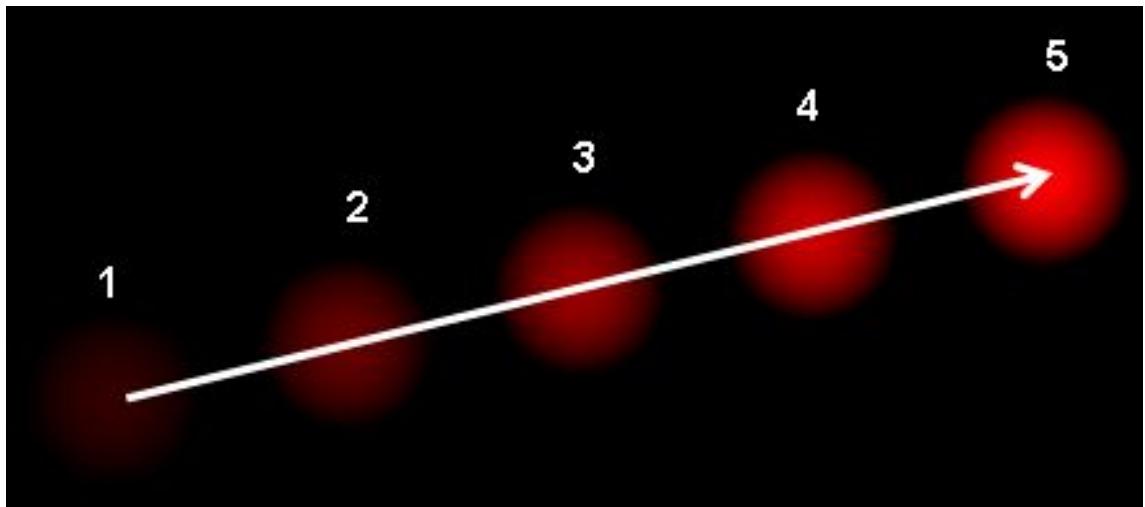


รูปที่ 2.19: UI ของโปรแกรม OpenLabeling

2.3 ทฤษฎีเกี่ยวข้อง

2.3.1 Optical flow

Optical flow^[7] คือการแปลงการเคลื่อนที่ของวัตถุในระหว่างสองรูปภาพซึ่งอาจจะเกิดจากการเคลื่อนที่ของวัตถุหรือตัวกล้องอุปกรณ์ในรูปแบบของเวกเตอร์ 2 มิติ โดยที่เวกเตอร์แต่ละตัวจะแสดงถึงทิศทางการเคลื่อนที่ของวัตถุหรือบุคคลระหว่างภาพดังรูปด้านล่าง



รูปที่ 2.20: ตัวอย่างการเคลื่อนที่ของลูกบอล

จากรูปภาพจะแสดงให้เห็นถึงการเคลื่อนที่ของลูกบอลในภาพที่ต่อเนื่องกัน 5 ภาพโดยที่ลูกครรภ์แสดงถึงทิศทางการเคลื่อนที่ของเวกเตอร์

การทำงานของ optical flow อยู่บนสมมติฐานหลายประการได้แก่

- ความเข้มพิกเซล (pixel) ของวัตถุจะไม่เปลี่ยนแปลงระหว่างภาพที่ต่อเนื่องกัน
- พิกเซลที่อยู่ใกล้กันจะมีการเคลื่อนไหวที่คล้ายกัน

เมื่อพิจารณาพิกเซล $I(x,y,t)$ จากภาพแรกจะเคลื่อนไหวเป็นระยะทาง (dx,dy) ไปยังภาพต่อไปหลังจากผ่านไปแล้ว dt เวลา ดังนั้นเนื้องจากพิกเซลเหล่านี้เหมือนกัน และความเข้มไม่มีการเปลี่ยนแปลง จึงทำให้พูดได้ว่า

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (2.8)$$

โดยที่

I = พิกเซลจากภายในภาพ

x = ตำแหน่งของพิกเซลในแกน x

dx = ระยะทางที่เคลื่อนที่ในแกน x

y = ตำแหน่งของพิกเซลในแกน y

dy = ระยะทางที่เคลื่อนที่ในแกน y

t = เวลา

dt = ระยะเวลาที่เปลี่ยนไประหว่างภาพ

จากนั้นใช้การประมาณค่าของ taylor series ทางฝั่งขวาเมื่อและลบค่า common term แล้วหารด้วย dt เพื่อให้ได้สมการดังต่อไปนี้

$$f_x u + f_y v + f_t \quad (2.9)$$

$$f_x = \frac{\delta f}{\delta x}; f_y = \frac{\delta f}{\delta y} \quad (2.10)$$

$$u = \frac{\delta x}{\delta t}; v = \frac{\delta y}{\delta t} \quad (2.11)$$

โดยที่

f_x = เกรเดียน (gradient) ในแกน x

f_y = เกรเดียนในแกน y

f_t = เกรเดียนของเวลา

u = เวกเตอร์การเคลื่อนที่ของแกน x

v = เวกเตอร์การเคลื่อนที่ของแกน y

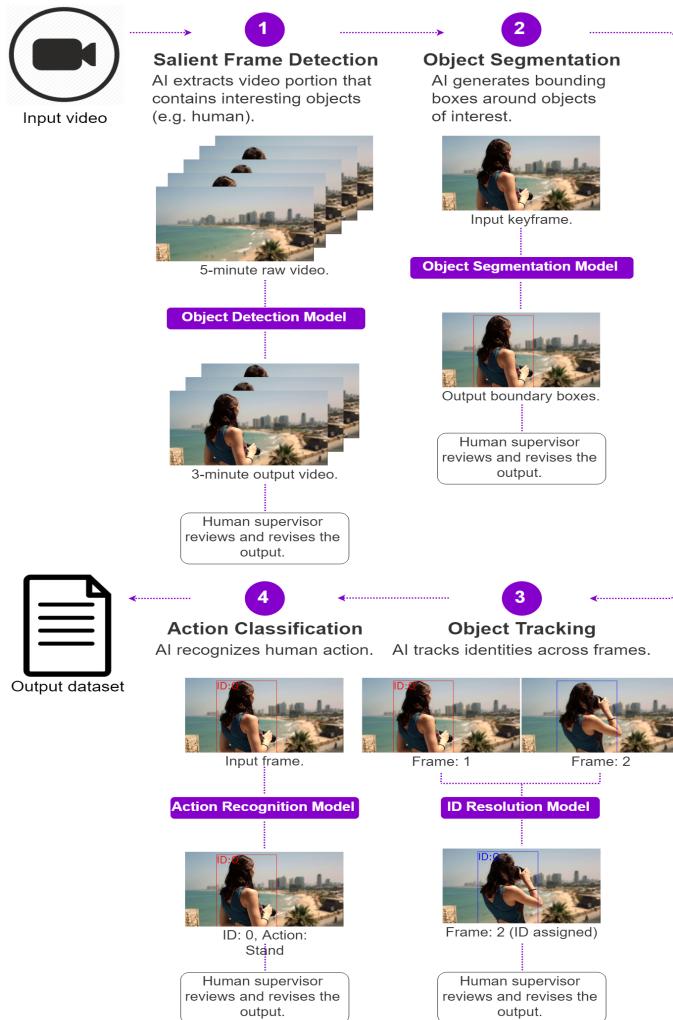
สมการข้างบนนี้จะเรียกว่าสมการ optical flow จากสมการทำให้สามารถหา f_x และ f_y โดยเป็นเกรเดียนของภาพ และ f_t เป็นเกรเดียนของเวลา แต่ n กับ b เป็นตัวแปรที่ไม่ทราบ ทำให้สมการนี้ไม่สามารถแก้ไขโดยมีตัวแปรที่ไม่ทราบถึง 2 ตัว จึงมีการนำวิธีการต่างๆเข้ามาใช้ในการแก้ปัญหานี้ โดยวิธีการที่นำเข้ามาใช้ในการแก้ปัญหาก็คือ dense optical flow ซึ่งใช้อัลกอริทึมของ Gunnar Farneback^[7] ซึ่งจะใช้วิธีการขยายพื้นที่ (polynomial expansion) เป็นวิธีการที่ทางคณะผู้วิจัยนำมาใช้งานในส่วนของการทำโมเดลปัญญาประดิษฐ์ Inflated 3D convolutional network

บทที่ 3

ระเบียบวิธีวิจัย

ในการทำโครงการวิจัยเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ จะมีการทำงานหลากหลายส่วนมาทำงานร่วมกัน ซึ่งต้องมีระเบียบวิธีวิจัยอธิบายถึงขั้นตอนการดำเนินงานตั้งแต่เริ่มศึกษาข้อมูลจนไปถึงสิ้นสุดกระบวนการวิจัย โดยใช้ภาษาไทยเป็นภาษาหลักในการเขียนโปรแกรม

3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.1: ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

3.2 ความต้องการของระบบ

3.2.1 ความต้องการเชิงการใช้งาน (functional requirements)

- เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถตัดวิดีโอช่วงเวลาที่ไม่มีมนุษย์อยู่ออกได้ อัตโนมัติโดยใช้ปัญญาประดิษฐ์
- เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์สามารถระบุตำแหน่งมนุษย์แต่ละคนในวิดีโอและจำแนก การกระทำของมนุษย์ในวิดีโอด้วยการกระทำที่กำหนดจะประกอบไปด้วย ยืน นั่ง นอน เล่นโทรศัพท์ เต้น กินข้าว พุดคุย
- ชุดข้อมูลที่ได้จากเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถนำไปใช้ในการพัฒนา โมเดลปัญญาประดิษฐ์ต่อได้
- สร้างระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ที่มีมนุษย์สามารถทำงานร่วมกับ ปัญญาประดิษฐ์ได้
- ระบบวิเคราะห์การกระทำการที่มนุษย์ต้องสามารถนำวิดีโอมาวิเคราะห์ข้อมูลการกระทำการและตำแหน่งของ มนุษย์แต่ละคน และนำข้อมูลเหล่านั้นไปสร้างรายงานที่มีคำกำกับบอกมาได้ โดยรายละเอียดรายงาน

จะมีดังนี้

- (a) เวลา (time stamp)
- (b) การกระทำ
- (c) ตำแหน่ง โดยจะบอกในลักษณะของกรอบสี่เหลี่ยมครอบพื้นที่ที่มนุษย์คนนั้นๆ ออู่

3.2.2 ความต้องการเชิงวิศวกรรม (non-functional requirements)

1. สร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์โดยใช้ภาษาไพธอน
2. ความละเอียดอย่างต่ำของวิดีโอต้องมากกว่า 640×480 (กว้าง x สูง)
3. วิดีโอจะต้องมีอัตราเฟรมต่อวินาที (fps) อย่างต่ำ 10 เฟรมต่อวินาที

3.3 หน้าที่ความรับผิดชอบ

ปฐมพงศ์ สินธุ์งาม สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดทำการกระทำมนุษย์ 3D รวมถึงออกแบบและสร้างระบบ Tracker

ศุภกร เบญจวิกรัย รวมฟังก์ชันและระบบต่างๆของแอพพลิเคชัน รวมถึงออกแบบและสร้างระบบ Select และ Detect

อุกฤษฎ์ เลิศวรรณาการ สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดทำการกระทำมนุษย์ Resnet-50 รวมถึงออกแบบและสร้างระบบ Person ReID

3.4 เครื่องมือที่ใช้ในงานวิจัย

ในหัวข้อนี้จะกล่าวถึงซอฟต์แวร์ ภาษา และ program library ที่ใช้ในการพัฒนาระบบ รวมถึงข้อมูลจำเพาะของคอมพิวเตอร์ที่ใช้ในการพัฒนาระบบ

Pycharm community 2017.1.2

เป็นโปรแกรมໄ่วยใช้สำหรับเขียนและแก้ไขโค้ดซึ่งข้อดีของโปรแกรมนี้ คือ มีคุณสมบัติต่างๆที่สามารถอำนวยความสะดวกในการเขียนโปรแกรมได้ เช่น syntax highlighting, auto-completion ฯลฯ และสามารถประมวลผล (compile) โปรแกรมทดสอบแอพพลิเคชันได้

Jupyter 2017.1.2

เป็นโปรแกรมสำหรับเขียนโปรแกรมที่เหมาะสมสำหรับใช้ในการทดสอบโปรแกรมแต่ละส่วนได้ ซึ่งมีข้อดีคือ หากมีการแก้ไขโปรแกรมเพียงแค่บางส่วน ก็สามารถปรับมาผลเฉพาะส่วนที่ต้องการได้มักจะใช้ในการสร้างโมเดลปัญญาประดิษฐ์

Qt Creator 4.9.2 (community)

เป็นเครื่องมือสำหรับออกแบบหน้าต่างแอพพลิเคชันของ library PyQt ซึ่งมีข้อดีคือ เรียกใช้ง่ายมีวิดเจ็ต (widget) ที่สามารถใช้ได้หลากหลายเหมาะสมสำหรับการออกแบบ

3.5 ภาษาที่ใช้ในการพัฒนาระบบ

ใช้ภาษาไพธอนในการพัฒนาเป็นหลัก เพราะเป็นภาษาที่ปัจจุบันมีการใช้กันอย่างแพร่ มีเครื่องมือและ library ที่อำนวยความสะดวกในการพัฒนาอย่างมาก ทั้งยังเป็นภาษาที่สามารถเข้าใจได้ง่าย โดยในการทำวิจัยครั้งนี้ได้เลือก python 3.6.8 มาใช้ในการพัฒนา เนื่องจากเป็นรุ่นที่รองรับการทำงานของ library Tensorflow 1.12 และ CUDA 9

3.6 Program library ที่ใช้ในการพัฒนาระบบและแอปพลิเคชัน

Library	Version	Description
numpy	1.16.4	library ใช้สำหรับการคำนวนและ array
pandas	0.24.2	library ใช้สำหรับการจัดการข้อมูลที่อยู่ในรูปแบบของ excel
opencv	4.1.0.25	library ใช้สำหรับการจัดการข้อมูลที่เป็นรูปภาพและวิดีโอ
pillow	6.0.0	library ใช้สำหรับการจัดการข้อมูลที่เป็นรูปภาพ
torchsummary	1.5.1	library ใช้สำหรับการวิเคราะห์โครงสร้างของโมเดล
pytorch	1.10.0	library ใช้สำหรับการสร้างปัญญาประดิษฐ์
torchvision	0.3.0	library ใช้สำหรับการสร้างปัญญาประดิษฐ์
scikit-learn	0.21.2	library ใช้สำหรับการสร้างปัญญาประดิษฐ์
scipy	1.3.0	library ใช้สำหรับการสร้างปัญญาประดิษฐ์
sklearn	0.0	library ใช้สำหรับการสร้างปัญญาประดิษฐ์
pickleshare	0.7.5	library ใช้สำหรับการทำรหัส (encoding) ไม่เดลปัญญาประดิษฐ์
tqdm	4.32.1	library ใช้สำหรับจัดการการทำงานซ้ำ (loop)
pyqt5	5.9.2	library ใช้สำหรับการทำแอปพลิเคชัน

3.7 แผนการดำเนินงาน

โดยจากที่กล่าวไปตอนต้นในบทนำการดำเนินงานและการออกแบบการสร้างเครื่องมือสำหรับกำกับข้อมูล ด้วยปัญญาประดิษฐ์ และระบบบุคลิค่าที่การกระทำการของมนุษย์ในวิดีโอ มีแผนการทำงานซึ่งถูกแบ่งออกเป็นสาม ขั้นตอนดังนี้ ขั้นตอนแรกคือ ขั้นตอนของการศึกษาหาความเป็นไปได้ รวมถึงเทคโนโลยีปัจจุบันที่เกี่ยวกับการ สร้างแอปพลิเคชัน และการจัดการกระทำการของมนุษย์ด้วยปัญญาประดิษฐ์ เพื่อนำมาประยุกต์ใช้กับงานวิจัย นี้ ขั้นตอนที่สองคือ ขั้นตอนของการออกแบบและสร้างแอปพลิเคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการเทรน ไม่เดลจากวิดีโอ ขั้นตอนที่สามคือ ขั้นตอนของการออกแบบและสร้างระบบบุคลิค่าที่การกระทำการของมนุษย์ได้โดย มีข้อกำหนดตามที่กล่าวไว้ในบทนำ ในการเริ่มทำงานวิจัยนี้นั้นสิ่งจำเป็นที่ต้องทำในอันดับแรกคือการศึกษาข้อมูล ในหัวข้อที่เกี่ยวข้อง หรืองานวิจัยอื่นที่ทำเอาระบบแล้ว เพื่อศึกษาและทำความเข้าใจ ข้อดี-ข้อเสีย ของเทคนิคหรือ กระบวนการต่างๆ เพื่อนำมาประยุกต์ใช้กับงานวิจัยนี้ ในการศึกษาเกี่ยวกับการออกแบบและสร้างแอปพลิ เคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการสร้างโมเดลจากวิดีโอ สิ่งที่ต้องให้ความสนใจคือฟังก์ชันการทำงาน การ ออกแบบและการจัดวางองค์ประกอบต่างๆในหน้าต่างแอปพลิเคชัน และความสะดวกในการใช้งาน จากนั้นจึงเริ่ม ศึกษาเกี่ยวกับ library ที่ใช้ในการสร้างแอปพลิเคชัน ส่วนการศึกษาเกี่ยวกับการสร้างระบบบุคลิค่าที่การกระทำ มนุษย์ จะมุ่งความสนใจไปที่ชุดข้อมูลสำหรับการวิเคราะห์วิดีโอ ไม่เดลสำหรับการวิเคราะห์วิดีโอ เทคนิคในการ สร้างโมเดล เทคโนโลยีในการระบบบุคลิค่าหัวใจ เพื่อใช้ในการออกแบบและสร้างระบบบุคลิค่าที่การกระทำ ของมนุษย์ในวิดีโอด้วยมีประสิทธิภาพ ในบทนี้จะกล่าวถึงกระบวนการออกแบบและการดำเนินการตามแผนที่วางแผนไว้

3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ ผู้วิจัยได้เลือกใช้ library PyQt และภาษา Python ในการพัฒนา เนื่องจาก PyQt นั้นเป็น library ที่มีผู้พัฒนาใช้กันอย่างแพร่หลาย จึงสะดวกในการศึกษา หาข้อมูลในการสร้างหรือแก้ไข อีกทั้งยังเป็น library ที่สามารถพัฒนาด้วยภาษา Python ได้ และใช้งานง่าย สามารถปรับปรุงแก้ไขได้สะดวก

3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

แอพพลิเคชันแบ่งการทำงานออกเป็นสี่ส่วนประกอบด้วยกระบวนการ Select, Detect, Track และ Label เพื่อช่วยแบ่งเบาภาระของผู้พัฒนาในการสร้างชุดข้อมูลสำหรับสร้างโมเดลจากข้อมูลประเภทวิดีโอ โดยกระบวนการ Select จะต้องสามารถตัดวิดีโอด้วยส่วนที่ไม่มีมนุษย์อยู่ออกจากวิดีโอด้วย กระบวนการ Detect จะต้องหาตำแหน่งของมนุษย์ภายในวิดีโอด้วย แล้วใช้กระบวนการ Track นำรายตำแหน่งต่อไปของมนุษย์ข้อมูลตำแหน่งของมนุษย์ที่ได้จากการกระบวนการ Detect และกระบวนการ Label นั้นต้องสามารถทำงานมุ่งเน้นที่การระบุที่พื้นฐานของมนุษย์ได้ เช่น ยืน เดิน นั่ง กินข้าว หรือ นอน เป็นต้น โดยทุกส่วนการทำงานมนุษย์ต้องสามารถทำงานร่วมกับปัญญาประดิษฐ์ได้ ดังรูปที่ 3.2

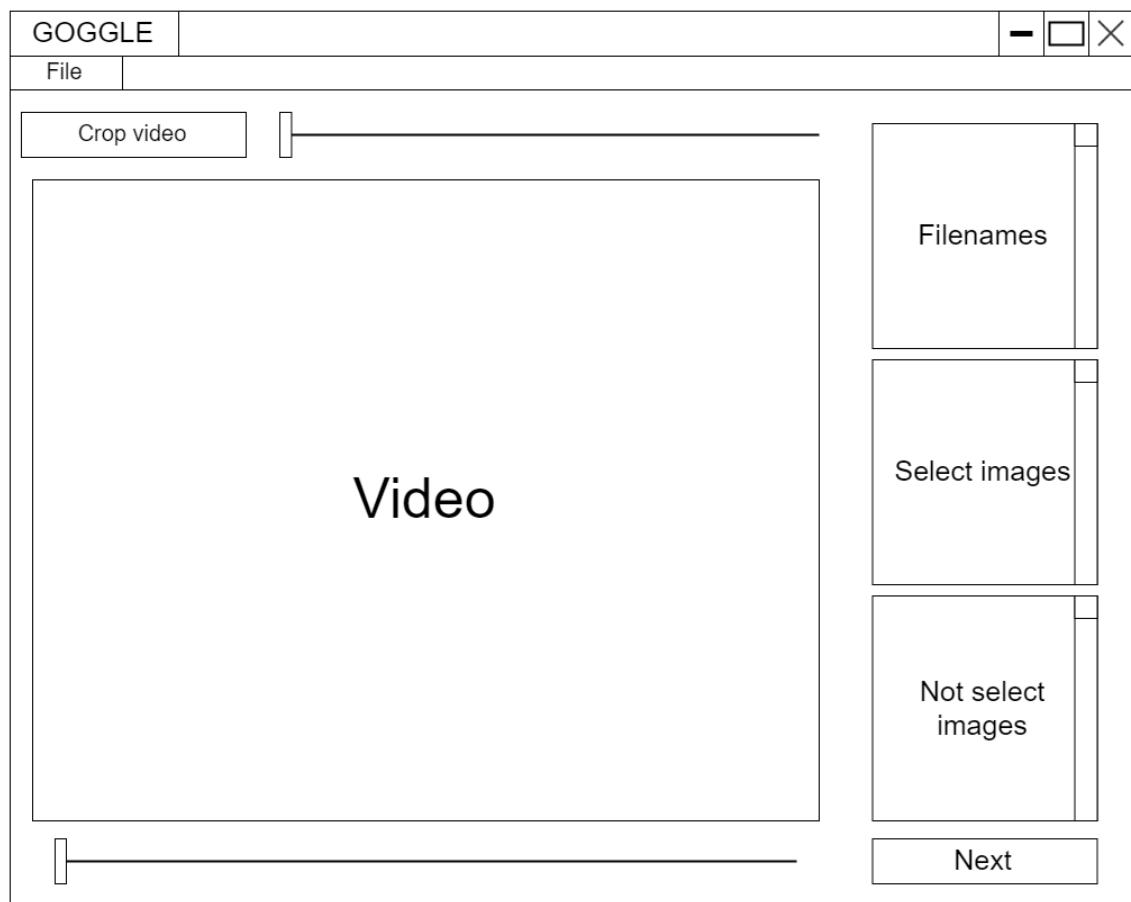


รูปที่ 3.2: กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

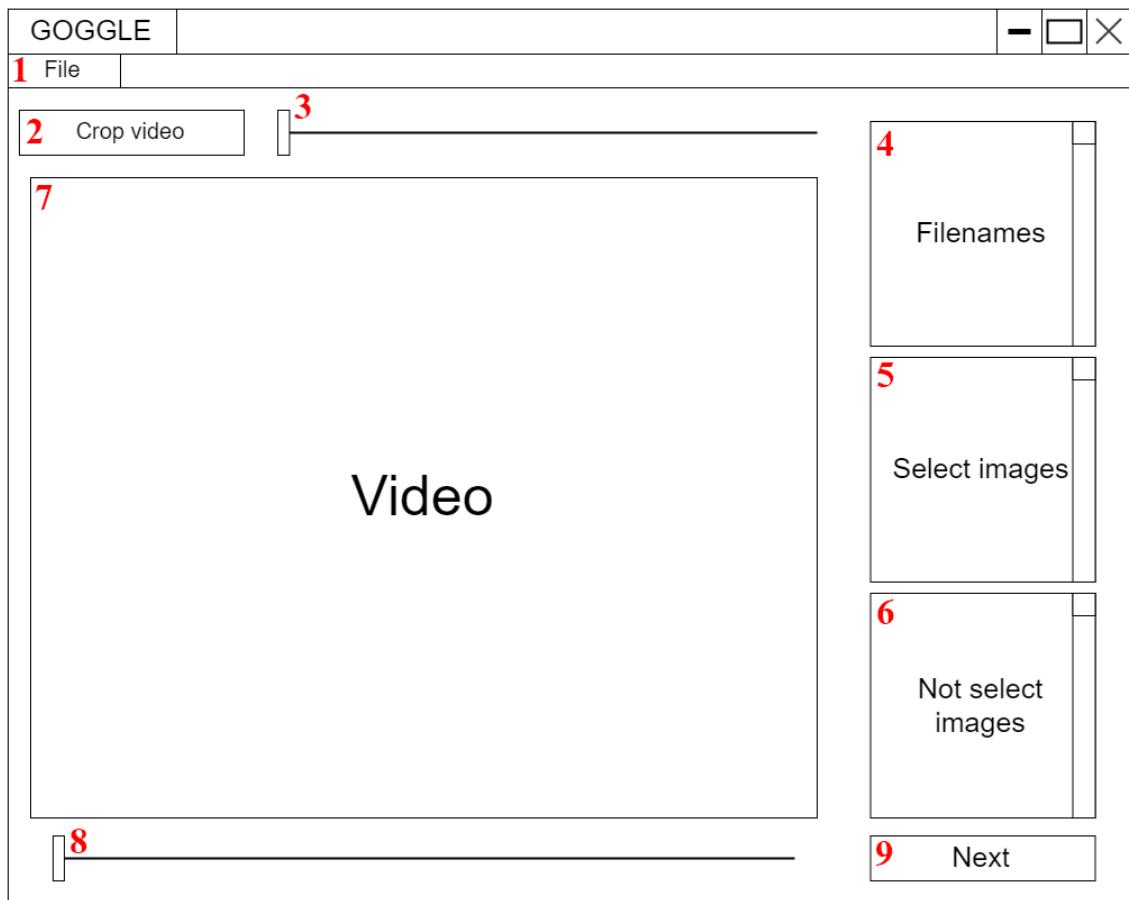
โดยแต่ละกระบวนการจะมีรายละเอียดดังนี้

3.8.1.1 Select

กระบวนการ Select จะต้องสามารถรับวิดีโอเข้ามา แล้วตัดวิดีโອในช่วงที่ไม่มีมนุษย์อยู่ในเฟรมออกได้ อัตโนมัติด้วยปัญญาประดิษฐ์ แต่เนื่องจากการประมวลผลทุกเฟรมในวิดีโอนั้นจะทำให้เสียเวลามากเกินไป จึงใช้วิธีการเลือกตัวอย่างเฟรมตัวอย่างตราชกที่ (สามารถกำหนดได้) ซึ่งเรียกว่าเฟรมเหล่านี้ว่า คีย์เฟรม (keyframe) จากนั้นใช้ปัญญาประดิษฐ์ประมวลผลคีย์เฟรมที่เหล่านั้น เพื่อลดระยะเวลาในการประมวลผลลง และมนุษย์จะต้องสามารถแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.3



รูปที่ 3.3: หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



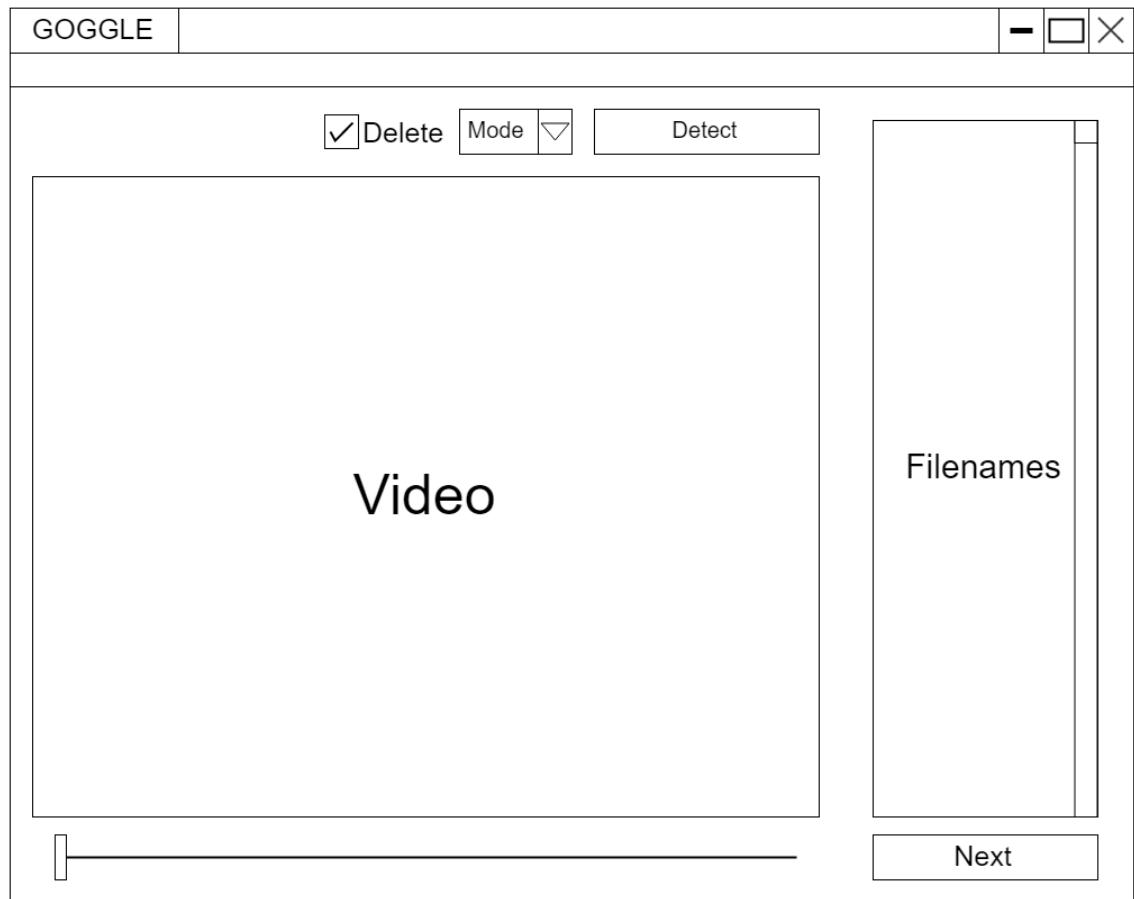
รูปที่ 3.4: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.4 มีรายละเอียดดังนี้

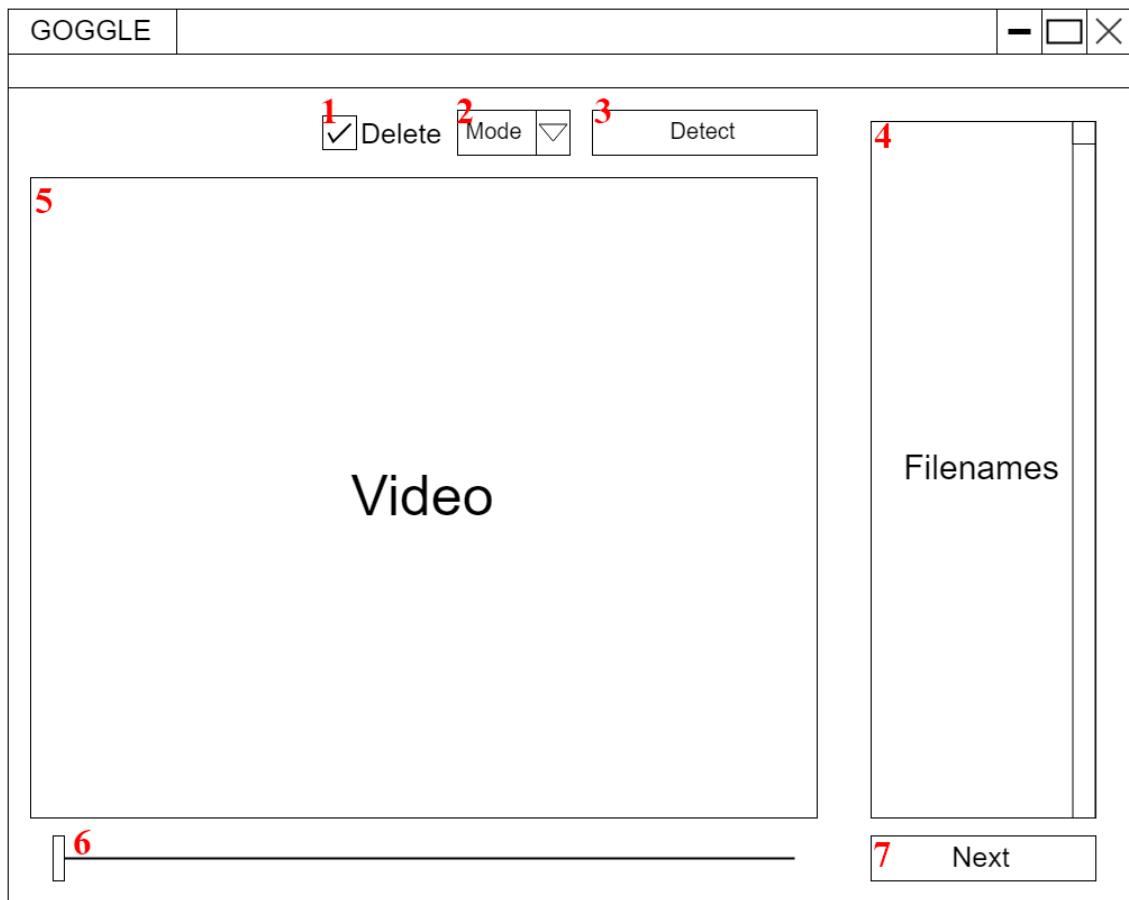
1. หมายเลข 1 คือปุ่มสำหรับเลือกไฟล์วิดีโอที่ต้องการจากในคอมพิวเตอร์เข้ามาในโปรแกรม
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบทำการสร้างคีย์เฟรมขึ้นมา แล้วใช้ปัญญาประดิษฐ์ประมวลผลเพื่อแยกคีย์เฟรมในเมมค่อนอยู่ และคีย์เฟรมใหม่เมมค่อนอยู่แบบอัตโนมัติ (Auto mode)
3. หมายเลข 3 คือแถบเลื่อนเพื่อกำหนดความถี่ในการหยิบคีย์เฟรม โดยจะมีช่วงอยู่ที่ 1 เฟรมต่อวินาที จนถึงอัตราเฟรมต่อวินาทีสูงสุดของวิดีโอิที่รับเข้ามา
4. หมายเลข 4 คือกล่องสำหรับแสดงชื่อวิดีโอที่รับเข้ามาในโปรแกรมเพื่อเลือกเข้ามาใช้ในการประมวลผล
5. หมายเลข 5 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
6. หมายเลข 6 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
7. หมายเลข 7 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 5 หมายเลข 6 หรือหมายเลข 8
8. หมายเลข 8 คือแถบเลื่อนสำหรับเลือนดูคีย์เฟรมทั้งหมดที่ระบบสร้างขึ้น
9. หมายเลข 9 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

3.8.1.2 Detect

กระบวนการ Detect จะต้องสามารถรับคีย์เฟรมจากการบันทึกการ Select มาประมวลผลด้วยปัญญาประดิษฐ์เพื่อหาตำแหน่งของมนุษย์ที่อยู่ในคีย์เฟรม และสร้างกรอบสีเหลี่ยมครอบบริเวณดังกล่าวได้ในแบบอัตโนมัติ เพื่อแบ่งเบาภาระผู้ใช้ในการที่ต้องสร้างกรอบสีเหลี่ยมครอบตำแหน่งของมนุษย์ด้วยตัวเอง และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสีเหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของปัญญาประดิษฐ์ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.5



รูปที่ 3.5: หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.6: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.6 มีรายละเอียดดังนี้

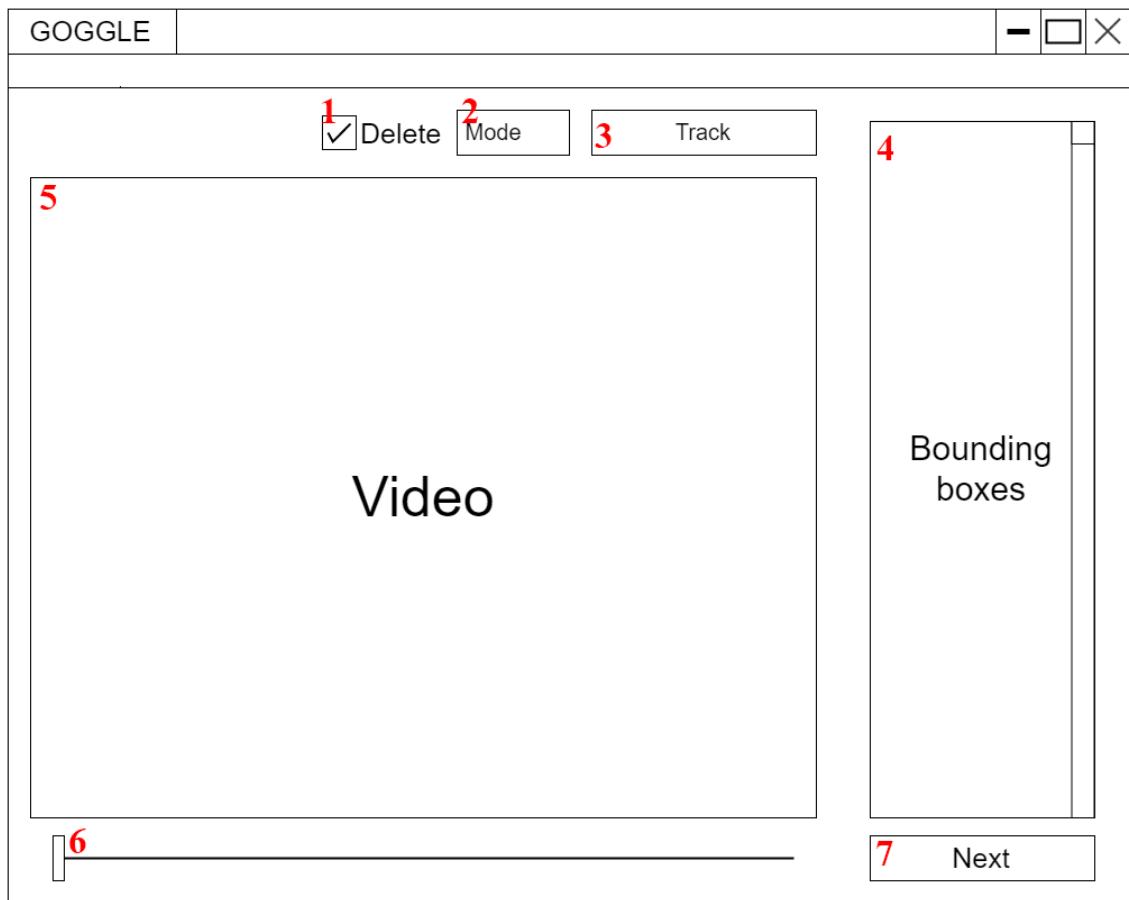
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเอง (Manual mode) เป็นลบรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจหาตำแหน่งของมนุษย์ในคิร์เฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
4. หมายเลข 4 คือกล่องสำหรับแสดงคิร์เฟรมทั้งหมด
5. หมายเลข 5 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 4 หรือหมายเลข 6
6. หมายเลข 6 คือแบบเลื่อนสำหรับเลื่อนดูคิร์เฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของปัญญาประดิษฐ์
7. หมายเลข 7 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

3.8.1.3 Track

เนื่องจากกระบวนการ Detect นั้นจะทำเฉพาะในคีย์เฟรมทำให้ในเฟรมอื่นๆ นอกเหนือจากนั้นจะไม่มีกรอบสี่เหลี่ยมอยู่ ดังนั้นกระบวนการ Track จึงต้องสามารถทำนายตำแหน่งต่อไปของมนุษย์แล้วสร้างกรอบสี่เหลี่ยมขึ้นมาบนเฟรมระหว่างคีย์เฟรมทั้งหมดได้โดยอัตโนมัติ เพื่อสร้างข้อมูลตำแหน่งของมนุษย์ในเฟรมเหล่านั้น และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสี่เหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของอัลกอริทึม จึงออกแบบหน้าต่างได้ดังรูปที่ 3.7



รูปที่ 3.7: หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



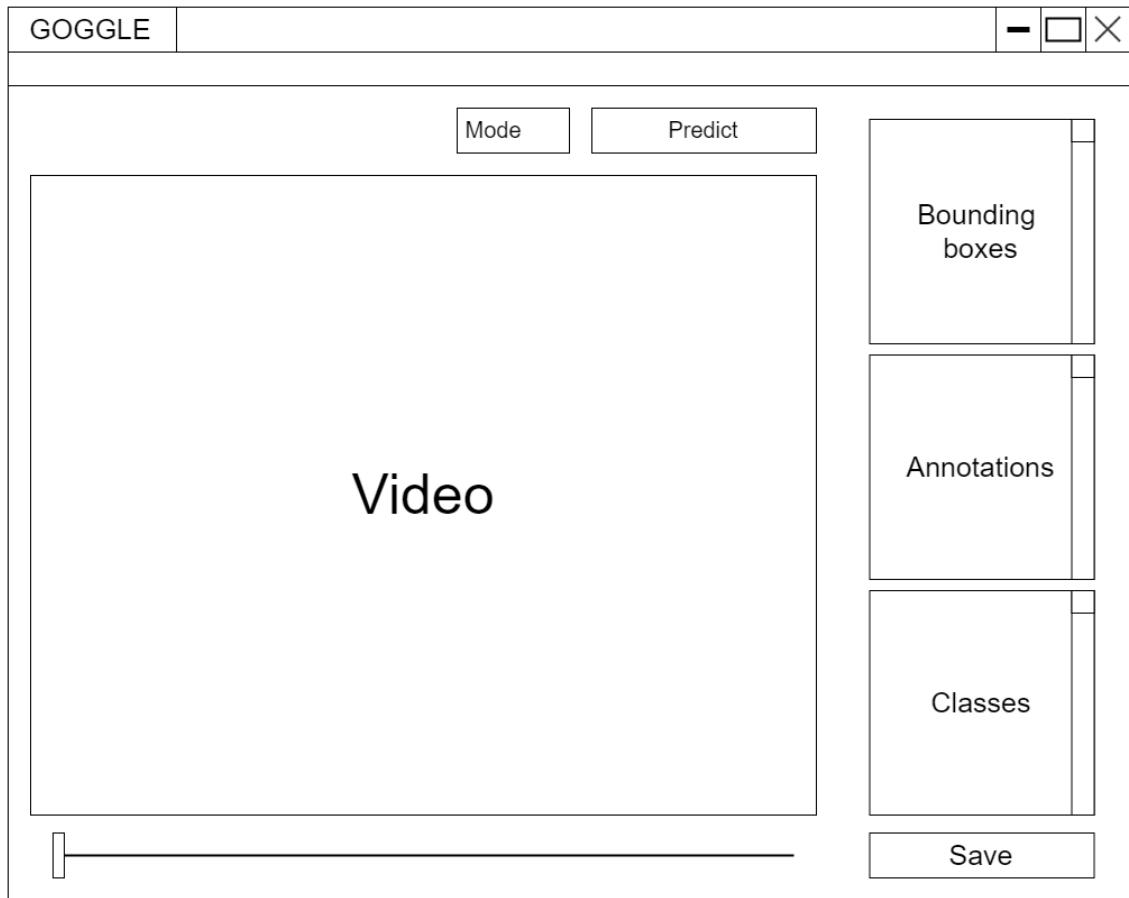
รูปที่ 3.8: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

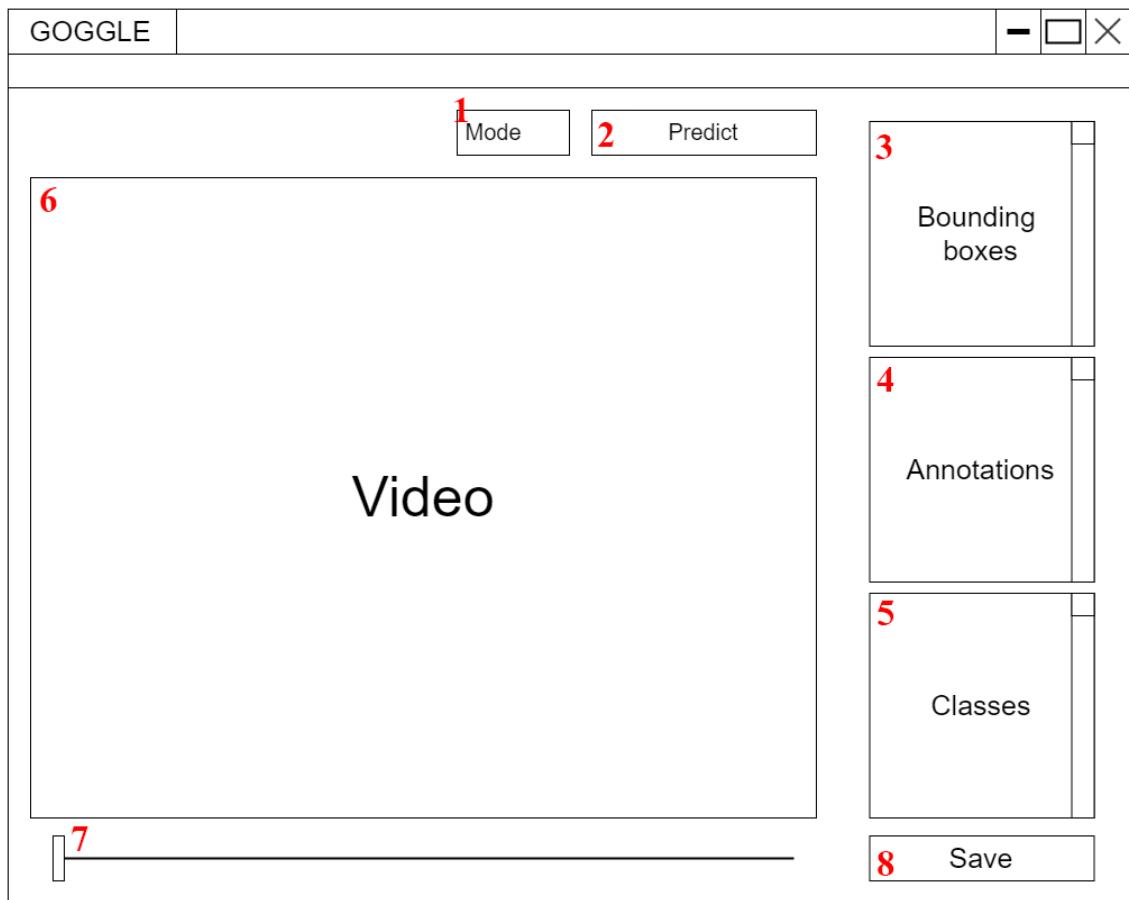
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเองเป็นลบกรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจสอบตำแหน่งของมนุษย์ในเฟรมระหว่างคิ้ยวเฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
4. หมายเลข 4 คือกล่องสำหรับแสดงกรอบสี่เหลี่ยมทั้งหมดที่อยู่ในเฟรม
5. หมายเลข 5 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 6
6. หมายเลข 6 คือแถบเลื่อนสำหรับเลื่อนดูเฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของอัลกอริทึม
7. หมายเลข 7 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

3.8.1.4 Label

กระบวนการ Label นั้นต้องสามารถทำนายว่าการกระทำของมนุษย์ที่อยู่ในแต่ละเฟรมว่าคืออะไรได้โดยอัตโนมัติด้วยปัญญาประดิษฐ์ และผู้ใช้จะต้องสามารถแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้หากมีการทำนายที่ผิดพลาดเกิดขึ้น หรือถ้าหากผู้ใช้ต้องการเพิ่มการกระทำที่ไม่ได้มีอยู่ในชุดการกระทำพื้นฐานที่มีอยู่แล้วของปัญญาประดิษฐ์ ผู้ใช้ก็สามารถเพิ่มการกระทำนั้นเข้ามาได้ จึงออกแบบหน้าต่างเดี๋งรูปที่ 3.9



รูปที่ 3.9: หน้าต่าง Label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.10: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Label

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

1. หมายเลข 1 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบคำนวณรูปแบบของมนุษย์ในทุกๆเฟรม
3. หมายเลข 3 คือกล่องสำหรับแสดงกรอบสี่เหลี่ยมทั้งหมดที่อยู่ในเฟรมที่เลือก
4. หมายเลข 4 คือกล่องสำหรับแสดงการกระทำของมนุษย์แต่ละคนที่อยู่ในเฟรมที่เลือก โดยจะเรียงลำดับคู่กับกรอบสี่เหลี่ยมที่อยู่ในข้อ หมายเหตุ 3
5. หมายเลข 5 คือกล่องสำหรับแสดงชุดการกระทำการที่ปัญญาประดิษฐ์มีอยู่แล้ว ซึ่งในการทำงานแบบแก้ไขด้วยตนเองนั้น จะสามารถค้นหาการกระทำการที่มีอยู่แล้วได้ และหากคำที่ใส่เขามานั้นมีอยู่ในชุดการกระทำการที่เป็นการเพิ่มการกระทำการที่มีอยู่แล้ว
6. หมายเหตุ 6 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 7
7. หมายเหตุ 7 คือແນບเลื่อนสำหรับเลื่อนดูเฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของปัญญาประดิษฐ์
8. หมายเหตุ 8 คือปุ่มสำหรับสร้างไฟล์ xml ของทุกๆเฟรมสำหรับใช้ในการสร้างโมเดลโดยรายละเอียดข้อมูลภายในไฟล์ xml จะอยู่ในหัวข้อ 3.8.1.5

3.8.1.5 รายละเอียดข้อมูลภายในไฟล์ xml

ไฟล์ xml นั้นเป็นรูปแบบที่นิยมใช้ในการเก็บข้อมูลสำหรับการสร้างโมเดลประเพณีตรวจสอบจับตุ๊ก โดยจะเก็บข้อมูลในรูปแบบของ PASCAL VOC ที่นิยมใช้ในการสร้างโมเดลด้วย library Tensorflow โดยภายในไฟล์จะมีข้อมูลดังรูปที่ 3.11 โดยข้อมูลส่วนสำคัญของรูปแบบนี้นั้นจะถูกใส่หมายเลขอ้างอิงแต่ละหมายเลขนั้นหมาย

```

<annotation>
    <folder>GeneratedData_Train</folder>1
    <filename>000001.png</filename>2
    <path>/my/path/GeneratedData_Train/000001.png</path>3
    <source>
        <database>Unknown</database>
    </source>
    <size> 4
        <width>224</width>
        <height>224</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
        <name>21</name> 5
        <pose>Frontal</pose>
        <truncated>0</truncated>
        <difficult>0</difficult>
        <occluded>0</occluded>
        <bndbox> 6
            <xmin>82</xmin>
            <xmax>172</xmax>
            <ymin>88</ymin>
            <ymax>146</ymax>
        </bndbox>
    </object>
</annotation>
```

รูปที่ 3.11: ตัวอย่างข้อมูลภายในไฟล์ xml

ถึง

1. หมายเลขอื่นๆ คือชื่อโฟลเดอร์ที่เก็บไฟล์รูปภาพที่เกี่ยวข้องกับไฟล์ xml นี้อยู่
2. หมายเลขอื่นๆ คือชื่อไฟล์ที่เกี่ยวข้องกับไฟล์ xml นี้
3. หมายเลขอื่นๆ คือเส้นทางในคอมพิวเตอร์ (directory path) ของไฟล์รูปภาพที่เกี่ยวข้องกับไฟล์ xml นี้
4. หมายเลขอื่นๆ คือขนาดและมิติของรูปภาพ ซึ่งจะประกอบด้วยความกว้าง (width) ความยาว (height) และจำนวนช่องสี (depth) โดยที่จำนวนช่องสีที่มีความลึก 3 มักจะหมายถึงภาพสี RGB และจำนวนช่องสีที่มีความลึก 2 จะหมายถึงภาพขาวดำ (gray scale)
5. หมายเลขอื่นๆ คือ label ของวัตถุหรืออย่างอื่น ที่อยู่ในกรอบสีเหลี่ยมที่ถูกกำหนดไว้ในส่วนของหมายเลขอื่นๆ
6. หมายเลขอื่นๆ คือ กรอบสีเหลี่ยมที่ครอบวัตถุที่สนใจ เช่นมนุษย์ เป็นต้น

3.9 การออกแบบการทดสอบการตรวจจับวัตถุ

3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล สิ่งที่ใช้ในการวัดผล

1. ความเร็วในการคำนวณ (มิลลิวินาที)
2. ความแม่นยำ โดยคำนึงถึงอัตราส่วนร่วมของกรอบที่เหลืออยู่ หรือ Intersection over Union (IoU)

จุดประสงค์

1. ผู้วิจัยได้ตั้งจุดประสงค์การทดลอง การใช้โมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับวัตถุ เพื่อวัดผลโมเดลปัญญาประดิษฐ์ที่ใช้ในปัจจุบัน และหาโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับวัตถุที่มีความเร็วมากที่สุดและแม่นยำสูงที่สุดเมื่อทดสอบกับชุดข้อมูลของผู้วิจัย

ตัวแปร

1. โมเดลปัญญาประดิษฐ์ ได้แก่
 - (a) SSD Mobilenet v1 ppn
 - (b) YOLO-v3 tiny
 - (c) YOLO-v3 spp
 - (d) YOLO-v3 320
 - (e) Faster RCNN inception v2

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลสำหรับทดสอบวัดผลที่ผู้วิจัยสร้างขึ้น (สูม 20 เฟรมจากวิดีโอที่ผู้วิจัยใช้สำหรับสร้างชุดข้อมูล)

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็นชุดข้อมูลสำหรับทดสอบ และชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ และเฟรม
 - (b) ชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ ประกอบด้วย : ชื่อของวิดีโอ เฟรม และตำแหน่งของกรอบสีเหลือง
2. เรียกชื่อและเฟรมของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่มีคำจำกัดเพื่อเป็นคำตอบ เก็บผลลัพธ์เป็นชุดข้อมูลผลลัพธ์จากการคำนวณ
 - (a) ชุดข้อมูลผลลัพธ์จากการคำนวณ ประกอบด้วย : ชื่อของวิดีโอ เฟรม และตำแหน่งของกรอบสีเหลือง
3. ประเมินผลค่าความแม่นยำในการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการคำนวณ และชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ (ตั้งค่า IoU = 0.5)
4. เปรียบเทียบผลลัพธ์จากแหล่งที่มา

3.10 การออกแบบการทดสอบการทำงานตามหน่วยต่อไปของมุขย์

3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำงานตามหน่วยต่อไปของวัตถุในวิดีโอ สิ่งที่ใช้ในการวัดผล

1. ความเร็วในการทำงานต่อวิดีโอ (วินาที)
2. ความแม่นยำ โดยคำนึงถึงอัตราส่วนร่วมของกรอบที่เหลือ

สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่า การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจสอบจับตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น จะทำให้ระบบสามารถทำงานได้เร็วขึ้น โดยที่ประสิทธิภาพจะลดลงเพียงเล็กน้อย

ตัวแปรควบคุม

1. วิดีโอสาระที่ไม่ติดลิขสิทธิ์ ความยาวประมาณ 10 - 30 วินาที หนึ่งวิดีโอ
2. ใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับตำแหน่งวัตถุ ResNet50 ในการสร้างชุดข้อมูลที่มีการกำกับตำแหน่งวัตถุไว้ (ground-truth) และใช้มุขย์ในการตรวจสอบความถูกต้อง เพื่อใช้เป็นค่าตอบของการทำงาน
3. โมเดลปัญญาประดิษฐ์สำหรับตรวจจับตำแหน่งที่ใช้ในการเปรียบเทียบ: YOLO-V3 320
4. อัลกอริทึมสำหรับระบบทำงานตามหน่วยต่อไปของวัตถุ: dlib
5. อัตราส่วนร่วมของกรอบที่เหลือ: มีส่วนที่หักกันมากกว่า 80% ขึ้นไปจึงจะนับว่าผลการทำงานถูกต้อง

วิธีการทดลอง

1. ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมในวิดีโอ และเปรียบเทียบผลลัพธ์กับชุดข้อมูลที่ถูกกำหนดตำแหน่งวัตถุไว้แล้ว เพื่อคำนวณหาความแม่นยำ
2. ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกๆ N เฟรมในวิดีโอ และใช้ระบบทำงานตามหน่วยต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น และเปรียบเทียบผลลัพธ์กับชุดข้อมูลที่ถูกกำหนดตำแหน่งวัตถุไว้แล้ว เพื่อคำนวณหาความแม่นยำ โดยที่ค่า N จะเท่ากับ 10 20 และ 25
3. เปรียบเทียบความเร็วในการประมวลผล และความแม่นยำ

3.11 การออกแบบการทดสอบการระบุตัวตนของบุคคลภายในภาพ

3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ สิ่งที่ใช้ในการวัดผล

1. ความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพ

สมมุติฐาน

ผู้จัดได้ตั้งสมมุติฐานว่า ผลลัพธ์ของการทดลองการใช้งานจริงของโมเดลปัญญาประดิษฐ์ ResNet50 ที่สร้างด้วยชุดข้อมูล Market1501 นั้นควรจะมีความแม่นยำในการระบุตัวตนของบุคคลภายในภาพมากที่สุดเมื่อเทียบกับโมเดลปัญญาประดิษฐ์ที่สร้างด้วยชุดข้อมูลอื่นๆ เพราะเมื่อเทียบกับโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลอื่นที่มาจากการแหล่งข้อมูลเดียวกัน โมเดลปัญญาประดิษฐ์ ResNet50 ที่สร้างด้วยชุดข้อมูล Market1501 นั้นจะมีความแม่นยำสูงสุด

ตัวแปร

1. โมเดลปัญญาประดิษฐ์ ซึ่งได้แก่
 - (a) ResNet50 ของชุดข้อมูล Market1501
 - (b) ResNet50 ของชุดข้อมูล DukeMTMCReID
 - (c) ResNet50 ของชุดข้อมูล CUHK03
 - (d) ResNet50 ของชุดข้อมูล MSMT17

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลที่ทางผู้จัดสร้างขึ้นสำหรับการทดสอบ
2. โมเดลปัญญาประดิษฐ์ : YOLO-V3 320 สำหรับการทำหนังของบุคคล

วิธีการทดลอง

1. ดาวน์โหลดโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลต่างๆ ได้แก่ Market1501, DukeMTMCReID, CUHK03 และ MSMT17
2. นำชุดข้อมูลที่ผู้จัดสร้างขึ้นมาผ่านโมเดลปัญญาประดิษฐ์ YOLO-V3 320 เพื่อหาตำแหน่งของบุคคล
3. นำโมเดลปัญญาประดิษฐ์แต่ละอันมาทดสอบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพ ด้วยตำแหน่งของบุคคลที่ได้มาจากการขั้นตอนก่อนหน้านี้
4. ประเมินผลการทำงานโดยเทียบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายในภาพของแต่ละโมเดล ปัญญาประดิษฐ์ เพื่อหาโมเดลปัญญาประดิษฐ์ที่ได้ผลลัพธ์ดีที่สุด

3.12 การออกแบบการทดสอบการจดจำการกระทำของมนุษย์

3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกтренร์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

สิ่งที่ใช้ในการวัดผล

1. ความเร็วในการทำงานต่อรูปภาพ (มิลลิวินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมติฐาน

ผู้จัยได้ตั้งสมมติฐานว่า ผลลัพธ์ของการทดลองจะมีความแม่นยำเทียบเท่ากับผลลัพธ์จากแหล่งที่มา แต่ความเร็วต่อรูปภาพจะมีความเร็วน้อยกว่าผลลัพธ์จากแหล่งที่มา เนื่องจากแหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่ากราฟิกการ์ดของผู้จัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

1. ชุดข้อมูล : The validation split of AVA v2.1
2. โมเดลปัญญาประดิษฐ์ : Faster RCNN ResNet101 AVA v2.1

วิธีการทดลอง

1. ดาวน์โหลดชุดข้อมูล The validation split of AVA v2.1
2. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ ประกอบด้วย : ชื่อของวิดีโอ เพริม ตำแหน่งของกรอบสีเหลี่ยม และรหัสของการกระทำ
3. เรียกชื่อของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่เคยฝึกอบรม จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ เพริม ตำแหน่งของกรอบสีเหลี่ยม รหัสของการกระทำ และความมั่นใจ
4. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
5. เปรียบเทียบผลลัพธ์กับแหล่งที่มา

3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง สิ่งที่ใช้ในการวัดผล

1. ความใน การทำงานเร็วต่อรูปภาพ (มิลลิวินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่าผลลัพธ์ของการทดลองจะมีความแม่นยำต่ำลงเมื่อเทียบกับความแม่นยำของการทดลองที่ผ่านมา เนื่องจากชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ได้มีการตัดหมวดหมู่บางอย่างออกไป ทำให้โมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วย AVA มีหมวดหมู่ของการกระทำไม่ตรงกับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ซึ่งมีผลทำให้ความแม่นยำลดลง ในส่วนของความเร็วต่อรูปภาพจะมีความเร็วน้อยกว่าผลลัพธ์จากแหล่งที่มา เนื่องจาก แหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X card ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่า กราฟิกการ์ดของผู้วิจัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลที่ผู้วิจัยสร้างด้วย AI-assisted labeling tool
2. โมเดลปัญญาประดิษฐ์ : Faster RCNN ResNet101 AVA v2.1

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม และรหัสของการกระทำ
2. เรียกชื่อของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม รหัสของการกระทำ และความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง สิ่งที่ใช้ในการวัดผล

1. ความเร็วในการทำงานต่อรูปภาพ (มิลลิวินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่าผลลัพธ์ของการทดลองจะมีความแม่นยำสูงขึ้นเมื่อเทียบกับความแม่นยำของการทดลองที่ผ่านมา เนื่องจากโมเดลปัญญาประดิษฐ์ในการทดลองนี้ เป็นโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยได้สร้างขึ้น ซึ่งจะมีหมวดหมู่ของการกระทำของโมเดลปัญญาประดิษฐ์และชุดข้อมูลทดสอบตรงกัน ในส่วนของความเร็วต่อรูปภาพจะมีความเร็วอ่อนกว่าผลลัพธ์จากแหล่งที่มา เนื่องจากแหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่ากราฟิกการ์ดของผู้วิจัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลที่ผู้วิจัยสร้างด้วย AI-assisted labeling tool
2. โมเดลปัญญาประดิษฐ์ : โมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม และรหัสของการกระทำ
2. เรียกชื่อของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นำมายังผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม รหัสของการกระทำ และความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และ ชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

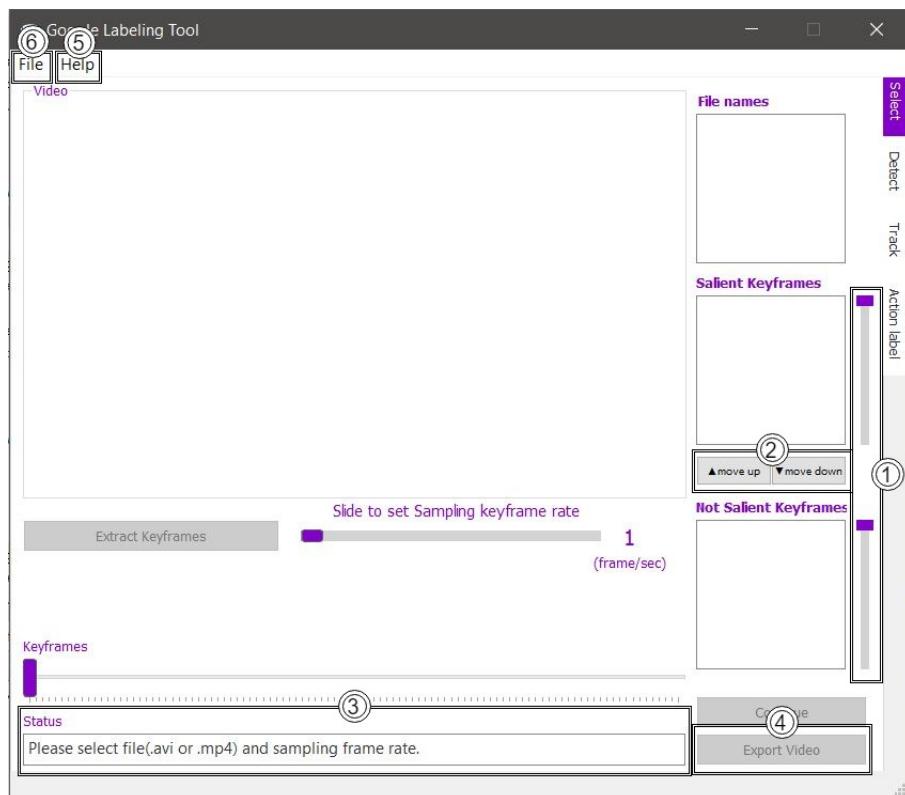
บทที่ 4

ผลการดำเนินงาน

4.1 เครื่องมือกำกับคุณลักษณะ

4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน

หน้าต่าง Select

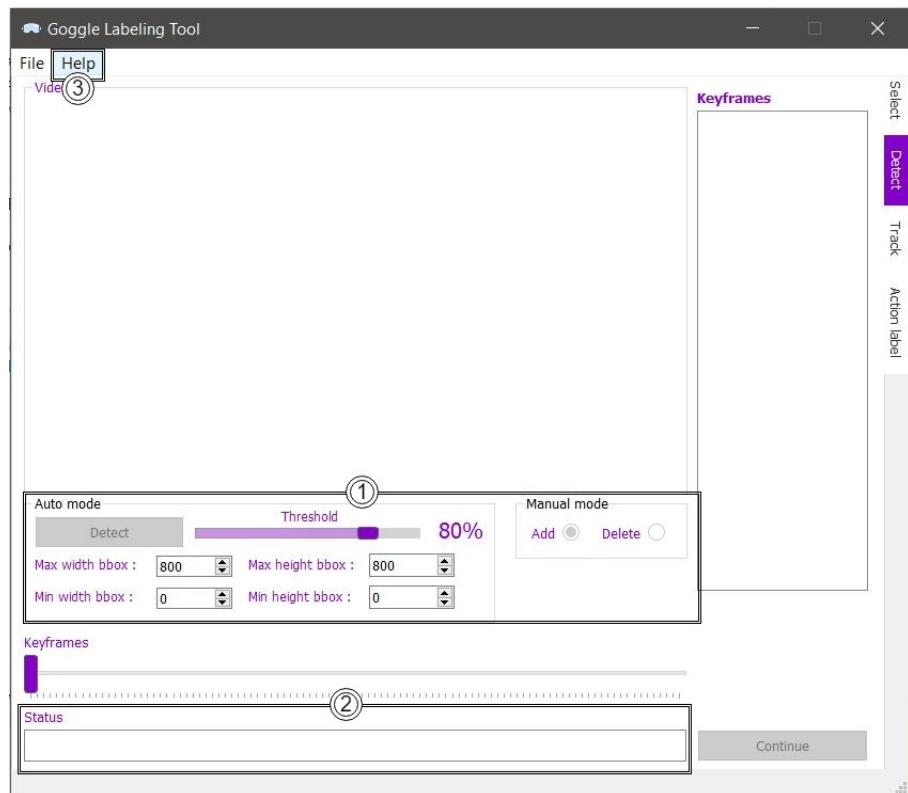


รูปที่ 4.1: รูปหน้าต่างแสดงผลของหน้าต่าง Select

จากรูปที่ 4.1 แสดงหน้าต่าง Select ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับรูปที่ 3.3 จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. แถบเลื่อนสำหรับเลื่อนคุณภาพที่มีมนุษย์หรือไม่มีมนุษย์ เพื่อเพิ่มความสะดวกในการเลือกคุณภาพ
2. ปุ่มสำหรับแก้ไขคุณภาพที่มีมนุษย์หรือไม่มีมนุษย์
3. แถบแสดงสถานะกระบวนการทำงาน
4. ปุ่มสำหรับนำผลลัพธ์ออกเป็นไฟล์วิดีโอด้วยไฟล์ในช่วงที่มีมนุษย์อยู่
5. แถบสำหรับคำแนะนำช่วยเหลือ
6. ปุ่มสำหรับเปิดไฟล์

หน้าต่าง Detect

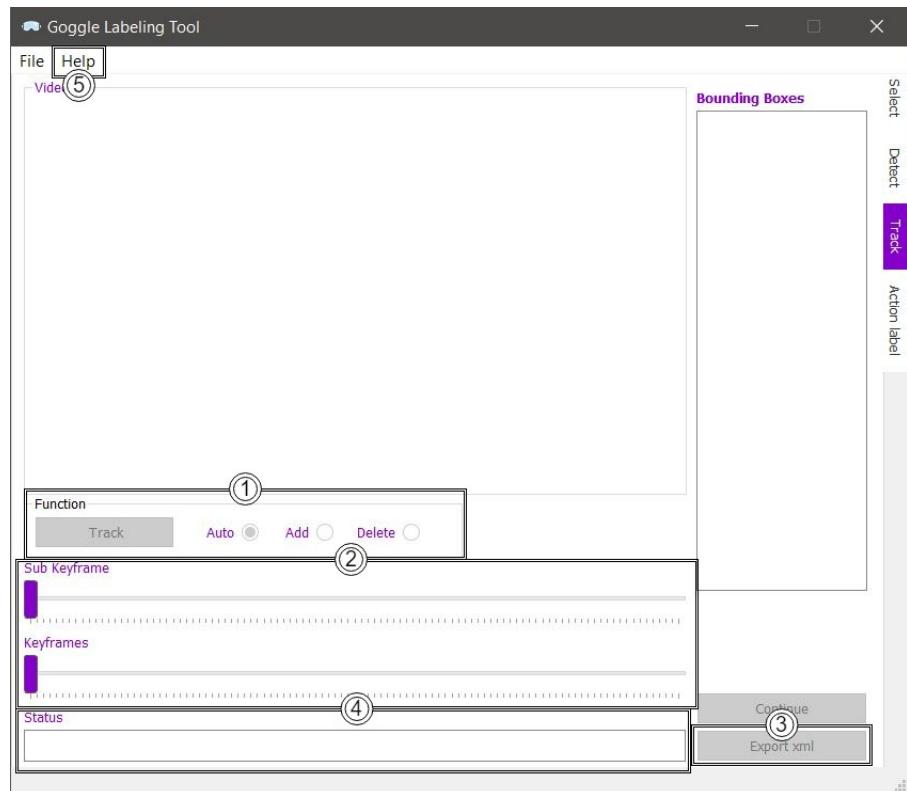


รูปที่ 4.2: รูปหน้าต่างแสดงผลของหน้าต่าง Detect

จากรูปที่ 4.2 แสดงหน้าต่าง Detect ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับรูปที่ (3.5) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่ด้วยการทำงานแบบอัตโนมัติ และกำหนดเองสามารถใช้งานได้สะดวกขึ้น และเพิ่มความหลากหลายในการปรับแก้ในการทำงานอัตโนมัติ
2. แถบแสดงสถานะกระบวนการทำงาน
3. แถบสำหรับคำแนะนำช่วยเหลือ

หน้าต่าง Track

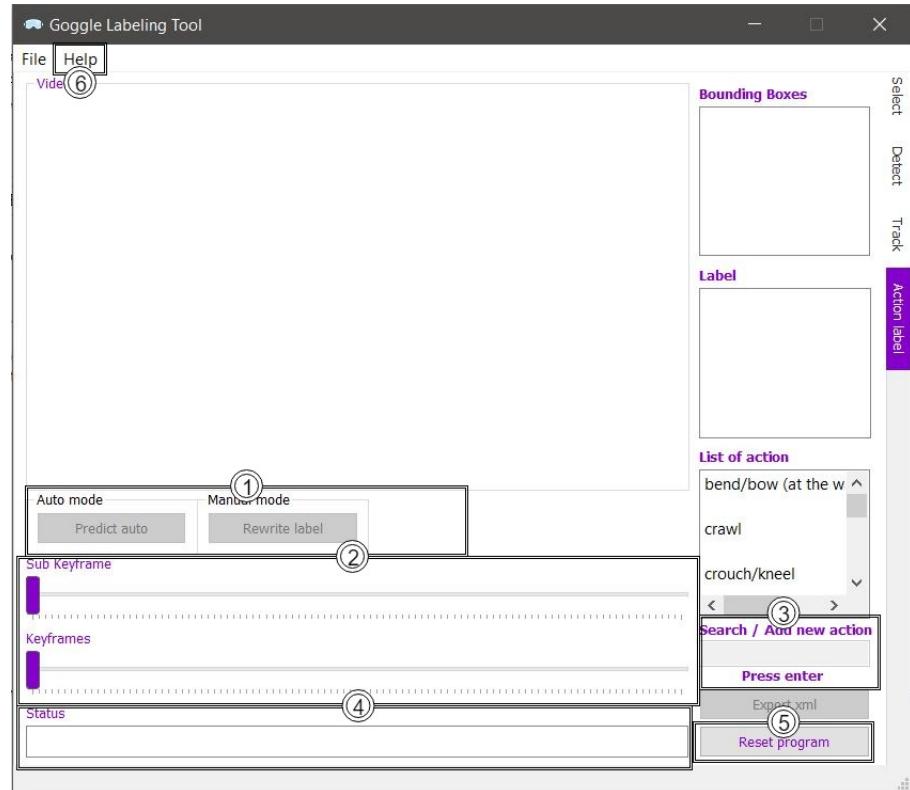


รูปที่ 4.3: รูปหน้าต่างแสดงผลของหน้าต่าง Track

จากรูปที่ 4.3 แสดงหน้าต่าง Track ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ (3.7) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແລບເລືອນເປັນ 2 ແລບເລືອນທາງໃຫ້ສາມາຄຸດຄີ່ຍິ່ນແລະ ເພີ້ມໄວ້ຮ່ວງໜ່ວຍຄີ່ຍິ່ນແລະ ເພີ້ມໄວ້ຮ່ວງໜ່ວຍຄີ່ຍິ່ນ
3. เพิ่ມປຸ່ມສໍາຫຼັບນຳພລັບໂພຣອອກເປັນໄຟລ് xml
4. ແຕບແສດງສະຖານະກະບວນການທຳງານ
5. ແຕບສໍາຫຼັບຄໍາແນະນຳໜ່ວຍເຫຼືອ

หน้าต่าง Label



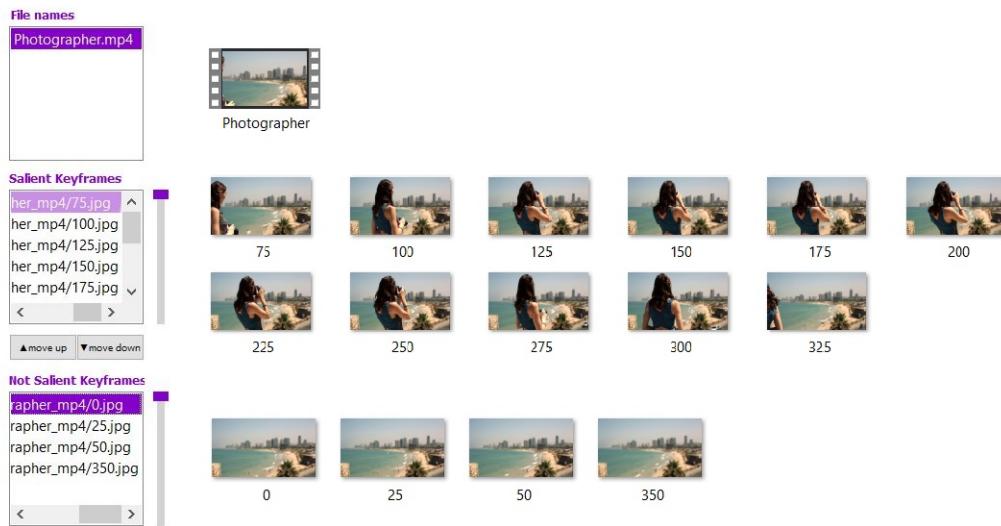
รูปที่ 4.4: รูปหน้าต่างแสดงผลของหน้าต่าง Label

จากรูปที่ 4.4 แสดงหน้าต่าง Label ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ (3.9) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແຕບເລື່ອນປັນ 2 ແຕບເລື່ອນທຳໃຫ້ສາມາດຄຸງຄິຍໝ່າງແລ້ວໄດ້ສະໜັກຂຶ້ນ
3. ເຄີຍອົງມື້ສໍາຫຼັບຄັນຫາຫຼືເພີ່ມໝາວດໜູ່ຂອງການກະທຳ
4. ແຕບແສດງສະຖານະກະບວນການກະທຳ
5. ປຸ່ມສໍາຫຼັບເຮີມຕົ້ນການກະທຳໃໝ່
6. ແຕບສໍາຫຼັບຄຳແນະນຳໜ່ວຍເຫຼືອ

4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอปพลิเคชัน

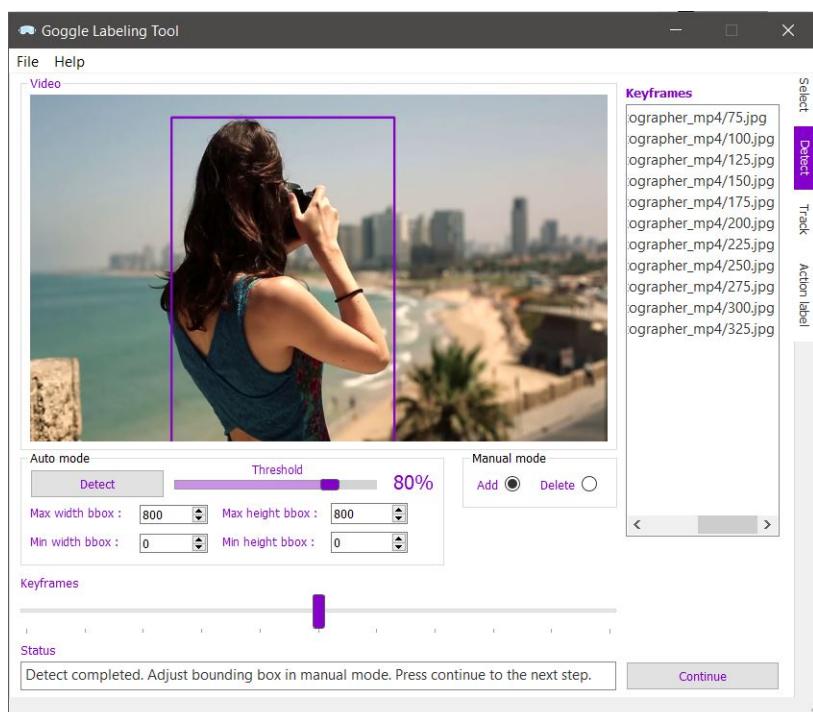
ผลลัพธ์การทำงานของหน้าต่าง Select



รูปที่ 4.5: รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ภายในเฟรม

ขั้นตอนแรกแอปพลิเคชัน จะสกัดแยกวิดีโอออกเป็นเฟรมทั้งหมด และทำการสั่มคีย์เฟรมอกรอบตามความถี่ที่ผู้ใช้งานตั้งไว้ จากนั้นแอปพลิเคชันจะนำโมเดล YOLO-v3 320 มาตรวจสอบว่าแต่ละคีย์เฟรมมีเฟรมใดบ้างที่มีมนุษย์อยู่ภายในเฟรม จากนั้นจะทำการแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ ดังรูปที่ 4.5

ผลลัพธ์การทำงานของหน้าต่าง Detect



รูปที่ 4.6: รูปคีย์เฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่

แอ��พลิเคชันจะนำคีย์เฟรมที่มีนุชย์ที่ได้จากหน้าต่าง Select นำมาตีกรอบสี่เหลี่ยมในส่วนของเฟรมที่มีมนุชย์อยู่โดยสามารถใช้โหมดการทำงานแบบบอตโน้มัติหรือแบบแก้ไขเองก็ได้ ซึ่งผลลัพธ์ที่ได้จะได้คีย์เฟรมที่มีกรอบสี่เหลี่ยม ดังรูปที่ 4.6 จากนั้นจะบันทึกข้อมูลในไฟล์ txt

ผลลัพธ์การทำงานของหน้าต่าง Track



(ก) ตัวอย่างเฟรมที่ถูกตีกรอบสี่เหลี่ยม

```
<?xml version="1.0"?>
<annotation>
  <folder>D:/Goggle/Goggle_team/out/Photographer_mp4/img</folder>
  <filename>75.jpg.txt</filename>
  <path>D:/Goggle/Goggle_team/out/Photographer_mp4/img/75.jpg</path>
  - <source>
    <database>Unknown</database>
  </source>
  - <size>
    <width>1280</width>
    <height>720</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  - <object>
    <name>person</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    - <bndbox>
      <xmin>2</xmin>
      <ymin>35</ymin>
      <xmax>368</xmax>
      <ymax>714</ymax>
    </bndbox>
  </object>
</annotation>
```

(ข) ตัวอย่างไฟล์ xml

รูปที่ 4.7: รูปผลลัพธ์การทำงานของหน้าต่าง Track

แอ��พลิเคชันจะนำคีย์เฟรมที่ถูกตีกรอบสี่เหลี่ยมจากหน้าต่าง Detect มาทำนายกรอบสี่เหลี่ยมในเฟรมที่เหลือระหว่างช่วงคีย์เฟรม ซึ่งผลลัพธ์ที่ได้จะได้เฟรมทุกเฟรมที่มีมนุชย์อยู่จะถูกตีกรอบสี่เหลี่ยม ดังรูปที่ 4.7ก จากนั้นสามารถบันทึกข้อมูลออกเป็นไฟล์ xml ได้ดังรูปที่ 4.7ข

ผลลัพธ์การทำงานของหน้าต่าง Label



(ก) ตัวอย่างเฟรมที่ถูกตีกรอบสีเหลืองและคำทำนายการกระทำ

```
<?xml version="1.0"?>
- <annotation>
  <folder>D:/Goggle/Goggle_team/out/Photographer_mp4/Photographer_mp4/img</folder>
  <filename>75.jpg.txt</filename>
  <path>D:/Goggle/Goggle_team/out/Photographer_mp4/Photographer_mp4/img/75.jpg</path>
  - <source>
    <database>Unknown</database>
  </source>
  - <size>
    <width>1280</width>
    <height>720</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  - <object>
    <name>carry/hold (an object)</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    - <bndbox>
      <xmin>2</xmin>
      <ymin>35</ymin>
      <xmax>368</xmax>
      <ymax>714</ymax>
    </bndbox>
  </object>
</annotation>
```

(ข) ตัวอย่างไฟล์ xml

รูปที่ 4.8: รูปผลลัพธ์การทำงานของหน้าต่าง Label

แอพพลิเคชันจะนำกรอบสีเหลืองของทุกเฟรมที่มีมนุษย์อยู่มาทำนายมนุษย์ในกรอบสีเหลืองนั้นกำลังมีการกระทำการอะไรอยู่ โดยสามารถทำงานได้ทั้งหมดอัตโนมัติหรือแบบแก้ไขเอง และสามารถบันทึกข้อมูลออกเป็นไฟล์ xml ได้ดังรูปที่ 4.8x

4.2 ผลการทดลองการตรวจจับวัตถุ

4.2.1 ข้อมูลรายละเอียดประกอบการทดสอบ

จำนวนเฟรมทั้งหมด: 20 เฟรม

จำนวนมนุษย์ที่อยู่ในเฟรม: 0-5 คน

ความละเอียดรูปภาพ: 1280×720 พิกเซล

ขอบเขตอัตราส่วนร่วมของกรอบที่เหลือที่จะนับว่าการทำนายถูกต้อง: 50% ขึ้นไป

4.2.2 ผลทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

ข้อมูลความแม่นยำของโมเดลปัญญาประดิษฐ์เมื่อทดสอบด้วยชุดข้อมูลของผู้วิจัย

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (0.5 IOU)
SSD Mobilenet v1 ppm	63.82	37.03
YOLOv3-320	65.00	64.91
YOLOv3-tiny	17.21	44.44
YOLOv3-spp	65.40	70.30
Faster rcnn inceptionv2	981.21	42.59

ตารางที่ 4.1: ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

จากตารางที่ 4.2 ผู้วิจัยได้ทำการทดสอบความแม่นยำและความเร็วในการประมวลผลของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล พบว่าโมเดลปัญญาประดิษฐ์ที่มีความแม่นยามากที่สุดคือ YOLOv3-spp และ โมเดลปัญญาประดิษฐ์ที่มีความเร็วในการทำนายต่อรูปภาพเร็วที่สุดคือ YOLO-tiny

4.3 ผลการทดสอบการติดตามทำนายตำแหน่งของมนุษย์

4.3.1 ข้อมูลรายละเอียดประกอบการทดสอบ

ชื่อวิดีโอ: Photographer beach photography

ความยาววิดีโอ: 15 วินาที

จำนวนเฟรมทั้งหมด: 374 เฟรม

อัตราเฟรมต่อวินาที: 24.9 เฟรมต่อวินาที

ความละเอียดของวิดีโอ: 1920 × 1080

ความละเอียดของวิดีโอที่ใช้ในการประมวลผลจริง: 1280 × 720

ขอบเขตอัตราส่วนร่วมของกรอบที่เหลือที่จะนับว่าการทำนายถูกต้อง: 80% ขึ้นไป

4.3.2 ผลทดสอบประสิทธิภาพ และความเร็วในการประมวลผล

วิธีการทดสอบ	ความแม่นยำ (%)	ความเร็วในการประมวลผล (วินาที)		
ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมในวิดีโอ	95	-	452	-
ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุก N เฟรมในวิดีโอ แล้วใช้ระบบทำนายตำแหน่งต่อไปของวัตถุในเฟรมระหว่างนั้น				
N = 10	85	-10	69	-383
N = 20	80	-15	41	-411
N = 25	75	-20	35	-417

ตารางที่ 4.2: ผลการทดสอบประสิทธิภาพของการตรวจจับกรอบสีเหลี่ยมภายในวิดีโอ

จากตารางที่ 4.2 ผู้วิจัยได้ทำการทดสอบความแม่นยำและความเร็วในการประมวลผลของการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรม แม้จะตั้งขอบเขตอัตราส่วนร่วมของกรอบที่เหลือที่จะนับว่าการทำนายถูกต้องสูงถึง 80% แต่ความแม่นยำสูงถึง 95% ใช้เวลาในการประมวลผล 452 วินาที เฉลี่ยเฟรมละ 1.2 วินาที ซึ่งถือเป็นความแม่นยำที่สูงมากเมื่อเทียบกับเวลาที่ใช้ในการประมวลผล

ต่อมาเป็นการทดสอบโดยใช้โมเดลปัญญาประดิษฐ์ประมวลผลเฉพาะบางเฟรมทุกๆ ห่วงหนึ่ง แล้วใช้ระบบทำนายตำแหน่งต่อไปของวัตถุในการสร้างกรอบสีเหลี่ยมในเฟรมระหว่างนั้น เพื่อเพิ่มความเร็วในการประมวลผล โดยระยะที่ใช้ในการทดสอบคือ ทุกๆ 10 เฟรม 20 เฟรม และ 25 เฟรม ซึ่งจากการทดสอบนั้นพบว่ามีการนี้มีความแม่นยำลดลงมาเหลือ 85% น้อยกว่าอยู่เพียง 10% เท่านั้น ถือเป็นความแม่นยำที่สูงเมื่อเทียบกันด้วยระยะเวลาในการประมวลผล ในขณะที่การใช้ระยะประมวลผลเป็น 10 เฟรมนั้นใช้เวลาในการประมวลผลเพียง 69 วินาที น้อยกว่าการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมถึง 383 วินาที ซึ่งเร็วกว่าถึง 6.5 เท่า และความแม่นยำลดลงมาเหลือ 85% น้อยกว่าอยู่เพียง 10% เท่านั้น ถือเป็นความแม่นยำที่สูงเมื่อเทียบกันด้วยระยะเวลาในการประมวลผล ในขณะที่การใช้ระยะประมวลผล 20 เฟรมนั้นจะประมวลผลเร็วกว่าการใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมถึง 11 เท่า และมีความแม่นยำต่ำกว่า 15% และเมื่อใช้ระยะประมวลผล 25 เฟรมจะเร็วกว่าประมาณ 13 เท่า และความแม่นยำต่ำลงถึง 20%

4.4 ผลการทดสอบระบบระบุตัวตนของมนุษย์

4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของบุคคล

ความแม่นยำของโมเดลปัญญาประดิษฐ์จากแหล่งที่มีมาตราฐานด้านล่างดังนี้

โมเดลปัญญาประดิษฐ์	rank1/mAP โดยใช้วิธีการทดสอบด้วย Global+DMLI
ResNet50 Market1501	91.0/77.6
ResNet50 DukeMTMCRID	80.7/68.0
ResNet50 CUHK03	60.9/59.7
ResNet50 MSMT17	66.3/40.6

ตารางที่ 4.3: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์

ต่อมานำโมเดลปัญญาประดิษฐ์แต่ละอันมาทดสอบกับตัวอย่างภาพชุดข้อมูลที่ทางคณะผู้วิจัยได้สร้างขึ้น โดยภาพชุดข้อมูลที่นำมาใช้จะผ่านการตรวจหาบุคคลภายในภาพด้วยโมเดลปัญญาประดิษฐ์ YOLO v3 320 และจะเป็นการทดลองในกรณีที่บุคคลในภาพนั้นเป็นบุคคลเดียวกัน



รูปที่ 4.9: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 1

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.4308
ResNet50 DukeMTMCRID	0.4827
ResNet50 CUHK03	0.4914
ResNet50 MSMT17	0.4668

ตารางที่ 4.4: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 1



รูปที่ 4.10: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 2

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.3035
ResNet50 DukeMTMCReID	0.3332
ResNet50 CUHK03	0.3042
ResNet50 MSMT17	0.3684

ตารางที่ 4.5: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 2



รูปที่ 4.11: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 3

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.3308
ResNet50 DukeMTMCReID	0.3296
ResNet50 CUHK03	0.3134
ResNet50 MSMT17	0.3968

ตารางที่ 4.6: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 3

ต่อมาจะเป็นการทดลองในกรณีที่บุคคลในภาพไม่เป็นบุคคลเดียวกัน



รูปที่ 4.12: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 4

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.7285
ResNet50 DukeMTMCReID	0.6882
ResNet50 CUHK03	0.6727
ResNet50 MSMT17	0.7408

ตารางที่ 4.7: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 4



รูปที่ 4.13: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 5

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.6098
ResNet50 DukeMTMCReID	0.6522
ResNet50 CUHK03	0.6 275
ResNet50 MSMT17	0.6155

ตารางที่ 4.8: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 5



รูปที่ 4.14: ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 6

โมเดลปัญญาประดิษฐ์	ค่าสำหรับการระบุบุคคล (Original distance)
ResNet50 Market1501	0.6159
ResNet50 DukeMTMCReID	0.5352
ResNet50 CUHK03	0.5888
ResNet50 MSMT17	0.6119

ตารางที่ 4.9: ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 6

ค่าความแม่นยำในการระบุบุคคลนั้นค่าอยู่ที่ 0 แสดงบุคคลใน 2 เฟรมนั้นเป็นบุคคลเดียวกัน จากการทดลองครั้งที่ 1 จะเป็นเฟรมที่ไม่ต่อเนื่องกัน การทดลองครั้งที่ 2 และ 3 นั้นจะเป็นเฟรมที่ต่อเนื่องกันมากขึ้นตามลำดับ และการทดลองที่ 4 5 และ 6 นั้นจะนำภาพที่แต่ละบุคคลที่ท่าทางใกล้เคียงกันมาใช้ ซึ่งจะแสดงให้เห็นว่า โมเดลปัญญาประดิษฐ์ที่สามารถให้ผลลัพธ์ที่มีประสิทธิภาพต่อเนื่องมากที่สุดคือ ResNet50 Market1501

4.5 ผลการทดสอบการจัดการกระทำของมนุษย์

4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกтренร์ผ่าน AVA เพียงผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (PASCAL mAP)
แหล่งอ้างอิง	0.93	11
ผลการทดสอบของผู้วิจัย	5,279	6.8

ตารางที่ 4.10: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์เทียบผลลัพธ์กับแหล่งอ้างอิง

ความเร็วของตอรูปภาพทางผู้วิจัยได้ใช้กราฟฟิกการ์ด GTX 2080 Ti ในการทดสอบซึ่งจะให้ความเร็วอยู่ที่ 5 วินาที ซึ่งทางแหล่งอ้างอิงนั้นใช้กราฟฟิกการ์ด Nvidia GeForce GTX TITAN X ในส่วนของค่าความแม่นยำที่ไม่เท่ากัน คาดว่าจะเป็นเพราะการประมวลผลของกราฟฟิกการ์ดของรุ่นที่ต่างกันจึงทำให้ค่า mAP ที่ออกมามีเท่ากัน

4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนร์ผ่าน AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนร์ผ่านชุดข้อมูลสำหรับการเทรนร์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์การทดสอบก่อนหน้า

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (PASCAL mAP)
แหล่งอ้างอิง	X	X
ผลการทดสอบของผู้วิจัย	X	X

ตารางที่ 4.11: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ เมื่อใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (PASCAL mAP)
ผลการทดสอบที่ผ่านมา	X	X
ผลการทดสอบของผู้วิจัย	X	X

ตารางที่ 4.12: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น ใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

ภาคผนวก

ภาคผนวก ก

ตัวอย่างชุดข้อมูลที่ผู้วิจัยสร้างขึ้น

ตัวอย่างชุดข้อมูลสำหรับการทดสอบโมเดลปัญญาประดิษฐ์ในการตรวจจับภาพบุคคล



รูปที่ ก.1: รูปผลลัพธ์การทำงานของหน้าต่าง Track