



Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562





Goggle : People Video Analytics and Deep Learning Platform

Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์

นายปฐมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

Google แพลตฟอร์มการเรียนรู้เชิงลึกและระบบบวิเคราะห์การกระทำของมนุษย์

นายปฐุมพงศ์ สินธุจาม

นายศุภกร เบญจวิกรัย

นายอุกฤษฎ์ เลิศวรณาการ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาชีวกรรมหุ่นยนต์และระบบอัตโนมัติ

สถาบันวิทยาการหุ่นยนต์ภาคสนาม

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ปีการศึกษา 2562

คณะกรรมการสอบวิทยานิพนธ์

ประธานกรรมการสอบวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

(ดร.วรารสิณี ฉายแสงมงคล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์

()

กรรมการสอบวิทยานิพนธ์

(อ.บวรศักดิ์ ศกุลเกื้อกูลสุข)

กรรมการสอบวิทยานิพนธ์

(ดร.บุญทริกา เกษมลันติธรรม)

ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

ชื่อวิทยานิพนธ์	Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์
หน่วยกิต	6
ผู้เขียน	นายปัจมพงศ์ สินธุจัน นายศุภกร เบญจวิกรัย นายอุกฤษฎ์ เลิศวรรณการ
อาจารย์ที่ปรึกษา	ทีปรึกษาวิทยานิพนธ์หลัก ดร.วราสินี ฉายแสงมงคล
หลักสูตร	วิศวกรรมศาสตรบัณฑิต
สาขาวิชา	วิศวกรรมที่นุนยนต์และระบบอัตโนมัติ
คณะ	สถาบันวิทยาการที่นุนยนต์ภาคสนาม
ปีการศึกษา	2562

---

## บทคัดย่อ

งานวิทยานิพนธ์นี้เป็นงานที่เกี่ยวกับการออกแบบและสร้างเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ รวมถึงระบบวิเคราะห์การกระทำของมนุษย์ โดยใช้ชื่อว่า Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบวิเคราะห์การกระทำของมนุษย์ ซึ่งมีจุดประสงค์เพื่อให้ผู้พัฒนาสามารถใช้งานเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ในการสร้างชุดข้อมูลสำหรับสร้างปัญญาประดิษฐ์ได้ง่ายและสะดวกขึ้น ภาพรวมของวิทยานิพนธ์นี้จะแบ่งออกเป็นหัวข้อหลักสองส่วน โดยที่ส่วนแรกเป็นการออกแบบและสร้างแอปพลิเคชันที่ใช้ในการสร้างชุดข้อมูลจากวิดีโอ สำหรับพัฒนาโมเดลปัญญาประดิษฐ์ และส่วนที่สองเป็นใช้ชุดข้อมูลที่ได้จากแอปพลิเคชันในการพัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการกระทำของมนุษย์ เพื่อทดสอบประสิทธิภาพของชุดข้อมูล

คำสำคัญ : ระบบวิเคราะห์การกระทำการกระทำของมนุษย์ / 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 และบุคลากรในสถาบันวิทยาการหุ่นยนต์ภาควิชานามทุกท่าน ที่ได้ให้คำปรึกษา และช่วยเหลือด้านสถานที่พร้อมทั้งส่งอำนวยความสะดวกต่างๆ ในระหว่างการทำวิทยานิพนธ์

ขอขอบคุณนักศึกษาปริญญาตรี สถาบันวิทยาการหุ่นยนต์ภาควิชานามทุกท่าน ที่ได้ให้คำแนะนำ ถามไถ่ และเป็นกำลังใจมาโดยตลอด

และสุดท้ายนี้ ขอน้อมรำลึกถึงพระคุณบิดา มารดา และครอบครัว ที่ส่งเสริมให้กำลังใจ และให้การสนับสนุนในเรื่องต่างๆ จนกระทั้งข้าพเจ้าประสบความสำเร็จในการศึกษา

นายปฐมพงศ์ สินธุรงาม  
นายศุภกร เบญจวิกรัย  
นายอุกฤษฎ์ เลิศวรรณาการ

## สารบัญ

เรื่อง	หน้า
บทคัดย่อ .....	๑
กิตติกรรมประกาศ .....	๔
สารบัญ .....	๕
รายการรูปภาพ .....	๗
รายการตาราง .....	๘
รายการสัญลักษณ์ .....	๙
ประมวลศัพท์และตัวย่อ .....	๙
<b>บทที่ 1 บทนำ .....</b>	<b>1</b>
1.1 ที่มาและความสำคัญ .....	1
1.2 วัตถุประสงค์ .....	1
1.3 ประโยชน์ที่คาดว่าจะได้รับ .....	1
1.4 ขอบเขตการดำเนินงาน .....	2
1.5 ขั้นตอนการดำเนินงาน .....	2
<b>บทที่ 2 ทฤษฎี/การวิจัยที่เกี่ยวข้อง .....</b>	<b>4</b>
2.1 ทฤษฎีที่เกี่ยวข้อง .....	5
2.1.1 Convolutional Neural Network .....	5
2.1.2 Optical flow .....	9
2.1.3 โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุที่เกี่ยวข้อง .....	11
2.2 การประมวลผลวิดีโอ .....	14
2.2.1 การตรวจจับวัตถุ .....	14
2.2.2 ระบบติดตามการเคลื่อนไหวของวัตถุ .....	15
2.2.3 ระบบระบุตัวตนของบุคคล .....	17
2.2.4 ระบบจำแนกการกระทำของมนุษย์ .....	18
2.3 ชุดข้อมูลสำหรับการประมวลผลวิดีโอ .....	20
2.4 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการของมนุษย์ .....	34
2.4.1 ResNet .....	34
2.4.2 Inflated 3D convolutional network .....	36

## สารบัญ (ต่อ)

เรื่อง	หน้า
2.4.3 เครื่องมือกำกับคุณลักษณะ .....	38
บทที่ 3 ระเบียบวิจัย .....	40
3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	40
3.2 ความต้องการของระบบ.....	41
3.2.1 ความต้องการใช้งาน (functional requirements).....	41
3.2.2 ความต้องการเชิงวิศวกรรม (non-functional requirements).....	41
3.3 หน้าที่ความรับผิดชอบ.....	42
3.4 เครื่องมือที่ใช้ในงานวิจัย .....	42
3.5 ภาษาที่ใช้ในการพัฒนาระบบ .....	43
3.6 Program library ที่ใช้ในการพัฒนาระบบและแอพพลิเคชัน.....	43
3.7 แผนการดำเนินงาน .....	43
3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	44
3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	44
3.9 การออกแบบการทดสอบการตรวจจับวัตถุ.....	54
3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล .....	54
3.10 การออกแบบการทดสอบการทำงานโดยทำให้แน่ใจว่าโมเดลปัญญาประดิษฐ์สามารถทำงานได้แม่นยำ .....	55
3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำให้แน่ใจว่าโมเดลปัญญาประดิษฐ์สามารถทำงานได้แม่นยำ .....	55
3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์ .....	56
3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ .....	56
3.12 การออกแบบการทดสอบการจำแนกประเภทของมนุษย์ .....	57
3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง .....	57
3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง .....	58
3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง .....	59

## สารบัญ (ต่อ)

เรื่อง	หน้า
บทที่ 4 ผลการดำเนินงาน .....	60
4.1 เครื่องมือกำกับคุณลักษณะ .....	60
4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน .....	60
4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอพพลิเคชัน .....	64
4.2 ผลการทดลองการตรวจจับวัตถุ .....	67
4.2.1 ข้อมูลรายละเอียดประกอบการทดสอบ .....	67
4.2.2 ผลทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำงาน ตรวจจับภาพบุคคล .....	67
4.3 ผลการทดสอบระบบติดตามตำแหน่งของมนุษย์ .....	68
4.3.1 ข้อมูลรายละเอียดประกอบการทดสอบ .....	68
4.3.2 ผลทดสอบประสิทธิภาพ และความเร็วในการประมวลผล .....	68
4.4 ผลการทดสอบระบบระบุตัวตนของมนุษย์ .....	69
4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตน ของบุคคล .....	69
4.5 ผลการทดสอบการจัดการกระทำของมนุษย์ .....	72
4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้ .....	72
4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง ...	72
4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่านชุด ข้อมูลสำหรับการเทรนด์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบ และเทียบผลลัพธ์การทดสอบก่อนหน้า .....	72
เอกสารอ้างอิง .....	75
ภาคผนวก ก ตัวอย่างชุดข้อมูลที่ผู้วิจัยสร้างขึ้น .....	76

## รายการรูปภาพ

รูป	หน้า
รูปที่ 2.1 ตัวอย่างโครงสร้างของ CNN ที่ใช้ในการจำแนกประเภทของรูปภาพ .....	5
รูปที่ 2.2 ตัวอย่างเครื่องเนล และภาพที่ใช้ในการประมวลผล .....	5
รูปที่ 2.3 ตัวอย่างการหาผังคุณลักษณะ .....	6
รูปที่ 2.4 ตัวอย่างการใช้ max pooling และ average pooling กับภาพ .....	6
รูปที่ 2.5 ภาพแสดงการหา IoU ของกรอบสี่เหลี่ยมจริงของเฟรม และ กรอบสี่เหลี่ยมที่ทำนายขึ้นมา .....	8
รูปที่ 2.6 ตัวอย่างการทำงานของ NMS .....	8
รูปที่ 2.7 ตัวอย่างการเคลื่อนที่ของลูกบอล .....	9
รูปที่ 2.8 กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO .....	11
รูปที่ 2.9 โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ YOLO .....	11
รูปที่ 2.10 โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO .....	12
รูปที่ 2.11 โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ Faster RCNN .....	13
รูปที่ 2.12 โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ SSD .....	14
รูปที่ 2.13 แนวคิดของระบบติดตามการเคลื่อนไหวของวัตถุ .....	16
รูปที่ 2.14 การแบ่งรูปออกเป็น 8 ส่วนของระบบบุตตันของบุคคล .....	17
รูปที่ 2.15 ตัวอย่างการประยุกต์ใช้งานระบบจำแนกการกระทำนุழย์ <sup>[12]</sup> .....	18
รูปที่ 2.16 โครงสร้างโมเดลปัญญาประดิษฐ์ของ deep bag of frames .....	22
รูปที่ 2.17 โครงสร้าง LSTM ที่ใช้การอ้างอิงในบทความนี้ .....	22
รูปที่ 2.18 หลักการของ Residual block ของ ResNet .....	34
รูปที่ 2.19 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D <sup>[5]</sup> .....	36
รูปที่ 2.20 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D <sup>[5]</sup> .....	36
รูปที่ 2.21 UI ของโปรแกรม DarkLabel .....	38
รูปที่ 2.22 UI ของโปรแกรม OpenLabeling .....	39
รูปที่ 3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ .....	40
รูปที่ 3.2 กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ .....	44
รูปที่ 3.3 หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ .....	45
รูปที่ 3.4 ตำแหน่งของแต่ละวิธีเจ็ตในหน้าต่าง Select .....	46
รูปที่ 3.5 หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ .....	47

## รายการรูปภาพ (ต่อ)

รูป	หน้า
รูปที่ 3.6 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect.....	48
รูปที่ 3.7 หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ .....	49
รูปที่ 3.8 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track.....	50
รูปที่ 3.9 หน้าต่าง Label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	51
รูปที่ 3.10 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Label.....	52
รูปที่ 3.11 ตัวอย่างข้อมูลภายในไฟล์ xml .....	53
รูปที่ 4.1 รูปหน้าต่างแสดงผลของหน้าต่าง Select .....	60
รูปที่ 4.2 รูปหน้าต่างแสดงผลของหน้าต่าง Detect .....	61
รูปที่ 4.3 รูปหน้าต่างแสดงผลของหน้าต่าง Track.....	62
รูปที่ 4.4 รูปหน้าต่างแสดงผลของหน้าต่าง Label.....	63
รูปที่ 4.5 รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่และไม่มีมนุษย์อยู่ภายในเฟรม .....	64
รูปที่ 4.6 รูปคิริ์เฟรมที่ถูกตีกรอบสีเหลี่ยมในส่วนที่มีมนุษย์อยู่ .....	64
รูปที่ 4.7 รูปผลลัพธ์การทำงานของหน้าต่าง Track.....	65
รูปที่ 4.8 รูปผลลัพธ์การทำงานของหน้าต่าง Label .....	66
รูปที่ 4.9 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 1 .....	69
รูปที่ 4.10 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 2.....	70
รูปที่ 4.11 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 3.....	70
รูปที่ 4.12 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 4.....	71
รูปที่ 4.13 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 5.....	71
รูปที่ 4.14 ภาพตัวอย่างชุดข้อมูลสำหรับการทดลองครั้งที่ 6.....	72
รูปที่ ก.1 รูปผลลัพธ์การทำงานของหน้าต่าง Track.....	76

## รายการตาราง

ตาราง	หน้า
ตารางที่ 1.1 แผนการดำเนินงาน .....	3
ตารางที่ 2.1 ตารางแสดงการเปรียบเทียบระหว่าง softmax function และ sigmoid function.....	7
ตารางที่ 2.2 ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ YouTube-8M.....	25
ตารางที่ 2.3 ผลการทดลองของวิธีต่างๆบนคุณลักษณะระดับเฟรม .....	29
ตารางที่ 2.4 ประสิทธิภาพของโมเดล Resnet50 I3D ที่ใช้ชุดข้อมูล Kinetics และ Moments in Time ..	31
ตารางที่ 2.5 อัตราเร้อยล์ของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet.....	34
ตารางที่ 2.6 ค่าความผิดพลาดที่ได้จากการทดลองจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10 .....	35
ตารางที่ 2.7 ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อใช้ข้อมูลจาก UCF-101, HMDB-51 และ Kinetics ในการสร้างและทดสอบด้วยเครื่องมือวัดผลแบบความแม่นยำจากการทำนายอันดับแรกสุด.....	37
ตารางที่ 4.1 ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจสอบจับภาพบุคคล .....	67
ตารางที่ 4.2 ผลการทดสอบประสิทธิภาพของการตรวจจับกรอบสี่เหลี่ยมภายในวิดีโอ.....	68
ตารางที่ 4.3 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์.....	69
ตารางที่ 4.4 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 1 .....	69
ตารางที่ 4.5 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 2 .....	70
ตารางที่ 4.6 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 3 .....	70
ตารางที่ 4.7 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 4 .....	71
ตารางที่ 4.8 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 5 .....	71
ตารางที่ 4.9 ผลการทดสอบความแม่นยำสำหรับการระบุบุคคลของโมเดลปัญญาประดิษฐ์ครั้งที่ 6 .....	72
ตารางที่ 4.10 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ยึดผลลัพธ์กับแหล่งอ้างอิง.....	72
ตารางที่ 4.11 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ เมื่อใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น.....	73
ตารางที่ 4.12 ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยสร้างขึ้น ใช้กับชุดข้อมูลที่ผู้วิจัยสร้างขึ้น .....	73

## รายการสัญลักษณ์

$\theta$	เซ็ต้า
$d$	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 โดยโครงการนี้เน้นศึกษาเกี่ยวกับการวิเคราะห์และจำแนกการกระทำของมนุษย์ (human action classification) ภายในสำนักงานจากวิดีโอด้วยหลัก

#### 1.2 วัตถุประสงค์

- เพื่อสร้างต้นแบบของเครื่องมือกำกับคุณลักษณะด้วยปัญญาประดิษฐ์ ที่มีนุบำรุงและปัญญาประดิษฐ์ทำงานร่วมกันเพื่อสร้างชุดข้อมูลในการนำไปพัฒนาปัญญาประดิษฐ์อื่นๆ ที่เหมาะสมกับปัญหาที่ต้องการ
- เพื่อออกแบบและสร้างต้นแบบของระบบประมวลผลวิดีโอที่สามารถตรวจจับมนุษย์และจำแนกการกระทำพื้นฐานของมนุษย์ภายในสำนักงาน ประกอบด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว นอน โดยใช้ปัญญาประดิษฐ์
- เพื่อสร้างเครื่องมือที่สามารถสร้างชุดข้อมูลสำหรับการจำแนกการกระทำการของมนุษย์ให้สามารถใช้งานได้ง่าย สะดวกสบายมากขึ้น และมีประสิทธิภาพที่สูงกว่าเครื่องมือตัวอื่นในปัจจุบัน

#### 1.3 ประโยชน์ที่คาดว่าจะได้รับ

- เพิ่มความสะดวกในการสร้างชุดข้อมูลสำหรับพัฒนาโมเดลปัญญาประดิษฐ์จากวิดีโอ
- ต้นแบบระบบประมวลผลวิดีโอที่สามารถจำแนกการกระทำการของมนุษย์

## 1.4 ขอบเขตการดำเนินงาน

1. สร้างต้นแบบของเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์ โดยระบบจะประกอบไปด้วยสี่ส่วนดังนี้
  - (a) หน้าต่างของแอพพลิเคชัน (user interface)
  - (b) ระบบตรวจจับมนุษย์ในภาพ (person detection)
  - (c) ระบบติดตามการเคลื่อนไหวของมนุษย์ในวิดีโอ (person tracker)
  - (d) ระบบจำแนกการกระทำการของมนุษย์ซึ่งประกอบไปด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว นอน
2. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการของมนุษย์กับชุดข้อมูลที่ได้จากการใช้งานจริงได้หรือไม่
3. พัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการของมนุษย์ภายในสำนักงานอย่างน้อย 2 โมเดล

## 1.5 ขั้นตอนการดำเนินงาน

การดำเนินงานวิจัยถูกแบ่งออกเป็นสามส่วน โดยส่วนแรกคือการศึกษาเทคโนโลยีในปัจจุบันเพื่อทำความเข้าใจและกำหนดขอบเขตของงาน ส่วนที่สองคือออกแบบและสร้างเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์ เพื่อช่วยผู้พัฒนาในการสร้างชุดข้อมูล และส่วนที่สุดท้ายคือการนำชุดข้อมูลที่ได้จากการใช้เครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์ไปพัฒนาโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง

1. ศึกษาเกี่ยวกับการประมวลผลวิดีโอ (video analytics)
2. ศึกษาเกี่ยวกับชุดข้อมูลสำหรับการประมวลผลวิดีโอ
3. ศึกษาเกี่ยวกับโมเดลปัญญาประดิษฐ์ที่ใช้ในการประมวลผลวิดีโอ
4. ศึกษาเครื่องมือที่ใช้ในการช่วยสร้างชุดข้อมูลจากวิดีโอ

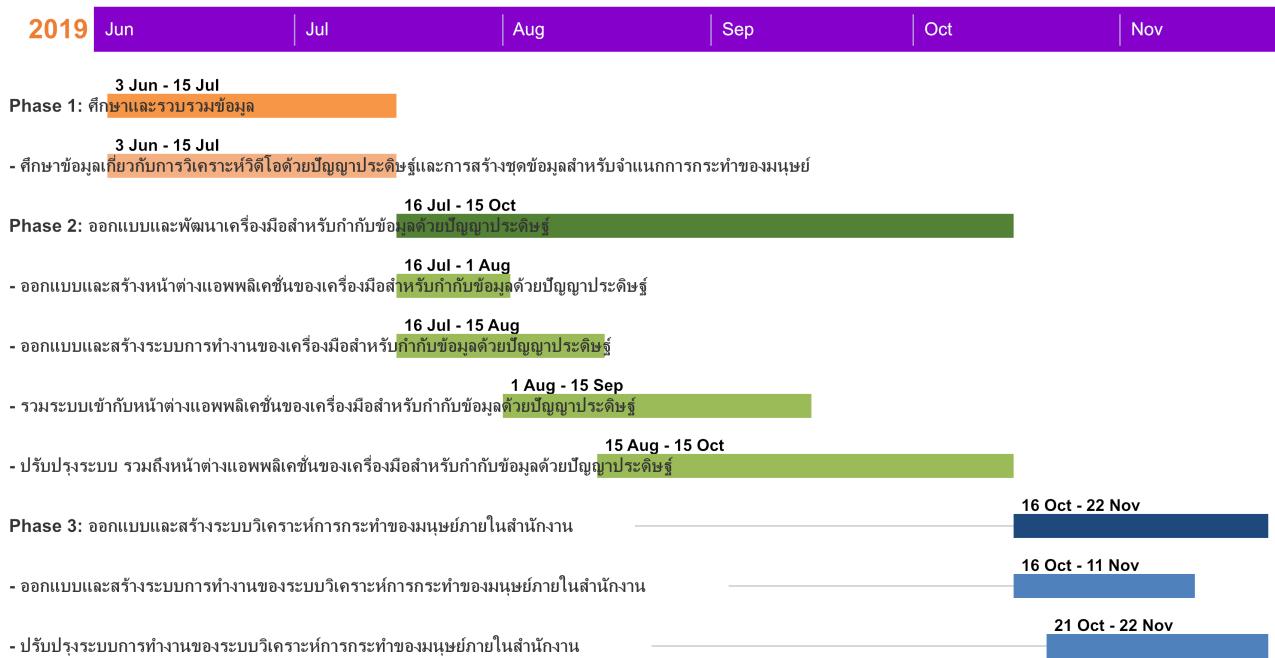
**เครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์**

1. ออกแบบและสร้างหน้าต่างแอพพลิเคชันของเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. ออกแบบและสร้างระบบของเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์
3. ทดสอบและปรับปรุงการทำงานของเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์

**โมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน**

1. สร้างชุดข้อมูลสำหรับสร้างโมเดลปัญญาประดิษฐ์จากเครื่องมือที่มีการทำงานร่วมกับคุณลักษณะด้วยปัญญาประดิษฐ์
2. สร้างโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน
3. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำการของมนุษย์ภายในสำนักงาน

## แผนการดำเนินงาน



ตารางที่ 1.1: แผนการดำเนินงาน

## บทที่ 2

### ทฤษฎี/การวิจัยที่เกี่ยวข้อง

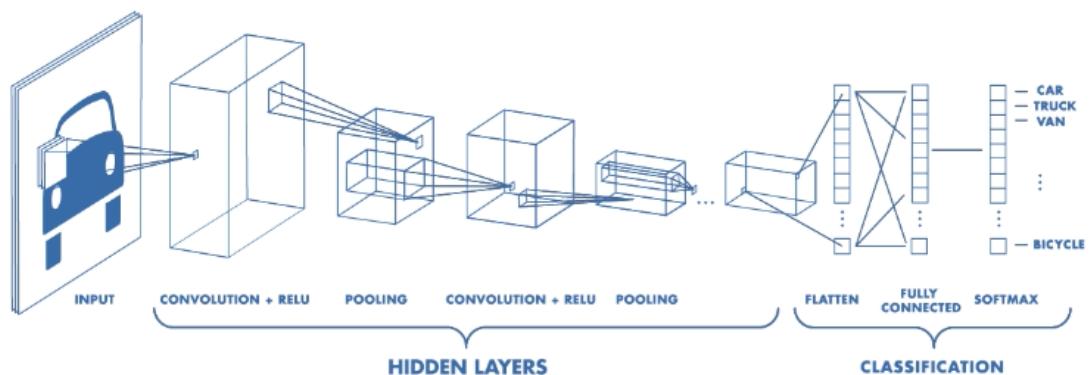
การประมวลผลวิดีโอในปัจจุบันนั้นมีวิธีการและเทคนิคมาหลาย ผู้วิจัยจึงต้องศึกษาองค์ความรู้และงานวิจัยที่เกี่ยวข้องกับวัตถุประสงค์ของงาน เพื่อศึกษาและใช้เป็นแนวทางในการประยุกต์สำหรับสร้างเครื่องมือกำกับข้อมูลด้วยปัญญาประดิษฐ์ และโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ ซึ่งหัวข้อที่ผู้วิจัยได้ไปศึกษามา มีดังต่อไปนี้

1. ทฤษฎีที่เกี่ยวข้อง
  - (a) Convolutional neural network
  - (b) Optical flow
  - (c) โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุที่เกี่ยวข้อง
2. การประมวลผลวิดีโอ
  - (a) การตรวจจับวัตถุ (object detection)
  - (b) การติดตามตำแหน่งของวัตถุ (object tracker)
  - (c) การระบุตัวตนของบุคคล (person re-identification)
  - (d) การจำแนกการกระทำของมนุษย์ (human action classification)
3. โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมนุษย์ (human action classification model)
4. เครื่องมือกำกับชุดข้อมูลภาพ

## 2.1 ทฤษฎีเกี่ยวกับ

### 2.1.1 Convolutional Neural Network

Convolution Neural Network (CNN) คือโมเดลปัญญาประดิษฐ์ประเภทหนึ่งมักจะนำมาใช้กับงานที่เกี่ยวกับการจำแนกวัตถุในภาพ เช่น แมว หมา มุขย์ รถ เป็นต้น การที่ CNN สามารถจำแนกภาพอ комากได้ว่าเป็นประเภทอะไรนั้นต้องผ่านชั้นตัวกรอง (filter layer) หรือเคอร์เนล (kernel), pooling layer, fully connected layer (FC) และใช้ softmax หรือ logistic function ในการจำแนกว่าเป็นวัตถุประเภทไร ดังรูปที่ 2.1



รูปที่ 2.1: ตัวอย่างโครงสร้างของ CNN ที่ใช้ในการจำแนกประเภทของรูปภาพ

### ตัวกรอง/เคอร์เนล (Filter/Kernel)

ตัวกรองหรือเคอร์เนล คือชั้นที่ใช้ในการสกัดคุณลักษณะของรูปภาพอ กมาด้วยสี่เหลี่ยมเล็กๆ ขนาด  $N \times N$  โดยที่  $N \in [1, 2, 3, \dots]$  ดังรูปที่ 2.2ก และสมมติให้ภาพที่ใช้ในการประมวลผลเป็นดังรูปที่ 2.2ข

1	0	1
0	1	0
1	0	1

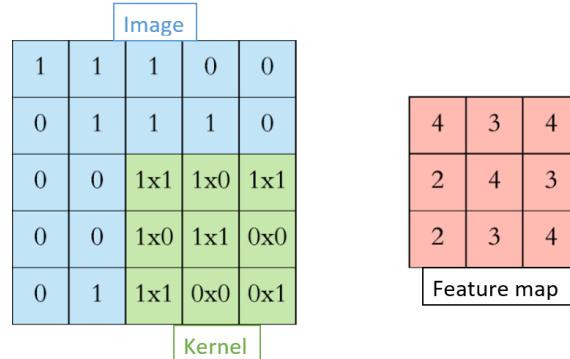
(ก) ตัวอย่างเคอร์เนลขนาด  $3 \times 3$

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	1	0	0

(ข) ตัวอย่างภาพที่ใช้ในการประมวลผล

รูปที่ 2.2: ตัวอย่างเคอร์เนล และภาพที่ใช้ในการประมวลผล

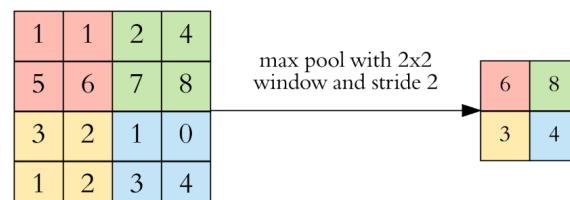
เมื่อนำเครื่องเนล (รูปที่ 2.2ก) ไปทาบกับภาพ (รูปที่ 2.2ข) แล้วคุณค่าในเครื่องเนลกับพิกเซล (pixel) ที่ทาบจะได้คุณลักษณะของช่องนั้นจากนั้นเลื่อนต่อไปจนครบทั้งรูป ซึ่งการเลื่อน (Stride) นั้นขึ้นอยู่กับผู้สร้างว่าต้องการจะให้เลื่อนครั้งละกี่ช่อง แต่ระยะการเลื่อนที่มากขึ้นจะทำให้สัมพันธ์ของคุณลักษณะที่ได้ออกมาแน่นอยู่ด้วยกัน 皱纹เครื่องเนลเทียบบนภาพนั้นจะวางไม่ให้เกินกรอบรูป แต่ถ้าต้องการทابกับทุกพิกเซลในภาพสามารถทำได้โดยการให้พื้นที่ที่เกินขอบภาพไปเท่ากับ 0 เทคนิคนี้เรียกว่า padding และคุณลักษณะที่ได้ออกมาทั้งหมดจะเรียกว่าผังคุณลักษณะ (features map) ตามรูปที่ 2.3



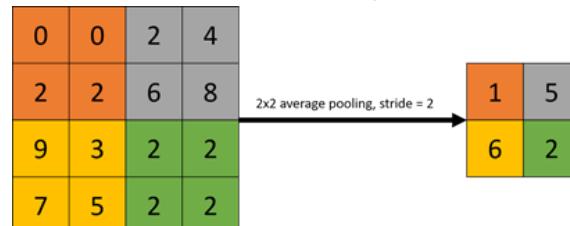
รูปที่ 2.3: ตัวอย่างการทำผังคุณลักษณะ

### Pooling

Pooling คือขั้นที่สามารถลดขนาดของภาพลงเพื่อลดข้อมูลที่ไม่จำเป็นลง ซึ่งมีหลายประเภทแต่นิยมใช้มีสองประเภทได้แก่ max pooling และ average pooling โดยที่ max pooling จะใช้ในการหาค่าที่มากที่สุดในเครื่องเนลที่ทาบอยู่ดังรูปที่ 2.4ก ในขณะที่ average pooling จะหาค่าเฉลี่ยของภายในเครื่องเนลออกม�다ังรูปที่ 2.4ข



(ก) ตัวอย่างการทำ max pooling



(ข) ตัวอย่างการทำ average pooling

รูปที่ 2.4: ตัวอย่างการใช้ max pooling และ average pooling กับภาพ

### Activation function

ในแต่ละชั้นของ CNN นั้นจะมี activation function เป็นสิ่งที่กำหนดว่าเอาท์พุตของชั้นนั้นว่าจะอยู่ในช่วงไหน เช่น -1 ถึง 1, 0 ถึง 1 ขึ้นอยู่กับฟังก์ชันที่ใช้ ซึ่งจะแบ่งออกเป็นสองแบบคือ แบบเป็นเส้นตรง และแบบไม่เป็นเส้นตรง

แบบเป็นเส้นตรงจะมีสมการของฟังก์ชันดังนี้  $f(x) = x$  โดยที่  $x$  คืออินพุตของฟังก์ชัน ทำให้อาท์พุตที่ได้จากฟังก์ชันนี้มีค่าอยู่ในช่วง  $-\infty$  ถึง  $\infty$

แบบไม่เป็นเส้นตรงจะมีหลายฟังก์ชัน แต่ฟังก์ชันที่เป็นที่นิยมใช้คือ Rectified Linear Unit หรือ ReLU, Leaky ReLU และ Parametric ReLU โดยทั้งสามฟังก์ชันมีสมการดังนี้

ReLU

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{otherwise} \end{cases}$$

Leaky ReLU

$$f(x) = \begin{cases} 0.01x & \text{if } x < 0, \\ x & \text{otherwise} \end{cases}$$

Parametric ReLU โดยที่  $\alpha \in \mathbb{R}^+$

$$f(x) = \begin{cases} \alpha x & \text{if } x < 0, \\ x & \text{otherwise} \end{cases}$$

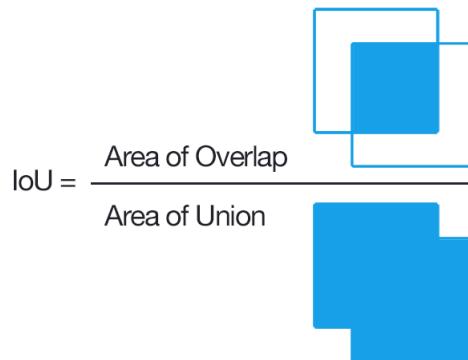
### Fully connected layer

เป็นชั้นที่จะรวมคุณลักษณะทั้งหมดของชั้นก่อนหน้าเป็นเวกเตอร์ ก่อนที่จะนำเวกเตอร์ที่ได้ไปผ่าน activation function เพื่อกำหนดคำตอบสำหรับการแยกประเภทของภาพ ซึ่งฟังก์ชันที่นิยมใช้จะมีสองฟังก์ชัน คือ softmax และ sigmoid (logistic) โดยทั้งสองมีความแตกต่างกันดังตารางที่

	Softmax	Sigmoid
1	สมการคือ $f(x_i) = \frac{\exp(x_i)}{\sum_{j=0}^k \exp(x_j)}$ โดยที่ k คือจำนวนข้อมูล และ i ∈ 0, 1, 2, ..., k	$f(x_i) = \frac{1}{1+\exp(-x_i)}$ โดยที่ i ∈ 0, 1, 2, ..., k
2	ผลรวมของความน่าจะเป็นจะเท่ากับ 1 เสมอ	ผลรวมของความน่าจะเป็นไม่จำเป็นต้องเท่ากับ 1
3	มักใช้ในการจำแนกประเภทมากกว่าสองประเภทขึ้นไป	มักใช้ในการจำแนกประเภทเพียงสองประเภท

ตารางที่ 2.1: ตารางแสดงการเปรียบเทียบระหว่าง softmax function และ sigmoid function

### Intersection Over Union (IoU)



รูปที่ 2.5: ภาพแสดงการหา IoU ของกรอบสี่เหลี่ยมจริงของเฟรม และ กรอบสี่เหลี่ยมที่ทำนายขึ้นมา

เป็นวิธีในการทดสอบประสิทธิภาพของการตรวจจับวัตถุ โดยค่า IoU นั้นสามารถหาได้จากการนำกรอบสี่เหลี่ยมจริงของวัตถุ และกรอบสี่เหลี่ยมที่ได้จากการทำนายมาหารด้วยจำนวนที่ต้องห้ามทั้งสองทับซ้อนกัน และหารด้วยพื้นที่ทั้งหมดของกรอบสี่เหลี่ยมทั้งสองรวมกัน ซึ่งสามารถเขียนในรูปสมการได้ดังนี้

$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|} \quad (2.1)$$

โดยที่

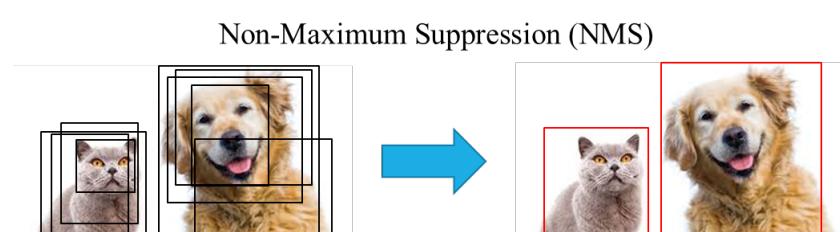
$\text{IoU}$  = ค่าที่ใช้สำหรับวัดผลความใกล้เคียงระหว่างสองกรอบสี่เหลี่ยม

$P$  = พื้นที่ของกรอบสี่เหลี่ยมที่ทำนายได้

$G$  = พื้นที่ของกรอบสี่เหลี่ยมจริงของรูปภาพ

### Non Maximum Suppression (NMS)

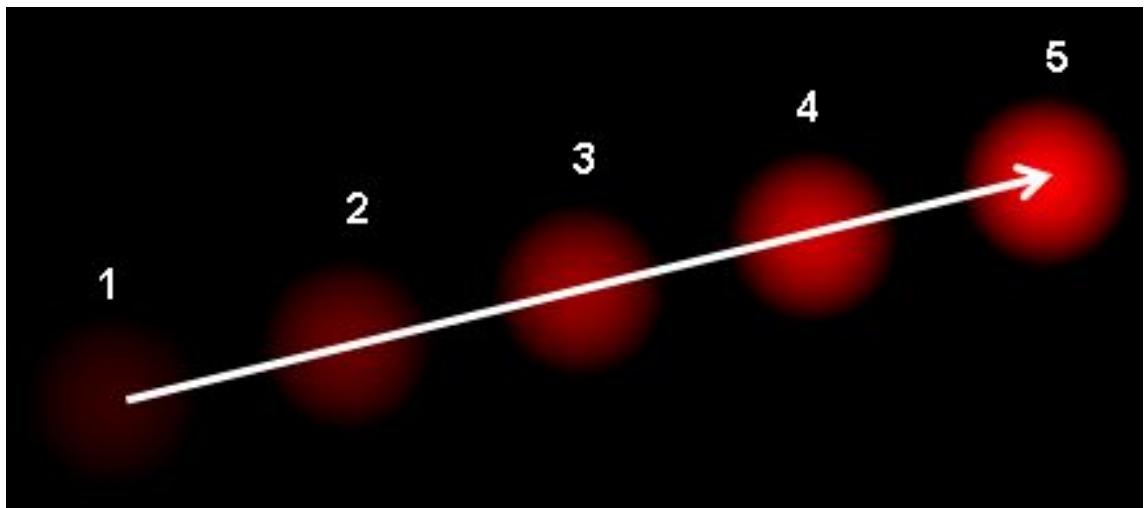
Non Maximum Suppression คือ อัลกอริทึมที่นิยมใช้เข้ามาช่วยจัดการปัญหากรอบสี่เหลี่ยมที่ซ้อนทับกันซึ่งเกิดจากการทำนายซ้ำ เพื่อให้ได้กรอบสี่เหลี่ยมที่บ่งบอกถึงตำแหน่งของวัตถุนั้นเพียงกรอบเดียว



รูปที่ 2.6: ตัวอย่างการทำงานของ NMS

### 2.1.2 Optical flow

Optical flow<sup>[1]</sup> คือการแปลงการเคลื่อนที่ของวัตถุในระหว่างสองรูปภาพซึ่งอาจเกิดจากการเคลื่อนที่ของวัตถุหรือตัวกล้องอุปกรณ์ในรูปแบบของเวกเตอร์ 2 มิติ โดยที่เวกเตอร์แต่ละตัวจะแสดงถึงทิศทางการเคลื่อนที่ของวัตถุหรือบุคคลระหว่างภาพตั้งรูปด้านล่าง



รูปที่ 2.7: ตัวอย่างการเคลื่อนที่ของลูกบอล

จากรูปที่ 2.7 แสดงให้เห็นถึงการเคลื่อนที่ของลูกบอลในภาพที่ต่อเนื่องกัน ภาพโดยที่ลูกศรแสดงถึงทิศทางการเคลื่อนที่ของเวกเตอร์

การทำงานของ optical flow อุปบนสมมติฐาน 2 ประการได้แก่

1. ความเข้มพิกเซลของวัตถุจะไม่เปลี่ยนแปลงระหว่างภาพที่ต่อเนื่องกัน
2. พิกเซลที่อยู่ใกล้กันจะมีลักษณะการเคลื่อนไหวที่คล้ายกัน

เมื่อพิจารณาพิกเซล  $I(x, y, t)$  จากภาพแรกจะเคลื่อนไหวเป็นระยะทาง  $(dx, dy)$  ไปยังภาพต่อไปหลังจากเวลาผ่านไปแล้ว  $dt$  ดังนั้นเนื่องจากพิกเซลเหล่านี้เหมือนกัน และความเข้มไม่มีการเปลี่ยนแปลง จึงทำให้พูดได้ว่า

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (2.2)$$

โดยที่

$I$  = พิกเซลจากภายในภาพ

$x$  = ตำแหน่งของพิกเซลในแกน  $x$

$dx$  = ระยะทางที่เคลื่อนที่ในแกน  $x$

$y$  = ตำแหน่งของพิกเซลในแกน  $y$

$dy$  = ระยะทางที่เคลื่อนที่ในแกน  $y$

$t$  = เวลา

$dt$  = ระยะเวลาที่เปลี่ยนไประหว่างภาพ

จากนั้นใช้การประมาณค่าของ taylor series ทางฝั่งขวามือ และลบค่า common term แล้วหารด้วย  $dt$  เพื่อให้ได้สมการดังต่อไปนี้

$$f_x u + f_y v + f_t \quad (2.3)$$

$$f_x = \frac{\delta f}{\delta x}; f_y = \frac{\delta f}{\delta y} \quad (2.4)$$

$$u = \frac{\delta x}{\delta t}; v = \frac{\delta y}{\delta t} \quad (2.5)$$

โดยที่

$f_x$  = เกรเดียน (gradient) ในแกน x

$f_y$  = เกรเดียนในแกน y

$f_t$  = เกรเดียนของเวลา

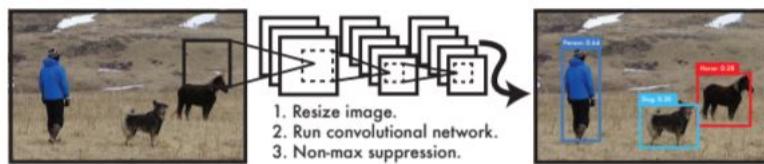
$u$  = เวกเตอร์การเคลื่อนที่ของแกน x

$v$  = เวกเตอร์การเคลื่อนที่ของแกน y

สมการข้างบนเรียกว่าสมการ optical flow จากสมการทำให้สามารถหา  $f_x$  และ  $f_y$  โดยเป็นเกรเดียนของภาพในแกน x และแกน y ตามลำดับ และ  $f_t$  เป็นเกรเดียนของเวลา แต่ n กับ v เป็นตัวแปรที่ไม่ทราบ ทำให้สมการนี้ไม่สามารถแก้ไขโดยมีตัวแปรที่ไม่ทราบถึง 2 ตัว จึงมีการนำวิธีการต่างๆเข้ามาใช้ในการแก้ปัญหานี้ โดยวิธีการที่นำเข้ามาใช้ในการแก้ปัญหาคือ dense optical flow ซึ่งใช้อัลกอริทึมของ Gunnar Farneback<sup>[7]</sup> ซึ่งจะใช้วิธีการขยายพหุนาม (polynomial expansion) เป็นวิธีการที่ทางคณะผู้วิจัยนำมาใช้งานในส่วนของการทำโมเดลปัญญาประดิษฐ์ Inflated 3D convolutional network

### 2.1.3 โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุที่เกี่ยวข้อง

#### 2.1.3.1 YOLO



รูปที่ 2.8: กระบวนการทำงานของโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO

โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เป็นโครงสร้างที่มีความเร็วในการประมวลผลถึง 45 เฟรมต่อวินาที ทำให้สามารถประมวลผลแบบเรียลไทม์ได้ นอกจากนี้ยังมีความแม่นยำ mAP มากกว่าโมเดลสำหรับตรวจจับวัตถุอื่นๆถึง 2 เท่า ซึ่งเหตุผลที่โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO เร็วกว่าโมเดลปัญญาประดิษฐ์ตัวอื่นๆ เนื่องจาก สำหรับการตรวจจับวัตถุในวิธีการก่อนหน้าจะใช้วิธีท่านายกรอบสี่เหลี่ยมก่อน แล้วจึงค่อยนำกรอบสี่เหลี่ยมไปท่านายว่าเป็นหมวดหมู่อะไร ซึ่ง YOLO มีวิธีการที่ต่างออกไป คือ ท่านายตำแหน่งของกรอบสี่เหลี่ยม และท่านายว่าเป็นหมวดหมู่อะไรพร้อมกัน โดยใช้โครงข่ายประสาทแบบคอนโวลูชัน ด้วยแนวคิดนี้จึงเป็นที่มาของชื่อ YOLO หรือ you only look once

#### โครงสร้างของโมเดลปัญญาประดิษฐ์ของ YOLO

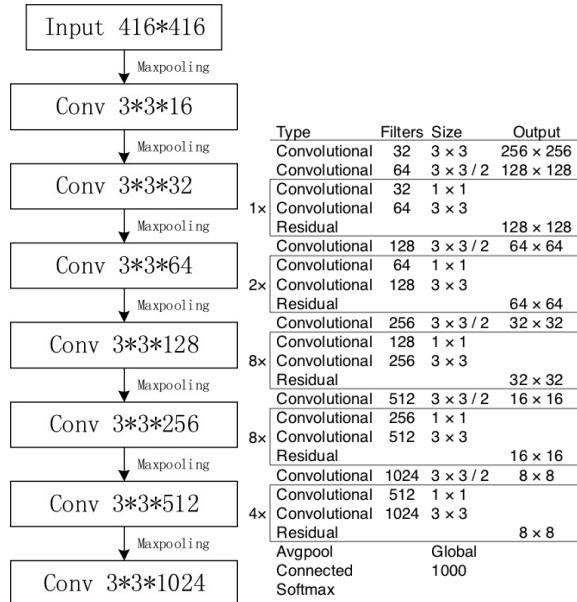


รูปที่ 2.9: โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ YOLO

จากรูปภาพที่ 2.9 จะเห็นได้ว่า YOLO ใช้โครงข่ายประสาทเทียมเพียงตัวเดียวซึ่งภายในโครงข่ายจะมีกระบวนการหลักๆอยู่ 3 อย่าง กระบวนการแรคคือการสกัดคุณลักษณะกระบวนการนี้จะมีจำนวนขั้นของเลยอร์ที่แตกต่างกันไปตามความลึกของการสกัดแล้วแต่โมเดล ซึ่งจะมีตัวอย่างอยู่ในบทความด้านล่าง และ ขั้นตอนถัดมา คือ การท่านายผล หลังจากที่ได้คุณลักษณะมาแล้วจะนำไปท่านายผลผ่าน Fully connected ซึ่งจะได้ผลลัพธ์ออกเป็นหมวดหมู่และตำแหน่งของกรอบสี่เหลี่ยม ขั้นตอนสุดท้ายคือ การทำ NMS เพื่อให้ได้ผลลัพธ์ที่ดีที่สุดออกมานะ

ซึ่งโครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO ที่ถูกใช้ในงานวิจัยนี้ประกอบไปด้วย 1) YOLO-v3 tiny 2) YOLO-v3 3) YOLO-v3 spp ซึ่งทั้ง 3 โครงสร้างจะมีความแตกต่างของโครงสร้างดังนี้

1. YOLO-v3 tiny ใช้ Max-Pooling layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
2. YOLO-v3 ใช้ Convolutional layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง
3. YOLO-v3 spp ใช้ Convolutional layers+คุณลักษณะที่ดีที่สุดของ Max-Pooling layers ในขั้นตอนของการลดจำนวนข้อมูลตัวอย่าง



(ก) โครงสร้างโมเดล

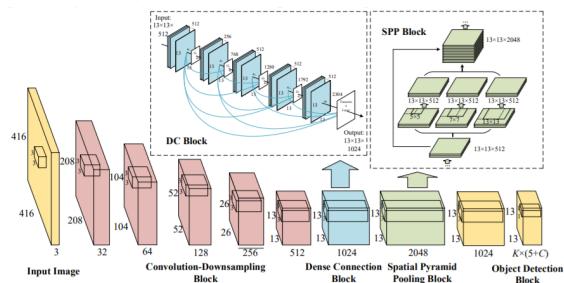
ปัญญาประดิษฐ์ของ

YOLO-v3 tiny

Table 1. Darknet-53.

(ข) โครงสร้างโมเดลปัญญาประดิษฐ์

ของ YOLO-v3

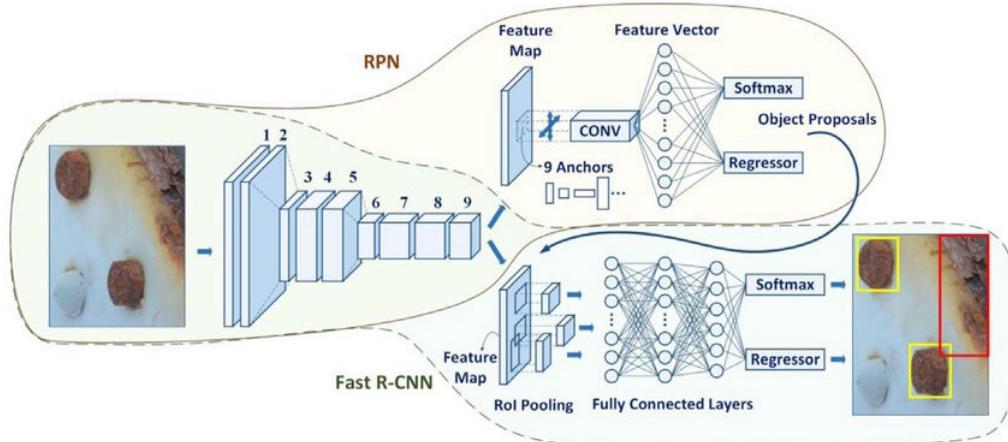


11

(ค) โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO-v3 spp

รูปที่ 2.10: โครงสร้างโมเดลปัญญาประดิษฐ์ของ YOLO

## Faster-RCNN



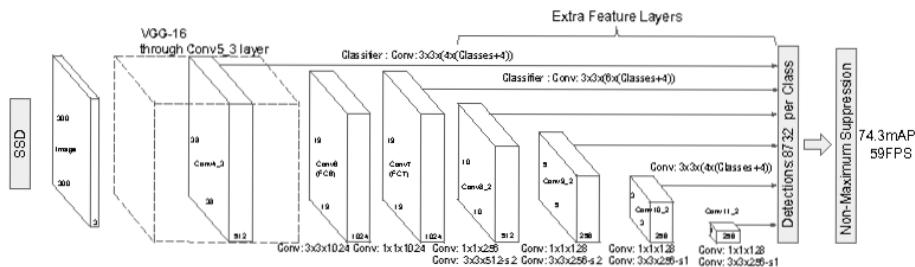
รูปที่ 2.11: โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ของ Faster RCNN

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) สุดท้ายจะได้ผลลัพธ์เป็นหมวดหมู่ของกรอบสี่เหลี่ยม และ ตำแหน่งของกรอบสี่เหลี่ยม

region proposal network (RPN) คือ โครงข่ายที่เสนอพื้นที่ที่คาดว่าจะมีวัตถุอยู่ จะถูกใช้หลังรูปภาพผ่านการสกัดคุณลักษณะ ซึ่ง RPN มีโครงสร้างที่มีองค์ประกอบ 2 อย่าง คือมีการระบุว่าบริเวณนั้นมีวัตถุอยู่หรือไม่ (classification layer) และ สำหรับการระบุพิกัดของกรอบสี่เหลี่ยมที่คาดว่าจะมีวัตถุอยู่ (regression layer) ซึ่งผลลัพธ์จะได้ ROI (พื้นที่บริเวณที่เราสนใจ)

## SSD



รูปที่ 2.12: โครงสร้างทั่วไปของโมเดลปัญญาประดิษฐ์ SSD

SSD เป็นโมเดลปัญญาประดิษฐ์ที่ใช้โครงข่ายประสาทเทียมตัวเดียวสำหรับการตรวจจับวัตถุ ซึ่งภายในโครงข่ายจะประกอบไปด้วยการทำงานหลัก 3 อย่าง คือ

1. การสกัดคุณลักษณะ  
นำรูปภาพทั่วไปมาผ่าน VGG-16 (เป็นโมเดล CNN ชนิดหนึ่ง) เพื่อการสกัดคุณลักษณะของรูปภาพ
2. การทำนายผล  
หลังจากที่ได้คุณลักษณะมาแล้วจะนำไปทำนายผลผ่าน Fully connected
3. การเลือกคัดกรองผลลัพธ์  
หลังจากได้ผลลัพธ์เป็นหมวดหมู่ของกรอบสี่เหลี่ยม และ ตำแหน่งของกรอบสี่เหลี่ยมจะนำไปผ่านกระบวนการ NMS เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด

## 2.2 การประมวลผลวิดีโอ

ในส่วนของงานวิจัยสิ่งที่สนใจ คือ ข้อมูลการกระทำของมนุษย์แต่ละคนภายในวิดีโอ เพื่อที่จะได้ผลลัพธ์ที่มีประสิทธิภาพอุปกรณ์เป็นข้อมูลของสิ่งที่สนใจ เช่น จำนวนคนที่เดินผ่านกัน หรือทิศทางการเดินของคนในวิดีโอ จึงจำเป็นต้องใช้การประมวลผลวิดีโอเพื่อที่จะสกัดสิ่งที่สนใจออกมาจากวิดีโอ ซึ่งการประมวลผลวิดีโอมีหลากหลายกระบวนการ โดยในแต่ละกระบวนการจะมีจุดประสงค์ของการทำและผลลัพธ์หลังการประมวลผลที่แตกต่างกัน ในหัวข้อนี้จะมาอธิบายถึงกระบวนการในการประมวลผลของวิดีโอและผลลัพธ์ของกระบวนการนั้น

### 2.2.1 การตรวจจับวัตถุ

การตรวจจับวัตถุนั้นเป็นหนึ่งในกระบวนการประมวลผลวิดีโอ กล่าวคือกระบวนการที่ผู้วิจัยจะต้องทำการระบุสิ่งที่สนใจว่า คืออะไร อยู่ที่ตำแหน่งใด การตรวจจับวัตถุก็ค้นพบเมื่อเวลาแล้ว และในปัจจุบันนี้สามารถทำได้หลากหลายวิธี โดยภายในบทความนี้จะสรุปให้ความสำคัญของวิธีการต่างในการตรวจจับวัตถุ เช่น Sliding Window, Brute Force Search, RCNN, Fast-RCNN, Faster-RCNN, YOLO, SSD

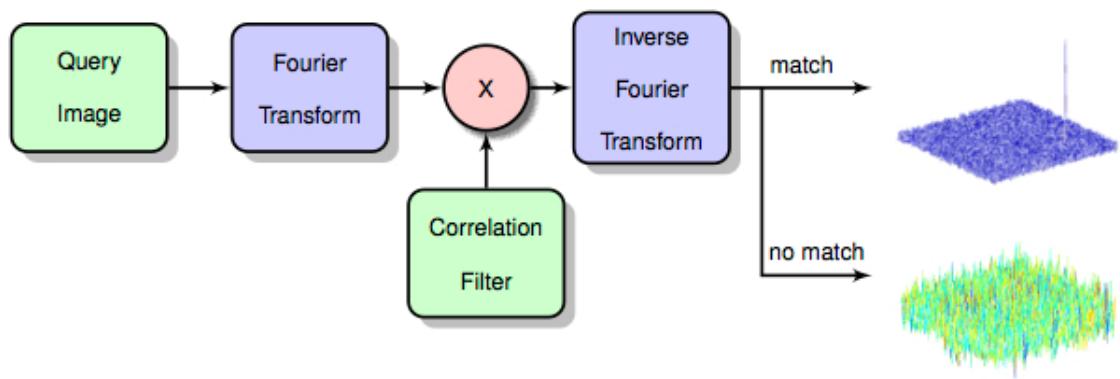
1. Sliding window วิธีการที่เปรียบเสมือนมีเครื่องนอล (kernel) ค่อยๆ เลื่อนไปยังแต่ละพิกเซลบนรูป ซึ่งก่อนการเลื่อนของเครื่องนอลแต่ละครั้งจะนำส่วนของรูปภาพที่ถูกเครื่องนอลทับอยู่ไปทำนายว่าใช้วัตถุที่เราต้องการหรือไม่ จากนั้นจึงค่อยเลื่อนต่อไปจนครบทั้งรูปภาพ
2. Brute force search คุกสร้างขึ้นมาเพื่อแก้ปัญหาขนาดของเครื่องนอลไม่ตรงกับขนาดของวัตถุที่อยู่ในภาพ ทำให้มีโอกาสที่จะไม่พบวัตถุ โดยหลักการของวิธีการนี้ คือ การย่อ-ขยาย รูปภาพและนำเข้าในหลาย ๆ อัตราส่วน ตั้งแต่ 0.1 เท่า จนถึง 2 เท่า แต่ข้อเสียของวิธีการนี้คือ มีการคำนวนพื้นที่ช้าๆ และใช้เวลานาน

3. RCNN ใช้อัลกอริทึม selective search เข้ามาช่วยในการเสนอพื้นที่ที่น่าจะมีวัตถุอยู่ท่ามกลางการค้นหา ทุกๆ ตำแหน่ง จำนวนนักนำรูปภาพในส่วนพื้นที่นั้นไปทำนายว่าวัตถุนั้นคืออะไร กรณีที่มีพื้นที่ที่อยู่ใกล้ๆ วัตถุก็เสนอเข้ามาเป็นจำนวนมากด้วย เราจะใช้ Non-Maximum Suppression (NMS) หรือการเลือก พื้นที่ที่ถูกหับซ้อนมากที่สุดในบริเวณนั้น
4. Fast-RCNN จากวิธีการ RCNN แต่ละพื้นที่จะถูกนำไปสกัดคุณลักษณะ และ ทำนายผลทีละพื้นที่ ทำให้ เสียเวลา โดย Faster-RCNN จะมีส่วนที่คล้ายกับ RCNN ในส่วนการทำ selective search หากพื้นที่ที่น่าจะ มีวัตถุเหมือนเดิม แต่ Faster-RCNN จะนำรูปภาพทั้งรูปภาพไปสกัดคุณลักษณะ หลังจากที่ได้คุณลักษณะ แล้ว นำพิกัดของพื้นที่ที่น่าจะมีวัตถุ บนรูปภาพที่ถูกสกัดคุณลักษณะแล้วของ ไปผ่าน ROI Pooling (การ ลดขนาดข้อมูลให้มีขนาดคงที่เพื่อเป็นอินพุตให้กับโมเดลในการทำนายผล)
5. Faster-RCNN พัฒนาจาก Fast-RCNN โดยวิธีของ Faster-RCNN จะรวมในส่วนของ selective search และ การทำงานอื่นๆ ให้อยู่ในโครงข่ายเดียวกัน สรุปคือการทำงานของโครงข่ายของ Faster-RCNN จะมี การทำงาน 3 อย่างหลักคือ 1) สกัดคุณลักษณะ 2) การเสนอส่วนที่น่าจะมีวัตถุอยู่ในรูปภาพ 3) หลังจาก ได้รูปภาพจากการสกัดคุณลักษณะ นำพิกัดของพื้นที่ที่น่าจะมีวัตถุ บนรูปภาพที่ถูกสกัดคุณลักษณะแล้ว ของ ไปผ่าน ROI Pooling
6. YOLO เป็นวิธีการที่ใช้โครงข่ายประสาทแบบconvโอลูชันเพียงตัวเดียวทำนายรูปภาพทั้งรูป โดยโครงข่าย จะแบ่งรูปภาพออกเป็นพื้นที่ และ ใช้ fully-connected (เป็นโครงข่ายประสาทเทียมที่นำเอาคุณลักษณะ มาทำนายผล) ทำนายตำแหน่งของกรอบสี่เหลี่ยมและหมวดหมู่ของกรอบสี่เหลี่ยมในแต่ละพื้นที่ไปพร้อม กัน
7. SSD ใช้โครงข่ายประสาทเทียมตัวเดียวเหมือนกับ YOLO แต่การออกแบบโครงสร้างแตกต่างกัน SSD จะ ใช้ VGG-16 (เป็นโมเดล CNN ชนิดหนึ่ง) ในการสกัดคุณลักษณะ และ ใช้ Convolution layer ต่อ กัน หลายชั้นเพื่อลดมิติและความละเอียดทำให้ตรวจจับวัตถุในหลายขนาด ซึ่งในแต่ละชั้นจะได้ผลลัพธ์ เป็น Convolution filter จำนวนจำนวน Convolution filter ไปทำนายผลต่อ

## 2.2.2 ระบบติดตามการเคลื่อนไหวของวัตถุ

การติดตามการเคลื่อนไหวของวัตถุ<sup>[1]</sup> คือระบบที่ใช้สำหรับการติดตามการเคลื่อนไหวของวัตถุที่สนใจ ที่อยู่ ในรูปภาพ โดยใช้การคำนวณทางคณิตศาสตร์ และการประมวลผลภาพ (image processing) ทำให้การประมวล ผลนั้นเร็กว่าการใช้โมเดลปัญญาประดิษฐ์ ซึ่งอัลกอริทึมติดตามการเคลื่อนไหวที่นิยมใช้มีสองอัลกอริทึม คือ correlation filter และ kalman filter ซึ่งหลักการของทั้งสองอัลกอริทึมนั้นจะแตกต่างกันโดยที่ correlation filter นั้นจะใช้พิเศษของวัตถุในการคำนวณตำแหน่งถัดไปของวัตถุ และ kalman filter จะใช้ข้อมูลการ เคลื่อนไหวในการคำนวณตำแหน่งถัดไปของวัตถุ ซึ่งจากการศึกษาในบทความ "Object Tracking using Correlation, Kalman Filter and Fast Means Shift Algorithms"<sup>[2]</sup> kalman filter มีประสิทธิภาพที่สูงนั้นจะขึ้นอยู่ กับข้อมูลที่ได้จากการรับ (measurement) และความซับซ้อนในการเคลื่อนไหวของวัตถุ ในขณะที่ correlation นั้นมีประสิทธิภาพที่ด้อยกว่า kalman filter เพียงเล็กน้อยและสามารถติดตามการเคลื่อนไหวที่ซับซ้อนของวัตถุ ได้ดีกว่า (การเคลื่อนไหวที่ซับซ้อนหมายถึง การเคลื่อนไหวที่เกิดการเปลี่ยนทิศทางฉับพลันบ่อย) ผู้วิจัยจึงตัดสินใจ เลือกใช้ correlation filter ในงานครั้งนี้

จากรูปที่ 2.13 เป็นหลักการในการติดตามการเคลื่อนไหวของวัตถุแบบ correlation filter โดยการนำรูป มาผ่านกระบวนการแปลงฟูรีเยร์ (fourier transform) และนำมาคูณกับ correlation filter ซึ่งเป็นตัวกรองที่ใช้ สำหรับการหาความสัมพันธ์กับวัตถุในภาพ จำนวนทำการแปลงฟูรีเยร์ผกผัน (inverse fourier transform) เพื่อ



รูปที่ 2.13: แนวคิดของระบบติดตามการเคลื่อนไหวของวัตถุ

ตรวจสอบว่าวัตถุในภาพนั้นอยู่ที่ตำแหน่งใด โดยมีการคำนวนเริ่มจากการหา correlation filter ที่ดีที่สุดโดยใช้ วิธีลดผลรวมของข้อผิดพลาดกำลังสองให้น้อยที่สุดดังนี้

$$\epsilon = \left\| \sum_{l=1}^d h^l * f^l - g \right\|^2 + \lambda \sum_{l=1}^d \|h^l\|^2 \quad (2.6)$$

โดยที่

$\epsilon$  = ค่าความคลาดเคลื่อน

$d$  = จำนวนมิติของผังคุณลักษณะ (feature map) ของภาพ

$h$  = correlation filter

$*$  = circular correlation

$f$  = พื้นที่สีเหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

$g$  = ผลลัพธ์ correlation ที่ต้องการของ  $f$

$\lambda$  = regularization term

เมื่อพิจารณาจากรูปภาพเดียวในกรณีที่เวลา ( $t$ ) เท่ากับ 1 จะสามารถจัดรูปสมการด้านบนได้ดังนี้

$$H^l = \frac{\bar{G}F^l}{\sum_{k=1}^d \bar{F}^k F^k + \lambda} \quad (2.7)$$

$$H_t^l = \frac{A_t^l}{B_t} \quad (2.8)$$

$$A_t^l = (1 - \eta)A_{t-1}^l + \eta \bar{G}_t F_t^l \quad (2.9)$$

$$B_t = (1 - \eta)B_{t-1} + \eta \sum_{k=1}^d \bar{F}_t^k F_t^k \quad (2.10)$$

โดยที่

$H$  = correlation filter

$\eta$  = อัตราการเรียนรู้

$\bar{G} = g$  ที่ผ่านการทำ complex conjugation

$F$  = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

$\bar{F} = f$  ที่ผ่านการทำ complex conjugation

$t$  = เวลา

จากสมการที่ได้มาจะสามารถทำให้หาตำแหน่งต่อไปของวัตถุที่สนใจได้ด้วยสมการต่อไปนี้

$$y = F^{-1} \left\{ \frac{\sum_{l=1}^d \bar{A}^l Z^l}{B + \lambda} \right\} \quad (2.11)$$

โดยที่

$y$  = correlation score

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

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

$Z$  = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการหาผังคุณลักษณะของภาพใหม่

โดยค่าของ  $y$  ที่ได้ออกมาจะทำให้รู้ถึงตำแหน่งของวัตถุที่สนใจได้ ณ ตำแหน่งที่  $y$  มีค่าสูงสุด

### 2.2.3 ระบบระบุตัวตนของบุคคล

ระบบระบุตัวตนของบุคคล<sup>[14]</sup> คือการระบุตัวตนของบุคคลภายในวิดีโอหรือระหว่างรูป 2 รูป สามารถนำมาประยุกต์ใช้ในด้านของการรักษาความปลอดภัย การตามหาบุคคล หรือการตรวจสอบการกระทำการของบุคคลนั้นในวิดีโอด้วยการระบุตัวตนของบุคคลนั้นเป็นปัญหาที่ท้าทาย เนื่องจากคุณลักษณะทั่วไปของบุคคลในรูปไม่เพียงพอต่อการระบุบุคคลภายในรูปว่าเป็นบุคคลคนเดียวกันได้ ซึ่งวิธีการที่ใช้ในการระบุตัวตนของบุคคลเรียกว่า Dynamically Matching Local Information (DMLI) ที่สามารถจัดแแนวรายละเอียดข้อมูลของรูป และเพิ่มประสิทธิภาพให้สูงขึ้น ถึงแม้ว่าวิธีการทำ Dynamically Matching Local Information นั้นจะไม่ใช้วิธีการที่มีประสิทธิภาพสูงสุดแต่มีประสิทธิภาพใกล้เคียงกัน และด้วยทางที่คณผู้วิจัยสามารถนำวิธีนี้มาประยุกต์เข้ากับงานวิจัยครั้งนี้ได้สะดวกที่สุด จึงนำวิธี Dynamically Matching Local Information มาใช้สำหรับงานวิจัยครั้งนี้



รูปที่ 2.14: การแบ่งรูปออกเป็น 8 ส่วนของระบบระบุตัวตนของบุคคล

การทำงานของระบบระบุตัวตนของบุคคลจะเริ่มจากการแบ่งรูปออกเป็น 8 ส่วนและนำคุณลักษณะของรูปมาผ่านกระบวนการ normalization เพื่อลดความช้ำช้อนของข้อมูล แล้วนำมาเปรียบเทียบความแตกต่าง

ของคุณลักษณะของรูป หลังจากนั้นาค่าเฉลี่ยของความแตกต่างของมา โดยค่าที่ได้ออกมาจะเรียกว่า original distance ถ้าค่าที่ออกมากใกล้เคียงกับ 0 จะหมายถึงบุคคลในรูปทั้งสองเป็นบุคคลเดียวกัน และมีการตั้งค่าเกณฑ์สำหรับ original distance เพื่อใช้สำหรับในการระบุบุคคลในรูปเป็นบุคคลเดียวกัน

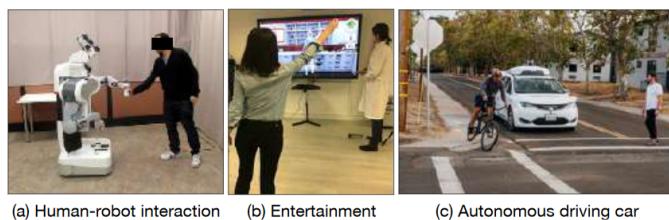
โดยชุดข้อมูลที่นำมาใช้สำหรับการทำโมเดลปัญญาประดิษฐ์ได้แก่

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

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

#### 2.2.4 ระบบจำแนกการกระทำการของมนุษย์

ระบบจำแนกการกระทำการที่เป็นหัวข้อที่มีการให้ความสนใจอย่างมากสำหรับการทำระบบประมวลผลวิดีโอในปัจจุบัน เนื่องจากระบบประมวลผลวิดีโอนั้นสามารถใช้งานได้หลากหลายสถานการณ์ เช่น ใช้สอดส่องการจราจรบนท้องถนน วิเคราะห์พฤติกรรมการเคลื่อนที่ของลูกค้าภายในห้าง การตามหาคนหายหรือพัດหลงภายในอาคาร เป็นต้น ซึ่งการที่ระบบสามารถจำแนกการกระทำการทั่งของมนุษย์ภายในวิดีโอด้วยความสามารถเพิ่มความสามารถของระบบประมวลผลวิดีโอด้วย เช่น สามารถแจ้งเตือนเมื่อพบบุคคลที่มีพฤติกรรมน่าสงสัยในวิดีโอด้วย หรือไม่ให้เกิดเหตุการณ์อันตรายขึ้น เป็นต้น



รูปที่ 2.15: ตัวอย่างการประยุกต์ใช้งานระบบจำแนกการกระทำการทำมนุษย์<sup>[12]</sup>

การสร้างโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำการทำมนุษย์ จำเป็นต้องมีชุดข้อมูลที่เหมาะสมกับการกระทำที่สนใจ และโครงสร้างโมเดลปัญญาประดิษฐ์ที่เหมาะสม (model architecture) ซึ่งในปัจจุบันนั้นมีชุดข้อมูลสาธารณะที่สามารถนำมาใช้งานได้หลากหลายชุดข้อมูล เช่น YouTube-8M ชุดข้อมูลสำหรับการประมวล

ผลวิดีโอที่มีขนาดใหญ่ที่สุดและมีจำนวนคำกำกับมากที่สุด, AVA ชุดข้อมูลของการกระทำที่มาการขยายเพียงเล็กน้อย, Sports-1M ชุดข้อมูลที่เกี่ยวกับกิจกรรมกีฬาต่างๆของมนุษย์ เป็นต้น ซึ่งชุดข้อมูลที่ผู้วิจัยเลือกนำมาศึกษาได้แก่ YouTube-8M, AVA และ Moment in Time โดยแต่ละชุดข้อมูลจะมีความแตกต่างกันในหลายๆด้านแต่จะมีสิ่งที่เหมือนกัน คือ เป็นชุดข้อมูลสำหรับการวิเคราะห์วิดีโอที่สนใจการกระทำและกิจกรรมของมนุษย์ โดยจะกล่าวถึงความแตกต่างในด้านต่างๆ เช่น เป้าหมายของแต่ละชุดข้อมูล วิธีการเก็บข้อมูลสำหรับชุดข้อมูล วิธีการสร้างคำกำกับคุณลักษณะ และรายละเอียดของชุดข้อมูล จากนั้นจะสรุปข้อมูลของแต่ละชุดข้อมูล

## 2.3 ชุดข้อมูลสำหรับการประมวลผลวิดีโอ

ในปัจจุบันมีชุดข้อมูลมากมายถูกสร้างขึ้นมาสำหรับใช้สร้างโมเดลสำหรับแก้ปัญหาในด้านต่างๆ เช่น การตรวจจับวัตถุภายในรูปภาพ การจดจำใบหน้าบุคคลภายในรูปภาพ การจำแนกการกระทำของมนุษย์ เป็นต้น ซึ่ง สิ่งที่ทำให้โมเดลปัญญาประดิษฐ์นั้นมีประสิทธิภาพสูงคือ จำนวนของข้อมูล โดยในปัจจุบันปัญหาด้านการทำความเข้าใจรูปด้วยปัญญาประดิษฐ์ (image understanding) สามารถพัฒนาให้มีประสิทธิภาพสูงนั้นเนื่องจากมีจำนวนข้อมูลที่มากมาย ในขณะที่ปัญหาด้านการทำความเข้าใจวิดีโอด้วยปัญญาประดิษฐ์ (video understanding) นั้น กำลังมีการให้ความสนใจเพิ่มขึ้นเรื่อยๆ ในช่วงระยะเวลาหลายปีที่ผ่านมาในหัวข้อนี้จะพูดถึงการศึกษาชุดข้อมูลที่ใช้สำหรับการทำความเข้าใจวิดีโอ โดยจะมุ่งเน้นไปที่การจำแนกการกระทำการของมนุษย์เป็นหลัก

### ชุดข้อมูล YouTube-8M

YouTube-8M คือชุดข้อมูลวิดีโอที่มีจำนวนวิดีโอยอดเยี่ยมที่สุดถึง 8 ล้านวิดีโอ (พ.ศ. 2559) โดยมีจุดมุ่งหมายหลักในการจำแนกสาระสำคัญของวิดีโอ (video theme) ด้วยคำสั้นๆ เช่น ถ้าวิดีโอนั้นมีมนุษย์กำลังปั่นจักรยานบนถนนดินริมหน้าผาชุดข้อมูลนี้จะกำกับวิดีโอนี้ว่า mountain biking ซึ่งทำให้ YouTube-8M แตกต่างจากชุดข้อมูลวิดีโອื่นๆ ส่วนใหญ่ที่จะเน้นการกระทำ หรือกิจกรรมของมนุษย์ ซึ่งข้อมูลโดยสรุปของชุดข้อมูลมีดังนี้

#### 1. รายละเอียดของชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : เพื่อจำแนกสาระสำคัญของวิดีโอ
- (b) จำนวนของวิดีโอ : 8,264,650 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 229.6 วินาที
- (d) จำนวนของหมวดหมู่ของคำกำกับคุณลักษณะ : 4,800 หมวดหมู่
- (e) กฎในการรับรวมวิดีโอดังนี้
  - i. ทุกคำกำกับต้องเป็นรูปธรรม
  - ii. ในแต่ละคำกำกับต้องมีจำนวนวิดีโอมากกว่า 200 วิดีโอ
  - iii. ความยาวของวิดีโอดังกล่าวอยู่ระหว่าง 120 - 500 วินาที

หลังจากได้กฎในการรับรวมวิดีโอล้วน ขั้นตอนต่อไปคือการสร้างคำศัพท์ที่ใช้ในการค้นหาข้อมูลวิดีโอดังใน YouTube

#### (f) ขั้นตอนในการสร้างคำศัพท์มีดังนี้

- i. กำหนดบัญชีขาว (whitelist) ของคำกำกับที่เป็นรูปธรรมมา 25 ชนิด เช่น กีฬา เป็นต้น
- ii. กำหนดบัญชีดำ (blacklist) ของคำกำกับที่คิดว่าไม่เป็นรูปธรรมไว้ เช่น software เป็นต้น
- iii. รวบรวมคำกำกับที่มีอยู่ในรายการที่อนุญาตอย่างน้อย 1 คำ และต้องไม่มีอยู่ในบัญชีดำ ซึ่งจะทำให้ได้คำกำกับที่ต้องการมาประมาณ 50,000 คำ
- iv. จากนั้นใช้ผู้ประเมินจำนวน 3 คน ในการคัดคำกำกับที่คิดว่าเป็นรูปธรรม และสามารถจำจำหรือเข้าใจได้ยากโดยไม่ต้องเชี่ยวชาญในด้านนั้น ซึ่งผู้ประเมิน ก็จะมีคำถามว่า “มันยากขนาดไหนถึงจะระบุได้ว่ามีคำกำกับดังกล่าวอยู่ในรูปหรือวิดีโอ โดยใช้เพียงแค่การมองเท่านั้น?” โดยแบ่งเป็นระดับดังนี้

A. บุคคลที่ไม่สามารถเข้าใจได้

B. บุคคลที่ไม่ได้ผ่านการอ่านบทความที่เกี่ยวข้องมาแล้วสามารถเข้าใจได้

C. ต้องเชี่ยวชาญในด้านใดซักด้านจึงจะเข้าใจได้

D. เป็นไปไม่ได้ ถ้าไม่มีความรู้ที่ไม่ได้เป็นรูปธรรม

E. ไม่เป็นรูปธรรม

- v. หลังจากคำนึงข้างบนและการให้คะแนน จะทำการเก็บไว้เฉพาะคำกำกับที่มีคะแนนเฉลี่ยมากที่สุดอยู่ที่ประมาณ 2.5 คะแนนหรือต่ำกว่าเท่านั้น

- vi. ทำให้สุดท้ายเหลือเพียงประมาณ 10,000 คำที่สามารถใช้ได้
- vii. หลังจากได้คำกำกับที่คิดว่าเป็นรูปธรรมแล้วก็นำไปค้นหาและรวบรวมด้วย YouTube annotation system โดยมีขั้นตอนดังนี้
  - A. สุ่มเลือกวิดีโอมาก 10 ล้านวิดีโอพร้อมกับคำกำกับของวิดีโอด้วยใช้กฎที่กำหนดไว้
  - 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 เนื่องจากมีข้อมูลที่มีขนาดใหญ่ทำให้ใช้เวลาในการประมวลผลนาน ซึ่งกระบวนการนี้จะมีการลดความเร็วเฟรมต่อวินาที เวกเตอร์ของคุณลักษณะ (feature vector) และแปลงข้อมูลจาก 32 บิต ให้เป็น 8 บิต
- ii. คุณลักษณะระดับวิดีโอ : การแยกเวกเตอร์คุณลักษณะระดับวิดีโອอกจากคุณลักษณะระดับเฟรมซึ่งการทำแบบนี้ทำให้ได้ประโยชน์ 3 ข้อ คือโมเดลทั่วไปที่ไม่ใช้โครงข่ายประสาทเทียมสามารถนำไปใช้งานได้ ขนาดข้อมูลเล็กลง และเหมาะสมกับการนำไปสร้างโมเดลในขอบเขตอื่นมากขึ้น

### (b) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับเฟรม

#### i. one vs all logistic regression classifier + average pooling

สร้างโมเดลปัญญาประดิษฐ์ของทุกคำกำกับแยกกัน จะได้โมเดลปัญญาประดิษฐ์ 4800 โมเดล ซึ่งในการทำนายผลจะใช้การเฉลี่ยความน่าจะเป็น (probability) ของแต่ละคำกำกับจากทุกๆเฟรมในวิดีโอ โดยคำกำกับที่มีความน่าจะเป็นมากที่สุดจะเป็นคำตอบของการทำนายโดยมีสมการคำนวนความน่าจะเป็นเฉลี่ยของแต่ละคำกำกับดังนี้

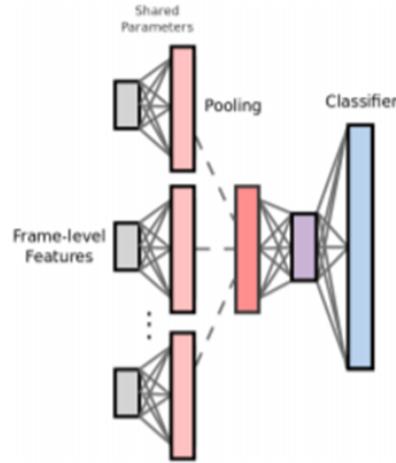
$$p_v(e|X_{1:F_v}^v) = \frac{1}{F_v} \sum_{j=1}^{F_v} p(e|X_j^v) \quad (2.12)$$

โดยที่

$v$	= วิดีโอที่ใช้ในการทำนายผล
$e$	= คำกำกับ
$F_v$	= จำนวนเฟรมสูงสุดของวิดีโอ $v$
$p_v(e X_{1:F_v}^v)$	= ความน่าจะเป็นของคำกำกับ $e$ บนวิดีโอ $v$

#### ii. Deep bag of frames

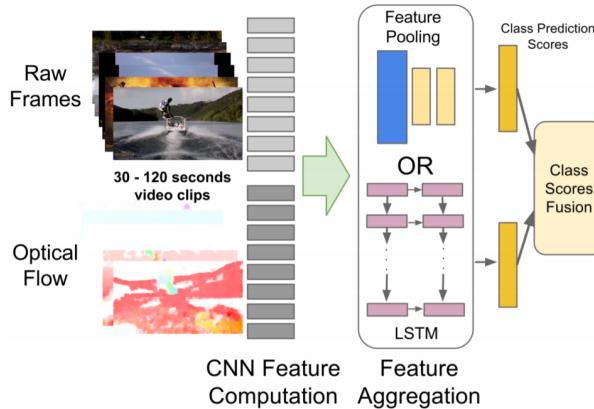
มีหลักการเหมือนกับ deep bag of words<sup>[13]</sup> คือการแยกคุณลักษณะของเฟรม (extract frame feature) ที่โมเดลคิดว่าสำคัญอ่อนมาทำนายผล ซึ่งโครงสร้างโมเดลปัญญาประดิษฐ์ เป็นดังนี้ดังรูปที่ 2.16 โดยที่จะสุมหิบ 20 เฟรมของวิดีโอมาก่อน RELU activation function จากนั้นทำการ batch normalization ก่อนจะใช้ max pooling ในการรวมคุณลักษณะที่ได้ให้เป็นคุณลักษณะระดับวิดีโอ สุดท้ายใช้ softmax layer ในการจำแนกว่าเป็นคำกำกับได้



รูปที่ 2.16: โครงสร้างโมเดลปัญญาประดิษฐ์ของ deep bag of frames

### iii. Long short-term memory (LSTM)

โมเดล LSTM ที่ใช้ในบทความนิ้นมีการอ้างอิงโครงสร้างมาจากบทความ "Beyond Short Snippets: Deep Networks for Video Classification"<sup>[20]</sup> ซึ่งมีโครงสร้างดังรูปที่ 2.17 แต่



รูปที่ 2.17: โครงสร้าง LSTM ที่ใช้การอ้างอิงในบทความนี้

เนื่องจากข้อมูลของ YouTube-8M นั้นไม่สามารถเข้าถึงเฟรมวิดีโอติด (raw video frame) ได้จึงทำให้สามารถใช้ได้เพียงชั้นของ LSTM และ softmax เท่านั้น ซึ่งจากการทดลองพบว่า การใช้ LSTM 2 ชั้นที่มี hidden unit 1024 หน่วย นั้นมีประสิทธิภาพมากที่สุด

(c) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับวิดีโอ

#### i. Logistic regression

สร้างโมเดลปัญญาประดิษฐ์ของทุกคำกำกับแยกกัน จะได้โมเดลปัญญาประดิษฐ์ 4800 โมเดล โดยที่พารามิเตอร์ Θ (parameter) ของแต่ละโมเดลหาได้จาก

$$\sum_{i=1}^N L(y_{i,e}, \sigma(w_e X_i)) \quad (2.13)$$

$$L(y_{i,e}, \sigma(w_e X_i)) = y_{i,e} \log(\sigma(w_e X_i)) + (1 - y_{i,e}) \log(1 - \sigma(w_e X_i)) \quad (2.14)$$

$$p(e|X) = \sigma(w_e^T X_i) \quad (2.15)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (2.16)$$

โดยที่

$N$  = จำนวนวิดีโอทั้งหมด

$X$  = คุณลักษณะระดับวิดีโอ

$y_{i,e}$  = คำตอบของคำกำกับ  $e$  ในวิดีโอที่  $i$

$w_e$  = weight ของคำกำกับ  $e$

$p(e|X)$  = ความน่าจะเป็นของคำกำกับ  $e$  ของคุณลักษณะระดับวิดีโอ  $X$

- ii. Support vector machine (SVM) สร้างโมเดลปัญญาประดิษฐ์ของทุกคำกำกับแยกกันทำให้จะได้โมเดล SVM 4800 โมเดล โดยได้ใช้ค่า -1 ถึง 1 ในการแสดงถึงคำกำกับด้านลบ (negative label) และคำกำกับค้านบวก (positive label) ตามลำดับ และใช้ hinge loss ในการคำนวณหา loss ( $L$ ) ซึ่งสมการจะเป็นดังนี้

$$L(y, \hat{y}) = \max(0, b - (2y - 1)\hat{y}) \quad (2.17)$$

โดยที่

$y$  = คือคำตอบของการทำนาย โดยสามารถเป็น 0 หรือ 1 เท่านั้น

$\hat{y}$  = คำผลการทำนายโดยจะมีค่าอยู่ในช่วง -1 ถึง 1

$b$  = Hinge-loss พารามิเตอร์

- iii. Mixture of Expert (MoE) Mixture of experts ที่ถูกนำมาใช้ในการอ้างอิงในบทความนิยมจากบทความ "Hierarchical mixtures of experts and the EM algorithm"<sup>[10]</sup> โดยการทำนายความน่าจะเป็นว่าคำกำกับ  $e$  ในวิดีโอ  $X$  ด้วยสมการ

$$p(e|X) = \sum_{h \in H_e} p(h|X) \sigma(u_h X) \quad (2.18)$$

ซึ่ง  $p(h|X)$  สามารถเขียนได้ในรูปสมการดังนี้

$$p(h|X) = \frac{\exp(w_h X)}{1 + \sum_{h' \in H_e} \exp(w'_h X)} \quad (2.19)$$

โดยที่

$H_e$  = hidden state หรือ expert ของโมเดล

$X$  = วิดีโอที่ใช้ในการทำนายผล

$u_h$  = logistic weight ของ expert  $h$

$w_h$  = softmax weight ของ expert  $h$

$p(h|X) = |H_e| + 1$  ที่ผ่าน softmax function

ให้ชุดข้อมูลสำหรับใช้ในการสร้างโมเดล  $(x_i, g_i)_{i=1\dots N}$  โดยที่  $x_i$  คือเวกเตอร์คุณลักษณะ (feature vector)  $g_i$  คือคำตอบของการทำนายซึ่งสามารถเป็นได้เพียง 0 และ 1 เท่านั้น

และ N คือจำนวนวิดีโอสูงสุด ซึ่งสมการ log-loss ระหว่างผลการทำนายกับคำตอบของการทำนายคือ

$$L(p, g) = -g \log p - (1 - g) \log(1 - p) \quad (2.20)$$

ซึ่งสามารถเขียนในรูปอนุพันธ์ของ softmax weight และ logistic weight ได้ดังนี้

$$\frac{\partial L[p_{y|x,g}]}{\partial w_h} = x \frac{p_{h|x}(p_{y|h,x} - p_{y|x})(P_{y|x} - g)}{p_{y|x}(1 - p_{y|x})} \quad (2.21)$$

$$\frac{\partial L[p_{y|x,g}]}{\partial u_h} = x \frac{p_{h|x}p_{y|h,x}(1 - p_{y|x})(P_{y|x} - g)}{p_{y|x}(1 - p_{y|x})} \quad (2.22)$$

(d) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือ

- i. Mean Average Precision (mAP) ในแต่ละคำทำกับได้ทำการปัดเศษแนบทั้ง  $10^{-4}$  แล้วเรียงลำดับคะแนนทั้งหมดที่ไม่ใช่ 0 จากนั้นให้ค่า  $\tau$  เป็นค่าแบ่งเกณฑ์ (threshold) โดยที่มี  $P(\tau)$  คือ precision และ  $R(\tau)$  คือ recall ซึ่งหาได้จาก

$$P(\tau) = \frac{\sum_{t \in T} \mathbb{I}(y_t \geq \tau) g_t}{\sum_{t \in T} \mathbb{I}(y_t \geq \tau)} \quad (2.23)$$

$$R(\tau) = \frac{\sum_{t \in T} \mathbb{I}(y_t \geq \tau) g_t}{\sum_{t \in T} g_t} \quad (2.24)$$

โดยที่

- |                             |   |
|-----------------------------|---|
| $y_t$                       | = ค่าความน่าจะเป็นในการทำนาย ซึ่งมีค่าอยู่ในช่วง 0 ถึง 1                                  |
| $g_t$                       | = คำตอบของการทำนาย โดยจะมีค่าเป็นได้แค่ 0 และ 1 เท่านั้น                                  |
| $\mathbb{I}(y_t \geq \tau)$ | = พังก์ชันซึ่งจะมีค่าเป็น 1 ถ้าหาก $y_t$ มีค่ามากกว่า $\tau$ นอกเหนือจากนั้นจะมีค่าเป็น 0 |

สามารถหา average precision ได้จากการนี้

$$AP = \sum_{j=1}^{10000} P(\tau_j) [R(\tau_j) - R(\tau_j + 1)] \quad (2.25)$$

โดยที่  $\tau = \frac{j}{10000}$

ii. Hit@k

เหมือนกันกับ Top@k คือการจัดลำดับความน่าจะเป็นของแต่ละคำทำกับจำนวน k อันดับแรก ถ้าหากมีคำทำกับที่ถูกต้องอยู่ในลำดับเหล่านั้น จะถือว่าการทำนายถูกต้อง ซึ่งสามารถเขียนเป็นสมการได้ดังนี้

$$\frac{1}{|V|} \sum_{v \in V} V_{e \in G_v} \mathbb{I}(rank_{v,e} \leq k) \quad (2.26)$$

โดยที่

$$\begin{aligned}
 V &= \text{วิดีโอที่ใช้ในการทดสอบทั้งหมด} \\
 G_v &= \text{คำตอบของวิดีโอ } v \\
 rank_{v,e} &= \text{อันดับของคำตอบที่ถูกต้อง } e \text{ ของวิดีโอ } v \text{ ที่ได้จากการทำนาย} \\
 k &= \text{อันดับที่ใช้เป็นเกณฑ์}
 \end{aligned}$$

iii. Precision at equal recall rate (PERR)

สำหรับแต่ละวิดีโอจะดูความแม่นยำของผลการทำนาย  $k$  อันดับแรก โดยที่  $k$  คือจำนวนคำตอบทั้งหมดของวิดีโอนั้น จากนั้นเฉลี่ยค่าเหล่านั้นด้วยจำนวนวิดีโอทั้งหมด สามารถเขียนได้ในรูปสมการดังนี้ โดยใช้ตัวแปรเดียวกันกับของ Hit@ $k$

$$\frac{1}{|V : |G_v| > 0|} \sum_{v \in V : |G_v| > 0} \left[ \frac{1}{|G_v|} \sum_{e \in G_v} \mathbb{I}(rank_{v,e} \leq |G_v|) \right] \quad (2.27)$$

- (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 2.2: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลสำหรับทดสอบของ YouTube-8M

- (f) โมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างแล้วปรับโมเดลด้วยชุดข้อมูลของ Sports-1M เมื่อนำไปทดสอบกับชุดข้อมูลของ Sports-1M พบร่วงประสิทธิภาพเมื่อทดสอบด้วย Hit@1 และ Hit@5 สูงสุดอยู่ที่ 65.7% และ 86.2% ตามลำดับ ซึ่งใกล้เคียงกับสถิติสูงสุดในตอนนั้นที่ทำไว้ 73.0% และ 91.0% (พ.ศ. 2559) โดยที่ใช้เพียงคุณลักษณะจากการเรียนรู้ข้อมูลไว้แล้ว เมื่อเทียบกับสถิติเก่าที่มีการใช้ optical flow เข้ามาช่วย จึงเป็นการพิสูจน์ให้เห็นว่าจำนวนข้อมูลนั้นมีผลต่อการพัฒนาโมเดลปัญญาประดิษฐ์
- (g) โมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างแล้วปรับโมเดลด้วยชุดข้อมูลของ ActivityNet เมื่อนำไปทดสอบกับชุดข้อมูลของ ActivityNet พบร่วงประสิทธิภาพเมื่อทดสอบด้วย mAP สูงสุดอยู่ที่ 77.6% ในขณะที่สถิติเดิมทำไว้เพียง 53.8% เป็นการพิสูจน์ว่าชุดข้อมูลนี้นั้นมีความครอบคลุม (generalize) พอดีจะนำไปใช้กับงานประเภทนี้ในหมวดอื่นๆ เนื่องจากว่าจำนวนข้อมูลที่มีขนาดใหญ่และมีความหลากหลาย
- (h) ปัญหาที่พบ  
เนื่องจากว่า YouTube-8M นั้นมีจำนวนข้อมูลที่เยอะมาก ทำให้ไม่สามารถตรวจสอบความถูกต้องของชุดข้อมูลได้ทั้งหมดว่ามีความถูกต้องมากน้อยขนาดไหน ทำให้อาจเกิดข้อผิดพลาดได้ (ปัจจุบันปี 2019 YouTube-8M ได้มีการตรวจสอบข้อมูลอีกครั้ง เพื่อเพิ่มประสิทธิภาพของชุดข้อมูลซึ่งทำให้ปัจจุบันจำนวนข้อมูล และจำนวนคำจำกัดน้อยลงจากข้อมูลที่ใช้อ้างอิงในบทความข้างต้นที่ได้กล่าวมา)

## ชุดข้อมูล Atomic visual action (AVA)

AVA<sup>[8]</sup> 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 วิดีโอด้วยคัดเลือกจะมีเกณฑ์ดังนี้
  - A. วิดีโอด้วยอยู่ในหมวด ภาพยนต์ และละครโทรทัศน์
  - B. วิดีโอด้วยมีความยาวมากกว่า 30 นาที
  - C. เผยแพร่มาแล้วเป็นระยะเวลาอย่างน้อย 1 ปี
  - D. มีจำนวนยอดคนดูมากกว่า 1,000 คน
  - E. ลงทะเบียนวิดีโอบางประเภท เป็นภาพขาว-ดำ มีความละเอียดต่ำ การ์ตูน หรือวิดีโอกลเม
- 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
 

การสร้างคำจำกัดของกิจกรรมทำจะถูกสร้างขึ้นโดยมนุษย์ ซึ่งผู้วิจัยจะใช้โปรแกรมสำหรับช่วยเหลือในการสร้างคำจำกัดความลักษณะ โดยสามารถกำหนดคำจำกัดของกิจกรรมทำได้สูงสุดถึง 7 คำต่อ 1 กรอบสี่เหลี่ยม นอกจากนี้สามารถตั้งสถานะเนื้อหาที่ไม่เหมาะสม 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 (FCN) คือ โครงข่ายประสานภาพเทียมที่นำมาใช้แสดงลักษณะของวัตถุ(ตำแหน่งวัตถุ, ความตื้นลึกวัตถุ) ที่ปรากฏบนรูป RGB1
    - A. Appearance-FCN (A-FCN) คือ โครงข่ายประสานภาพเทียมที่นำมาระยะห่างมาใช้แสดงลักษณะของวัตถุ(ตำแหน่งวัตถุ, ความตื้นลึกวัตถุ) ที่ปรากฏบนรูป RGB1
    - B. MotionFCN (M-FCN) คือ โครงข่ายประสานภาพเทียมที่แยกการเคลื่อนไหว 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 ใช้คอนโวลูชันในการสกัดคุณลักษณะในแต่ละเฟรม 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 2.3: ผลการทดลองของวิธีต่างๆบนคุณลักษณะระดับเฟรม

- (d) ปัญหาที่พบ ในปัจจุบันยังไม่มีโมเดลปัญญาประดิษฐ์ที่ทดสอบด้วยชุดข้อมูล AVA และได้ผลการทำงานที่ดี เนื่องจากชุดข้อมูลนี้สนใจการกระทำของมนุษย์ที่มีรายละเอียดเล็กๆน้อยๆ ทำให้ยากต่อการทำนายสำหรับโมเดลปัญญาประดิษฐ์

## ชุดข้อมูล Moments in Time

Moments in time<sup>[16]</sup> คือชุดข้อมูลที่ใช้มนุษย์ในการกำกับข้อมูล ทั้งหมดให้กับวิดีโอด้วยสิ่งที่ 1 ล้านวิดีโอ และมีจำนวนกิจกรรมหรือกระทำต่างกัน 339 หมวดหมู่ โดยแต่ละวิดีโอมีความยาวอยู่ที่ 3 วินาที เนื่องจากเป็นเวลาเฉลี่ยที่มนุษย์ใช้ในการเข้าใจกับเหตุการณ์ที่เกิดขึ้น (human working memory) รูปแบบของชุดข้อมูลจะมีอยู่ทั้งหมดอยู่ 3 รูปแบบ ได้แก่ ภายนอก (spatial) เสียง (auditory) และการเคลื่อนไหว (temporal) นอกจากนี้ชุดข้อมูลนี้นั้นไม่รวมเพียงแค่การกระทำของมนุษย์เท่านั้น ยังรวมไปถึง สัตว์ สิ่งของ และ ปรากฏการณ์ธรรมชาติ ทำให้ ชุดข้อมูลนี้เป็นการท้าทายรูปแบบใหม่ เพราะด้วยข้อมูลที่มีความซับซ้อนมากขึ้น เช่น การสร้างโมเดลที่สามารถอภิปรายการกระทำ ได้ถึงแม้ว่าสิ่งที่เราสนใจ (มนุษย์ สัตว์ สิ่งของ หรือปรากฏการณ์ธรรมชาติ) จะแตกต่างกัน เป็นต้น

### 1. รายละเอียดชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนับสนุนให้การเรียนรู้ในวิดีโอด้วยการกระทำของคนหรือสัตว์ 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<sup>[17]</sup> เว็บไซต์ที่เก็บรวบรวมคำกริยาภาษาอังกฤษขนาดใหญ่ จากนั้นนำมาแบ่งกลุ่มคำที่มีความหมายใกล้เคียงกันโดยใช้คุณลักษณะจาก Propbank<sup>[21]</sup> และ FrameNet<sup>[3]</sup> โดยเก็บข้อมูลเป็นแบบเวกเตอร์คุณลักษณะฐานสอง (binary feature vector) ซึ่งถ้าคำใดมีความเกี่ยวข้องกับคุณลักษณะนี้จะให้ค่าเป็น 1 ถ้าไม่เกี่ยวข้องกันจะให้ค่าเป็น 0 จากนั้นจึงใช้วิธี k-means clustering ในการแบ่งกลุ่ม เมื่อแบ่งกลุ่มแล้วจากนั้นจะเลือกคำจากในแต่ละกลุ่มนั้น โดยคำที่เลือกมาจะเป็นคำที่ใช้บ่อยที่สุดในกลุ่มนั้น และลบคำน้ออกจากกลุ่มอื่นๆทั้งหมด (คำๆหนึ่งสามารถอยู่ได้หลายกลุ่ม) จากนั้นจะทำการวนการนี้ไปเรื่อยๆ แต่คำที่เลือกมาจะต้องไม่มีความหมายคลุมเครือ หรือเป็นสิ่งที่ไม่สามารถมองเห็นหรือได้ยินได้ และต้องไม่มีความหมายเหมือนกับคำที่เคยเลือกมาก่อน จนสุดท้ายแล้วได้ออกมาที่ 339 หมวดหมู่
  - ii. ต่อมาทำการหาชุดข้อมูลวิดีโอด้วยตัวตัดออกมาเพียง 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 โดยวิดีโอด้วยวิดีโอด้วยคำกำกับอยู่แล้วนั้นถ้าพนักงานของ AMT ตอบเหมือนกันเกิน 90% ถึงจะนำเข้าไปรวมกับชุดข้อมูลส่วนอีก 64 วิดีโอ ถ้าเป็นชุดข้อมูลสำหรับสร้างโมเดลจะต้องผ่านพนักงานของ AMT อย่างน้อย 3 ครั้ง และต้องมีคำกำกับเหมือนกัน 75% ขึ้นไปถึงจะถือว่าเป็นคำกำกับที่ถูกต้อง ถ้าเป็นชุดข้อมูลสำหรับตรวจสอบ และชุดข้อมูลสำหรับทดสอบ จะต้องผ่านพนักงานของ AMT อย่างน้อย 4 ครั้ง และต้องมีคำกำกับเหมือนกัน 85% ขึ้นไป เพื่อผลที่ไม่ต่างกันทั้งหมดทั้ง 100% เพราะจะทำให้วิดีโอนั้นยากเกินไปที่จะทำให้สามารถจำการกระทำได้

## 2. การเตรียมข้อมูล

- (a) ชุดข้อมูลสำหรับสร้างโมเดลจะมี 802,264 วิดีโอ และมีวิดีโອนในแต่ละหมวดหมู่อยู่ที่ 500 ถึง 5,000 วิดีโอ
- (b) ชุดข้อมูลสำหรับตรวจสอบความต่อจะมี 33,900 วิดีโอ และมีวิดีโອนในแต่ละหมวดหมู่อยู่ที่ 100 วิดีโอ
- (c) แยกเฟรม RGB ออกมาจากวิดีโอ และทำการเปลี่ยนขนาดให้เป็น 340x256 pixel
- (d) ใช้อัลกอริทึม TVL1 optical flow จาก OpenCV เพื่อลดข้อมูลรบกวนที่จะเกิดขึ้น
- (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 2.4: 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

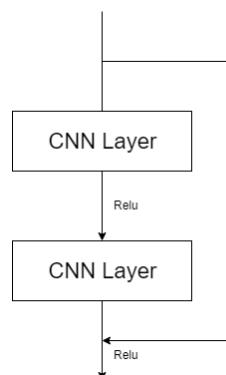
#### 4. ปัญหาที่พบ

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

## 2.4 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมนุษย์

### 2.4.1 ResNet

ในการสร้างโมเดลปัญญาประดิษฐ์นั้นการใช้จำนวนชั้น (layer) เยอะนั้นจะทำให้ได้คุณลักษณะของข้อมูลที่อุ่นมาเยื่อตามไปด้วย แต่การที่คุณลักษณะของข้อมูลเยื่อไม่ได้หมายความว่าโมเดลปัญญาประดิษฐ์จะให้ประสิทธิภาพที่ดีเสมอไป ซึ่งสามารถแก้ปัญหานี้ได้โดยใช้ residual network (ResNet) ซึ่งเป็น convolution neuron network (CNN) ประเภทหนึ่ง ที่ส่วนใหญ่จะนำมาใช้กับข้อมูลที่เป็นรูปภาพ เช่น การจดจำวัตถุ เป็นต้น โดย ResNet นี้จะสามารถทำการข้ามชั้นที่ไม่จำเป็นได้ การข้ามชั้นที่ไม่จำเป็นจะช่วยลดเวลาที่ใช้ในการสร้างโมเดลปัญญาประดิษฐ์ และทำให้ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ดีขึ้น



รูปที่ 2.18: หลักการของ Residual block ของ ResNet

การทดลองของโมเดลปัญญาประดิษฐ์ ResNet ด้วยการทำจำแนกรูปภาพโดยใช้ชุดข้อมูลทดสอบ ImageNet ที่มีหมวดหมู่มากกว่า 1,000 หมวดหมู่ มาเทียบกับโมเดลปัญญาประดิษฐ์ทั่วไป (plain model) ที่จำนวนชั้น 18 ชั้น และ 34 ชั้น โดยโครงสร้างพื้นฐานของโมเดลปัญญาประดิษฐ์ ResNet และโมเดลปัญญาประดิษฐ์ทั่วไปเหมือนกัน ซึ่งผลลัพธ์อัตรา誤率ของความผิดพลาดจะได้ออกมาตามตารางที่ 2.5

จำนวนชั้นของโมเดลปัญญาประดิษฐ์	Training error	
	plain	ResNet
18	27.94	27.88
34	28.54	25.03

ตารางที่ 2.5: อัตรา誤率ของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet

จากตาราง 2.5 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

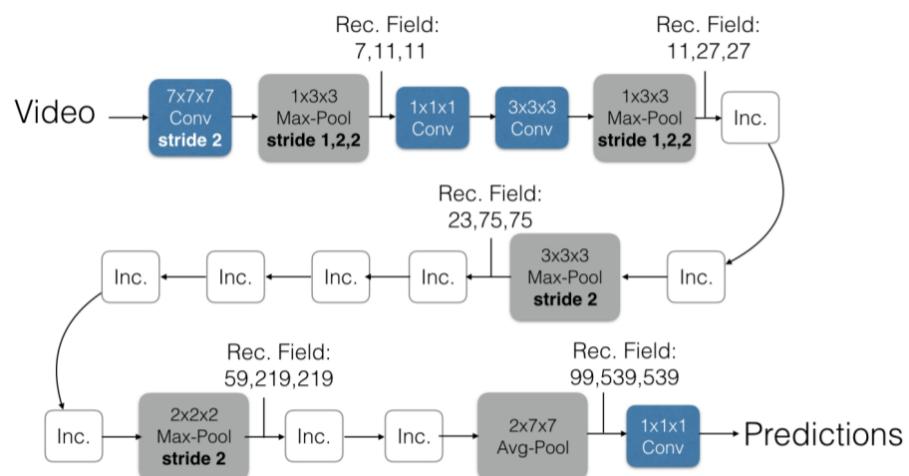
ตารางที่ 2.6: ค่าความผิดพลาดที่ได้จากการทดลองจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10

จากตาราง 2.6 จะเห็นได้ว่าที่โมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1,202 นั้นมีค่าความผิดพลาดเกิดขึ้นมากกว่าจำนวนชั้น 110 ซึ่งอาจจะเป็นไปได้ว่าขนาดของโมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1,202 นั้นมากเกินไปสำหรับชุดข้อมูลขนาดเล็กนี้

## 2.4.2 Inflated 3D convolutional network

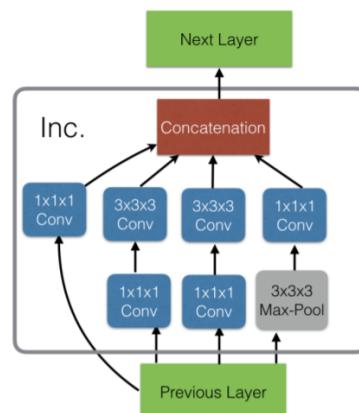
ในการพัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์นั้นมีพื้นฐานจากการจำแนกวัตถุ (object classification) หมายถึงการใช้รูปภาพหนึ่งรูปในการประมวลผลและทำนายอุกมาดาว่ายในรูปนั้นมีบริบทการกระทำอย่างไร โดยไม่ได้คำนึงถึงข้อมูลเชิงต่อเนื่อง (spatio-temporal information) จากบทความ ”Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”<sup>[5]</sup> นั้นได้พัฒนาโครงสร้างของโมเดลปัญญาประดิษฐ์ (architecture) ที่มีประสิทธิภาพในการประมวลผลภาพเคลื่อนไหวได้ชื่อว่า I3D หรือ inflated 3D-convolution network โดยโครงสร้างพื้นฐานของ I3D นั้นมาจากการนำ Inception-v1<sup>[18]</sup> ที่ถูกพัฒนาโดย Google ซึ่งเป็นโครงสร้างที่มีประสิทธิภาพสูงในการจำแนกวัตถุในรูปภาพ แล้ว I3D นั้นได้ทำการขยายมิติของเครื่องเรนาลจาก 2 มิติ เป็น 3 มิติ เพื่อให้โมเดลปัญญาประดิษฐ์สามารถเรียนรู้ข้อมูลเชิงต่อเนื่องได้

### Inflated Inception-V1



รูปที่ 2.19: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D<sup>[5]</sup>

### Inception Module (Inc.)



รูปที่ 2.20: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D<sup>[5]</sup>

จากรูปที่ 2.19 ในส่วนของชั้น Inception (Inc.) จะมีลักษณะโครงสร้างดังรูปที่ 2.20 เนื่องจากว่าการที่โมเดลนั้นซับชั้อนมากขึ้นก็ต้องใช้ทรัพยากรในการประมวลผลมากขึ้น Google จึงออกแบบโครงสร้างที่สามารถลดความซับซ้อนของโมเดลลงด้วยการใช้เครอร์เนลขนาด  $1 \times 1$  ในเครอร์เนล 2 มิติ ( $1 \times 1 \times 1$  ใน 3 มิติ) เพื่อลดจำนวน channel ของเครอร์เนลลง ตัวอย่างเช่น หากในชั้นก่อนหน้าเป็นเอาร์พุตที่มีขนาด  $14 \times 14 \times 480$  (480 คือจำนวน channel ของเครอร์เนล) หากในชั้นถัดไปใช้เครอร์เนลที่มีขนาด  $5 \times 5 \times 48$  จะทำให้มีจำนวนพารามิเตอร์ถึง  $(14 \times 14 \times 480) \times (5 \times 5 \times 48) = 112.9$  ล้าน แต่ถ้าหากใช้เครอร์เนลขนาด  $1 \times 1 \times 16$  มาคั่นระหว่างชั้นนั้นจะทำให้จำนวนพารามิเตอร์กลายเป็น  $(14 \times 14 \times 480) \times (1 \times 1 \times 16) = 1.5$  ล้าน และเอาร์พุตของชั้นนี้จะมีขนาด  $14 \times 14 \times 16$  ก่อนจะนำไปคำนวณในชั้นถัดไป  $(14 \times 14 \times 16) \times (5 \times 5 \times 48) = 3.8$  ล้าน เมื่อนำจำนวนพารามิเตอร์มารวมกันจะได้พารามิเตอร์เพียง  $3.8 + 1.5 = 5.3$  ล้านเท่านั้น ซึ่งน้อยกว่าการใช้เครอร์เนลขนาด  $5 \times 5 \times 48$  โดยตรงซึ่งทำให้การพัฒนาโมเดลนั้นเป็นไปได้เร็วขึ้นมาก ทั้งยังสามารถลดปัญหาการเกิด overfit ได้ด้วย ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อเทียบกับ long-short term memory (LSTM), 3D-convolution network, two-stream และ 3D-fused โดยใช้เครื่องมือในการวัดผลคือ Top@1 accuracy ตามตารางที่ 2.7

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

### 2.4.3 เครื่องมือกำกับคุณลักษณะ

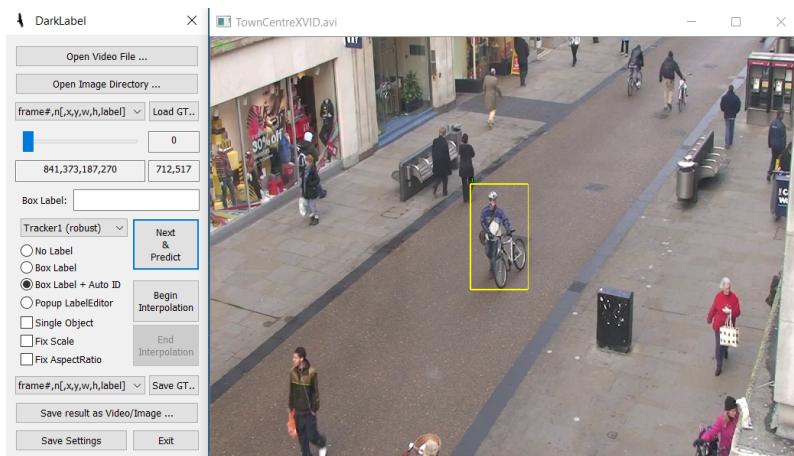
จากการค้นคว้าหาเครื่องมือในการสร้างคำกำกับข้อมูลเพื่อใช้เป็นแนวทางในการออกแบบเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ พบรเครื่องมือที่เปิดให้ใช้งานสาธารณะ (open source) 2 เครื่องมือ คือ DarkLabel และ OpenLabeling โดยสรุปข้อสำคัญได้ดังนี้

#### โปรแกรม DarkLabel

เป็นโปรแกรมที่ช่วยในการทำนายคำกำกับและบันทึกในรูปแบบต่างๆ รองรับข้อมูลป้อนเข้าในรูปแบบไฟล์วิดีโอ avi, mp4 หรือกลุ่มรูปภาพ มีขั้นตอนการสร้างคำกำกับดังนี้

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

หลังจากที่ผู้ใช้จัดตั้งค่าและเริ่มใช้โปรแกรม DarkLabel พบรฯ เป็นโปรแกรมที่การทำงานส่วนใหญ่เป็นการสร้างคำกำกับแบบใช้มนุษย์เป็นคนทำด้วยตัวเอง ซึ่งทำให้ใช้เวลาในการทำงาน



รูปที่ 2.21: UI ของโปรแกรม DarkLabel

## โปรแกรม OpenLabeling

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

### 1. การทำงานแบบอัตโนมัติ

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดนิยมดังนี้

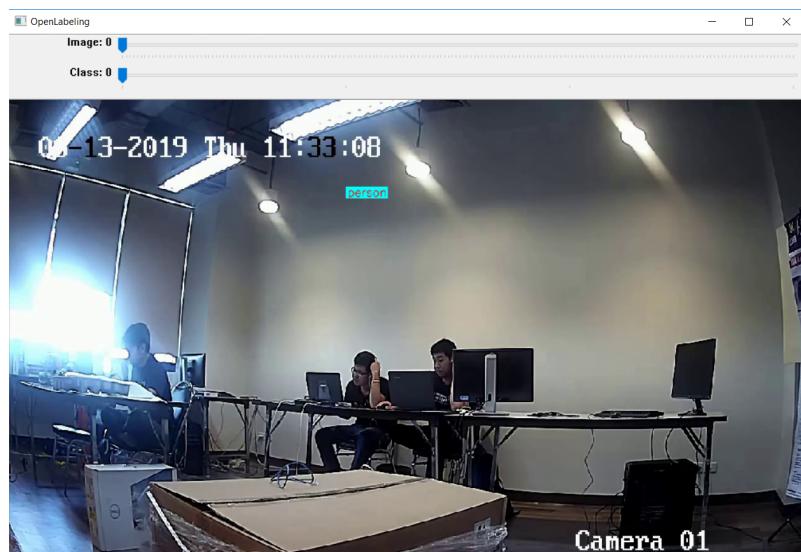
- โปรแกรมจะทำงานอัตโนมัติโดยใช้โมเดลปัญญาประดิษฐ์ในการทำนายคีย์เฟรม (keyframe) และทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดไปด้วยอัลกอริทึมที่ใช้การคำนวนคณิตศาสตร์และการประมวลผลภาพในภาพที่เหลือ ผลลัพธ์ที่ได้คือรูปภาพและไฟล์คำจำกัดนิยมคุณลักษณะ

### 2. การทำงานแบบทำด้วยตัวเอง

หลังจากป้อนวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการสร้างคำจำกัดนิยมดังนี้

- สร้างกรอบสี่เหลี่ยมขึ้นมาโดยใช้มนุษย์เป็นคนสร้าง
- กดปุ่มเพื่อทำนายตำแหน่งต่อไปของกรอบสี่เหลี่ยมในเฟรมถัดไป จนกว่าทั้งเกิดข้อผิดพลาด
- ลบกรอบสี่เหลี่ยมที่พลาด และเริ่มทำขั้นตอนที่ 1 อีกครั้งจนครบทุกเฟรมในวิดีโอ

หลังจากที่ได้ทดลองใช้โปรแกรม OpenLabeling ทั้ง 2 รูปแบบการทำงานแล้วพบว่า การทำงานแบบอัตโนมัติ ไม่สามารถปรับแก้ไขสิ่งใดในระหว่างกระบวนการนั้น ทำให้หากเกิดกรณีที่โมเดลทำนายกรอบสี่เหลี่ยมพลาดหรือเกินมา จะไม่สามารถแก้ไขได้ และการทำงานแบบทำด้วยตัวเองไม่มีระบบตรวจสอบกรอบสี่เหลี่ยม ทำให้ผู้ใช้งานจะต้องสร้างกรอบสี่เหลี่ยมขึ้นมาเอง



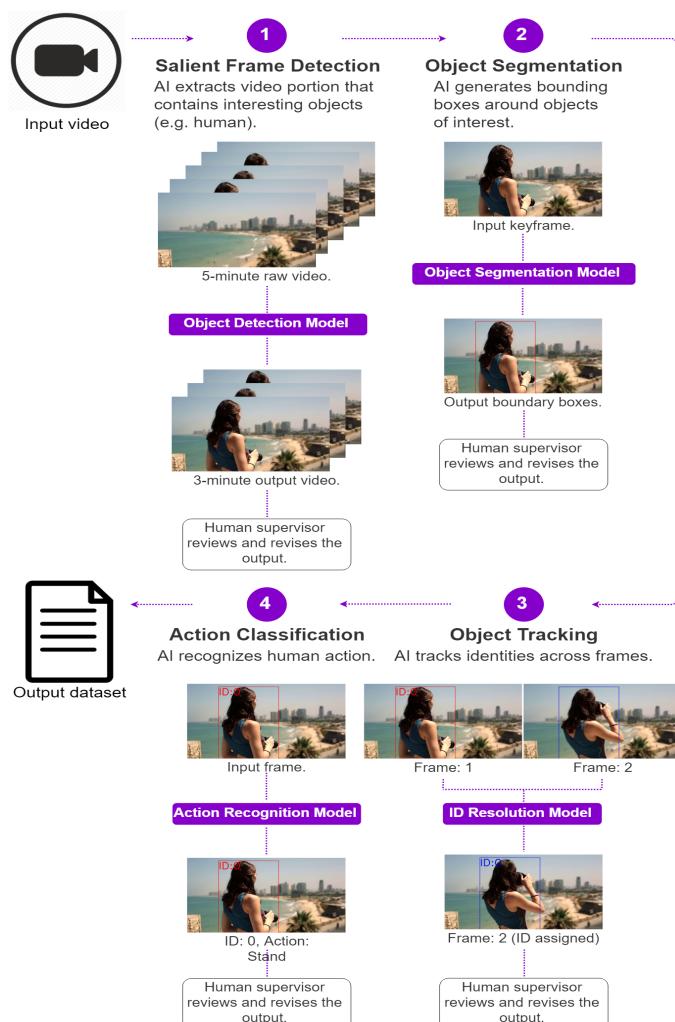
รูปที่ 2.22: UI ของโปรแกรม OpenLabeling

## บทที่ 3

### ระเบียบวิธีวิจัย

ในการทำโครงการวิจัยเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ จะมีการทำงานหลากหลายส่วนมาทำงานร่วมกัน ซึ่งต้องมีระเบียบวิธีวิจัยอธิบายถึงขั้นตอนการดำเนินงานตั้งแต่เริ่มศึกษาข้อมูลจนไปถึงสิ้นสุดกระบวนการวิจัย โดยใช้ภาษาไทยเป็นภาษาหลักในการเขียนโปรแกรม

#### 3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.1: ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

## 3.2 ความต้องการของระบบ

### 3.2.1 ความต้องการเชิงการใช้งาน (functional requirements)

1. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถตัดวิดีโอช่วงเวลาที่ไม่มีมนุษย์อยู่ออกได้อัตโนมัติโดยใช้ปัญญาประดิษฐ์
2. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์สามารถระบุตำแหน่งมนุษย์แต่ละคนในวิดีโอด้วย ยืน นั่ง นอน เล่นโทรศัพท์ เดิน กินข้าว พูดคุย
3. ชุดข้อมูลที่ได้จากเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถนำไปใช้ในการพัฒนาโมเดลปัญญาประดิษฐ์ต่อได้
4. สร้างระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ที่มีมนุษย์สามารถทำงานร่วมกับปัญญาประดิษฐ์ได้
5. ระบบวิเคราะห์การกระทำมนุษย์ต้องสามารถนำวิดีโอมาวิเคราะห์ข้อมูลการกระทำและตำแหน่งของมนุษย์แต่ละคน แล้วนำข้อมูลเหล่านั้นไปสร้างรายงานที่มีคำกำกับอອกมาได้ โดยรายละเอียดรายงานจะมีดังนี้
  - (a) เวลา (time stamp)
  - (b) การกระทำ
  - (c) ตำแหน่ง โดยจะบอกในลักษณะของกรอบสี่เหลี่ยมครอบพื้นที่ที่มีมนุษย์คนนั้นๆอยู่

### 3.2.2 ความต้องการเชิงวิศวกรรม (non-functional requirements)

1. สร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์โดยใช้ภาษาไฟรอน
2. ความละเอียดอย่างต่ำของวิดีโอต้องมากกว่า  $640 \times 480$  (กว้าง x สูง)
3. วิดีโอจะต้องมีอัตราเฟรมต่อวินาที (fps) อย่างต่ำ 10 เฟรมต่อวินาที

### 3.3 หน้าที่ความรับผิดชอบ

ปฐมพงศ์ สินธุ์งาม สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดการกระทำมนุษย์ 3D รวมถึงออกแบบและสร้างระบบ Tracker

ศุภกร เบญจวิกรัย รวมฟังก์ชันและระบบต่างๆของแอพพลิเคชัน รวมถึงออกแบบและสร้างระบบ Select และ Detect

อุกฤษฎ์ เลิศวรรณาการ สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดการกระทำมนุษย์ 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 ใช้สำหรับการคำนวณและ 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 และระบบบิเคราะห์การกระทำการของมนุษย์ในวิดีโอ มีแผนการทำงานซึ่งถูกแบ่งออกเป็นสาม ขั้นตอนดังนี้ ขั้นตอนแรกคือ ขั้นตอนของการศึกษาทำความเป็นไปได้ รวมถึงเทคโนโลยีปัจจุบันที่เกี่ยวกับการ สร้างแอปพลิเคชัน และการจัดการกระทำการของมนุษย์ด้วยปัญญาประดิษฐ์ เพื่อนำมาประยุกต์ใช้กับงานวิจัย นี้ ขั้นตอนที่สองคือ ขั้นตอนของการออกแบบและสร้างแอปพลิเคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการเทรน ไมเดลจากวิดีโอ ขั้นตอนที่สามคือ ขั้นตอนของการออกแบบและสร้างระบบบิเคราะห์การกระทำการของมนุษย์ได้โดย มีข้อกำหนดตามที่กล่าวไว้ในบทนำ ในการเริ่มทำงานวิจัยนี้นั้นสิ่งจำเป็นที่ต้องทำในอันดับแรกคือการศึกษาข้อมูล ในหัวข้อที่เกี่ยวข้อง หรืองานวิจัยอื่นที่ทำเอ้าไว้แล้ว เพื่อศึกษาและทำความเข้าใจ ข้อดี-ข้อเสีย ของเทคนิคหรือ กระบวนการต่างๆ เพื่อนำมาประยุกต์ใช้กับงานวิจัยนี้ ในการศึกษาเกี่ยวกับการออกแบบและ การสร้างแอปพลิ เคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการสร้างไมเดลจากวิดีโอ สิ่งที่ต้องให้ความสนใจคือฟังก์ชันการทำงาน การ ออกแบบและการจัดวางองค์ประกอบต่างๆในหน้าต่างแอปพลิเคชัน และความสะดวกในการใช้งาน จากนั้นจึงเริ่ม ศึกษาเกี่ยวกับ library ที่ใช้ในการสร้างแอปพลิเคชัน ส่วนการศึกษาเกี่ยวกับการสร้างระบบบิเคราะห์การกระทำ มนุษย์ จะมุ่งความสนใจไปที่ชุดข้อมูลสำหรับการวิเคราะห์วิดีโอ ไมเดลสำหรับการวิเคราะห์วิดีโอ เทคนิคในการ สร้างไมเดล เทคโนโลยีในการระบบบิเคราะห์วิดีโอ เพื่อใช้ในการออกแบบและสร้างระบบบิเคราะห์การกระทำ ของมนุษย์ในวิดีโอด้วยมีประสิทธิภาพ ในบทนี้จะกล่าวถึงกระบวนการออกแบบและการดำเนินการตามแผนที่วาง เอาไว้

### 3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

การออกแบบเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ ผู้วิจัยได้เลือกใช้ library PyQt และภาษา Pythonในการพัฒนา 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 จะต้องสามารถตัดวิดีโอ์ส่วนที่ไม่มีมนุษย์อยู่ออกจากวิดีโอด้วย กระบวนการ Detect จะต้องหาตำแหน่งของมนุษย์ภายในวิดีโอด้วย แล้วใช้กระบวนการ Track นำรายตำแหน่งต่อไปของมนุษย์ข้อมูลตำแหน่งของมนุษย์ที่ได้จากการกระบวนการ Detect และกระบวนการ Label นั้นต้องสามารถทำงานมุ่งเน้นที่การระบุพื้นฐานของมนุษย์ได้ เช่น ยืน เดิน นั่ง กินข้าว หรือ นอน เป็นต้น โดยทุกส่วนการทำงานมนุษย์ต้องสามารถทำงานร่วมกับปัญญาประดิษฐ์ได้ ดังรูปที่ 3.2

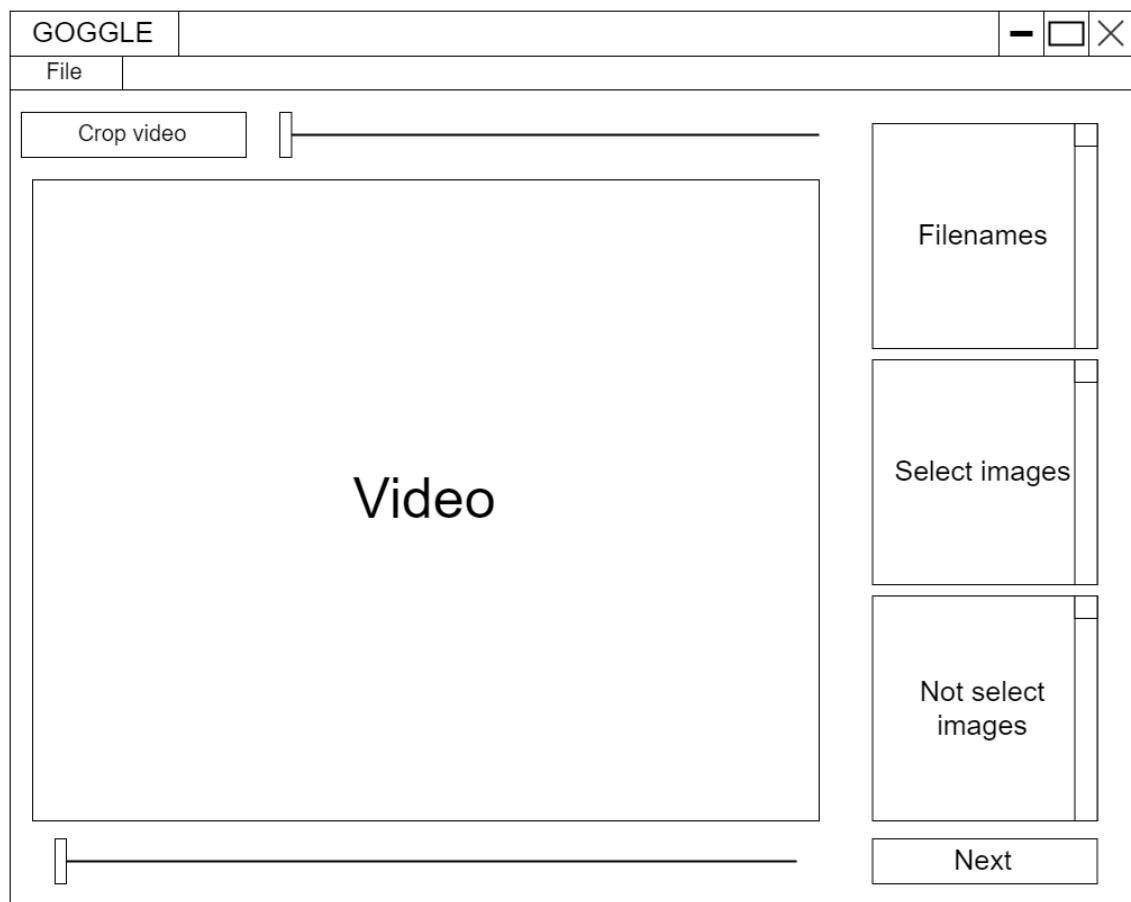


รูปที่ 3.2: กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

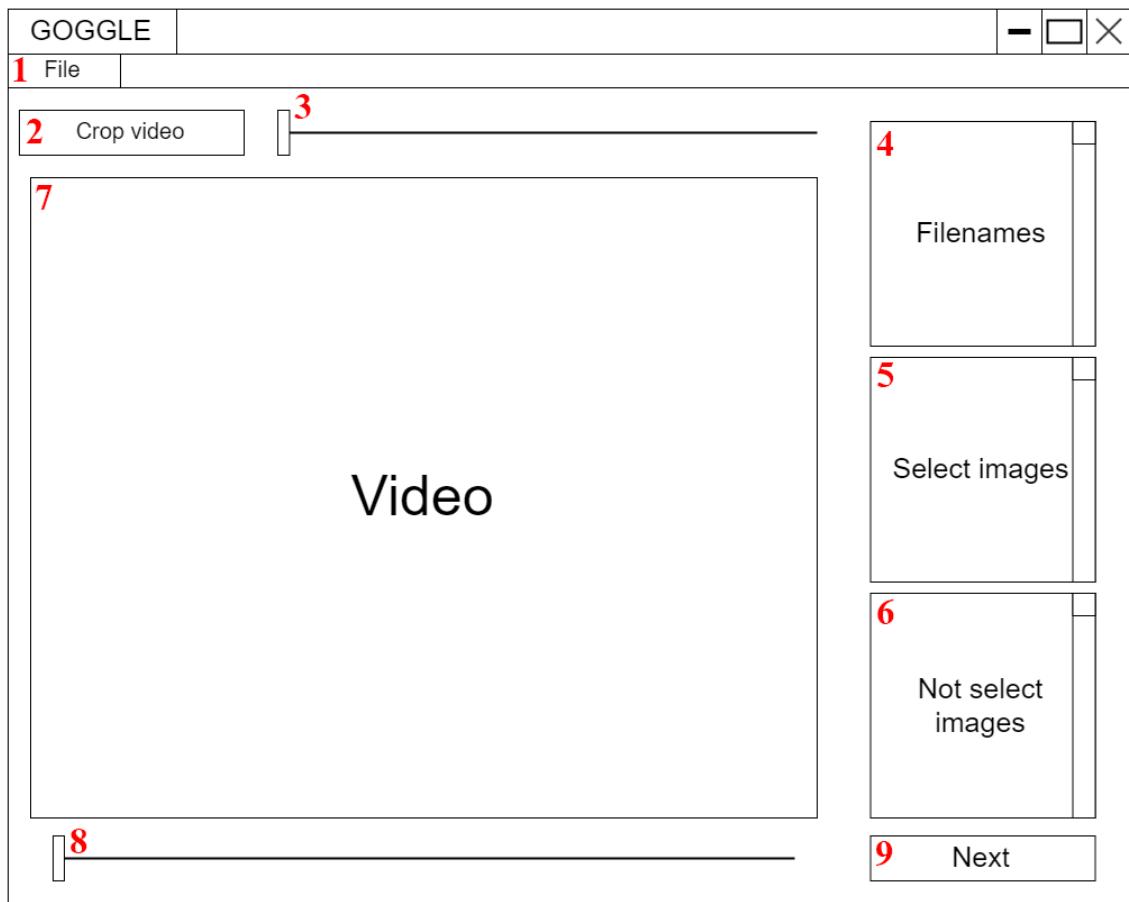
โดยแต่ละกระบวนการจะมีรายละเอียดดังนี้

#### 3.8.1.1 Select

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



รูปที่ 3.3: หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



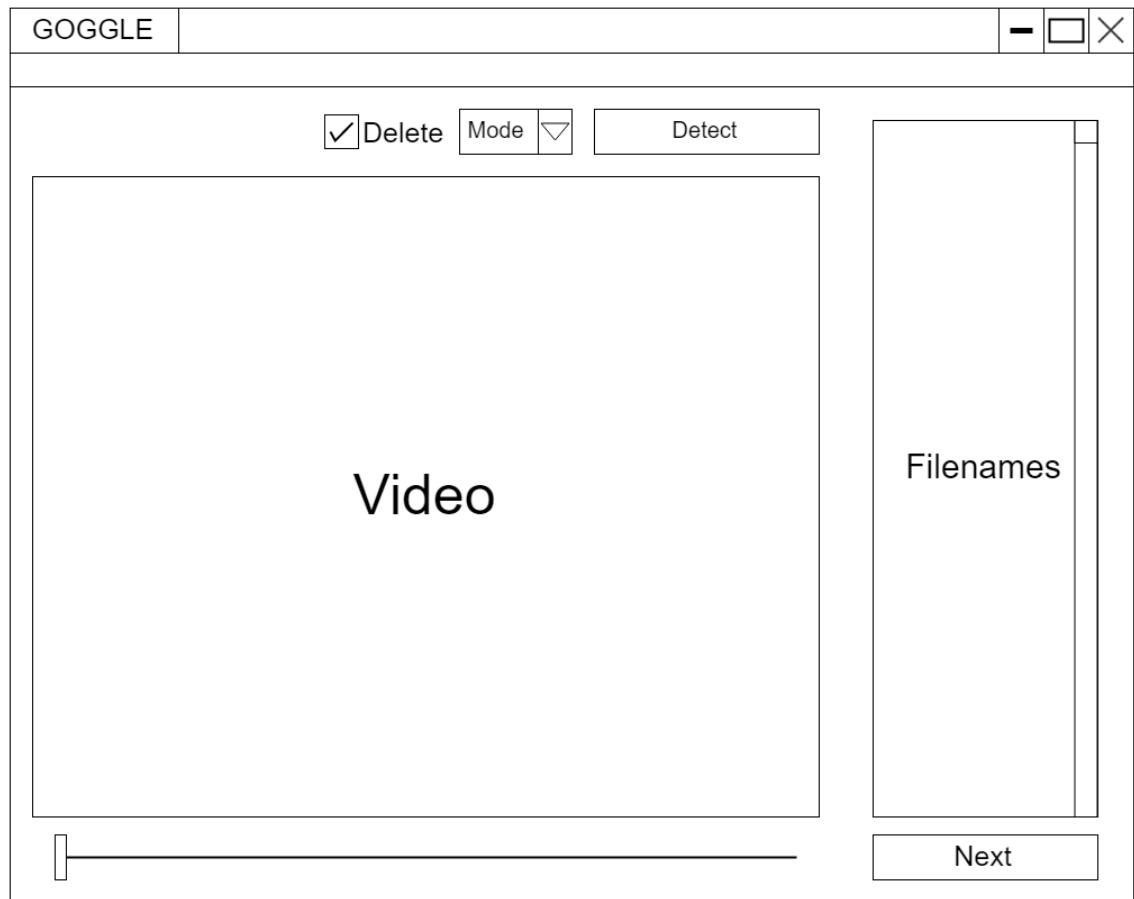
รูปที่ 3.4: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.4 มีรายละเอียดดังนี้

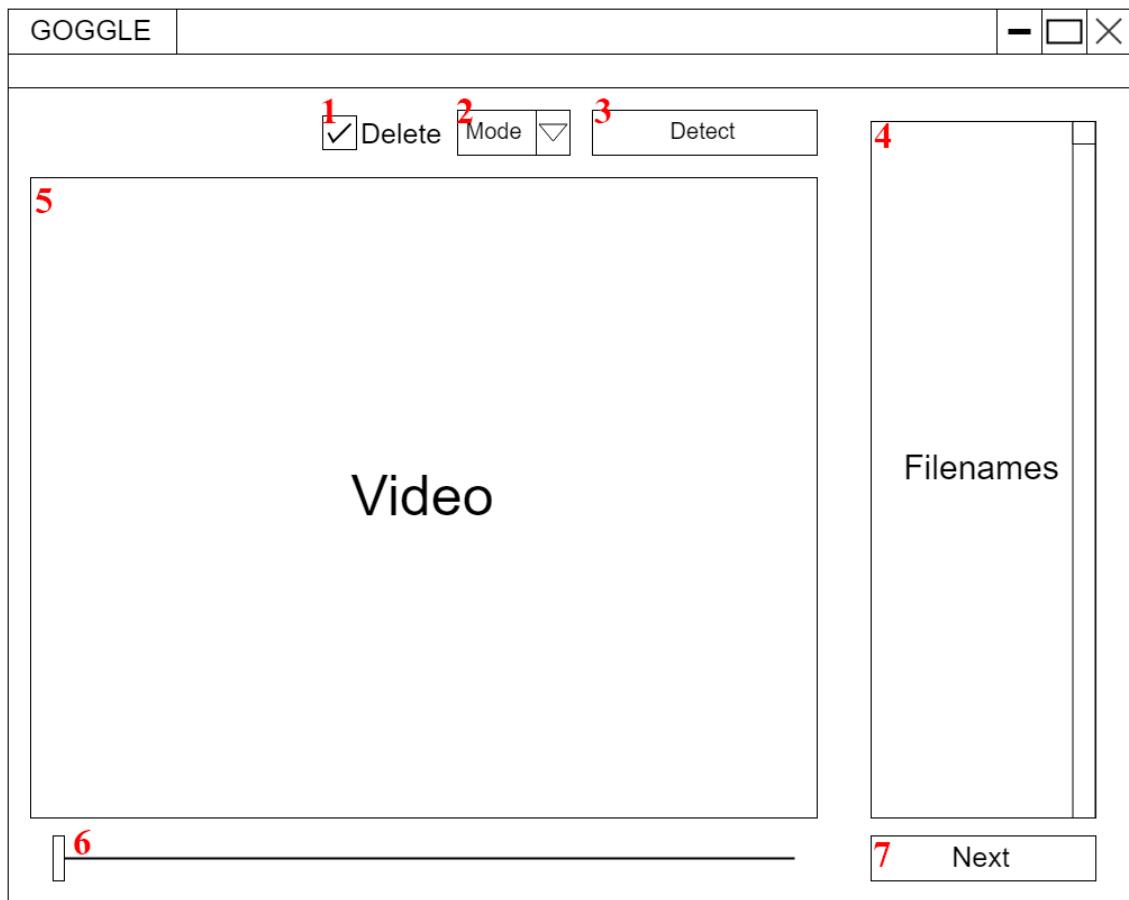
1. หมายเลข 1 คือปุ่มสำหรับเลือกไฟล์วิดีโอที่ต้องการจากในคอมพิวเตอร์เข้ามาในโปรแกรม
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบทำการสร้างคีย์เฟรมขึ้นมา แล้วใช้ปัญญาประดิษฐ์ประมวลผลเพื่อแยกคีย์เฟรมในเมื่อน้อย และคีย์เฟรมในเมื่อน้อยแบบอัตโนมัติ (Auto mode)
3. หมายเลข 3 คือแถบเลื่อนเพื่อกำหนดความถี่ในการหยิบคีย์เฟรม โดยจะมีช่วงอยู่ที่ 1 เฟรมต่อวินาที จนถึงอัตราเฟรมต่อวินาทีสูงสุดของวิดีโอิที่รับเข้ามา
4. หมายเลข 4 คือกล่องสำหรับแสดงชื่อวิดีโอิที่รับเข้ามาในโปรแกรมเพื่อเลือกเข้ามาใช้ในการประมวลผล
5. หมายเลข 5 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
6. หมายเลข 6 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
7. หมายเลข 7 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 5 หมายเลข 6 หรือหมายเลข 8
8. หมายเลข 8 คือแถบเลื่อนสำหรับเลือนคุณคีย์เฟรมทั้งหมดที่ระบบสร้างขึ้น
9. หมายเลข 9 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

### 3.8.1.2 Detect

กระบวนการ Detect จะต้องสามารถรับคีย์เพرمจากกระบวนการ Select มาประมวลผลด้วยปัญญาประดิษฐ์เพื่อหาตำแหน่งของมนุษย์ที่อยู่ในคีย์เพرم และสร้างกรอบสีเหลี่ยมครอบบริเวณดังกล่าวได้ในแบบอัตโนมัติ เพื่อแบ่งเบาภาระผู้ใช้ในการที่ต้องสร้างกรอบสีเหลี่ยมครอบตำแหน่งของมนุษย์ด้วยตัวเอง และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสีเหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของปัญญาประดิษฐ์ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.5



รูปที่ 3.5: หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



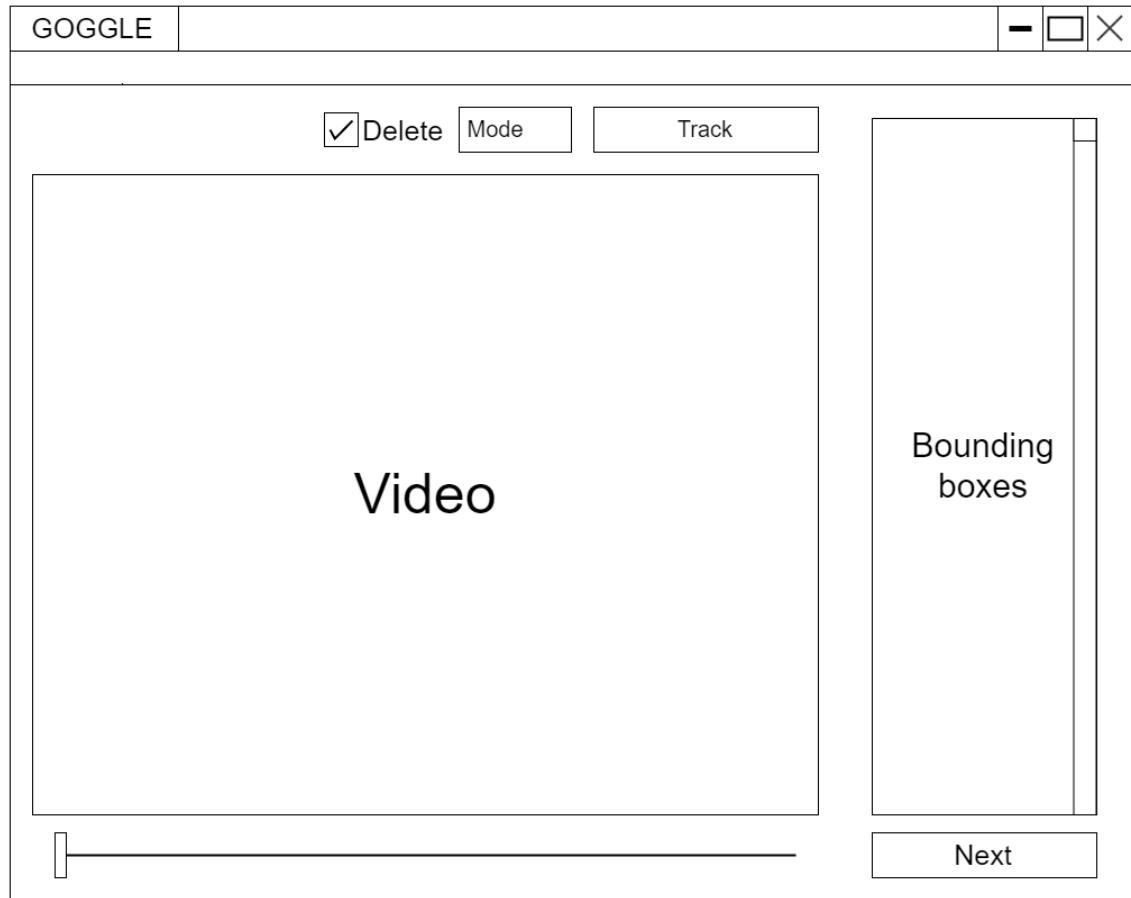
รูปที่ 3.6: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.6 มีรายละเอียดดังนี้

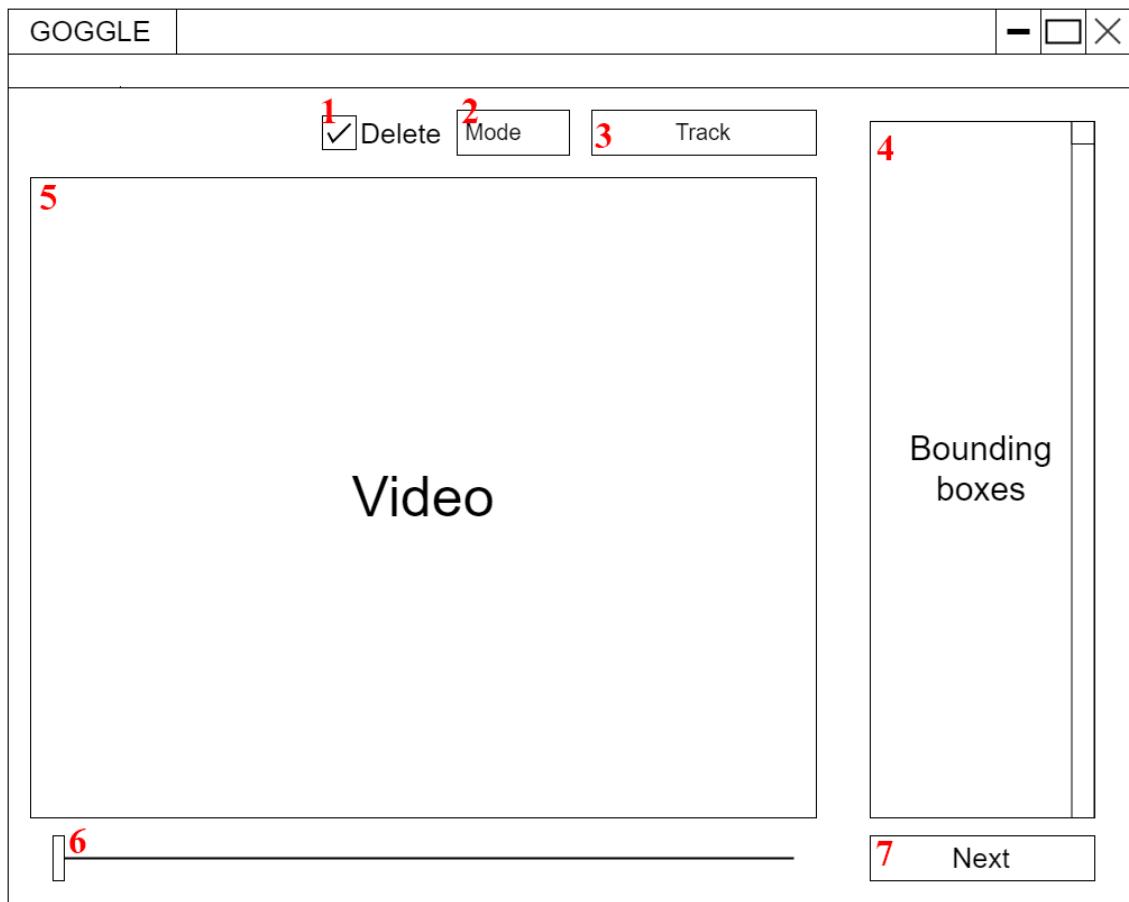
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเอง (Manual mode) เป็นลบรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจหาตำแหน่งของมนุษย์ในคิร์เฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
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



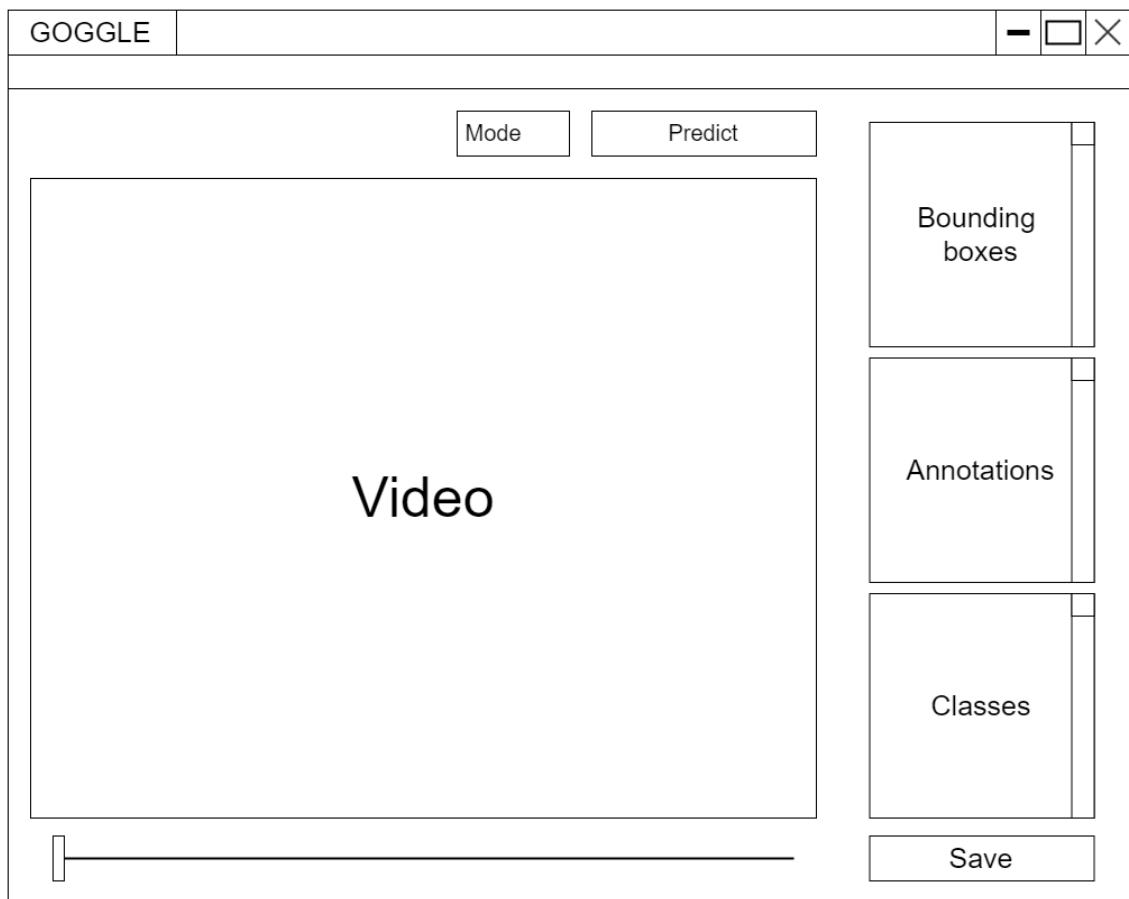
รูปที่ 3.8: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

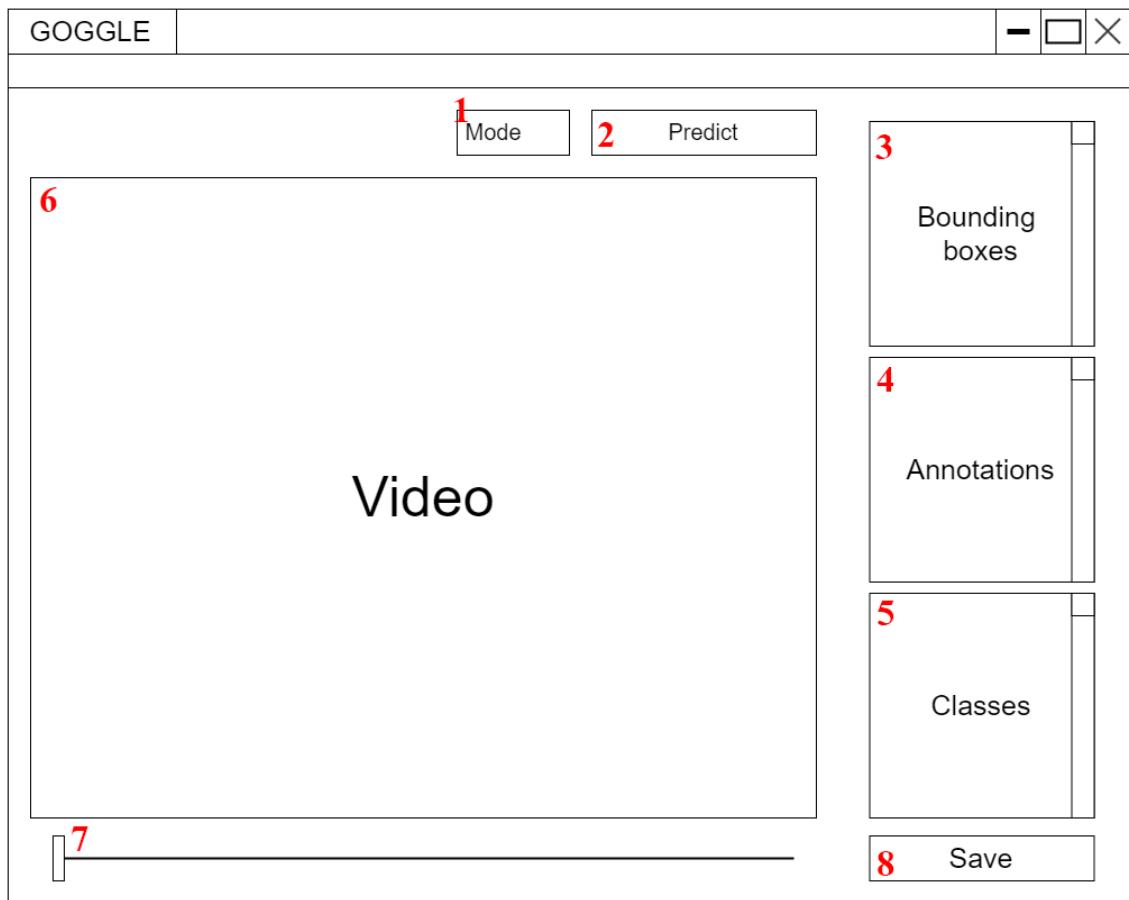
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเองเป็นลบกรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจสอบตำแหน่งของมนุษย์ในเฟรมระหว่างคิ้ยวเฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
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 จึงออกแบบหน้าต่างเดดงรูปที่ 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 โดยจะเรียงลำดับคู่กับกรอบสี่เหลี่ยมที่อยู่ในช่องหมายเลข 3
5. หมายเลข 5 คือกล่องสำหรับแสดงชุดการกระทำที่ปัญญาประดิษฐ์มีอยู่แล้ว ซึ่งในการทำงานแบบแก้ไขด้วยตนเองนั้น จะสามารถค้นหาการกระทำที่มีอยู่แล้วได้ และหากคำที่ใส่เขามานั้นมีอยู่ในชุดการกระทำก็จะเป็นการเพิ่มการกระทำนั้นเข้ามาแทน
6. หมายเลข 6 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 7
7. หมายเลข 7 คือແນບเลื่อนสำหรับเลื่อนดูเฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของปัญญาประดิษฐ์
8. หมายเลข 8 คือปุ่มสำหรับสร้างไฟล์ xml ของทุกๆเฟรมสำหรับใช้ในการสร้างโมเดลโดยรายละเอียดข้อมูลภายในไฟล์ xml จะอยู่ในหัวข้อ 3.8.1.5

### 3.8.1.5 รายละเอียดข้อมูลภายในไฟล์ xml

ไฟล์ xml นั้นเป็นรูปแบบที่นิยมใช้ในการเก็บข้อมูลสำหรับการสร้างโมเดลประเพณีตรวจสอบจับตัวตุ๊ก 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 การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุ 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 และตำแหน่งของกรอบสี่เหลี่ยม
2. เรียกชื่อและเฟรมของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่มีคำจำกัดเพื่อเป็นคำตอบ เก็บผลลัพธ์เป็นชุดข้อมูลผลลัพธ์จากการคำนวณ
  - (a) ชุดข้อมูลผลลัพธ์จากการคำนวณ ประกอบด้วย : ชื่อของวิดีโอ เฟรม และตำแหน่งของกรอบสี่เหลี่ยม
3. ประเมินผลค่าความแม่นยำในการคำนวณโดยเทียบระหว่างชุดผลลัพธ์จากการคำนวณ และชุดข้อมูลที่มีคำจำกัดเพื่อเป็นคำตอบ (ตั้งค่า IoU = 0.5)
4. เปรียบเทียบผลลัพธ์จากแหล่งที่มา

### 3.10 การออกแบบการทดสอบการทำงานของระบบนำทางตำแหน่งต่อไปของมนุษย์

#### 3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบนำทางตำแหน่งต่อไปของวัตถุในวิดีโอ สิ่งที่ใช้ในการวัดผล

1. ความเร็วในการนำทางต่อวิดีโอ (วินาที)
2. ความแม่นยำ โดยคำนึงถึงอัตราส่วนร่วมของกรอบที่เหลือ

#### สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่า การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุและสร้างกรอบสี่เหลี่ยมทุกๆ N เฟรม แล้วใช้ระบบนำทางตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น จะทำให้ระบบสามารถทำงานได้เร็วขึ้น โดยที่ประสิทธิภาพจะลดลงเพียงเล็กน้อย

#### ตัวแปรควบคุม

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

#### วิธีการทดลอง

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

### 3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์

#### 3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายนอกในภาพสิ่งที่ใช้ในการวัดผล

- ความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอก

#### สมมุติฐาน

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

#### ตัวแปร

- โมเดลปัญญาประดิษฐ์ ซึ่งได้แก่
  - ResNet50 ของชุดข้อมูล Market1501
  - ResNet50 ของชุดข้อมูล DukeMTMCReID
  - ResNet50 ของชุดข้อมูล CUHK03
  - ResNet50 ของชุดข้อมูล MSMT17

#### ตัวแปรควบคุม

- ชุดข้อมูล : ชุดข้อมูลที่ทางผู้จัดสร้างขึ้นสำหรับการทดสอบ
- โมเดลปัญญาประดิษฐ์ : YOLO-V3 320 สำหรับการหาตำแหน่งของบุคคล

#### วิธีการทดลอง

- ดาวน์โหลดโมเดลปัญญาประดิษฐ์ที่ถูกสร้างด้วยชุดข้อมูลต่างๆ ได้แก่ Market1501, DukeMTMCReID, CUHK03 และ MSMT17
- นำชุดข้อมูลที่ผู้จัดสร้างขึ้นมาผ่านโมเดลปัญญาประดิษฐ์ YOLO-V3 320 เพื่อหาตำแหน่งของบุคคล
- นำโมเดลปัญญาประดิษฐ์แต่ละอันมาทดสอบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอกในภาพ ด้วยตำแหน่งของบุคคลที่ได้มาจากการบันทึกหน้าบันทึกหน้า
- ประเมินผลการทำงานโดยเทียบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอกของแต่ละโมเดล ปัญญาประดิษฐ์ เพื่อหาโมเดลปัญญาประดิษฐ์ที่ได้ผลลัพธ์ดีที่สุด

### 3.12 การออกแบบการทดสอบการจดจำการกระทำของมนุษย์

3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน 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 ตำแหน่งของกรอบสีเหลือง และรหัสของการกระทำ
3. เรียกชื่อของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่เคยถูกฝึกอบรม จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
  - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ เพริม ตำแหน่งของกรอบสีเหลือง รหัสของการกระทำ และความมั่นใจ
4. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
5. เปรียบเทียบผลลัพธ์กับแหล่งที่มา

3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกสร้างด้วย AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

สิ่งที่ใช้ในการวัดผล

1. ความในการทำนายเร็วต่อรูปภาพ (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 ต่อไปนี้

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

#### สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่าผลลัพธ์ของการทดลองจะมีความแม่นยำสูงขึ้นเมื่อเทียบกับความแม่นยำของการทดลองที่ผ่านมา เนื่องจากโมเดลปัญญาประดิษฐ์ในการทดลองนี้ เป็นโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยได้สร้างขึ้น ซึ่งจะมีหมวดหมู่ของการกระทำของโมเดลปัญญาประดิษฐ์และชุดข้อมูลทดสอบตรงกัน ในส่วนของความเร็วต่อรูปภาพจะมีความเร็วอีกกว่าผลลัพธ์จากแหล่งอ้างอิง 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 และนำโมเดลปัญญาประดิษฐ์ทำนายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำนาย
  - (a) ชุดข้อมูลผลลัพธ์จากการทำนาย ประกอบด้วย : ชื่อของวิดีโอ เพرم ตำแหน่งของกรอบสีเหลี่ยม รหัสของการกระทำ และความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำนาย และ ชุดข้อมูลที่มีคำกำกับเพื่อเป็นคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

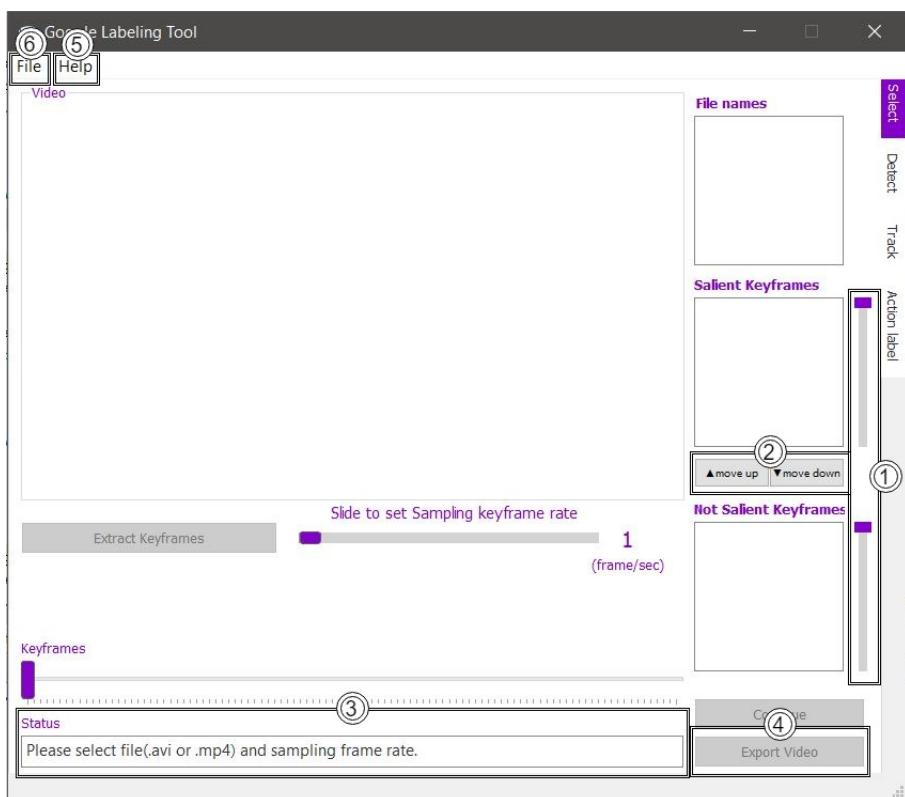
## บทที่ 4

### ผลการดำเนินงาน

#### 4.1 เครื่องมือกำกับคุณลักษณะ

##### 4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน

หน้าต่าง Select

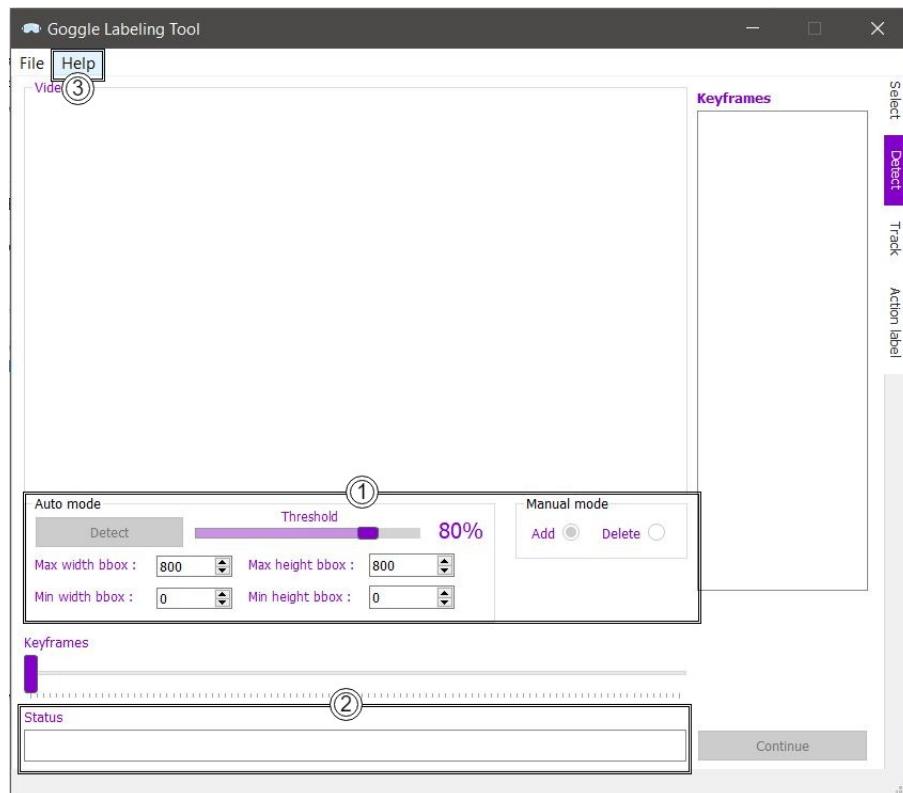


รูปที่ 4.1: รูปหน้าต่างแสดงผลของหน้าต่าง Select

จากรูปที่ 4.1 แสดงหน้าต่าง Select ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับรูปที่ 3.3 จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. แถบเลื่อนสำหรับเลื่อนคูเพรมที่มีมนุษย์หรือไม่มีมนุษย์ เพื่อเพิ่มความสะดวกในการเลือกคูเพรม
2. ปุ่มสำหรับแก้ไขเพรมที่มีมนุษย์หรือไม่มีมนุษย์
3. แถบแสดงสถานะกระบวนการทำงาน
4. ปุ่มสำหรับนำผลลัพธ์ออกเป็นไฟล์วิดีโอด้วยไฟล์ในช่วงที่มีมนุษย์อยู่
5. แถบสำหรับคำแนะนำช่วยเหลือ
6. ปุ่มสำหรับเปิดไฟล์

### หน้าต่าง Detect

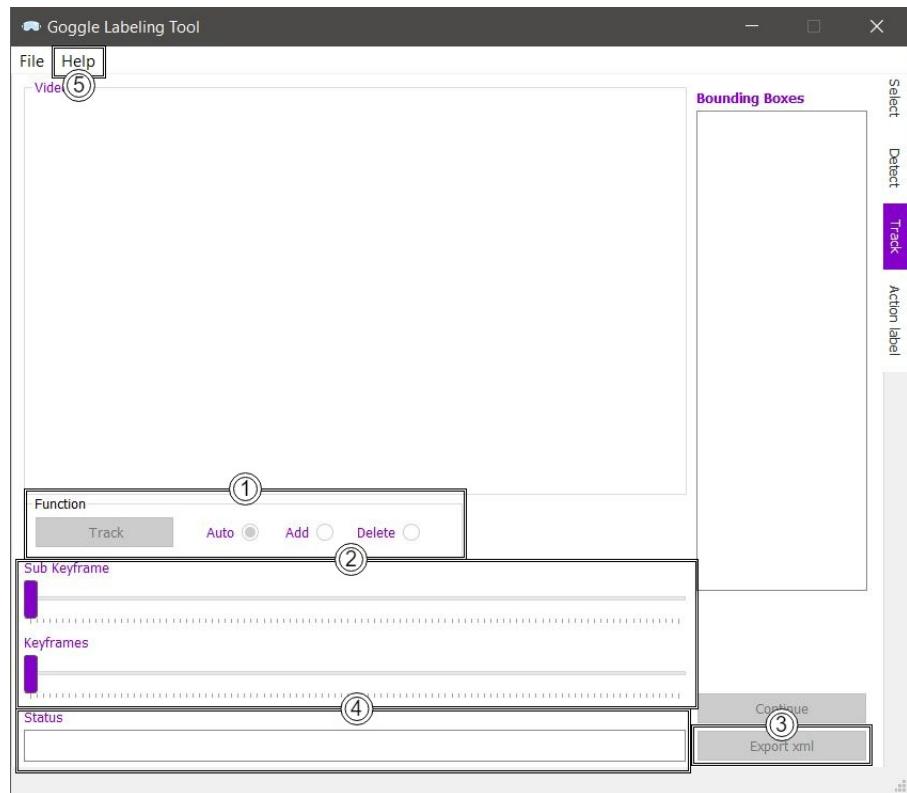


รูปที่ 4.2: รูปหน้าต่างแสดงผลของหน้าต่าง Detect

จากรูปที่ 4.2 แสดงหน้าต่าง Detect ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับรูปที่ (3.5) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่ด้วยการทำงานแบบอัตโนมัติ และกำหนดเองสามารถใช้งานได้สะดวกขึ้น และเพิ่มความหลากหลายในการปรับแก้ในการทำงานอัตโนมัติ
2. แถบแสดงสถานะกระบวนการทำงาน
3. แถบสำหรับคำแนะนำช่วยเหลือ

### หน้าต่าง Track

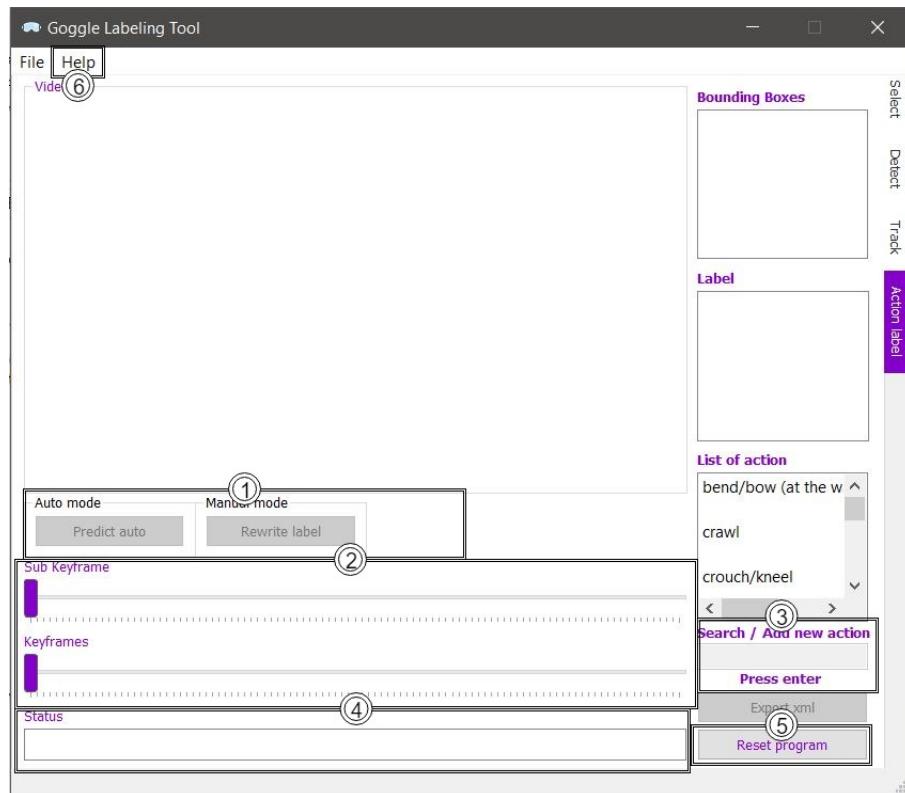


รูปที่ 4.3: รูปหน้าต่างแสดงผลของหน้าต่าง Track

จากรูปที่ 4.3 แสดงหน้าต่าง Track ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ (3.7) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແລບເລືອນເປັນ 2 ແລບເລືອນທຳໃຫ້ສາມາຄຸດຄີ່ຍິ່ງແຮມແລະແຮມທີ່ອຸ່ຽ່ງຫວ່າງໜຶ່ງຄີ່ຍິ່ງແຮມໄດ້ສະດວກขື້ນ
3. เพิ่ມປຸມສໍາຫຼັບນຳພລັດພົບອອກເປັນໄຟລ് xml
4. ແຕບແສດງສະຖານະກະບວນການທຳຈານ
5. ແຕບສໍາຫຼັບຄໍາແນະນຳໜ່ວຍເຫຼືອ

### หน้าต่าง Label



รูปที่ 4.4: รูปหน้าต่างแสดงผลของหน้าต่าง Label

จากรูปที่ 4.4 แสดงหน้าต่าง Label ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับกับฉบับร่างตามรูปที่ (3.9) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແຕບເລື່ອນປັນ 2 ແຕບເລື່ອນທຳໃຫ້ສາມາຄດຸກີຍິ່ພຣມແລະພຣມທີ່ອຸ່ຽ່ຮ່ວ່າງຊ່ວງຄີຍິ່ພຣມໄດ້ສະດວກຂຶ້ນ
3. ເຄື່ອງມືສໍາຫຼັບຄັນຫາຫຼືເພີ່ມໜ້ວດໜູ່ຂອງການກະທຳ
4. ແຕບແສດງສະຖານະກະບວນການກະທຳ
5. ປຸ່ມສໍາຫຼັບເຮີມຕັ້ນການກະທຳໃໝ່
6. ແຕບສໍາຫຼັບຄຳແນະນຳໜ້ວຍເຫຼືອ

#### 4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอปพลิเคชัน

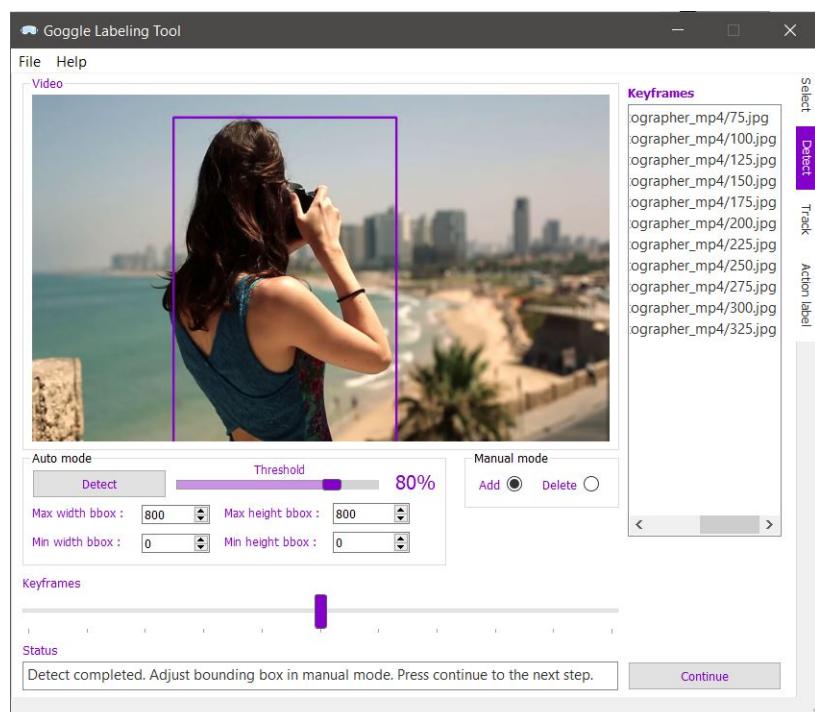
##### ผลลัพธ์การทำงานของหน้าต่าง Select



รูปที่ 4.5: รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ภายในเฟรม

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

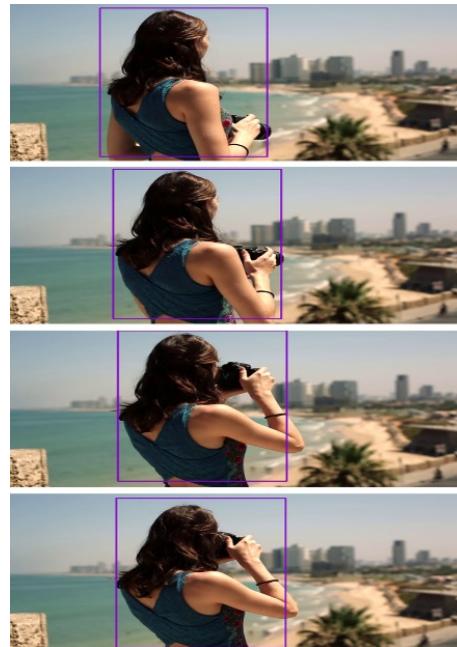
##### ผลลัพธ์การทำงานของหน้าต่าง Detect



รูปที่ 4.6: รูปคีย์เฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่

แอ��พลิเคชันจะนำคีย์เฟรมที่มีนุชย์ที่ได้จากหน้าต่าง Select นำมาตีกรอบสี่เหลี่ยมในส่วนของเฟรมที่มีมนุชย์อยู่โดยสามารถใช้โหมดการทำงานแบบบอตโนมัติหรือแบบแก้ไขเองก็ได้ 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:  $1280 \times 720$  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 ผลการทดสอบระบบติดตามตำแหน่งของมนุษย์

#### 4.3.1 ข้อมูลรายละเอียดประกอบการทดสอบ

ชื่อวิดีโอ: Photographer beach photography

ความยาววิดีโอ: 15 วินาที

จำนวนเฟรมทั้งหมด: 374 เฟรม

อัตราเฟรมต่อวินาที: 24.9 เฟรมต่อวินาที

ความละเอียดของวิดีโอ: 1920 ×1080

ความละเอียดของวิดีโอที่ใช้ในการประมวลผลจริง: 1280 ×720

ขอบเขตอัตราส่วนร่วมของกรอบที่เหลืออยู่ที่จะนับว่าการทำนายถูกต้อง: 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 แม้จะตั้งขอบเขตอัตราส่วนร่วมของกรอบที่เหลืออยู่ที่จะนับว่าการทำนายถูกต้องสูงถึง 80% แต่ความแม่นยำสูงถึง 95% ใช้เวลาในการประมวลผล 452 วินาที เฉลี่ยเฟรมละ 1.2 วินาที ซึ่งถือเป็นความแม่นยำที่สูงมากเมื่อเทียบกับเวลาที่ใช้ในการประมวลผล

ต่อมาเป็นการทดสอบโดยใช้โมเดลปัญญาประดิษฐ์ประมวลผลเฉพาะบางเฟรมทุกๆ ช่วงหนึ่ง 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 ผลการทดสอบระบบบุคคลตัวตนของมนุษย์

### 4.4.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการระบุตัวตนของบุคคล

ความแม่นยำของโมเดลปัญญาประดิษฐ์จากแหล่งที่มีมาค่าดั้งตารางด้านล่างดังนี้

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

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

#### 4.5 ผลการทดสอบการจดจำการกระทำของมนุษย์

4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้

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

ตารางที่ 4.10: ผลการทดสอบความแม่นยำของโมเดลปัญญาประดิษฐ์เทียบผลลัพธ์กับแหล่งอ้างอิง

ความเร็วของตอรูปภาพทางผู้วิจัยได้ใช้กราฟฟิกการ์ด GTX 2080 Ti ใน การทดสอบซึ่งจะให้ความเร็วอยู่ที่ 5 วินาที ซึ่งทางแหล่งอ้างอิงนั้นใช้กราฟฟิกการ์ด Nvidia GeForce GTX TITAN X ในส่วนของค่าความแม่นยำที่ไม่เท่ากัน คาดว่าจะเป็นเพราะการประมวลผลของกราฟฟิกการ์ดของรุ่นที่ต่างกันจึงทำให้ค่า mAP ที่ออกมามีเมื่อเท่ากัน

4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่านชุดข้อมูลสำหรับการเทรนด์ที่ผู้วิจัยสร้างขึ้น 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

## ເອກສາຮອ້າງອີງ

- [1] Optical flow.
- [2] Ahmad Ali and Sikander Majid Mirza. Object tracking using correlation, kalman filter and fast means shift algorithms. In 2006 International Conference on Emerging Technologies, pages 174–178. IEEE, 2006.
- [3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In Proceedings of the 17th International Conference on Computational Linguistics - Volume 1, COLING '98, pages 86–90, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 961–970, 2015.
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.
- [6] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.
- [7] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis, pages 363–370. Springer, 2003.
- [8] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. A video dataset of spatio-temporally localized atomic visual actions, 2017.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
- [10] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. Neural computation, 6(2):181–214, 1994.
- [11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

- [12] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. arXiv preprint arXiv:1806.11230, 2018.
- [13] Xiaobin Liu, Shiliang Zhang, Tiejun Huang, and Qi Tian. E <sup>A</sup>2 bows: An end-to-end bag-of-words model via deep convolutional neural network. arXiv preprint arXiv:1709.05903, 2017.
- [14] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. Pattern Recognition, 94:53–61, 2019.
- [15] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. Pattern Recognition, 68:334–345, 2017.
- [16] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Yan Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence, 2019.
- [17] Karin Kipper Schuler. Verbnet: A Broad-coverage, Comprehensive Verb Lexicon. PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.
- [18] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.
- [19] Du Tran, Lubomir D Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3d: generic features for video analysis. CoRR, abs/1412.0767, 2(7):8, 2014.
- [20] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4694–4702, 2015.
- [21] Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. The revised arabic propbank. In Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV ’10, pages 222–226, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

ภาคผนวก

## ภาคผนวก ก

### ตัวอย่างชุดข้อมูลที่ผู้จัดสร้างขึ้น

ตัวอย่างชุดข้อมูลสำหรับการทดสอบโมเดลปัญญาประดิษฐ์ในการตรวจจับภาพบุคคล



รูปที่ ก.1: รูปผลลัพธ์การทำงานของหน้าต่าง Track