



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 เกษมสันติธรรม)

ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

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

บทคัดย่อ

งานวิทยานิพนธ์นี้เป็นงานที่เกี่ยวกับการออกแบบและจัดทำแอพพลิเคชัน labeling tool และระบบบุคลากรที่การกระทำของมนุษย์ โดยใช้ชื่อว่า Goggle แพลตฟอร์มการเรียนรู้เชิงลึกและระบบบุคลากรที่การกระทำของมนุษย์ ซึ่งมีจุดประสงค์เพื่อให้ผู้พัฒนาสามารถใช้งาน labeling tool ในการสร้างชุดข้อมูลสำหรับสร้างปัญญาประดิษฐ์ได้ง่ายและสะดวกขึ้น ภาพรวมของวิทยานิพนธ์นี้จะแบ่งออกเป็นทั้งหมดสองส่วน คือ ส่วนแรกเป็นส่วนของการออกแบบและสร้างแอพพลิเคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการ tren โนเมเดลจากวิดีโอ และส่วนที่สอง เป็นส่วนของการออกแบบและสร้างระบบบุคลากรที่การกระทำของมนุษย์ได้ในขอบเขตที่กำหนดไว้ในบทนำ

คำสำคัญ : ระบบบุคลากรที่การกระทำของมนุษย์ / labeling tool / 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	ค
กิตติกรรมประกาศ	๑
สารบัญ	๒
รายการรูปภาพ	๗
รายการตาราง	๘
รายการสัญลักษณ์	๙
ประมวลศัพท์และตัวย่อ	๑๐
บทที่ ๑ บทนำ	๑
1.1 ที่มาและความสำคัญ	๑
1.2 วัตถุประสงค์	๑
1.3 ประโยชน์ที่คาดว่าจะได้รับ	๑
1.4 ขอบเขตการดำเนินงาน	๒
1.5 ขั้นตอนการดำเนินงาน	๒
บทที่ ๒ ทฤษฎี/การวิจัยที่เกี่ยวข้อง	๔
2.1 การวิเคราะห์ผลวิดีโอ	๔
2.1.1 การตรวจจับวัตถุ	๔
2.1.2 การนำทางตำแหน่งถัดไปของวัตถุ	๗
2.1.3 การระบุตัวตนของบุคคล	๘
2.1.4 การจำแนกการกระทำ	๘
2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ	๑๗
2.2.1 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมุขย์	๑๗
2.2.2 เครื่องมือสำหรับสร้างชุดข้อมูล	๒๐
2.3 ทฤษฎีที่เกี่ยวข้อง	๒๒
2.3.1 Optical flow	๒๒
บทที่ ๓ ระเบียบวิธีวิจัย	๒๔
3.1 ความต้องการของระบบ	๒๔
3.1.1 ความต้องการเชิงการใช้งาน (Functional Requirements)	๒๔

สารบัญ (ต่อ)

เรื่อง	หน้า
3.1.2 ความต้องการเชิงวิศวกรรม (Non-Functional Requirements).....	24
3.2 หน้าที่ความรับผิดชอบ.....	25
3.3 เครื่องมือที่ใช้ในงานวิจัย	25
3.4 ภาษาที่ใช้ในการพัฒนาระบบ	26
3.5 Program library ที่ใช้ในการพัฒนาระบบและแอพพลิเคชัน.....	26
3.6 แผนการดำเนินงาน	26
3.7 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	27
3.8 การออกแบบหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	28
3.8.1 เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์.....	28
3.9 การออกแบบการทดสอบการตรวจจับวัตถุ.....	38
3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล	38
3.10 การออกแบบการทดสอบการทำงานของระบบทำงานตามกำหนดต่อไปของมนุษย์.....	39
3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบทำงานตามกำหนดต่อไปของวัตถุในวิดีโอ	39
3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์	40
3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายในภาพ	40
3.12 การออกแบบการทดสอบการจดจำการกระทำการของมนุษย์	41
3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง	41
3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกトレน์ผ่านชุด AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง	42
3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกトレน์ผ่านชุด ข้อมูลสำหรับการเทรน์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบ และเทียบผลลัพธ์กับแหล่งอ้างอิง	43
บทที่ 4 ผลการดำเนินงาน.....	44
4.1 Labeling tool	44
4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน	44
4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอพพลิเคชัน.....	48

สารบัญ (ต่อ)

เรื่อง	หน้า
4.2 ผลการทดลองการตรวจจับวัตถุ	51
4.2.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล	51
4.2.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล	51
4.3 ผลการทดสอบการทำนายตำแหน่งต่อไปของมนุษย์.....	51
4.4 ผลการทดสอบการระบุตัวตนของมนุษย์.....	51
4.5 ผลการทดสอบการจำจารกรรมทำของมนุษย์.....	51
4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้.....	51
4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่าน AVA และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง...	51
4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกトレน์ผ่านชุดข้อมูลสำหรับการเทรน์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์การทดสอบก่อนหน้า	52
เอกสารอ้างอิง.....	54

รายการรูปภาพ

รูป	หน้า
รูปที่ 2.1 แนวคิดของระบบทำนายตำแหน่งตัวแทน่ตัดไปของวัตถุ	7
รูปที่ 2.2 ResNet	17
รูปที่ 2.3 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[3]	18
รูปที่ 2.4 โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D ^[3]	19
รูปที่ 2.5 UI ของโปรแกรม DarkLabel	20
รูปที่ 2.6 UI ของโปรแกรม OpenLabeling	21
รูปที่ 2.7 ตัวอย่างการเคลื่อนที่ของกลุ่มบล็อก	22
รูปที่ 3.1 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	27
รูปที่ 3.2 กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	28
รูปที่ 3.3 หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	29
รูปที่ 3.4 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select	30
รูปที่ 3.5 หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	31
รูปที่ 3.6 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect	32
รูปที่ 3.7 หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	33
รูปที่ 3.8 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track	34
รูปที่ 3.9 หน้าต่าง Action label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์	35
รูปที่ 3.10 ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Action label	36
รูปที่ 3.11 ตัวอย่างข้อมูลภายในไฟล์ XML	37
รูปที่ 4.1 รูปหน้าต่างแสดงผลของหน้าต่าง Select	44
รูปที่ 4.2 รูปหน้าต่างแสดงผลของหน้าต่าง Detect	45
รูปที่ 4.3 รูปหน้าต่างแสดงผลของหน้าต่าง Track	46
รูปที่ 4.4 รูปหน้าต่างแสดงผลของหน้าต่าง Label	47
รูปที่ 4.5 รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่และไม่มีมนุษย์อยู่ภายในเฟรม	48
รูปที่ 4.6 รูปคิริ์เฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่	48
รูปที่ 4.7 รูปผลลัพธ์การทำงานของหน้าต่าง Track	49
รูปที่ 4.8 รูปผลลัพธ์การทำงานของหน้าต่าง Label	50

รายการตาราง

ตาราง	หน้า
ตารางที่ 1.1 แผนการดำเนินงาน	3
ตารางที่ 2.1 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M.....	10
ตารางที่ 2.2 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M.....	11
ตารางที่ 2.3 ผลการทดสอบโมเดลต่างๆ บนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M.....	11
ตารางที่ 2.4 ผลการทดลองของวิธีต่างๆ บน Frame Level.....	13
ตารางที่ 2.5 Data transfer performance ของโมเดล Resnet50 I3D.....	15
ตารางที่ 2.6 Top-1 ของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet	17
ตารางที่ 2.7 ค่าความผิดพลาดที่ได้จากการทดลองจำนวนขั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10	18
ตารางที่ 2.8 ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อใช้ข้อมูลจาก UCF-101, HMDB-51 และ Kinetics ในการสร้างและทดสอบด้วยเครื่องมือวัดผลแบบความแม่นยำจากการทำนายอันดับแรกสุด	19

รายการสัญลักษณ์

θ	เซ็ต้า
d	distance
kg	Kilogram
m^2	Square Metre

ประมวลศัพท์และตัวย่อ

AVA	Atomic Visual Actions
Machine learning	การเรียนรู้ของเครื่อง
Artificial intelligence	ปัญญาประดิษฐ์
Label	คำอธิบายที่บ่งบอกถึงคุณลักษณะของสิ่งที่เราสนใจ
Labeling	การสร้างคำอธิบายคุณลักษณะ
Action recognition	การจดจำการกระทำ
Video labeling	การสร้างคำอธิบายคุณลักษณะภายในวิดีโอ
Video analytics	การวิเคราะห์ผลวิดีโอ
Video analytics platform	ระบบปฏิบัติการสำหรับช่วยวิเคราะห์ผลวิดีโอ
Goggle labeling tool	เครื่องมือของโครงการ Goggle สำหรับช่วยสร้างคำอธิบายที่บ่งบอกถึงคุณลักษณะของสิ่งที่เราสนใจ
Uniform label distribution	การที่มีจำนวนตัวอย่างภายในคลาสเท่ากันทุกคลาส
KMUTT	King Mongkut's University of Technology Thonburi

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

บริษัท เพอเซปตรา ดำเนินธุรกิจเกี่ยวกับการให้บริการและคำปรึกษาเกี่ยวกับปัญญาประดิษฐ์ (artificial intelligence) เนื่องจากปัจจุบันนี้ความสามารถและประสิทธิภาพของปัญญาประดิษฐ์มีความก้าวหน้าขึ้นจนสามารถก้าวข้ามความสามารถของมนุษย์ในงานหลายประเภท ทำให้ลูกค้าต้องมีความต้องการที่จะให้ทางบริษัทสร้างปัญญาประดิษฐ์เพื่อนำไปใช้งานหรือแก้ปัญหาที่ต่างกันออกไป เช่น ใช้ปัญญาประดิษฐ์มาช่วยประมวลผลภาพจากกล้องวงจรปิด เพื่อหาบุคคลที่มีท่าทางน่าสงสัย เป็นต้น ซึ่งการจะสร้างปัญญาประดิษฐ์ที่เหมาะสมกับการแก้ปัญหาเหล่านี้ จำเป็นต้องมีชุดข้อมูล (dataset) ที่เหมาะสม บางครั้งอาจต้องใช้มนุษย์ในการสร้างขึ้นมาโดยการเก็บข้อมูลวิดีโอ และลงมือสร้างชุดข้อมูลจากวิดีโอที่ได้ด้วยตัวเอง ซึ่งหนึ่งในปัจจัยสำคัญในการพัฒนาโมเดลปัญญาประดิษฐ์ให้มีประสิทธิภาพสูงคือจำนวนข้อมูล ซึ่งหากมีจำนวนวิดีโอเป็นจำนวนมาก การใช้มนุษย์ในการสร้างชุดข้อมูลนั้นอาจจะต้องใช้มนุษย์เป็นจำนวนมาก และใช้เวลานาน

ทางคณะผู้วิจัยจึงมีความต้องการที่จะออกแบบและพัฒนาระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ (A.I.-assisted video labeling tool) สำหรับสร้างชุดข้อมูลจากวิดีโอ เพื่อช่วยแบ่งเบาภาระของผู้พัฒนาในการสร้างชุดข้อมูลสำหรับการพัฒนาโมเดลปัญญาประดิษฐ์ในการใช้แก้ปัญหาที่ลูกค้าต้องการ โดยโครงการสหกิจนี้เน้นศึกษาเกี่ยวกับการวิเคราะห์และจัดการกระทำของมนุษย์ภายในสำนักงานจากภาพเคลื่อนไหวเป็นหลัก

1.2 วัตถุประสงค์

- เพื่อพัฒนาระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ ที่ทำให้มนุษย์กับปัญญาประดิษฐ์ทำงานร่วมกันเพื่อสร้างชุดข้อมูลในการนำมาพัฒนาปัญญาประดิษฐ์อื่นๆ ที่เหมาะสมกับปัญหาที่ต้องการ
- เพื่อออกแบบและสร้างระบบต้นแบบวิเคราะห์วิดีโอด้วยความสามารถตรวจจับมนุษย์และจำแนกการกระทำพื้นฐานของมนุษย์ภายในสำนักงาน ประกอบด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว พูดคุย นอน โดยใช้ปัญญาประดิษฐ์
- เพื่อพัฒนาเครื่องมือที่สามารถสร้างชุดข้อมูลสำหรับการจัดการกระทำของมนุษย์ให้สามารถใช้งานได้ง่าย สะดวกสบายมากขึ้น และมีประสิทธิภาพที่สูงกว่าเครื่องมือตัวอื่นในปัจจุบัน

1.3 ประโยชน์ที่คาดว่าจะได้รับ

- เพิ่มความสะดวกในการสร้างชุดข้อมูลสำหรับพัฒนาโมเดลปัญญาประดิษฐ์จากวิดีโอ
- ต้นแบบระบบวิเคราะห์วิดีโอด้วยความสามารถจำแนกการกระทำของมนุษย์ภายในสำนักงานได้

1.4 ขอบเขตการดำเนินงาน

1. สร้างระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ โดยระบบจะประกอบไปด้วยสีส่วนต่อไปนี้
 - (a) หน้าต่างของแอพพลิเคชัน (user interface)
 - (b) ระบบตรวจจับมนุษย์ในภาพ
 - (c) ระบบท่านายตำแหน่งต่อไปของมนุษย์ในภาพเคลื่อนไหว
 - (d) ระบบจำแนกการกระทำของมนุษย์ ซึ่งประกอบไปด้วย ยืน นั่ง เดิน เล่นโทรศัพท์ กินข้าว พูดคุย นอน
2. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์กับชุดข้อมูลที่ได้จากเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ เพื่อที่จะทดสอบว่าชุดข้อมูลที่ได้สามารถใช้งานจริงได้หรือไม่
3. พัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์ภายในสำนักงานอย่างน้อย 2 โมเดล

1.5 ขั้นตอนการดำเนินงาน

การดำเนินงานวิจัยถูกแบ่งออกเป็นสามส่วน โดยส่วนแรกคือการศึกษาเทคโนโลยีในปัจจุบันเพื่อหาความเป็นไปได้และกำหนดขอบเขตของงาน ส่วนที่สองคือเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ เป็นส่วนที่ออกแบบและพัฒนาเครื่องมือสำหรับช่วยผู้พัฒนาในการสร้างชุดข้อมูล และส่วนที่สุดท้ายคือการนำชุดข้อมูลที่ได้จากการใช้เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ไปพัฒนาโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ภายในสำนักงาน

ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง

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

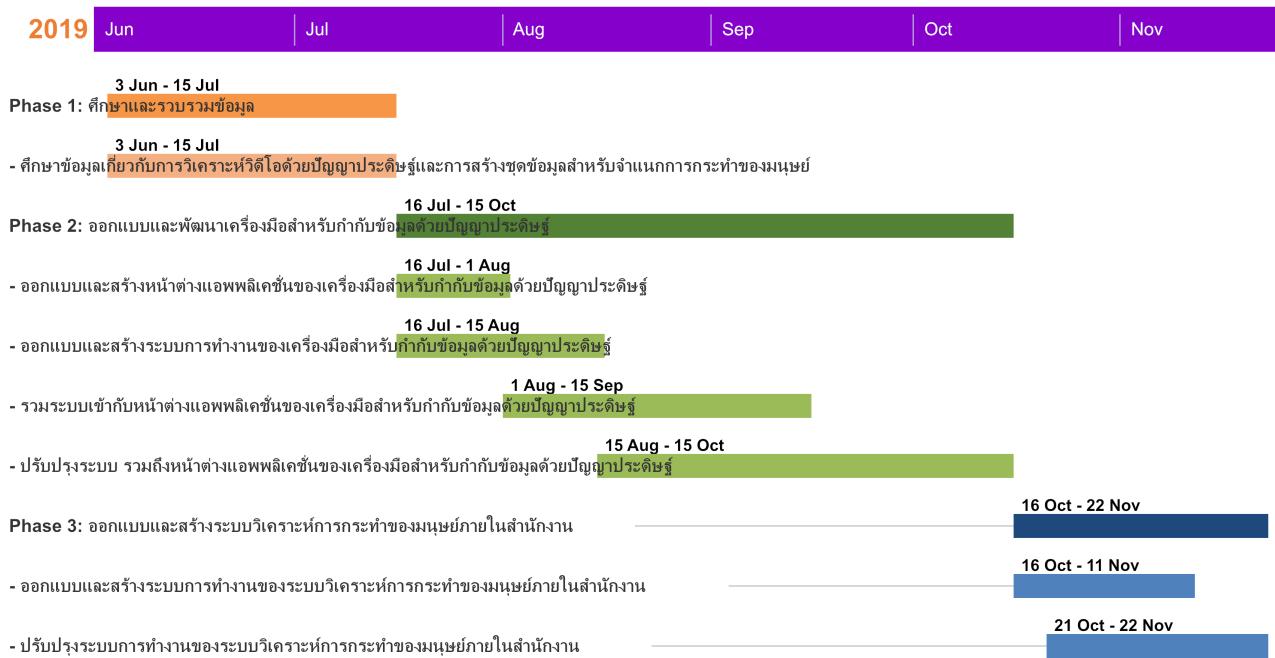
เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

1. ออกรูปแบบและสร้างหน้าต่างแอพพลิเคชันของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์
2. ออกรูปแบบและสร้างระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์
3. ทดสอบและปรับปรุงการทำงานของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

โมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ภายในสำนักงาน

1. สร้างชุดข้อมูลสำหรับสร้างโมเดลปัญญาประดิษฐ์จากเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์
2. สร้างโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ภายในสำนักงาน
3. ทดสอบโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ภายในสำนักงาน

แผนการดำเนินงาน



ตารางที่ 1.1: แผนการดำเนินงาน

บทที่ 2

ทฤษฎี/การวิจัยที่เกี่ยวข้อง

การวิเคราะห์วิดีโoinปัจจุบันนั้นมีวิธีและเทคนิคมากมาย ผู้วิจัยจึงต้องศึกษาองค์ความรู้และงานวิจัยที่เกี่ยวข้องกับวัตถุประสงค์ของงาน เพื่อศึกษาและใช้เป็นแนวทางในการประยุกต์สำหรับสร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ และโมเดลปัญญาประดิษฐ์สำหรับการจำแนกการกระทำของมนุษย์ ซึ่งหัวข้อที่ผู้วิจัยได้ไปศึกษามา มีดังต่อไปนี้

1. การวิเคราะห์ผลวิดีโอ

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

2. เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

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

3. ทฤษฎีที่เกี่ยวข้อง

- (a) Optical flow

2.1 การวิเคราะห์ผลวิดีโอ

ในส่วนของงานวิจัยสิ่งที่เราให้ความสนใจ คือ ข้อมูลการกระทำการของมนุษย์แต่ละคนภายในวิดีโอ เพื่อที่เราจะได้ผลลัพธ์ที่มีประสิทธิภาพมากเป็นข้อมูลของสิ่งที่เราสนใจ เช่น จำนวนคนที่เดินผ่านกล้อง หรือทิศทางการเดินของคนในวิดีโอ เราจึงจำเป็นต้องใช้การวิเคราะห์ผลวิดีโอเพื่อที่จะสกัดสิ่งที่เราสนใจออกมาจากวิดีโอด้วย การวิเคราะห์ผลวิดีโอมีหลากหลายกระบวนการ โดยในแต่ละกระบวนการจะมีจุดประสงค์ของการทำและผลลัพธ์หลังการประมวลผลที่แตกต่างกัน ในหัวข้อนี้จะมาอธิบายถึงกระบวนการในการวิเคราะห์ผลของวิดีโอและผลลัพธ์ของกระบวนการนั้น

«««< HEAD

2.1.1 การตรวจจับวัตถุ

การตรวจจับวัตถุนั้นเป็นกระบวนการที่นิยมใช้ในการวิเคราะห์ผลของวิดีโอ กล่าวคือกระบวนการที่ผู้วิจัยจะต้องทำคือระบุสิ่งที่สนใจว่าอยู่ตำแหน่งใด ซึ่งในปัจจุบันการทำการตรวจจับวัตถุมักนำปัญญาประดิษฐ์มาใช้เนื่องจากมีประสิทธิภาพสูง ซึ่งโมเดลปัญญาประดิษฐ์ที่เราเลือกใช้คือ YOLO v3 โดยเหตุผลที่เราเลือกใช้ YOLO v3 เนื่องจากสามารถตรวจจับวัตถุได้อย่างรวดเร็วและแม่นยำ โดยหลักการของ YOLO v3 คือ นำรูปภาพที่ต้องการผ่านโครงข่ายประสาทเทียม (neural network) โดยโครงข่ายจะแบ่งรูปภาพเป็นพื้นที่หลายๆ ส่วน และทำงานความน่าจะเป็นของวัตถุในพื้นที่ว่าเป็นวัตถุใด สุดท้ายจะเลือกรอบสี่เหลี่ยมที่มีค่าคะแนนน่าจะเป็นมากที่สุด =====

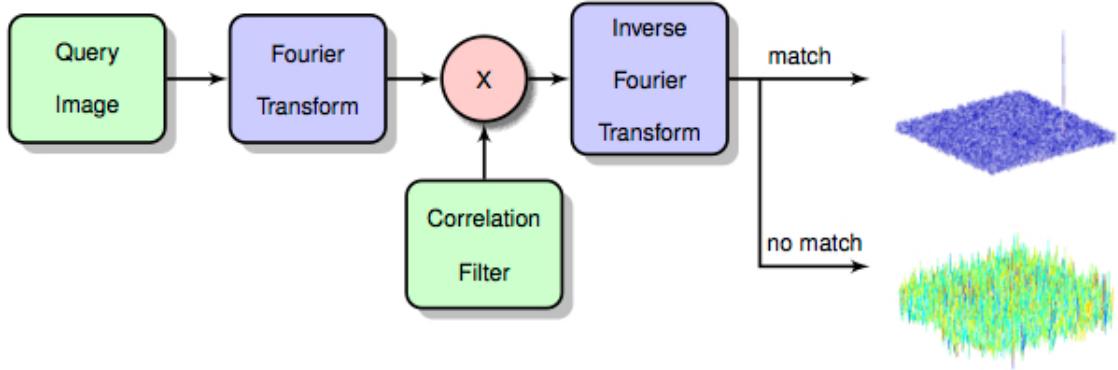
2.1.1 การตรวจจับวัตถุ

การตรวจจับวัตถุเป็นสิ่งที่สำคัญเป็นอันดับต้นๆของการวิเคราะห์ผลของวิดีโอ คือ การตรวจจับวัตถุ กล่าวคือกระบวนการที่ผู้วิจัยจะต้องทำการคำนวณว่ามีอะไรอยู่ในภาพ ทำให้เราสามารถนำ Machine learning model มาใช้เพื่อช่วยตรวจจับวัตถุที่เราสนใจ ซึ่ง Machine learning model ที่เราเลือกใช้คือ YOLO v3 โดยเหตุผลที่เราเลือกใช้ Machine learning model YOLO v3 จะถูกกล่าวไว้อยู่ในหัวข้อ Machine learning model ในหัวข้อถัดไป

YOLO v3 เป็น Machine learning model ที่ในปัจจุบันนิยมนำมาใช้ตรวจจับวัตถุในงานวิเคราะห์ผลของวิดีโอ เนื่องจากสามารถตรวจจับวัตถุได้แบบเรียลไทม์ และมีความแม่นยำ โดยหลักการของ YOLO v3 คือ นำรูปภาพที่ต้องการตรวจจับตัวแทนของวัตถุผ่าน neural network โดยโครงข่ายจะแบ่งรูปภาพเป็นพื้นที่ และจะทำนายกรอบสี่เหลี่ยมพร้อมกับทำนายความน่าจะเป็นของแต่ละหมวดหมู่ในแต่ละพื้นที่ สุดท้ายจะเลือกรอบสี่เหลี่ยมและหมวดหมู่ที่มีค่าความน่าจะเป็นมากที่สุด »»> 9c8c7a53f798b96980c97b09600822b759aa8362

2.1.2 การทำนายตำแหน่งถัดไปของวัตถุ

การทำนายตำแหน่งถัดไปของวัตถุ^[4] คือระบบที่ใช้สำหรับการติดตามการเคลื่อนไหวของวัตถุที่สนใจที่อยู่ในรูปภาพ โดยใช้การคำนวณทางคณิตศาสตร์ และการประมวลผลภาพ (image processing) ทำให้การประมวลผลนั้นเร็วมากขึ้น



รูปที่ 2.1: แนวคิดของระบบทำนายตำแหน่งถัดไปของวัตถุ

จากรูปที่ 2.1 จะเป็นหลักการในการทำนายตำแหน่งต่อไป โดยการนำรูปมาผ่านกระบวนการแปลงฟูรีเยร์ (fourier transform) และนำมามุ่งกับ correlation filter ซึ่งเป็นตัวรองที่ใช้สำหรับการหาความสัมพันธ์กับวัตถุในภาพ จากนั้นทำการแปลงฟูรีเยร์กลับ (inverse fourier transform) เพื่อตรวจสอบว่าวัตถุในภาพนั้นอยู่ที่ตำแหน่งใด โดยมีการคำนวณเริ่มจากการหา correlation filter ที่ดีที่สุดโดยใช้วิธีลดผลรวมของข้อผิดพลาดกำลังสองให้อยู่ที่สุดดังนี้

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

โดยที่

ϵ = ค่าความคลาดเคลื่อน

d = จำนวนมิติของผังคุณลักษณะ (feature map) ของภาพ

h = correlation filter

\star = circular correlation

f = พื้นที่สีเหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

g = ผลลัพธ์ correlation ที่ต้องการของ f

λ = regularization term

เมื่อพิจารณาจากรูปภาพเดี่ยวในกรณีที่เวลา (t) เท่ากับ 1 จะสามารถจัดรูปสมการด้านบนได้ดังนี้

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

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

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

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

โดยที่

H = correlation filter

η = จัตตราการเรียนรู้

\bar{G} = คือ g ที่ผ่านการทำ complex conjugation

F = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการทำผังคุณลักษณะ

\bar{F} = f ที่ผ่านการทำ complex conjugation

t = เวลา

จากสมการที่ได้มาจะสามารถทำให้หาตำแหน่งต่อไปของวัตถุที่สนใจได้ด้วยสมการต่อไปนี้

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

โดยที่

y = correlation score

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

Z = พื้นที่สี่เหลี่ยมของวัตถุที่สนใจที่ได้จากการหาผังคุณลักษณะของภาพใหม่

โดยค่าของ y ที่ได้ออกมาจะทำให้รู้ถึงตำแหน่งของวัตถุที่สนใจได้ ณ ตำแหน่งที่ y มีค่าสูงสุด

2.1.3 การระบุตัวตนของบุคคล

การระบุตัวตนของบุคคล คือการระบุตัวตนของบุคคลภายในวิดีโอหรือระหว่างรูปภาพ สามารถนำมาประยุกต์ใช้ในด้านของการรักษาความปลอดภัย การตามหาบุคคล หรือการตรวจสอบการกระทำการของบุคคลนั้นในวิดีโอด้วย

การระบุตัวตนของบุคคลนั้นเป็นปัญหาที่ท้าทาย เนื่องจากคุณลักษณะทั่วไปของบุคคลในรูปภาพไม่เพียงพอต่อการระบุบุคคลภายในภาพว่าเป็นบุคคลคนเดียวกันได้ ซึ่งวิธีการที่ใช้สำหรับการระบุตัวตนของบุคคล คือวิธีการที่เรียกว่า Dynamically Matching Local Information (DMLI) ที่สามารถจัดแนวรายละเอียดข้อมูลของภาพ และให้ประสิทธิภาพที่สูงอกรมา

การระบุตัวตนของบุคคล จะเริ่มจากการแบ่งภาพออกเป็น 8 ส่วนและใช้คุณลักษณะของภาพมาทำ normalize ซึ่งจะช่วยในการลดความซ้ำซ้อนของข้อมูล ต่อมาข้อมูลที่ทำการ normalize แล้วมาใช้เปรียบเทียบความแตกต่างของคุณลักษณะของรูป หลังจากนั้นหากค่าเฉลี่ยของความแตกต่างออกมาก ถ้าค่าที่ออกมากใกล้เคียงกับ 0 จะหมายถึงบุคคลในรูปทั้งสองเป็นบุคคลเดียวกัน

2.1.4 การจำแนกการกระทำ

การจำแนกการกระทำเป็นกระบวนการในการทำนายการกระทำการของมนุษย์หรือสิ่งที่สนใจอื่นๆที่เกิดการกระทำขึ้นภายในวิดีโอ โดยในหัวข้อนี้จะกล่าวถึงตั้งแต่ขั้นตอนการได้มาซึ่งชุดข้อมูลมีกระบวนการอย่างไร การนำ

โมเดลปัญญาประดิษฐ์มาใช้ในการจำแนกการกระทำ และการวัดผลของโมเดลปัญญาประดิษฐ์ โดยชุดข้อมูลที่ผู้วิจัยได้เลือกนำมาศึกษาจากชุดข้อมูลที่ถูกเป็นที่กล่าวถึงในปัจจุบัน และมีขนาดของชุดข้อมูลที่ใหญ่

จากบทความข้างต้นชุดข้อมูลที่เราได้เลือกนำมาใช้ได้แก่ YouTube-8M ,AVA ,Moment in Time โดยแต่ละชุดข้อมูลจะมีความแตกต่างกันในหลายๆด้าน แต่จะมีสิ่งที่เหมือนกัน คือ เป็นชุดข้อมูลสำหรับการวิเคราะห์ผลวิดีโอด้วยการสนับสนุนจากการกระทำของมนุษย์ โดยในบทความนี้จะกล่าวถึงความแตกต่างในด้านต่างๆ เช่น เป้าหมายของแต่ละชุดข้อมูล ,วิธีการเก็บข้อมูลสำหรับชุดข้อมูล ,วิธีการสร้างคำกำกับ และรายละเอียดของชุดข้อมูล จากนั้นจะสรุปข้อมูลของแต่ละชุดข้อมูล

YouTube-8M

1. ชุดข้อมูล

(a) เป้าหมายของชุดข้อมูล : ใช้ทำนายรีมของวิดีโอ

(b) จำนวนของวิดีโอ : 8,264,650 วิดีโอ

(c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 229.6 วินาที

(d) จำนวนของหมวดหมู่ของคำกำกับ : 4800 หมวดหมู่

(e) กฎในการรวบรวมข้อมูลดังนี้

i. ทุกๆหัวข้อต้องเป็นรูปธรรม

ii. ในแต่ละหัวข้อต้องมีจำนวนวิดีโอมากกว่า 200 วิดีโอ

iii. ความยาวของวิดีอottองอยู่ระหว่าง 120 - 500 วินาที

หลังจากได้กฎในการรวบรวมข้อมูลแล้ว ขั้นตอนต่อไปคือการสร้างคำศัพท์ที่ใช้ในการค้นหาข้อมูลวิดีโอด้วยใน YouTube

(f) ขั้นตอนในการสร้างคำศัพท์มีดังนี้

i. กำหนดรายการที่อนุญาตหัวข้อที่เป็นรูปธรรมมา 25 ชนิด เช่น เกมส์ เป็นต้น

ii. กำหนดบัญชีคำหัวข้อที่คิดว่าไม่เป็นรูปธรรมไว้ เช่น software เป็นต้น

iii. รวบรวมหัวข้อที่มีอยู่ในรายการที่อนุญาตอย่างน้อย 1 หัวข้อ และต้องไม่มีอยู่ในบัญชีคำซึ่งจะทำให้ได้หัวข้อที่ต้องการมาประมาณ 50,000 หัวข้อ

iv. จากนั้นใช้ผู้ประเมินจำนวน 3 คน ในการคัดหัวข้อที่คิดว่าเป็นรูปธรรม และสามารถจัดลำดับเรื่องเข้าใจได้ง่ายโดยไม่ต้องใช้ข้อมูลในด้านนั้นๆ ซึ่งผู้ประเมิน ก็จะมีความว่า “มันยกขานด้วยเห็นว่ามีหัวข้อดังกล่าวอยู่ในรูปหรือวิดีโอ โดยใช้เพียงแค่การมองเท่านั้น?” โดยแบ่งเป็นระดับดังนี้

A. บุคคลทั่วไปสามารถเข้าใจได้

B. บุคคลทั่วไปที่ผ่านการอ่านบทความที่เกี่ยวข้องมาแล้วสามารถเข้าใจได้

C. ต้องใช้ข้อมูลในด้านใดด้านนึงจะเข้าใจได้

D. เป็นไปไม่ได้ ถ้าไม่มีความรู้ที่ไม่ได้เป็นรูปธรรม

E. ไม่เป็นรูปธรรม

v. หลังจากคำตามข้างบนและการให้คะแนน จะทำการเก็บไว้เฉพาะหัวข้อที่มีคะแนนเฉลี่ยมากที่สุดอยู่ที่ประมาณ 2.5 คะแนนหรือต่ำกว่าเท่านั้น

vi. ทำให้สุดท้ายเหลือเพียงประมาณ 10,000 หัวข้อที่สามารถใช้ได้

vii. หลังจากได้หัวข้อที่คิดว่าเป็นรูปธรรมแล้วก็นำไปค้นหาและรวบรวมด้วย YouTube annotation system โดยมีขั้นตอนดังนี้

A. สุ่มเลือกวิดีโามา 10 ล้านวิดีโอ พร้อมกับหัวข้อของวิดีโอ โดยใช้กฎที่กำหนดไว้ เอาหัวข้อที่มีจำนวนวิดีโอมากกว่า 200 วิดีโอออก

B. ทำให้เหลือจำนวนวิดีโอยู่ 8,264,650 วิดีโอ

C. แยกออกเป็น 3 ส่วน Train set, Validate set และ Test set ในอัตราส่วน 70:20:10 ตามลำดับ

2. โมเดลปัญญาประดิษฐ์

(a) การเตรียมข้อมูล

i. คุณลักษณะระดับเฟรม : การลดขนาดของข้อมูล เนื่องจากมีข้อมูลที่มีขนาดใหญ่ทำให้ใช้เวลาในการเปิดนาน ซึ่งกระบวนการนี้จะมีการลดความเร็วเฟรมต่อวินาที เวกเตอร์ของคุณลักษณะ และแปลงข้อมูลจาก 32 บิต ให้เป็น 8 บิต

ii. คุณลักษณะระดับวิดีโอ : การแยกเวกเตอร์คุณลักษณะระดับวิดีโอจากคุณลักษณะระดับเฟรมซึ่งการทำแบบนี้ทำให้ได้ประโยชน์ 3 ข้อ คือโมเดลทั่วไปที่ไม่ใช้โครงข่ายประสาทเทียบสามารถนำไปใช้งานได้ ขนาดข้อมูลเล็กลง และเหมาะสมกับการนำไปสร้างโมเดล domain adaptive มากขึ้น

(b) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับเฟรม

- i. one vs all logistic regression classifier + average pooling
- ii. Deep bag of frames (แนวคิดแบบเดียวกับ deep bag of words)
- iii. Long short-term memory (LSTM)

(c) โมเดลปัญญาประดิษฐ์ที่ใช้ในการทดสอบชุดข้อมูลแบบที่เป็นคุณลักษณะระดับวิดีโอ

- i. Logistic regression
- ii. Support vector machine (SVM)
- iii. Mixture of Expert (MoE)

(d) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือ

- i. Mean Average Precision (mAP)
- ii. Hit@k (เหมือนกันกับ Top@k)
- iii. Precision at equal recall rate (PERR)

(e) ความสามารถของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ YouTube-8M

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

ตารางที่ 2.1: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M

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

(g) ความสามารถของโมเดลปัญญาประดิษฐ์ที่ใช้ชุดข้อมูลของ YouTube-8M ในการสร้างเทียบกับชุดข้อมูลสำหรับทดสอบของ ActivityNet

Approach	mAP	Hit@1	PERR
Logistic regression	58.0	60.1	79.6
Mixture-of-2-experts	61.3	63.2	82.6
LSTM	67.6	65.7	86.2
Hierarchical 3D convolutions ^[7]	-	61.0	80.0
Stacked 3D convolutions ^[12]	-	61.0	85.0
LSTM with optical flow and pixels ^[11]	-	73.0	91.0

ตารางที่ 2.2: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M

Approach	mAP	Hit@1	Hit@5
Mixture-of-2-experts	77.6	74.9	91.6
LSTM	57.9	63.4	81.0
Ma, Bargal et al. ^[9]	53.8	-	-
Heilbron et al. ^[2]	43.0	-	-

ตารางที่ 2.3: ผลการทดสอบโมเดลต่างๆบนชุดข้อมูลที่ใช้ในการทดสอบของ YouTube-8M

- (h) ปัญหาที่พบ เนื่องจากว่า YouTube-8M นั้นมีจำนวนข้อมูลที่เยอะมาก ทำให้ไม่สามารถตรวจสอบความถูกต้องของชุดข้อมูลได้ทั้งหมด ว่ามีความถูกต้องมากน้อยขนาดไหน ทำให้อาจเกิดข้อผิดพลาดได้ (ปัจจุบันปี 2019 YouTube-8M ได้มีการตรวจสอบข้อมูลอีกครั้ง เพื่อเพิ่มประสิทธิภาพของชุดข้อมูลซึ่งทำให้ปัจจุบันจำนวนข้อมูล และจำนวนหัวข้อลดน้อยลงจากข้อมูลที่ใช้อ้างอิงในบทความข้างต้นที่ได้กล่าวมา)

AVA

1. ชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนับสนุนการกระทำของมนุษย์เป็นศูนย์กลาง
- (b) จำนวนของวิดีโอ : 640 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 15 นาที และ ถูกสุมตัวอย่างด้วยความถี่ 1 hz
- (d) จำนวนของหมวดหมู่ : 80 หมวดหมู่
- (e) ขั้นตอนการเก็บข้อมูลสำหรับการทำชุดข้อมูลมีขั้นตอนการทำ 5 ขั้น คือ
 - i. การสร้างคำศัพท์การกระทำ จะมีหลัก 3 ข้อในการรวมคำศัพท์ คือ
 - A. เก็บรวบรวมคำศัพท์ทั่วไปที่เกิดขึ้นในชีวิตประจำวัน
 - B. จะต้องมีเอกสารชนิดสามารถเห็นได้ชัดเจน เช่น การถือของ
 - C. กำหนดรูปแบบของคำศัพท์ขึ้นมาและใช้ความรู้จากชุดข้อมูลอื่น ในการทำให้ได้หมวดหมู่ของการกระทำของมนุษย์ที่ครอบคลุมของชุดข้อมูล AVA
 - ii. หนังและส่วนที่เลือกมาใช้วิดีโอที่ใช้ทำชุดข้อมูล AVA ทั้งหมดจะถูกนำมาจาก YouTube โดยเริ่มจากการรวมรายการข้อของนักแสดงที่มีชื่อเสียง ซึ่งจะมีความหลากหลายของเชื้อชาติรวมกันอยู่ ซึ่งวิดีโอดังนี้คือ
 - A. วิดีโอต้องอยู่ในหมวด หนัง และ ละครโทรทัศน์
 - B. จะต้องมีความยาวมากกว่า 30 นาที
 - C. อัพโหลดเป็นเวลาอย่างน้อย 1 ปี
 - D. มียอดวิวคนดูมากกว่า 1000 วิว
 - E. ลงทะเบียนวิดีโอของประเภท เช่น ขาว-ดำ , ความละเอียดต่ำ , การ์ตูน , วิดีโอเกม
 - iii. การตีกรอบบุคคลที่อยู่ภายในภาพ ประกอบด้วย 2 ขั้นตอน
 - A. สร้างกรอบสี่เหลี่ยม โดยใช้โมเดล Faster R-CNN สำหรับการตรวจจับมนุษย์
 - B. นำมนุษย์มาใช้ในการตรวจสอบและแก้ไขกรอบสี่เหลี่ยมที่พลาดไป หรือ ตรวจจับผิด
 - iv. การเขียนของบุคคลในช่วงระยะเวลาสั้นๆของเฟรม
 - ทำการเชื่อมกรอบสี่เหลี่ยมที่อยู่ในช่วงเวลาเดียวกัน ซึ่งใช้วิธีการ track โดยยึดมนุษย์เป็นศูนย์กลาง ซึ่งจะนำมารวบรวมไว้ใน簇กลุ่มเดียวกันโดยการจับคู่กรอบสี่เหลี่ยม และ ใช้ person embedding จากนั้นจะใช้ Hungarian algorithm ในการหาตัวเลือกที่ดีที่สุด
 - v. การสร้างคำอธิบาย
 - การสร้างคำอธิบายของการกระทำจะถูกสร้างจากเหล่าคนที่เป็นผู้สร้างคำอธิบาย ซึ่งจะใช้หน้าต่างโปรแกรมสำหรับช่วยเหลือในการสร้างซึ่งใน 1 กรอบสี่เหลี่ยม สามารถมีคำอธิบายของการกระทำได้สูงสุดถึง 7 labels นอกจากนั้นสามารถตั้งสถานะบล็อกเนื้อหาที่ไม่เหมาะสม หรือ กรอบสี่เหลี่ยมที่ผิดพลาดได้อีกด้วย ในทางปฏิบัติจะสังเกตได้ว่ามันมีโอกาสผิดอย่างหลีกเลี่ยงไม่ได้ เมื่อต้องได้รับคำสั่งให้หาคำอธิบายของการกระทำที่ถูกต้องจาก 80 หมวดหมู่ จึงแบ่งขั้นตอนออกเป็น 2 ขั้นตอน คือ
 - A. ข้อเสนอของการกระทำสอบถามเหล่าผู้สร้างคำอธิบาย เพื่อสร้างข้อเสนอสำหรับคำอธิบายของการกระทำจากนั้นจับกลุ่มเข้าด้วยกัน ซึ่งจะทำให้มีโอกาสสูงต้องมากกว่าเป็นข้อเสนอแยกเดี่ยว

B. ผู้ตรวจสอบข้อเสนอจะตรวจสอบข้อเสนอที่ได้จากขั้นตอนแรก ซึ่งในแต่ละวิดีโอกลิปจะใช้มุชยในการตรวจสอบ 3 คน เมื่อคำอธิบายของการกระทำ ถูกตรวจสอบด้วยผู้ตรวจสอบข้อเสนออย่างน้อย 2 คน คำอธิบายของการกระทำนี้จะถูกยึดเป็นคำอธิบายหลัก

2. Machine learning model

- (a) Machine learning model ที่งานวิจัยนี้ใช้ two stream variant ซึ่งจะทำการประมวลผลทั้ง RGB flow และ optical flow และ เป็นโครงสร้างของ Faster RCNN ที่นำ Inception network เข้ามาใช้
- (b) เครื่องมือที่ใช้วัดผลสำหรับงานวิจัยนี้ คือ ค่า IOU และ 3D IOUs
 - i. ค่า IOU คือ ค่าที่ใช้วัดความสอดคล้องระหว่างสองเฟรม ซึ่งใช้สำหรับการวัดผลระดับเฟรม โดยจะเป็นการเทียบกันของกรอบสี่เหลี่ยมที่ตรวจเจอและกรอบสี่เหลี่ยมจริงของวัตถุ
 - ii. ค่า 3D IOUs คือ ค่าที่ใช้วัดความสอดคล้องระหว่างสองวิดีโอ ซึ่งใช้สำหรับการวัดผลระดับวิดีโอโดยเทียบกันของ ground truth tubes และ linked detection tubes ซึ่งก็คือ การนำเอกสารสี่เหลี่ยมจริงของวัตถุในเฟรมที่ติดตอกันมาเรียงต่อกันเป็น tube และ linked detection tube คือ การนำเอกสารสี่เหลี่ยม (bounding box) ที่ตรวจเจอมาระเบิดต่อกันเป็น tube
- (c) ความสามารถของ Machine learning model ในปัจจุบัน
 - i. จากการทดสอบการเทียบ Machine learning model ของงานวิจัยนี้และวิธีการอื่นๆ โดยนำไปทดสอบกับชุดข้อมูลวิดีโอ JHMDB และ UCF101-24 ได้ผลลัพธ์ออกมาดังนี้

Frame-mAP	JHMDB	UCF101-24
Actionness	39.9	-
Peng w/o MR	56.9	64.8
Peng w/ MR	58.5	65.7
ACT	65.7	69.5
Out approach	73.3	76.3

ตารางที่ 2.4: ผลการทดสอบของวิธีต่างๆ บน Frame Level

(d) ปัญหาที่พบ

- i. ในปัจจุบันยังไม่มีโมเดลปัญญาประดิษฐ์ที่เมื่อทดสอบกับชุดข้อมูล AVA และได้ผลการทำงานที่ดี เนื่องจาก ชุดข้อมูล AVA สนใจการกระทำของมนุษย์ที่มีรายละเอียดเล็กๆน้อยๆ ทำให้ยากต่อการทำนายสำหรับชุดข้อมูล AVA

Moment in Time

1. ชุดข้อมูล

- (a) เป้าหมายของชุดข้อมูล : สนใจการกระทำทุกการกระทำในวิดีโอ เช่น การกระทำการของ คน สัตว์ สิ่งของ และ ปรากฏการณ์ธรรมชาติ
- (b) จำนวนของวิดีโอ : >1,000,000 วิดีโอ
- (c) ความยาวเฉลี่ยของแต่ละวิดีโอ : 3 วินาที
- (d) จำนวนของหมวดหมู่ : 339 หมวดหมู่
- (e) วิธีการเก็บรวบรวมข้อมูล :
 - i. เริ่มจากการรวบรวมคำ (verb) ที่มีการใช้อยู่ทั่วไปในชีวิตประจำวันมา 4,500 คำ จาก Verb-Net จากนั้นนำมาแบ่งกลุ่มคำ(verb) ที่มีความหมายใกล้เคียงกันโดยใช้ features จาก Propbank และ FrameNet โดยเก็บข้อมูลเป็นแบบ binary feature vector ซึ่งถ้าคำ (verb) ในนี้มีความเกี่ยวข้องกับ feature ก็จะให้ค่าเป็น 1 ถ้าไม่เกี่ยวข้องกันจะให้ค่าเป็น 0 จากนั้นจึงใช้ วิธี k-means clustering ในการแบ่งกลุ่ม เมื่อแบ่งกลุ่มแล้วจากนั้นจะเลือกคำ (verb) จากในแต่ละกลุ่มนั้น โดยคำ (verb) ที่เลือกมานั้นจะเป็นที่ที่บอยที่สุดในกลุ่มนั้น และลบคำ (verb) นั้นออกจากกลุ่มทั้งหมด (คำ ๆ หนึ่งสามารถอยู่ได้หลายกลุ่ม) จากนั้นจะทำการวนการนี้ไปเรื่อย ๆ แต่คำ (verb) ที่เลือกมาจะต้องไม่มีความหมายคลุมเครือ ไม่สามารถมองเห็นหรือได้ยินได้ และต้องไม่มีความหมายเหมือนกับคำ (verb) ที่เคยเลือกมาก่อน จนสุดท้ายแล้วได้ออกมาที่ 339 class
 - ii. ต่อมาทำการหาชุดข้อมูลวิดีโอด้วยจะตัดออกมาเพียง 3 วินาทีที่เกี่ยวข้องกับคำ (verb) ใน 339 class ที่เลือกมา จากวิดีโอ แหล่งต่างกัน 10 แหล่ง การตัดวิดีโอนั้นจะไม่ใช้พวก Video2Gif (โมเดลที่ระบุตำแหน่งของสิ่งที่น่าสนใจในวิดีโอ) เพราะจะทำให้เกิด bias ขึ้นจะเกิดขึ้นตอนสร้างโมเดลจากนั้นจะทำการส่งข้อมูลของคำ (verb) และวิดีโอที่ตัดไปยัง Amazon Mechanical Turk (AMT หรือตลาดแรงงาน) เพื่อทำการ label โดยพนักงานแต่ละคนของ AMT จะได้ 64 วิดีโอซึ่งเกี่ยวข้องกับคำ (verb) หนึ่ง และอีก 10 วิดีโอที่มีการทำ label อยู่แล้ว โดยวิดีโอที่มีการทำ label ถ้ามีพนักงานของ AMT ตอบเหมือนกันกับที่ทำ label ไว้เกิน 90% ถึงจะนำเข้าไปรวมกับชุดข้อมูลส่วนอีก 64 วิดีโอีก 1 ที่จะเป็นของ training set จะต้องผ่านพนักงานของ AMT อย่างน้อย 3 ครั้ง และต้อง label เมื่อเทียบกับ 75% ขึ้นไปถึงจะถือว่าเป็น label ที่ถูกต้อง ถ้าเป็นของ validation และ test set จะต้องผ่านพนักงานของ AMT อย่างน้อย 4 ครั้ง และต้อง label เมื่อเทียบกับ 85% ขึ้นไป ที่ไม่ต่างกันที่กว่า 100% เพราะจะทำให้วิดีโอนั้นยากเกินไปที่จะทำให้สามารถจำการกระทำได้

2. การเตรียมข้อมูล

- (a) training set จะมี 802,264 วิดีโอ และมีวิดีโoinแต่ละ class อยู่ที่ 500 ถึง 5,000 วิดีโอ
- (b) validation set จะมี 33,900 วิดีโอ และมีวิดีโoinแต่ละ class อยู่ที่ 100 วิดีโอ
- (c) เริ่มการ preprocess จากแยกภาพRGB ออกแบบจากวิดีโอ และทำการเปลี่ยนขนาดของภาพให้เป็น 340x256 pixel
- (d) ใช้ TVL1 optical flow algorithm จาก opencv เพื่อลดข้อมูลรบกวนที่จะเกิดขึ้น
- (e) ทำการแปลงค่าที่อยู่ใน optical flow ให้เป็นเลขจำนวนเต็ม(integer) เพื่อทำให้การคำนวณนั้นเร็วขึ้น

- (f) ปรับค่า displacement ใน optical flow ให้ค่าสูงสุดเป็น 15 ต่ำสุดเป็น 0 และทำการปรับขนาดให้เป็นช่วง 0-255
- (g) เก็บข้อมูลอุปกรณ์ในรูปแบบของ grayscale image เพื่อลดพื้นที่ ๆ ใช้เก็บข้อมูล
- (h) แก้ปัญหาเรื่องการเคลื่อนไหวของกล้อง(camera motion) โดยการนำค่าเฉลี่ยของ เวกเตอร์(vector) ไปลบกับ displacement
- (i) สุดท้ายจะเป็นสุ่มตัดภาพอุปกรณ์เพื่อเพิ่มจำนวนข้อมูล

3. Machine learning model

(a) ในงานวิจัยนี้มีการทดสอบ Machine learning model หลายอัน ซึ่ง Machine learning model ที่มีประสิทธิภาพการทำงานที่ดีที่สุดตาม 5 ลำดับแรกดังนี้

- i. SVM มีรูปแบบข้อมูลอินพุท คือ Spatial+Temporal+Auditory
- ii. I3D มีรูปแบบข้อมูลอินพุท คือ Spatial+Temporal
- iii. TRN-Multiscale มีรูปแบบข้อมูลอินพุท คือ Spatial+Temporal
- iv. TSN-2stream มีรูปแบบข้อมูลอินพุท คือ Spatial+Temporal
- v. ResNet50-ImageNet มีรูปแบบข้อมูลอินพุท คือ Spatial

(b) เครื่องมือที่ใช้วัดผลงานวิจัยนี้

- i. Classification accuracy Top-1 , Top-5

(c) ความสามารถของ Machine learning model ในปัจจุบัน

- i. ทำการทดสอบ cross dataset transfer โดยการนำโมเดล ResNet50 I3D pretrained ลงทั้งบน Kinetics และ Moments in time และนำมาเทียบกับชุดข้อมูลอื่น โดยชุดข้อมูลแต่ละชุดจะมีการปรับ frame rate ของวิดีโอให้เป็น 5 fps เมื่อเทียบกัน

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

ตารางที่ 2.5: Data transfer performance ของโมเดล Resnet50 I3D

- ii. จะเห็นได้ว่า Kinetics ให้ผลลัพธ์ที่ดีกว่าใน UCF เพราะว่ามีการแชร์ class ด้วยกันอยู่หลายอย่าง ในขณะที่ HMDB นั้นมีการรวม source จากหลายแหล่ง และมีจำนวน class ที่หลากหลายจึงทำให้มีความใกล้เคียงกับตัวข้อมูลของ Moments in time ดังนั้นจึงเทียบผลลัพธ์จาก Something ซึ่งจะทำให้เห็นว่า Moments in time มีประสิทธิภาพที่ดีกว่าและวิดีโอด้วยความยาวมากกว่า 3 วินาทีจะไม่ส่งผลกระทบกับประสิทธิภาพของ Moments in time

4. ปัญหาที่พบ

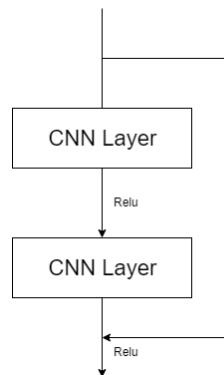
(a) ผลลัพธ์จากการทำนายด้วยโมเดลถ้าผ่านรูปภาพที่มีรายละเอียดเยอะจะทำให้การ ทำนายโอกาส ผิดนั่นค่อนข้างสูง ซึ่งปัญหานี้สามารถทำให้เกิดน้อยลงด้วยการนำวิธี Class Activation Mapping(CAM) จะเป็นการเน้นรูปภาพในส่วนที่มีข้อมูลมากที่สุดและ ทำนายผลออกมา แต่ก็ยังมีจุดที่ เป็นปัญหาอยู่ เช่น การกระที่เกิดขึ้นเร็วมาก (การลื่นล้ม) จะทำให้การทำนาย นั้นมีโอกาสผิดพลาดสูงขึ้น

2.2 เครื่องมือสำหรับการวิเคราะห์ผลวิดีโอ

2.2.1 โมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำมนุษย์

2.2.1.1 ResNet

ในการสร้างโมเดลปัญญาประดิษฐ์นั้นการใช้จำนวนชั้น (layer) เยอะนั้นจะทำให้ได้คุณลักษณะของข้อมูลที่ออกมากเยอะตามไปด้วย แต่การที่คุณลักษณะของข้อมูลเยอะไม่ได้หมายความว่าโมเดลปัญญาประดิษฐ์จะให้ประสิทธิภาพที่ดีเสมอไป ซึ่งสามารถแก้ปัญหานี้ได้โดยใช้ Residual Network (ResNet) ที่เป็น Convolution Neuron Network (CNN) ประเภทหนึ่ง ที่ส่วนใหญ่จะนำมาใช้กับข้อมูลที่เป็นรูปภาพ เช่น การจดจำตุ๊กตา เป็นต้น โดย ResNet นี้จะสามารถทำการข้ามชั้นของ CNN ที่ไม่จำเป็นได้ โดยในชั้นที่ไม่จำเป็นจะมีการปรับ weight ให้เข้าใกล้ 0 ในขณะที่ train ข้อมูล การข้ามชั้น CNN ที่ไม่จำเป็นจะช่วยลดเวลาที่ใช้ในการ train และทำให้ประสิทธิภาพของโมเดลปัญญาประดิษฐ์ดีขึ้น



รูปที่ 2.2: ResNet

การทดลองโมเดลปัญญาประดิษฐ์ ResNet ด้วยการทำจำแนกรูปภาพโดยใช้ชุดข้อมูลของ ImageNet ที่ประกอบไป class มากกว่า 1,000 class มาเทียบกับโมเดลปัญญาประดิษฐ์ทั่วไป (plain) ที่จำนวนชั้น 18 ชั้น และ 34 ชั้น ผลลัพท์จะได้ออกมาตามตารางด้านล่างดังนี้ (โครงสร้างพื้นฐานของโมเดลปัญญาดิษฐ์ ResNet และโมเดลปัญญาประดิษฐ์ทั่วไปเหมือนกัน)

จำนวนชั้นของ	training error	
	plain	ResNet
18	27.94	27.88
34	28.54	25.03

ตารางที่ 2.6: Top-1 ของความผิดพลาดของชุดข้อมูลทดสอบ ImageNet

จากตาราง 2.6 จะเห็นได้ว่าโมเดลปัญญาประดิษฐ์ทั่วไป 34 ชั้นมีค่าเบื้องตนต์ความผิดพลาดสูงกว่าโมเดลปัญญาประดิษฐ์ ResNet แบบชัดเจน ในขณะที่โมเดลปัญญาประดิษฐ์ทั่วไปจะมีเบื้องตนต์ความผิดพลาดสูงขึ้นเมื่อเทียบกันระหว่าง 18 ชั้นและ 34 ชั้น

ต่อมาจะนำโมเดลปัญญาประดิษฐ์ ResNet มาทดสอบกับชุดข้อมูล CIFAR-10 ซึ่งเป็นชุดข้อมูลที่มีภาพสำหรับ train 50,000 ภาพ ภาพสำหรับทดสอบ 10,000 ภาพ และมีจำนวน class ทั้งหมด 10 class โดยจะมีการออกแบบของจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet ตามจำนวนของชั้น Convolution ที่มีผังคุณลักษณะเท่ากัน 6 ชั้นติดกันและการข้ามชั้นที่ละ 2 จึงทำให้ได้รูปแบบการคิดชั้นดังนี้ $6n + 2$ สำหรับการทดสอบจะให้ค่า $n = [3, 5, 7, 9, 200]$ ดังตารางต่อไปนี้

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

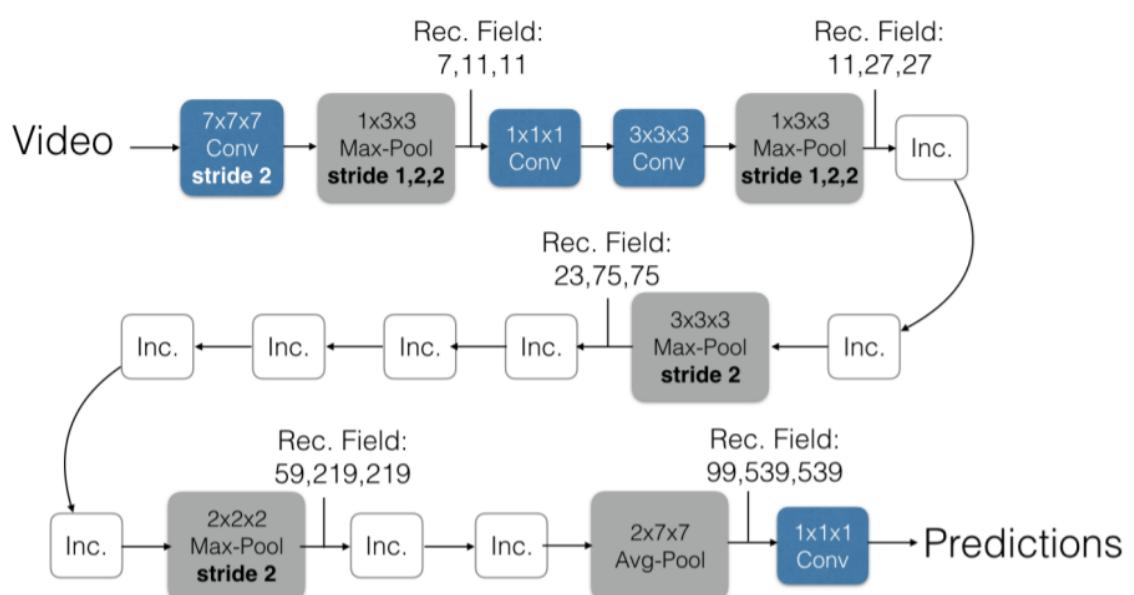
ตารางที่ 2.7: ค่าความผิดพลาดที่ได้จากการทดลองจำนวนชั้นของโมเดลปัญญาประดิษฐ์ ResNet บนชุดของข้อมูล CIFAR-10

จากตาราง 2.7 จะเห็นได้ว่าที่โมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1202 นั้นมีค่าความผิดพลาดเกิดขึ้นมากกว่าจำนวนชั้น 110 ซึ่งอาจจะเป็นไปได้ว่าขนาดของโมเดลปัญญาประดิษฐ์ ResNet ที่มีจำนวนชั้น 1202 นั้นมากเกินไปสำหรับชุดข้อมูลขนาดเล็กนี้

2.2.1.2 Inflated 3D convolutional network

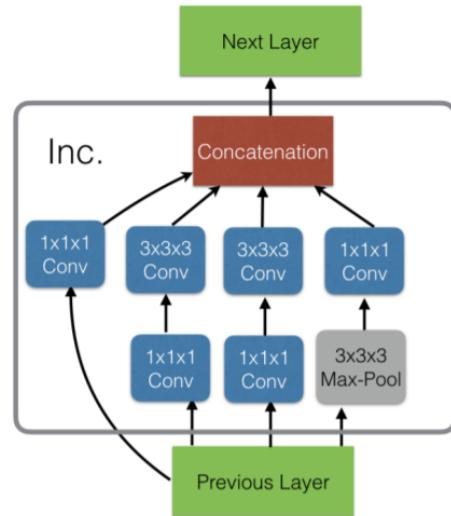
ในการพัฒนาโมเดลปัญญาประดิษฐ์สำหรับจำแนกการกระทำของมนุษย์นั้นมีพื้นฐานมาจากการจำแนกวัตถุ (object classification) หมายถึงการใช้รูปภาพหนึ่งรูปในการประมวลผลและนำয่อออกมาว่าภายในรูปนั้นมีบริบทการกระทำอย่างไร โดยไม่ได้คำนึงถึงข้อมูลเชิงต่อเนื่อง (spatio-temporal information) จากบทความ "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset"^[3] นั้นได้พัฒนาโครงสร้างของโมเดลปัญญาประดิษฐ์ (architecture) ที่มีประสิทธิภาพในการประมวลผลภาพเคลื่อนไหวได้ชื่อว่า I3D หรือ inflated 3D-convolution network โดยโครงสร้างพื้นฐานของ I3D นั้นมาจากการนำ Inception-v1^[10] ที่ถูกพัฒนาโดย Google ซึ่งเป็นโครงสร้างที่มีประสิทธิภาพสูงในการจำแนกวัตถุในรูปภาพ และ I3D นั้นได้ทำการขยายมิติของโครงสร้างจาก 2 มิติ เป็น 3 มิติ เพื่อให้โมเดลปัญญาประดิษฐ์สามารถเรียนรู้ข้อมูลเชิงต่อเนื่องได้

Inflated Inception-V1



รูปที่ 2.3: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[3]

Inception Module (Inc.)



รูปที่ 2.4: โครงสร้างของโมเดลปัญญาประดิษฐ์ I3D^[3]

ประสิทธิภาพของโมเดล I3D แบบ two-stream เมื่อเทียบกับ long-short term memory (LSTM), 3D-convolution network, two-stream และ 3D-fused โดยใช้เครื่องมือในการวัดผลคือความแม่นยำจากการทำนายอันดับแรกสุด (Top@1 accuracy) และความแม่นยำจากการทำนาย 5 อันดับแรก (Top@5 accuracy) ตามตารางที่ 2.8

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

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

2.2.2 เครื่องมือสำหรับสร้างชุดข้อมูล

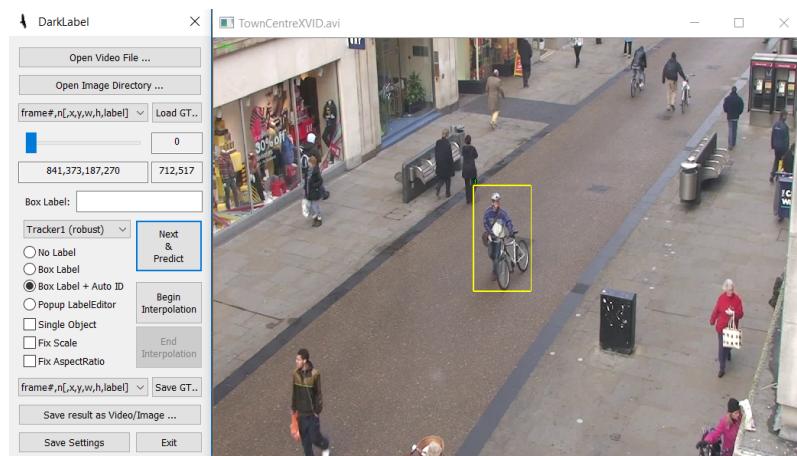
จากการค้นคว้าหาเครื่องมือในการ labeling เพื่อใช้เป็นแนวทางในการทำ Goggle labeling tool พบเครื่องมือที่เป็น open source เปิดให้ทดลองใช้อยู่ 2 เครื่องมือ คือ DarkLabel และ OpenLabeling โดยสรุปข้อสำคัญได้ดังนี้

โปรแกรม DarkLabel

เป็นโปรแกรมที่ช่วยในการทำงานคำขอธิบายและบันทึกในรูปแบบต่างๆ รองรับข้อมูลอินพุทในรูปแบบไฟล์วิดีโอ avi , mpq หรือ กลุ่มรูปภาพ มีขั้นตอนการ labeling ดังนี้

1. สร้างกรอบสี่เหลี่ยม(boundary box)ครอบบริเวณวัตถุที่สนใจ โดยใช้มุชย์เป็นคนสร้าง
2. กดปุ่ม Next และ Predict อย่างต่อเนื่อง เพื่อ track กรอบสี่เหลี่ยม ในเฟรมถัดๆไป จนกระทั่งการ track เกิดพลาดไป
3. ลบกรอบสี่เหลี่ยมที่พลาด และเริ่มทำขั้นตอนที่ 1 ใหม่ อีกรังส์จนครบทุกเฟรมในวิดีโอ

หลังจากที่ผู้วิจัยได้ทดลองใช้โปรแกรม DarkLabel พบร่วม เป็นโปรแกรมที่ค่อนข้างมีการทำงานส่วนใหญ่ที่เป็นการสร้างคำขอธิบายแบบใช้มุชย์เป็นคนกำหนดเองเป็นส่วนใหญ่ ซึ่งทำให้ใช้เวลาในการทำงาน และเสียพลังงานในการทำเป็นอย่างมาก



รูปที่ 2.5: UI ของโปรแกรม DarkLabel

โปรแกรม OpenLabeling

ที่ช่วยในการทำนายคำอธิบาย โดยโปรแกรมจะมีการทำงานอยู่ 2 โหมดการทำงาน คือ แบบทำด้วยมือ และ แบบอัตโนมัติ ซึ่งมีการทำงานแยกกันอย่างชัดเจน

1. Mode Auto

หลังจากอินพุตวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการ labeling ดังนี้

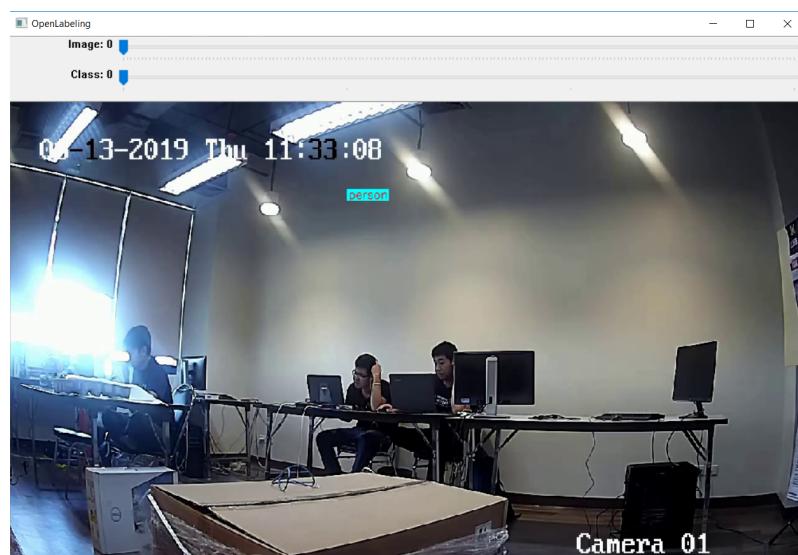
- (a) โปรแกรมจะทำงานอัตโนมัติ โดยใช้โมเดลในการทำนายคีย์เฟรม (predict keyframe) และ track ในภาพที่เหลือ ผลลัพธ์ที่ได้คือ ข้อมูลของชุดข้อมูล

2. Mode Manual

หลังจากอินพุตวิดีโอเข้าไปในโปรแกรมแล้วมีขั้นตอนการ labeling ดังนี้

- (a) สร้างกรอบสี่เหลี่ยม (bounding box) ขึ้นมาโดยใช้มนุษย์เป็นคนสร้าง
- (b) กดปุ่มเพื่อแทร็กกรอบสี่เหลี่ยม (track bounding box) ในเฟรมถัดๆไป จนกระทั่งการแทร็กกรอบสี่เหลี่ยม (track bounding box) เกิดพลาดไป
- (c) ลบกรอบสี่เหลี่ยม (bounding box) ที่พลาด และ เริ่มทำขั้นตอนที่ 1 อีกครั้งจนครบทุกเฟรมในวิดีโอ

หลังจากที่ได้ทดลองใช้โปรแกรม OpenLabeling ทั้ง 2 โหมดการทำงานแล้วพบว่า การทำงานแบบ mode auto การที่เรายังสามารถปรับแก้ไขสิ่งใดในระหว่างกระบวนการ labeling นั้น ทำให้หากเกิดกรณีที่ไม่เดลทำนายกรอบสี่เหลี่ยม (predict bounding) พลาด หรือ เกินมา เราจะไม่สามารถแก้ไขได้ และ การทำงานแบบ mode manual ไม่มีระบบตรวจสอบกรอบสี่เหลี่ยม (detect bounding box) ทำให้ผู้ใช้งานจะต้องสร้างกรอบสี่เหลี่ยม (bounding box) ขึ้นมาเอง

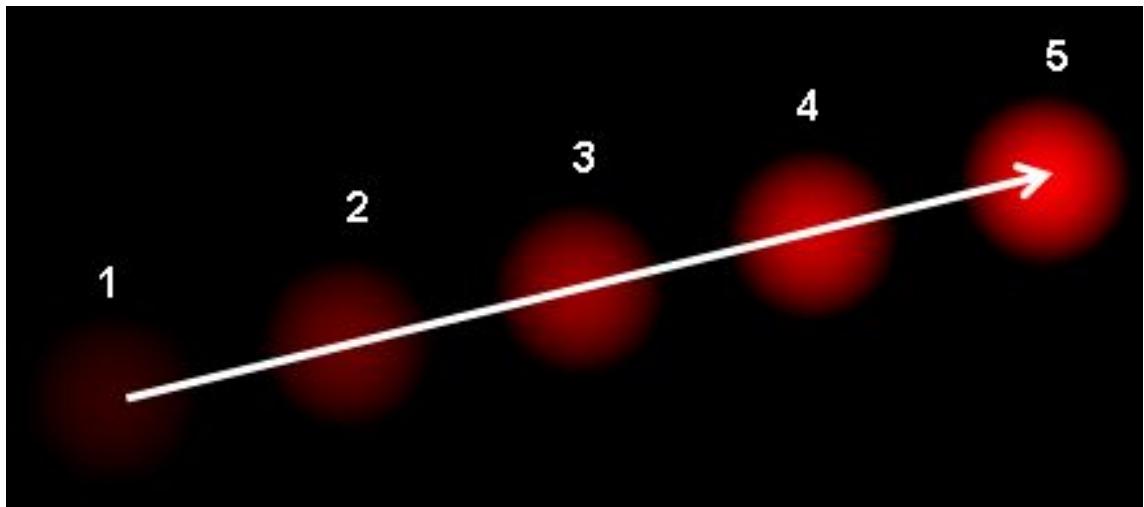


รูปที่ 2.6: UI ของโปรแกรม OpenLabeling

2.3 ทฤษฎีเกี่ยวข้อง

2.3.1 Optical flow

Optical flow¹ คือรูปแบบของการเคลื่อนที่ของวัตถุในรูปภาพระหว่างภาพซึ่งอาจจากการจากเคลื่อนที่ของวัตถุหรือตัวกล้อง ออกแบบมาในรูปแบบของ เวกเตอร์(vector) 2 มิติ โดยที่เวกเตอร์แต่ละตัวจะแสดงถึงทิศทาง การเคลื่อนที่ระหว่างภาพดังรูปด้านล่าง



รูปที่ 2.7: ตัวอย่างการเคลื่อนที่ของลูกบอล

จากรูปภาพจะแสดงให้เห็นถึงการเคลื่อนที่ของลูกบอลของภาพที่ต่อเนื่องกัน 5 ภาพโดยที่ลูกคระแสดงถึงทิศทางการเคลื่อนที่ของเวกเตอร์

การทำงานของ optical flow อยู่บนสมมติฐานหลายประการได้แก่

1. ความเข้มของพิกเซล(pixel) ของวัตถุจะไม่เปลี่ยนแปลงระหว่างภาพที่ต่อเนื่องกัน
2. พิกเซลที่อยู่ใกล้กันจะมีการเคลื่อนไหวที่คล้ายกัน

เมื่อพิจารณาพิกเซล $I(x,y,t)$ จากภาพแรกจะเคลื่อนไหวเป็นระยะทาง (dx,dy) ไปยังภาพต่อไปหลังจากผ่านไปแล้ว dt เวลา ดังนั้นเนื่องจาก พิกเซล เหล่านี้เหมือนกันและความเข้มไม่มีการเปลี่ยนแปลง จึงทำให้พูดได้ว่า

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (2.7)$$

โดยที่

¹Optical flow, shorturl.at/mrtEZ

- I = พิกเซลจากภาพ
 x = ตำแหน่งของพิกเซล ในแกน x
 dx = ระยะทางที่เคลื่อนที่ในแกน x
 y = ตำแหน่งของพิกเซล ในแกน y
 dy = ระยะทางที่เคลื่อนที่ในแกน y
 t = เวลา
 dt = ระยะเวลาที่เปลี่ยนไประหว่างภาพ

จากนั้นใช้การประมาณค่าของ taylor series ทางฝั่งขวามือและ ลบค่า common term และหารด้วย dt เพื่อให้ได้สมการดังต่อไปนี้

$$f_x u + f_y v + f_t \quad (2.8)$$

$$f_x = \frac{\delta f}{\delta x}; f_y = \frac{\delta f}{\delta y} \quad (2.9)$$

$$u = \frac{\delta x}{\delta t}; v = \frac{\delta y}{\delta t} \quad (2.10)$$

โดยที่

- f_x = เกรเดียน(gradient) ในแกน x
 f_y = เกรเดียนในแกน y
 f_t = เกรเดียนของเวลา
 u = เวกเตอร์การเคลื่อนที่ของแกน x
 v = เวกเตอร์การเคลื่อนที่ของแกน y

สมการข้างบนนี้จะเรียกว่าสมการ optical flow จากสมการทำให้สามารถหา f_x และ f_y โดยเป็น เกรเดียนของภาพ และ f_t เป็นเกรเดียน(gradient)ของเวลา แต่ u กับ v เป็นตัวแปรที่ไม่ทราบ ทำให้สมการนี้ไม่สามารถแก้ไขโดยมีตัวแปรที่ไม่ทราบถึง 2 ตัว จึงมีการนำวิธีการต่าง ๆ เข้ามาใช้ในการแก้ปัญหานี้ โดยวิธีการที่น่าเข้ามาใช้ในการแก้ปัญหาคือ dense optical flow ซึ่งใช้อัลกอริทึมของ Gunnar Farneback ซึ่งจะใช้วิธีการขยายพื้นที่² (polynomial expansion)

²polynomial expansion file: <http://www.diva-portal.org/smash/get/diva2:273847/FULLTEXT01.pdf>

บทที่ 3

ระเบียบวิธีวิจัย

ในการทำโครงการวิจัยแอพพลิเคชันสำหรับวิเคราะห์วิดีโอ จะมีการทำงานหลากหลายส่วนมาทำงานร่วมกัน ซึ่งต้องมีระเบียบวิธีวิจัยอธิบายถึงขั้นตอนการดำเนินงานตั้งแต่เริ่มศึกษาข้อมูลจนไปถึงสิ้นสุดกระบวนการวิจัยโดยใช้ภาษาไทยเป็นภาษาหลักในการเขียนโปรแกรม

3.1 ความต้องการของระบบ

3.1.1 ความต้องการเชิงการใช้งาน (Functional Requirements)

1. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถตัดวิดีโอช่วงเวลาที่ไม่มีมนุษย์อยู่ออกได้อัตโนมัติโดยใช้ปัญญาประดิษฐ์
2. เครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์สามารถระบุตำแหน่งมนุษย์แต่ละคนในวิดีโอด้วยการกรอกข้อมูลนี้ในวิดีโอด้วยการกระทำที่กำหนดจะประกอบไปด้วยการทำได้แก่ ยืน นั่ง นอน เล่นโทรศัพท์ เดิน กินข้าว พูดคุย
3. ชุดข้อมูลที่ได้จากการเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ต้องสามารถนำไปใช้ในการพัฒนาโมเดลปัญญาประดิษฐ์ต่อได้
4. สร้างระบบต้นแบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ที่มีมนุษย์สามารถทำงานร่วมกับปัญญาประดิษฐ์ได้
5. ระบบวิเคราะห์การกระทำการที่มีมนุษย์ต้องสามารถนำวิดีโอมาวิเคราะห์ข้อมูลการกระทำการและตำแหน่งของมนุษย์แต่ละคน และนำข้อมูลเหล่านั้นไปสร้างรายงานออกมากได้ โดยรายละเอียดรายงานจะมีดังนี้
 - (a) เวลา (Time stamp)
 - (b) รหัสระบุตัวตน (ID)
 - (c) การกระทำ
 - (d) ตำแหน่ง โดยจะบอกในลักษณะของกรอบสี่เหลี่ยมครอบพื้นที่ที่มีมนุษย์คนนั้นๆอยู่

3.1.2 ความต้องการเชิงวิศวกรรม (Non-Functional Requirements)

1. สร้างเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์ด้วยภาษาไทย
2. ความละเอียดอย่างต่ำของวิดีโอด้วยมากกว่า 640×480 (กว้าง x สูง)
3. วิดีโอจะต้องมีเฟรมเรทต่อวินาที(fps) อย่างต่ำ 10 fps

3.2 หน้าที่ความรับผิดชอบ

ปฐมพงศ์ สินธุ์งาม สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดการกระทำมนุษย์ 3D รวมถึงออกแบบและสร้างระบบ Tracker

ศุภกร เบญจวิกรัย รวมฟังก์ชันและระบบต่างๆของแอพพลิเคชัน รวมถึงออกแบบและสร้างระบบ Select และ Detect

อุกฤษฎ์ เลิศวรรณาการ สร้างและทดสอบโมเดลปัญญาประดิษฐ์สำหรับจัดการกระทำมนุษย์ Resnet-50 รวมถึงออกแบบและสร้างระบบ Person ReID

3.3 เครื่องมือที่ใช้ในงานวิจัย

ในหัวข้อจะกล่าวถึงซอฟต์แวร์ ภาษาและ program library ที่ใช้ในการพัฒนาระบบ รวมถึงข้อมูลจำเพาะของคอมพิวเตอร์ที่ใช้ในการพัฒนาระบบ

Pycharm community 2017.1.2

เป็นโปรแกรมໄwakeใช้สำหรับเขียนและแก้ไขโค้ดซึ่งข้อดีของโปรแกรมนี้ คือ มีคุณสมบัติต่างๆที่สามารถอำนวยความสะดวกในการเขียนโปรแกรมได้ เช่น syntax highlighting, Auto-completion ฯลฯ และสามารถประมวลผล (compile) โปรแกรมทดสอบแอพพลิเคชันได้

Jupyter 2017.1.2

เป็นโปรแกรมสำหรับเขียนโปรแกรม ที่เหมาะสมสำหรับใช้ในการทดสอบโปรแกรมแต่ละส่วนได้ ซึ่งมีข้อดีคือ หากมีการแก้ไขโปรแกรมเพียงแค่บางส่วน ก็สามารถประมวลผลเฉพาะส่วนที่ต้องการได้ มักจะใช้ในการสร้างโมเดล

Qt Creator 4.9.2 (Community)

เป็นเครื่องมือสำหรับออกแบบหน้าต่างแอพพลิเคชันของ library PyQt ซึ่งมีข้อดีคือ เรียกใช้ง่ายมีวิดเจ็ต(widget)ที่สามารถใช้ได้หลากหลายหมายเหตุสำหรับการออกแบบ

3.4 ภาษาที่ใช้ในการพัฒนาระบบ

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

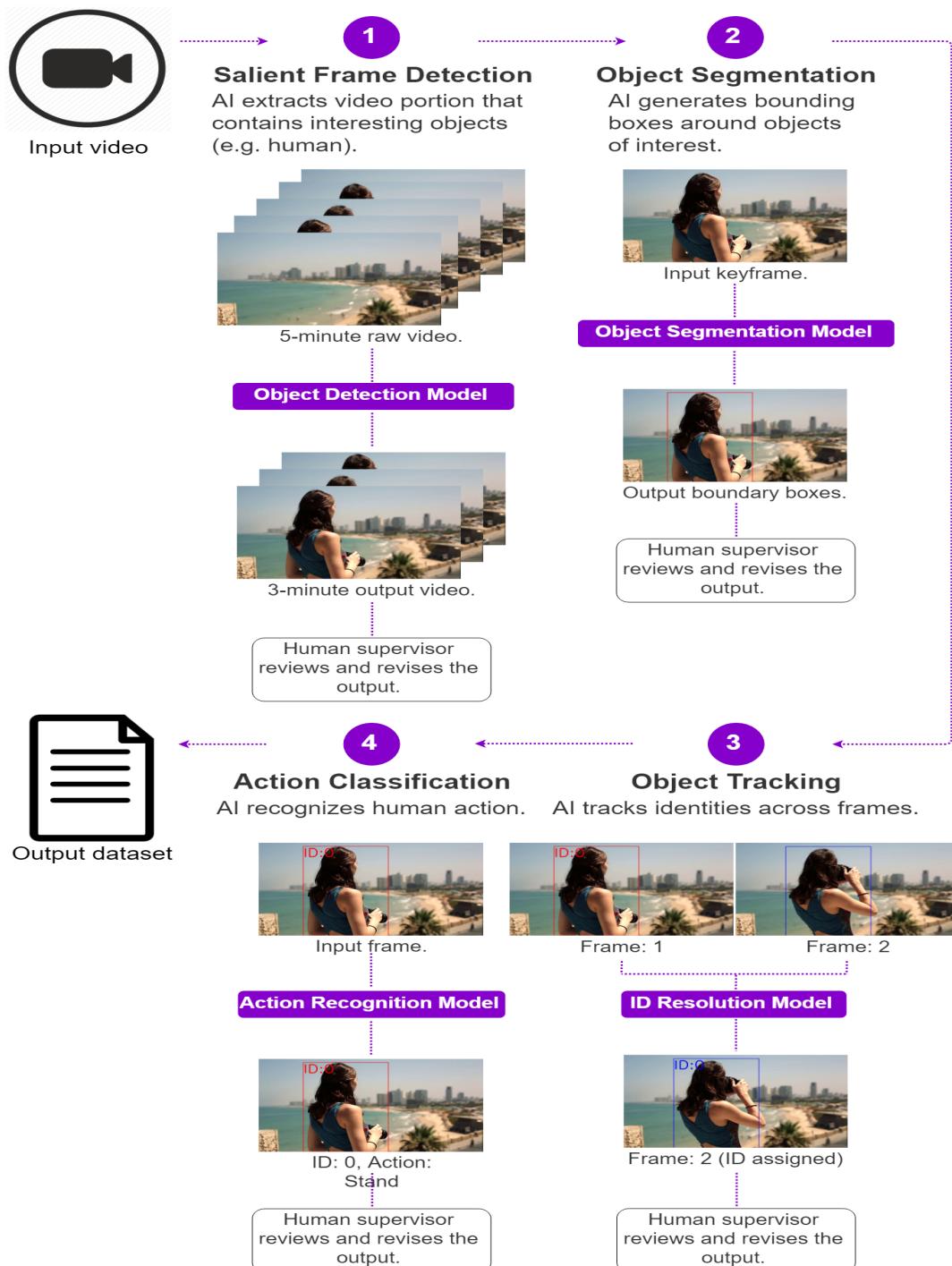
3.5 Program library ที่ใช้ในการพัฒนาระบบและแอปพลิเคชัน

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 ใช้สำหรับการทำ encoding โมเดล
tqdm	4.32.1	library ใช้สำหรับจัดการการทำงานซ้ำ(Loop)
pyqt5	5.9.2	library ใช้สำหรับการทำแอปพลิเคชัน

3.6 แผนการดำเนินงาน

โดยจากที่กล่าวไปตอนต้นในบทนำการดำเนินงานและการออกแบบการสร้างเครื่องมือสำหรับกำกับข้อมูล ด้วยปัญญาประดิษฐ์ และระบบวิเคราะห์การกระทำการของมนุษย์ในวิดีโอ มีแผนการทำงานซึ่งถูกแบ่งออกเป็นสาม ขั้นตอนดังนี้ ขั้นตอนแรกคือ ขั้นตอนของการศึกษาทำความเป็นไปได้ รวมถึงเทคโนโลยีปัจจุบันที่เกี่ยวกับการ สร้างแอปพลิเคชัน และการจัดการกระทำการของมนุษย์ด้วยปัญญาประดิษฐ์ เพื่อนำมาประยุกต์ใช้กับงานวิจัย นี้ ขั้นตอนที่สองคือ ขั้นตอนของการออกแบบและสร้างแอปพลิเคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการเทรน โมเดลจากวิดีโอ ขั้นตอนที่สามคือ ขั้นตอนของการออกแบบและสร้างระบบวิเคราะห์การกระทำการของมนุษย์ได้โดย มีข้อกำหนดตามที่กล่าวไว้ในบทนำ ในการเริ่มทำงานวิจัยนี้นั้นสิ่งจำเป็นที่ต้องทำในอันดับแรกคือการศึกษาข้อมูล ในหัวข้อที่เกี่ยวข้อง หรืองานวิจัยอื่นที่ทำเอาระบบแล้ว เพื่อศึกษาและทำความเข้าใจ ข้อดี-ข้อเสีย ของเทคนิคหรือ กระบวนการต่างๆ เพื่อนำมาประยุกต์ใช้กับงานวิจัยนี้ ในการศึกษาเกี่ยวกับการออกแบบและ การสร้างแอปพลิ เคชันที่ใช้ในการสร้างชุดข้อมูลสำหรับการสร้างโมเดลจากวิดีโอ สิ่งที่ต้องให้ความสนใจคือฟังก์ชันการทำงาน การ ออกแบบและการจัดวางองค์ประกอบต่างๆในหน้าต่างแอปพลิเคชัน และความสะดวกในการใช้งาน จากนั้นจึงเริ่ม ศึกษาเกี่ยวกับ library ที่ใช้ในการสร้างแอปพลิเคชัน ส่วนการศึกษาเกี่ยวกับการสร้างระบบวิเคราะห์การกระทำ มนุษย์ จะมุ่งความสนใจไปที่ชุดข้อมูลสำหรับการวิเคราะห์วิดีโอ โมเดลสำหรับการวิเคราะห์วิดีโอ เทคนิคในการ สร้างโมเดล เทคโนโลยีในการระบบวิเคราะห์วิดีโอ เพื่อใช้ในการออกแบบและสร้างระบบวิเคราะห์การกระทำ ของมนุษย์ในวิดีโอด้วยมีประสิทธิภาพ ในบทนี้จะกล่าวถึงกระบวนการออกแบบและการดำเนินการตามแผนที่วาง เอาไว้

3.7 ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.1: ภาพรวมระบบของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

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 Action label เพื่อช่วยแบ่งเบ้าภาระของผู้พัฒนาในการสร้าง label สำหรับสร้างโมเดลจากข้อมูลประเทวิดีโอ โดยกระบวนการ Select จะต้องสามารถตัดวิดีโอ่ว่าที่ไม่มีมนุษย์อยู่ออกจากวิดีโอด้วย กระบวนการ Detect จะต้องหาตำแหน่งของมนุษย์ภายในวิดีโอด้วย แล้วใช้กระบวนการ Track นำรายตำแหน่งต่อไปของมนุษย์ข้อมูลตำแหน่งของมนุษย์ที่ได้จากการกระบวนการ Detect และกระบวนการ Action label นั้นต้องสามารถทำงานร่วมกับปัญญาประดิษฐ์ได้ ดังรูปที่ 3.2

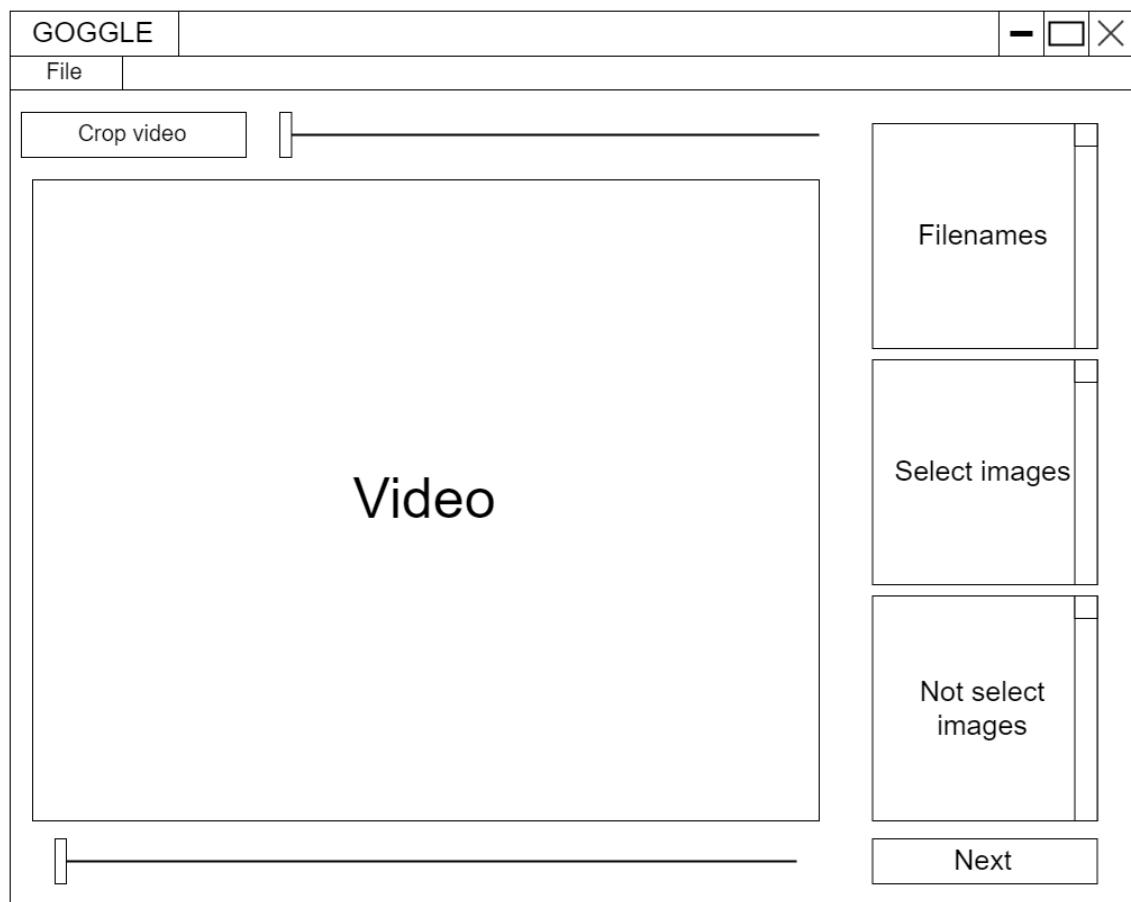


รูปที่ 3.2: กระบวนการหลักของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์

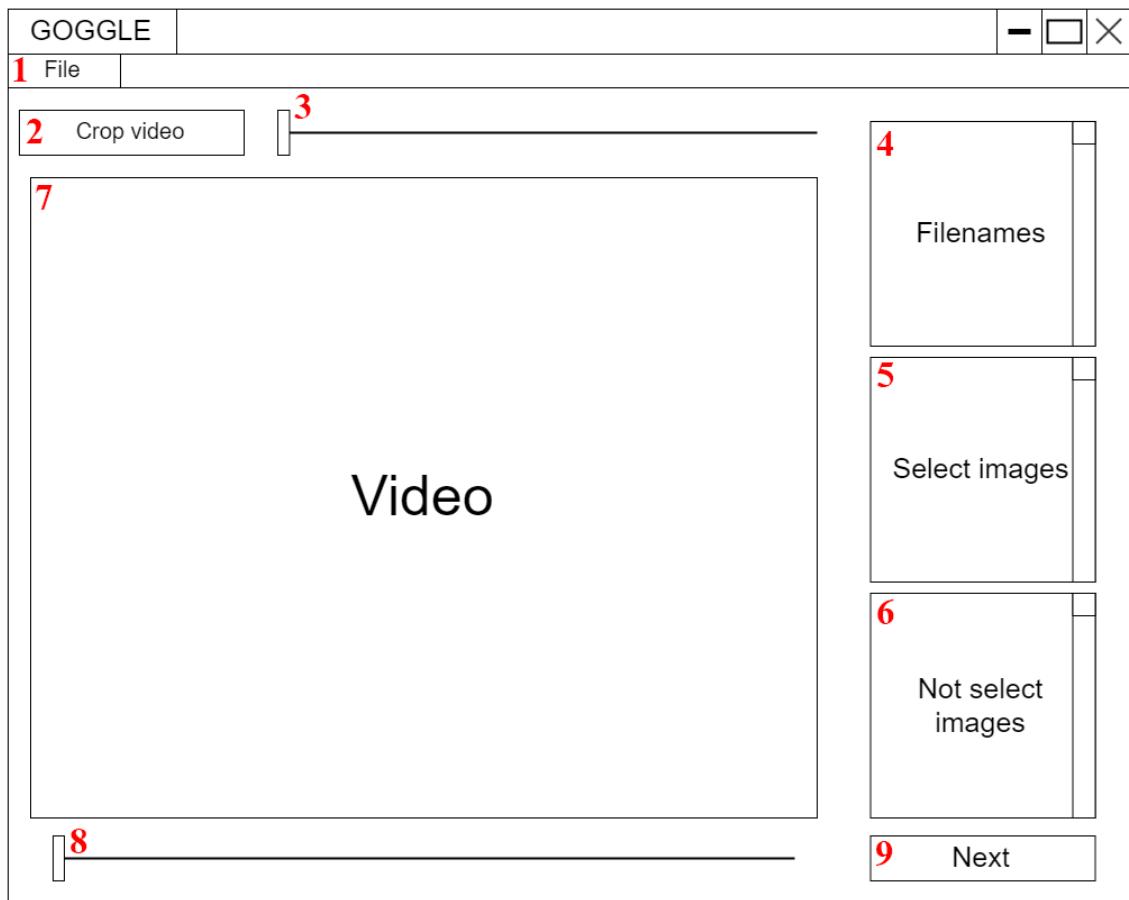
โดยแต่ละส่วนจะมีรายละเอียดดังนี้

3.8.1.1 Select

กระบวนการ Select จะต้องสามารถรับวิดีโอเข้ามา แล้วตัดวิดีโອนในช่วงที่ไม่มีมนุษย์อยู่ในเฟรม(frame)ออกได้อัตโนมัติด้วยปัญญาประดิษฐ์ แต่เนื่องจากการประมวลผลทุกเฟรมในวิดีโอนั้นจะทำให้เสียเวลามากเกินไป จึงใช้วิธีการเลือกตัวอย่างเฟรมด้วยอัตราคงที่(สามารถกำหนดได้) ซึ่งเรียกว่าเฟรมเหล่านี้ว่า คีย์เฟรม(keyframe) จากนั้นใช้ปัญญาประดิษฐ์ประมวลผลคีย์เฟรมที่เหล่านั้น เพื่อลดระยะเวลาในการประมวลผลลง และมนุษย์จะต้องสามารถแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.3



รูปที่ 3.3: หน้าต่าง Select ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



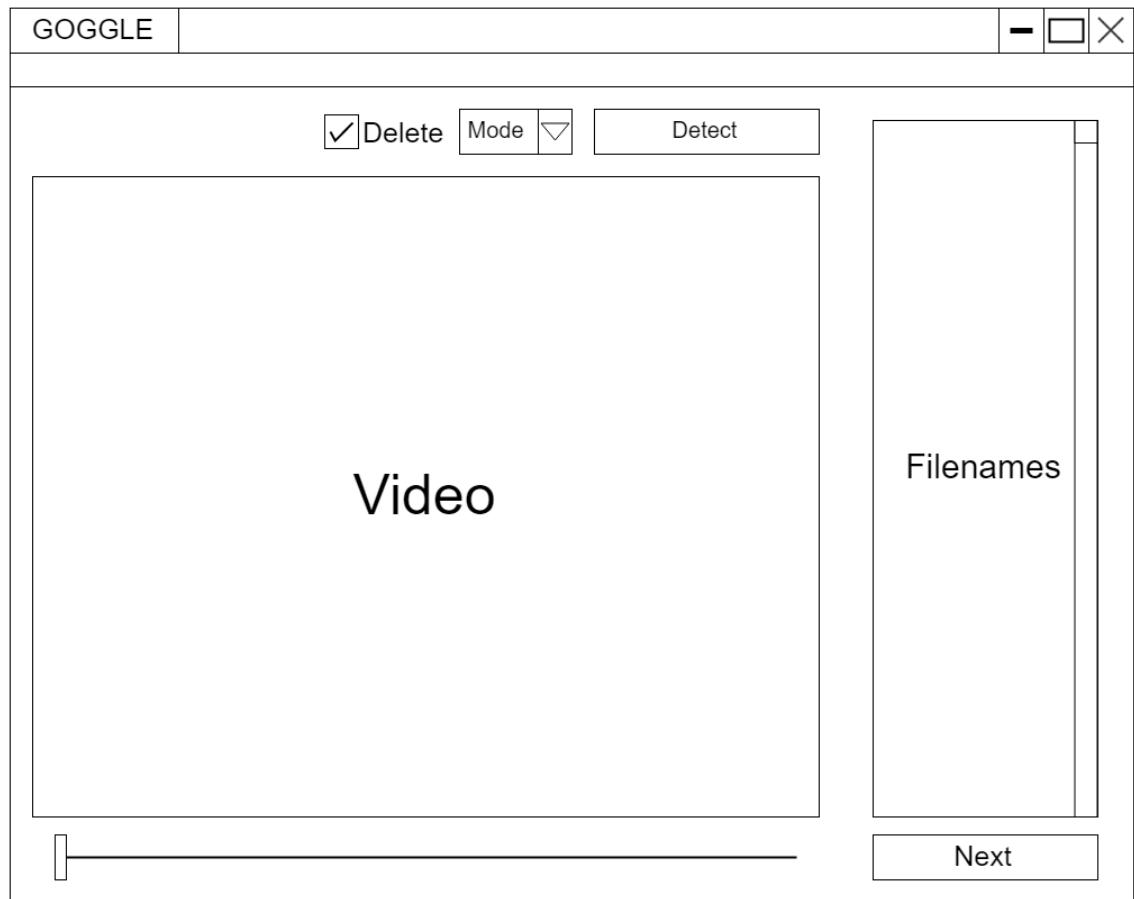
รูปที่ 3.4: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Select

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.4 มีรายละเอียดดังนี้

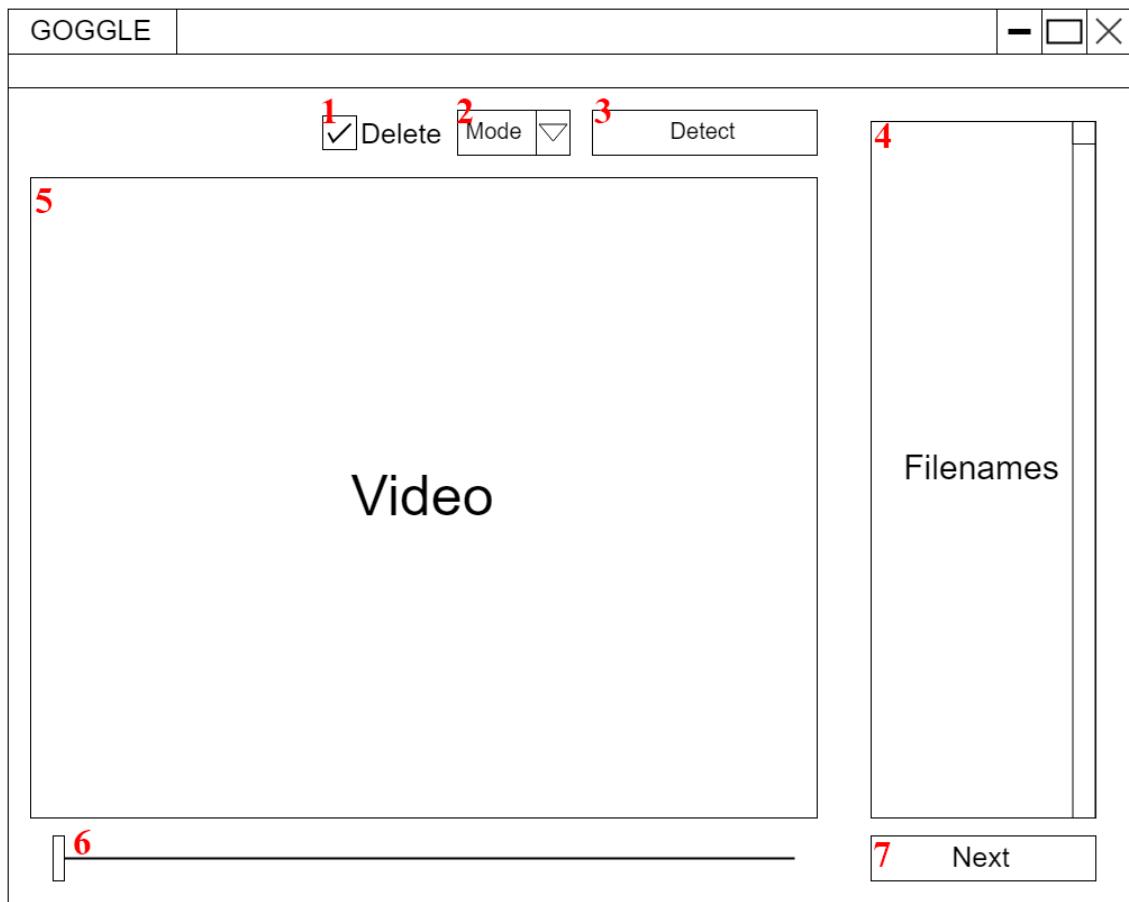
1. หมายเลข 1 คือปุ่มสำหรับเลือกไฟล์วิดีโอที่ต้องการจากในคอมพิวเตอร์เข้ามาในโปรแกรม
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบทำการสร้างคีย์เฟรมขึ้นมา แล้วใช้ปัญญาประดิษฐ์ประมวลผลเพื่อแยกคีย์เฟรมในหนีมีคนอยู่ และคีย์เฟรมไม่มีคนอยู่ แบบอัตโนมัติ(Auto mode)
3. หมายเลข 3 คือแถบเลื่อนเพื่อกำหนดความถี่ในการหยิบคีย์เฟรม โดยจะมีช่วงอยู่ที่ 1 เฟรมต่อวินาที จนถึงเฟรมต่อวินาทีสูงสุดของวิดีโอที่รับเข้ามา
4. หมายเลข 4 คือกล่องสำหรับแสดงชื่อวิดีโอที่รับเข้ามาในโปรแกรมเพื่อเลือกเข้ามาใช้ในการประมวลผล
5. หมายเลข 5 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
6. หมายเลข 6 คือกล่องสำหรับแสดงว่าคีย์เฟรมได้มีมนุษย์อยู่ในเฟรม โดยที่ผู้ใช้งานสามารถตรวจสอบความถูกต้องและแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้
7. หมายเลข 7 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 5 หมายเลข 6 หรือหมายเลข 8
8. หมายเลข 8 คือแถบเลื่อนสำหรับเลือนดูคีย์เฟรมทั้งหมดที่ระบบสร้างขึ้น
9. หมายเลข 9 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

3.8.1.2 Detect

กระบวนการ Detect จะต้องสามารถรับคีย์เฟรมจากการบันทึกการ Select มาประมวลผลด้วยปัญญาประดิษฐ์เพื่อหาตำแหน่งของมนุษย์ที่อยู่ในคีย์เฟรม และสร้างกรอบสีเหลี่ยมครอบบริเวณดังกล่าวได้ในแบบอัตโนมัติ เพื่อแบ่งเบาภาระผู้ใช้ในการที่ต้องสร้างกรอบสีเหลี่ยมครอบตำแหน่งของมนุษย์ด้วยตัวเอง และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสีเหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของปัญญาประดิษฐ์ เพื่อเพิ่มคุณภาพของชุดข้อมูล จึงออกแบบหน้าต่างได้ดังรูปที่ 3.5



รูปที่ 3.5: หน้าต่าง Detect ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



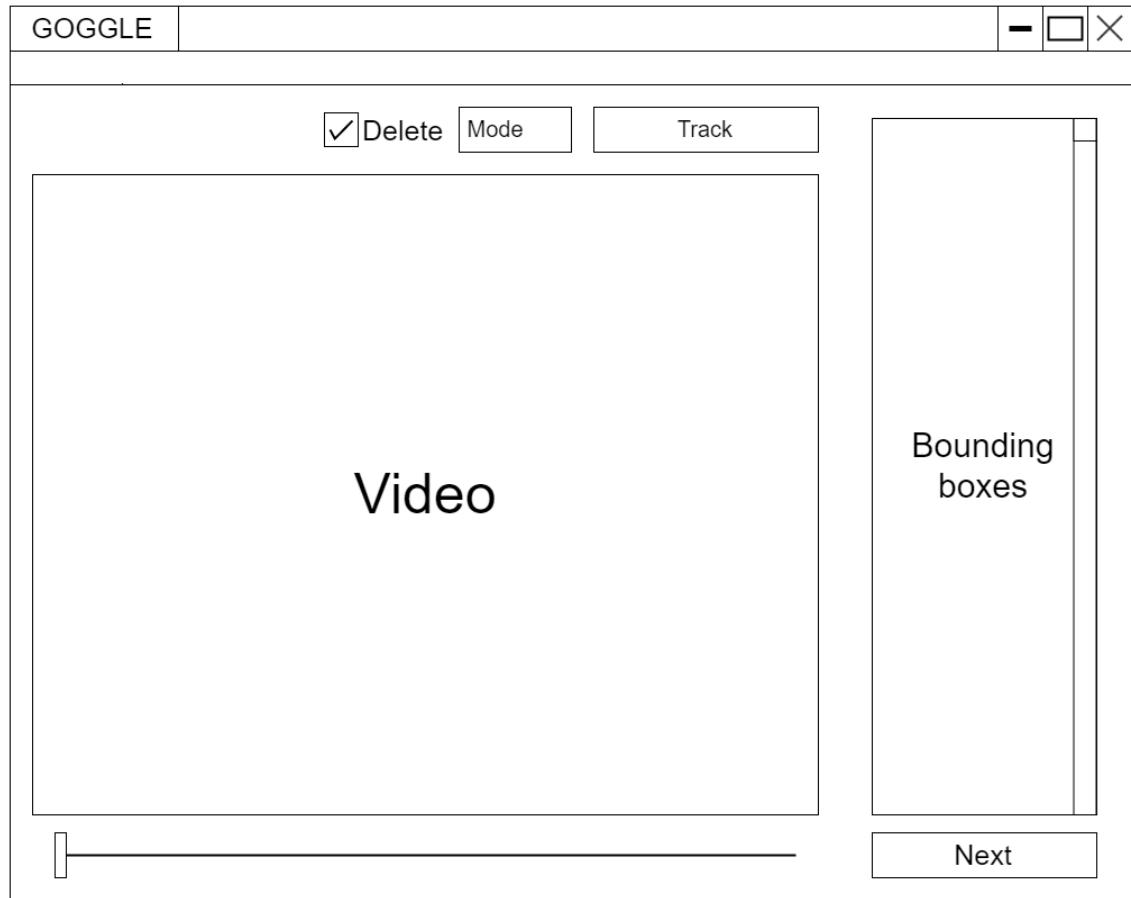
รูปที่ 3.6: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Detect

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.6 มีรายละเอียดดังนี้

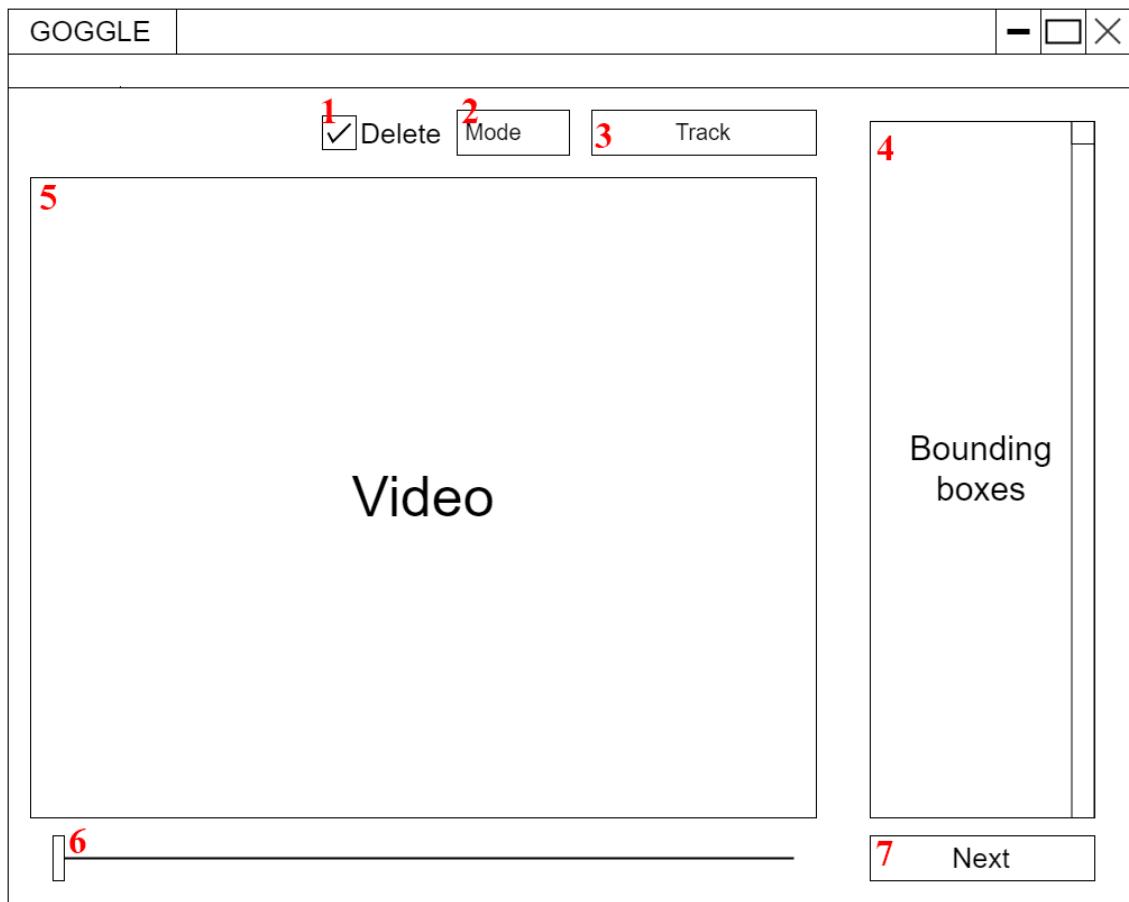
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเอง(Manual mode) เป็นลบรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจหาตำแหน่งของมนุษย์ในคิร์เฟรมทั้งหมดแล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด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

3.8.1.3 Track

เนื่องจากกระบวนการ Detect นั้นจะทำเฉพาะในคีย์เฟรมทำให้ในเฟรมอื่นๆ นอกเหนือจากนั้นจะไม่มีกรอบสี่เหลี่ยมอยู่ ดังนั้นกระบวนการ Track จึงต้องสามารถทำนายตำแหน่งต่อไปของมนุษย์แล้วสร้างกรอบสี่เหลี่ยมขึ้นมาบนเฟรมระหว่างคีย์เฟรมทั้งหมดได้โดยอัตโนมัติ เพื่อสร้างข้อมูลตำแหน่งของมนุษย์ในเฟรมเหล่านั้น และผู้ใช้ต้องสามารถสร้างหรือลบกรอบสี่เหลี่ยมได้ด้วยตัวเองสำหรับแก้ไขความผิดพลาดของอัลกอริทึม จึงออกแบบหน้าต่างได้ดังรูปที่ 3.7



รูปที่ 3.7: หน้าต่าง Track ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



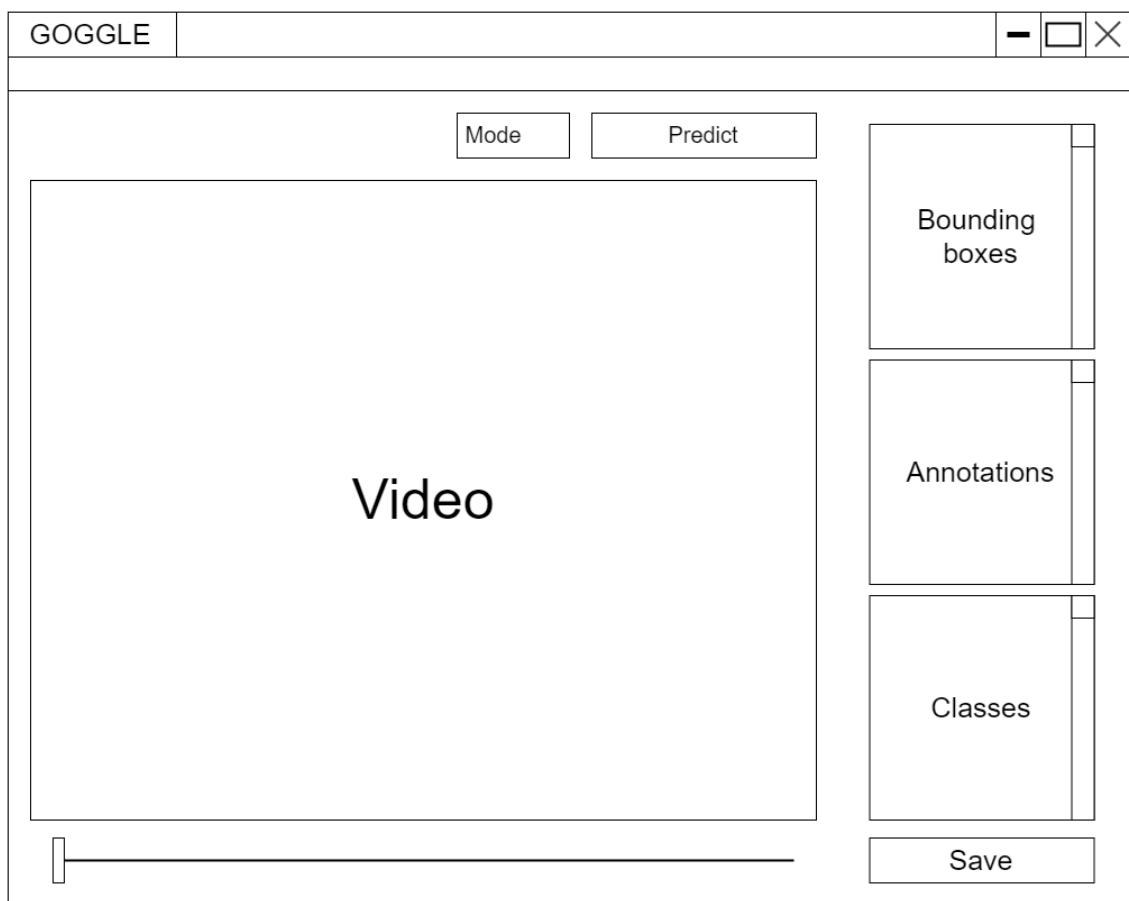
รูปที่ 3.8: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Track

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

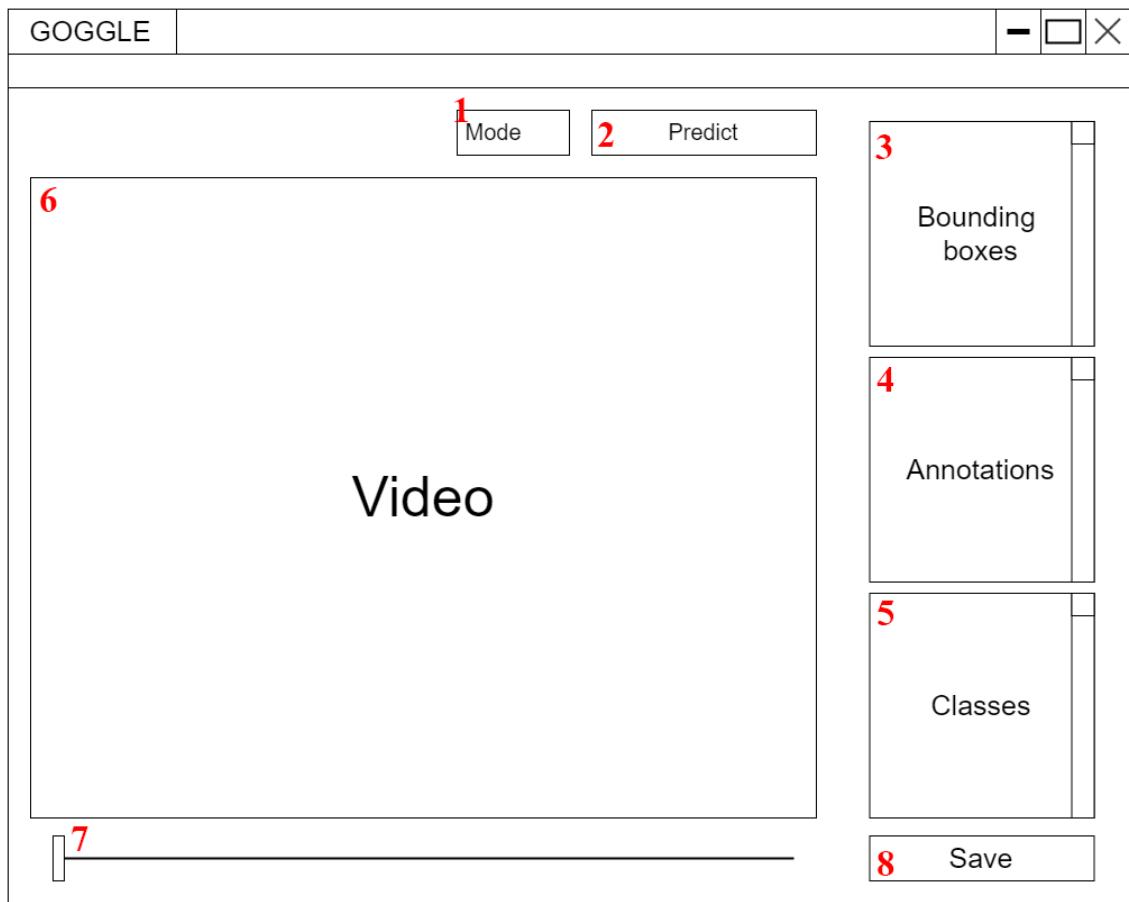
1. หมายเลข 1 คือช่องสำหรับกดเพื่อเปลี่ยนระบบจากสร้างกรอบสี่เหลี่ยมในแบบแก้ไขด้วยตนเอง(Manual mode) เป็นลับกรอบสี่เหลี่ยมแทน
2. หมายเลข 2 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
3. หมายเลข 3 คือปุ่มสำหรับสั่งให้ระบบทำการตรวจสอบตำแหน่งของมนุษย์ในเฟรมระหว่างคิ้ยวเฟรมทั้งหมด แล้วสร้างกรอบสี่เหลี่ยมขึ้นมาครอบบริเวณที่กำหนด
4. หมายเลข 4 คือกล่องสำหรับแสดงกรอบสี่เหลี่ยมทั้งหมดที่อยู่ในเฟรม
5. หมายเลข 5 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 6
6. หมายเลข 6 คือแถบเลื่อนสำหรับเลื่อนดูเฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของอัลกอริทึม
7. หมายเลข 7 คือปุ่มสำหรับไปกระบวนการต่อไปหลังจากระบบประมวลผลเสร็จแล้ว

3.8.1.4 Action label

กระบวนการ Action label นั้นต้องสามารถทำนายว่าการกระทำ(Action)ของมนุษย์ที่อยู่ในแต่ละเฟรมว่าคืออะไร ได้โดยอัตโนมัติด้วยปัญญาประดิษฐ์ และผู้ใช้จะต้องสามารถแก้ไขข้อผิดพลาดของปัญญาประดิษฐ์ได้หากมีการทำนายที่ผิดพลาดเกิดขึ้น หรือถ้าหากผู้ใช้ต้องการเพิ่มการกระทำที่ไม่ได้มีอยู่ในชุดการกระทำพื้นฐานที่มีอยู่แล้วของปัญญาประดิษฐ์ ผู้ใช้ก็สามารถเพิ่มการกระทำนั้นเข้ามาได้ จึงออกแบบหน้าต่างเดี๋ยวๆ รูปที่ 3.9



รูปที่ 3.9: หน้าต่าง Action label ของเครื่องมือสำหรับกำกับข้อมูลด้วยปัญญาประดิษฐ์



รูปที่ 3.10: ตำแหน่งของแต่ละวิดเจ็ตในหน้าต่าง Action label

โดยที่แต่ละวิดเจ็ตตามหมายเลขที่กำหนดตามรูปที่ 3.8 มีรายละเอียดดังนี้

1. หมายเลข 1 คือช่องสำหรับเลือกว่าจะใช้ระบบแบบใด ระหว่างแบบอัตโนมัติและแบบแก้ไขด้วยตนเอง
2. หมายเลข 2 คือปุ่มสำหรับสั่งให้ระบบคำนวณการทำงานของมนุษย์ในทุกๆเฟรม
3. หมายเลข 3 คือกล่องสำหรับแสดงกรอบสี่เหลี่ยมทั้งหมดที่อยู่ในเฟรมที่เลือก
4. หมายเลข 4 คือกล่องสำหรับแสดงการกระทำการของมนุษย์แต่ละคนที่อยู่ในเฟรมที่เลือก โดยจะเรียงลำดับคู่กับกรอบสี่เหลี่ยมที่อยู่ในช่องหมายเลข 3
5. หมายเลข 5 คือกล่องสำหรับแสดงชุดการกระทำการที่ปัญญาประดิษฐ์มีอยู่แล้ว ซึ่งในการทำงานแบบแก้ไขด้วยตนเองนั้น จะสามารถค้นหาการกระทำการที่มีอยู่แล้วได้ และหากคำที่ใส่เขามานั้นมีอยู่ในชุดการกระทำการที่จะเป็นการเพิ่มการกระทำนั้นเข้ามาแทน
6. หมายเลข 6 คือหน้าต่างสำหรับแสดงเฟรมที่เลือกจากหมายเลข 7
7. หมายเลข 7 คือແຄบเลื่อนสำหรับเลื่อนดูเฟรมทั้งหมดที่มี เพื่อตรวจสอบความถูกต้องของปัญญาประดิษฐ์
8. หมายเลข 8 คือปุ่มสำหรับสร้างไฟล์ XML ของทุกๆเฟรมสำหรับใช้ในการสร้างโมเดลโดยรายละเอียดข้อมูลภายในไฟล์ XML จะอยู่ในหัวข้อ 3.8.1.5

3.8.1.5 รายละเอียดข้อมูลภายในไฟล์ XML

ไฟล์ XML นั้นเป็นรูปแบบที่นิยมใช้ในการเก็บข้อมูลสำหรับการสร้างโมเดลประมวลผลจับวัตถุ(object detection) โดยจะเก็บข้อมูลในรูปแบบของ PASCAL VOC ที่นิยมใช้ในการสร้างโมเดลด้วย library Tensorflow โดยภายในไฟล์จะมีข้อมูลดังรูปที่ 3.11 โดยข้อมูลส่วนสำคัญของรูปแบบนี้นั้นจะถูกใส่หมายเลขกำกับไว้ซึ่งแต่ละ

```

<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

หมายเลขนั้นหมายถึง

1. หมายเลข 1 คือชื่อโฟลเดอร์ที่เก็บไฟล์รูปภาพที่เกี่ยวข้องกับไฟล์ XML นี้อยู่
2. หมายเลข 2 คือชื่อไฟล์ที่เกี่ยวข้องกับไฟล์ XML นี้
3. หมายเลข 3 คือเส้นทางในคอมพิวเตอร์(directory path)ของไฟล์รูปภาพที่เกี่ยวข้องกับไฟล์ XML นี้
4. หมายเลข 4 คือขนาดและมิติของรูปภาพ ซึ่งจะประกอบด้วยความกว้าง(width) ความยาว(height) และ จำนวนช่องสี(depth) โดยที่จำนวนช่องสีที่มีความลึก 3 มักจะหมายถึงภาพสี RGB และจำนวนช่องสีที่มี ความลึก 2 จะหมายถึงภาพขาวดำ(gray scale)
5. หมายเลข 5 คือ label ของวัตถุหรืออย่างอื่น ที่อยู่ในกรอบสีเหลี่ยมที่ถูกกำหนดไว้ในส่วนของหมายเลข 6
6. หมายเลข 6 คือ กรอบสีเหลี่ยมที่ครอบวัตถุที่สนใจ เช่นมนุษย์ เป็นต้น

3.9 การออกแบบการทดสอบการตรวจจับวัตถุ

3.9.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
2. ความแม่นยำของกรอบสี่เหลี่ยม (IOU)

สมมุติฐาน

1. ผู้วิจัยได้ตั้งสมมุติฐานว่า การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุ จะได้ผลลัพธ์จะใกล้เคียงกับ แหล่งข้อมูล (ลำดับของโมเดลปัญญาประดิษฐ์ที่เร็วที่สุด หรือแม่นยำที่สุดตามลำดับ) ดังนั้นผู้วิจัยจึงได้ตั้ง สมมุติฐาน ว่า โมเดลปัญญาประดิษฐ์ YOLOv3-320 จะเหมาะกับงานวิจัยนี้มากที่สุด เนื่องจากมีความเร็ว ในการทำงาน และมีความแม่นยำในระดับที่ผู้วิจัยยอมรับได้

ตัวแปร

1. โมเดลปัญญาประดิษฐ์ ซึ่งได้แก่
 - (a) Tiny YOLO
 - (b) YOLOv3-tiny
 - (c) SSD300
 - (d) YOLOv3-320
 - (e) YOLOv2 608x608

ตัวแปรควบคุม

1. ชุดข้อมูล : The validation split of AVA v2.1

วิธีการทดลอง

1. ดาวน์โหลดชุดข้อมูล The validation split of AVA v2.1
2. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และ ชุดข้อมูลที่มีคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ,เฟรม
 - (b) ชุดข้อมูลที่มีคำตอบ ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสี่เหลี่ยม
3. เรียกชื่อและเฟรมของวิดีโอจากชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่มีคำตอบ ผลงาน จากนั้น เก็บผลลัพธ์เป็น ชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสี่เหลี่ยม
4. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และ ชุดข้อมูลที่มีคำตอบ ผ่านฟังก์ชัน คำนวนค่า IOU
5. เปรียบเทียบผลลัพธ์จากแหล่งที่มา

3.10 การออกแบบการทดสอบการทำงานของระบบนำทางตำแหน่งต่อไปของมนุษย์

3.10.1 ทดสอบประสิทธิภาพการทำงานของระบบนำทางตำแหน่งต่อไปของวัตถุในวิดีโอ สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อวิดีโอ (วินาที)
2. ความแม่นยำ (อัตราส่วนร่วมของกรอบที่เหลืออยู่ หรือ Intersection of Union)

สมมุติฐาน

ผู้วิจัยได้ตั้งสมมุติฐานว่า การใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับวัตถุและสร้างกรอบสี่เหลี่ยมทุกๆ N เฟรม แล้วใช้ระบบนำทางตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น จะทำให้ระบบสามารถทำงานได้เร็วขึ้น โดยที่ประสิทธิภาพจะลดลงเพียงเล็กน้อย

ตัวแปรควบคุม

1. วิดีโอสารบันทุณฑ์ไม่ติดลิขสิทธิ์ ความยาวประมาณ 120 - 180 วินาที หนึ่งวิดีโอ
2. ใช้โมเดลปัญญาประดิษฐ์สำหรับตรวจจับตำแหน่งวัตถุ ResNet-50 ในการสร้างชุดข้อมูลที่มีการกำกับตำแหน่งวัตถุไว้ (ground-truth) เพื่อใช้เป็นคำตوبของการนำทาง
3. โมเดลปัญญาประดิษฐ์สำหรับตรวจจับตำแหน่งที่ใช้ในการเปรียบเทียบ: YOLO-V3 320
4. อัลกอริทึมสำหรับระบบนำทางตำแหน่งต่อไปของวัตถุ: dlib

วิธีการทดลอง

แบ่งการทดลองออกเป็น 3 รูปแบบ เพื่อหารูปแบบที่เหมาะสม (ความเร็วที่ได้เหมาะสมกับความแม่นยำ) ดังนี้

1. ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกเฟรมในวิดีโอ และเปรียบเทียบผลลัพธ์กับชุดข้อมูลที่ถูกกำกับตำแหน่งวัตถุไว้แล้ว เพื่อคำนวณหาความแม่นยำ
2. ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลเพียงเฟรมแรกของวิดีโอ แล้วใช้ระบบนำทางตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมที่เหลือ และเปรียบเทียบผลลัพธ์กับชุดข้อมูลที่ถูกกำกับตำแหน่งวัตถุไว้แล้ว เพื่อคำนวณหาความแม่นยำ
3. ใช้โมเดลปัญญาประดิษฐ์ YOLO-v3 320 ประมวลผลทุกๆ N เฟรมในวิดีโอ แล้วใช้ระบบนำทางตำแหน่งต่อไปของวัตถุในการสร้างกรอบสี่เหลี่ยมในเฟรมระหว่างนั้น และเปรียบเทียบผลลัพธ์กับชุดข้อมูลที่ถูกกำกับตำแหน่งวัตถุไว้แล้ว เพื่อคำนวณหาความแม่นยำ
4. เปรียบเทียบทั้ง 3 รูปแบบ และสรุปผลการทดลอง

3.11 การออกแบบการทดสอบการระบุตัวตนของมนุษย์

3.11.1 ทดสอบประสิทธิภาพการทำงานของระบบระบุตัวตนของบุคคลภายนอกในภาพสิ่งที่ใช้ในการวัดผล

- ความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอก

สมมุติฐาน :

ผู้จัดได้ตั้งสมมุติฐานว่า ผลลัพธ์ของการทดลองการใช้งานจริงของโมเดลปัญญาประดิษฐ์ ResNet50 ที่ผ่านการ train มาด้วยชุดข้อมูล Market1501 นั้นควรจะมีความแม่นยามากที่สุดในการระบุตัวตนของบุคคลภายนอกเมื่อเทียบกับโมเดลปัญญาประดิษฐ์ เพราะเมื่อเทียบกับโมเดลปัญญาประดิษฐ์อื่นที่มาจากการแหล่งข้อมูลเดียวกันโมเดลปัญญาประดิษฐ์ ResNet50 ที่ผ่านการ train มาด้วยชุดข้อมูล Market1501 นั้นมีความแม่นยามาก

สูงสุด

ตัวแปร

- โมเดลปัญญาประดิษฐ์ซึ่งได้แก่
 - ResNet50 ของชุดข้อมูล Market1501
 - ResNet50 ของชุดข้อมูล DukeMTMCReID
 - ResNet50 ของชุดข้อมูล CUHK03
 - ResNet50 ของชุดข้อมูล MSMT17

ตัวแปรควบคุม

- ชุดข้อมูล : ชุดข้อมูลที่ทางผู้จัดสร้างขึ้นสำหรับการทดสอบ
- โมเดลปัญญาประดิษฐ์ : YoLo-V3 320 สำหรับการทำหน้าของบุคคล

วิธีการทดลอง

- ดาวน์โหลดโมเดลปัญญาประดิษฐ์ที่ผ่านการ train ด้วยชุดข้อมูลต่างๆได้แก่ Market1501 , DukeMTMCReID, CUHK03 และ MSMT17
- นำชุดข้อมูลที่ผู้จัดสร้างขึ้นมาผ่านโมเดลปัญญาประดิษฐ์ YoLo-V3 320 เพื่อหาตำแหน่งของบุคคล
- นำโมเดลปัญญาประดิษฐ์แต่ละอันมาทดสอบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอกในภาพด้วยตำแหน่งของบุคคลที่ได้มาจากการก่อนหน้านี้
- ประเมินผลการทำงานโดยเทียบความแม่นยำสำหรับการระบุตัวตนของบุคคลภายนอกในภาพของแต่ละโมเดลปัญญาประดิษฐ์ เพื่อหาโมเดลปัญญาประดิษฐ์ที่ได้ผลลัพธ์ดีที่สุด

3.12 การออกแบบการทดสอบการจดจำการกระทำของมนุษย์

3.12.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA โดยใช้ชุดข้อมูลของ AVA ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมติฐาน :

ผู้จัยได้ตั้งสมมติฐานว่า ผลลัพธ์ของการทดลองจะมีความแม่นยำเทียบเท่ากับผลลัพธ์จากแหล่งที่มา แต่ความเร็วต่อรูปภาพจะมีความเร็วน้อยกว่าผลลัพธ์จากแหล่งที่มา เนื่องจาก แหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X card ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่า กราฟิกการ์ดของผู้จัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

1. ชุดข้อมูล : The validation split of AVA v2.1
2. Machine learning model : Faster rcnn resnet101 ava v2.1

วิธีการทดลอง

1. ดาวน์โหลดชุดข้อมูล The validation split of AVA v2.1
2. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และ ชุดข้อมูลที่มีคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำตอบ ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสี่เหลี่ยม,ไอเดียของการกระทำ
3. เรียกชื่อของวิดีโอจากชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน

 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสี่เหลี่ยม,ไอเดียของการกระทำ,ความมั่นใจ

4. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และ ชุดข้อมูลที่มีคำตอบ
5. เปรียบเทียบผลลัพธ์จากแหล่งที่มา

3.12.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

สิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมุติฐาน :

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

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลที่ผู้วิจัยสร้าง ถูกสร้างด้วย A.I. assisted labeling tool
2. Machine learning model : Faster rcnn resnet101 ava v2.1

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และ ชุดข้อมูลที่มีคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำตอบ ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสีเหลี่ยม,โฉดีของการกระทำ
2. เรียกชื่อของวิดีโอจากชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสีเหลี่ยม,โฉดีของการกระทำ,ความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และ ชุดข้อมูลที่มีคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

3.12.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรน์ผ่านชุดข้อมูลสำหรับการ tren์ที่ผู้วิจัยสร้างขึ้น และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้นในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิงสิ่งที่ใช้ในการวัดผล

1. ความเร็วต่อรูปภาพ (วินาที)
2. ความแม่นยำ (PASCAL mAP)

สมมุติฐาน :

ผู้วิจัยได้ตั้งสมมุติฐานว่าผลลัพธ์ของการทดลองจะมีความแม่นยำสูงขึ้นเมื่อเทียบกับความแม่นยำของการทดลองที่ผ่านมา เนื่องจาก โมเดลปัญญาประดิษฐ์ในการทดลองนี้ เป็นโมเดลปัญญาประดิษฐ์ที่ผู้วิจัยได้สร้างขึ้นซึ่งจะมีหมวดหมู่ของกรอบทำของโมเดลปัญญาประดิษฐ์และชุดข้อมูลทดสอบตรงกัน ในส่วนของความเร็วต่อรูปภาพจะมีความเร็วน้อยกว่าผลลัพธ์จากแหล่งที่มา เนื่องจาก แหล่งที่มาของข้อมูลได้ทำการทดสอบโดยใช้กราฟิกการ์ดรุ่น Nvidia GeForce GTX TITAN X card ซึ่งเป็นกราฟิกการ์ดที่มีประสิทธิภาพการทำงานดีกว่ากราฟิกการ์ดของผู้วิจัย จึงทำให้สามารถทดสอบด้วยความเร็วที่มากกว่า

ตัวแปรควบคุม

1. ชุดข้อมูล : ชุดข้อมูลที่ผู้วิจัยสร้าง ถูกสร้างด้วย A.I. assisted labeling tool
2. Machine learning model : Faster rcnn resnet101 ava v2.1

วิธีการทดลอง

1. แบ่งชุดข้อมูลออกเป็น ชุดข้อมูลสำหรับทดสอบ และ ชุดข้อมูลที่มีคำตอบ
 - (a) ชุดข้อมูลสำหรับทดสอบ ประกอบด้วย : ชื่อของวิดีโอ
 - (b) ชุดข้อมูลที่มีคำตอบ ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสีเหลี่ยม,oids ของกรากะทำ
2. เรียกชื่อของวิดีโอด้วยชุดข้อมูลทดสอบ และนำโมเดลปัญญาประดิษฐ์ที่นายผลลัพธ์ จากนั้นเก็บผลลัพธ์ เป็นชุดข้อมูลผลลัพธ์จากการทำงาน
 - (a) ชุดข้อมูลผลลัพธ์จากการทำงาน ประกอบด้วย : ชื่อของวิดีโอ,เฟรม,ตำแหน่งของกรอบสีเหลี่ยม,oids ของกรากะทำ,ความมั่นใจ
3. ประเมินผลการทำงานโดยเทียบระหว่างชุดผลลัพธ์จากการทำงาน และ ชุดข้อมูลที่มีคำตอบ
4. เปรียบเทียบผลลัพธ์กับผลการทดลองที่ผ่านมา

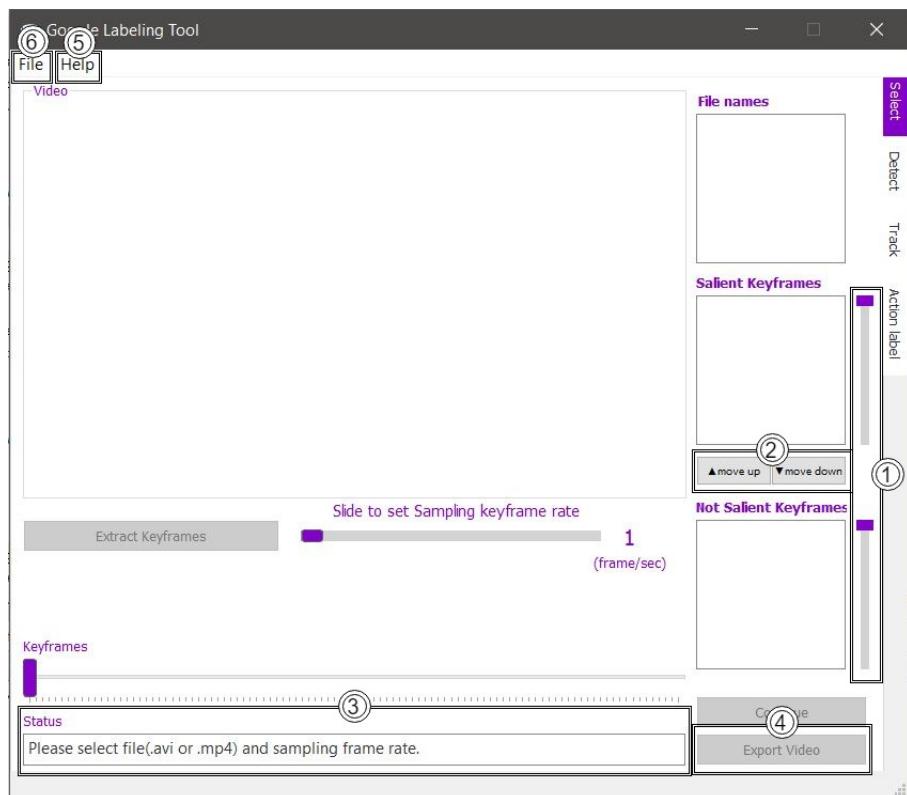
บทที่ 4

ผลการดำเนินงาน

4.1 Labeling tool

4.1.1 หน้าต่างแสดงผลของแอพพลิเคชัน

หน้าต่าง Select

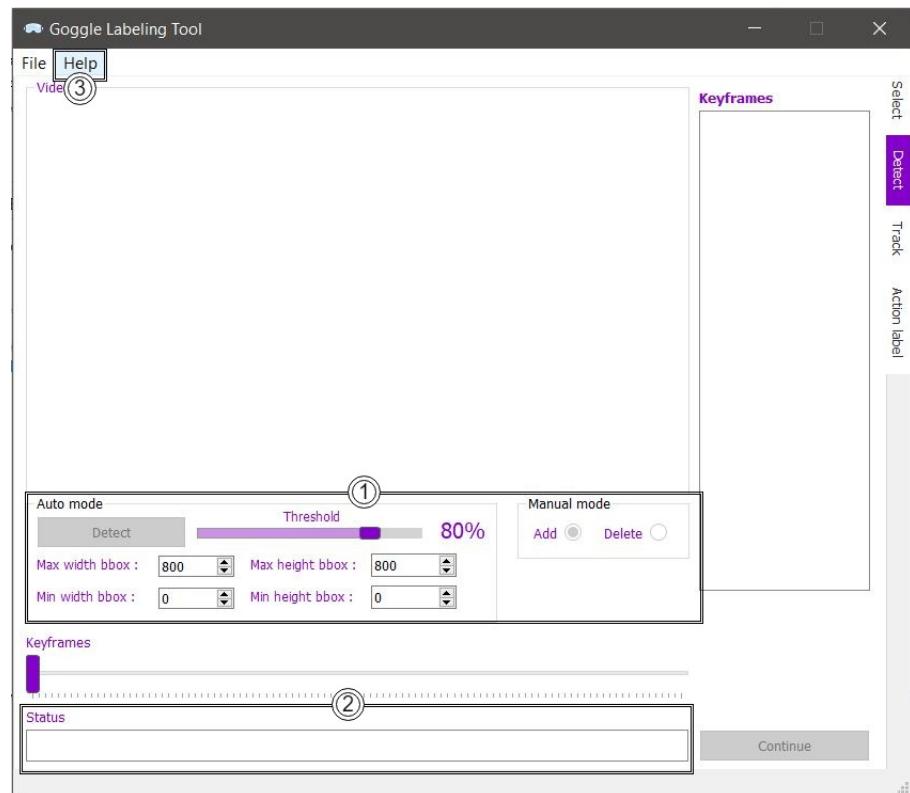


รูปที่ 4.1: รูปหน้าต่างแสดงผลของหน้าต่าง Select

จากรูปที่ 4.1 แสดงหน้าต่าง Select ของแอพพลิเคชัน ซึ่งเมื่อเทียบกันกับหน้าต่าง Select ในฉบับร่าง (??) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. แถบเลื่อนสำหรับเลื่อนคุณูปกรณ์ที่มีมนุษย์ หรือ ไม่มีมนุษย์ เพื่อเพิ่มความสะดวกในการเลือกคุณูปกรณ์
2. ปุ่มสำหรับแก้ไขคุณูปกรณ์ที่มีมนุษย์หรือไม่มีมนุษย์
3. แถบแสดงสถานะกระบวนการทำงาน
4. ปุ่มสำหรับนำผลลัพธ์ออกเป็นไฟล์วิดีโอด้วยไฟล์ในช่วงที่มีมนุษย์อยู่
5. แถบสำหรับคำแนะนำช่วยเหลือ
6. ปุ่มสำหรับเปิดไฟล์

หน้าต่าง Detect

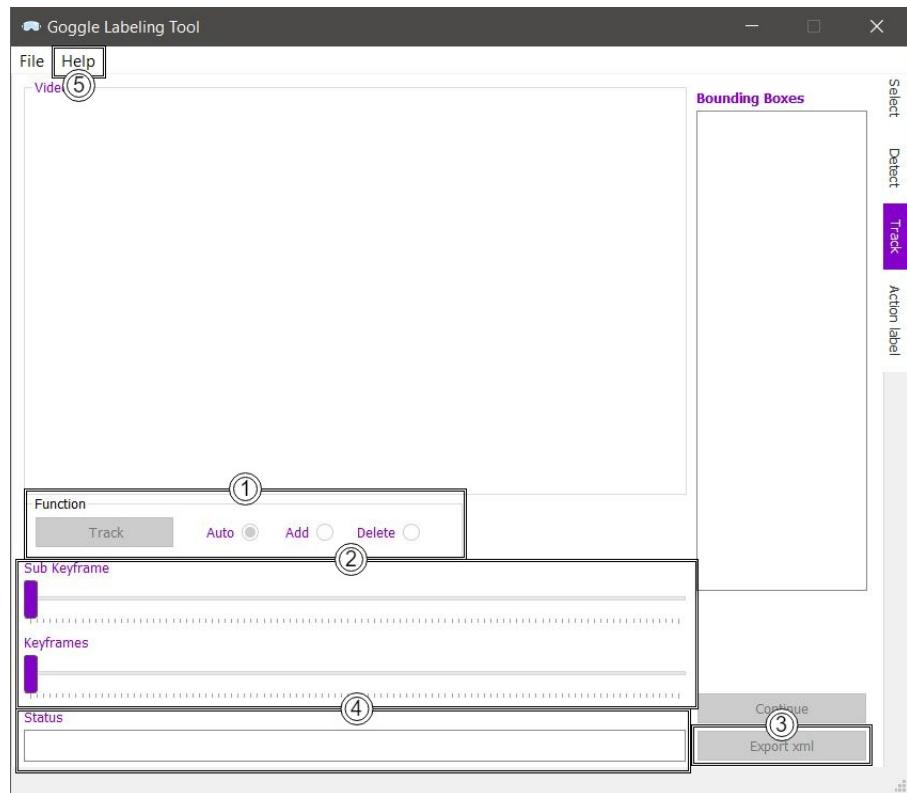


รูปที่ 4.2: รูปหน้าต่างแสดงผลของหน้าต่าง Detect

จากรูปที่ 4.2 แสดงหน้าต่าง Detect ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับหน้าต่าง Detect ในฉบับร่าง (??) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่ด้วยการทำงานแบบอัตโนมัติและกำหนดเองสามารถใช้งานได้สะดวกขึ้น และ เพิ่มความหลากหลายในการปรับแก้จากการทำงานอัตโนมัติ
2. ແຕบแสดงสถานะกระบวนการทำงาน
3. ແຕบสำหรับคำแนะนำช่วยเหลือ

หน้าต่าง Track

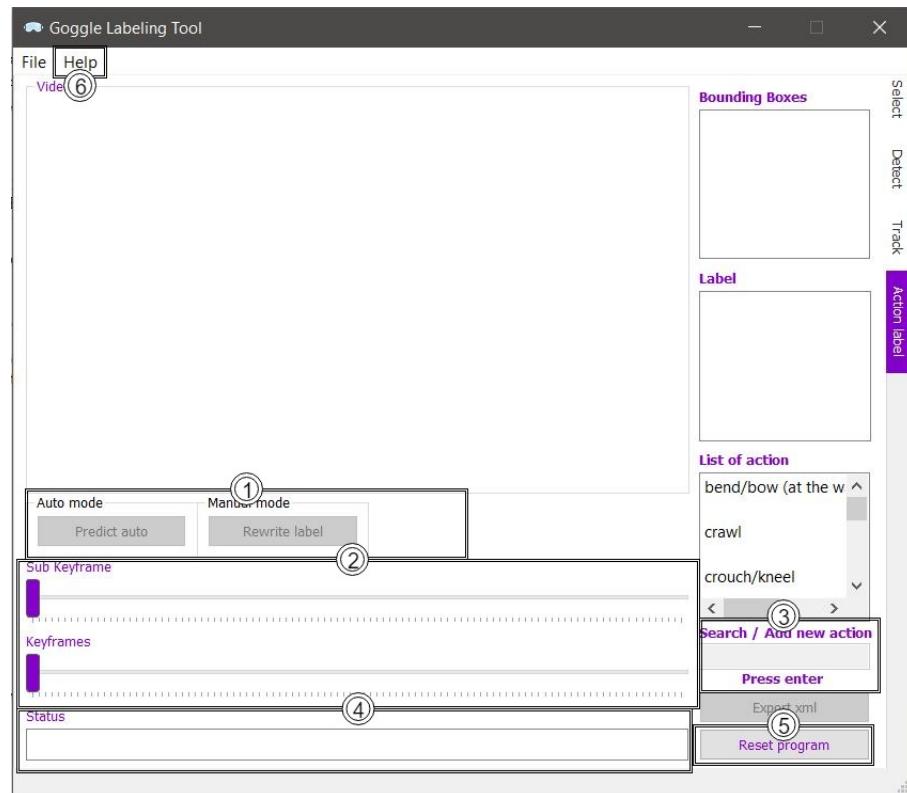


รูปที่ 4.3: รูปหน้าต่างแสดงผลของหน้าต่าง Track

จากรูปที่ 4.3 แสดงหน้าต่าง Track ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับหน้าต่าง Track ในฉบับร่าง (??) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແຄບເລືອນ ເປັນ 2 ແຄບເລືອນທຳໃຫ້ສາມາດຄຸ້ມືກິງໄຟຣ໌ແລະແຄບເລືອນທີ່ອູ່ຮ່ວ່າງໜີ່ຈຳກັດໄຟຣ໌ໄດ້ສະດວກขື້ນ
3. เพิ่ມປຸ່ມສໍາຫຼັບນຳພລັດພົບອອກເປັນໄຟລ໌ XML
4. ແຄບແສດງສຕານະກະບວນການທຳງານ
5. ແຄບສໍາຫຼັບຄໍາແນະນຳໜ່ວຍເຫຼືອ

หน้าต่าง Label



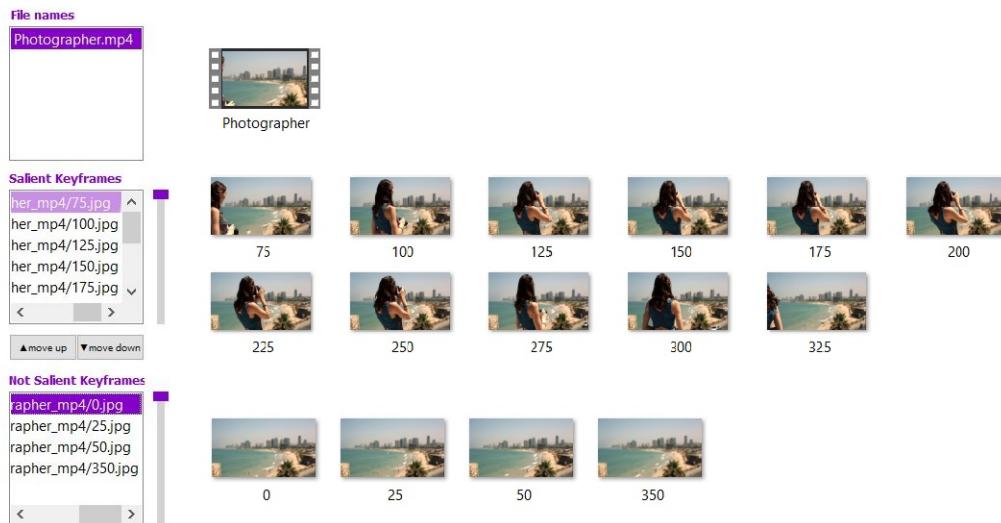
รูปที่ 4.4: รูปหน้าต่างแสดงผลของหน้าต่าง Label

จากรูปที่ 4.4 แสดงหน้าต่าง Label ของแอพพลิเคชัน ซึ่งเมื่อเทียบกับหน้าต่าง Label ในฉบับร่าง (??) จะมีส่วนที่เพิ่มเติมขึ้นมาดังนี้

1. ปรับหน้าตาใหม่จากการทำงานแบบอัตโนมัติและกำหนดเองจากฉบับร่างเพื่อให้สามารถใช้งานได้สะดวกขึ้น
2. เพิ่มແຕບเลื่อน เป็น 2 ແຕບເລື່ອນທາໃຫ້ສາມາດຄຸ້ມືກິງໄຟມີເພື່ອໃຊ້ກຳນົດທີ່ອູ່ຮ່ວມມືນ
3. ເພີ້ມສຳຫຼັບຄົນຫາຫຼືເພີ້ມໜົວດໝູ່ຂອງການກະທຳ
4. ແຕບແສດງສະຖານະກະບວນການກະທຳ
5. ປຸ່ມສຳຫຼັບເຮີ່ມຕົ້ນການກະທຳໃໝ່
6. ແຕບສຳຫຼັບຄຳແນະນຳໜໍ່ຢ່າງເລື້ອ

4.1.2 ผลลัพธ์การทำงานในแต่ละหน้าต่างของแอปพลิเคชัน

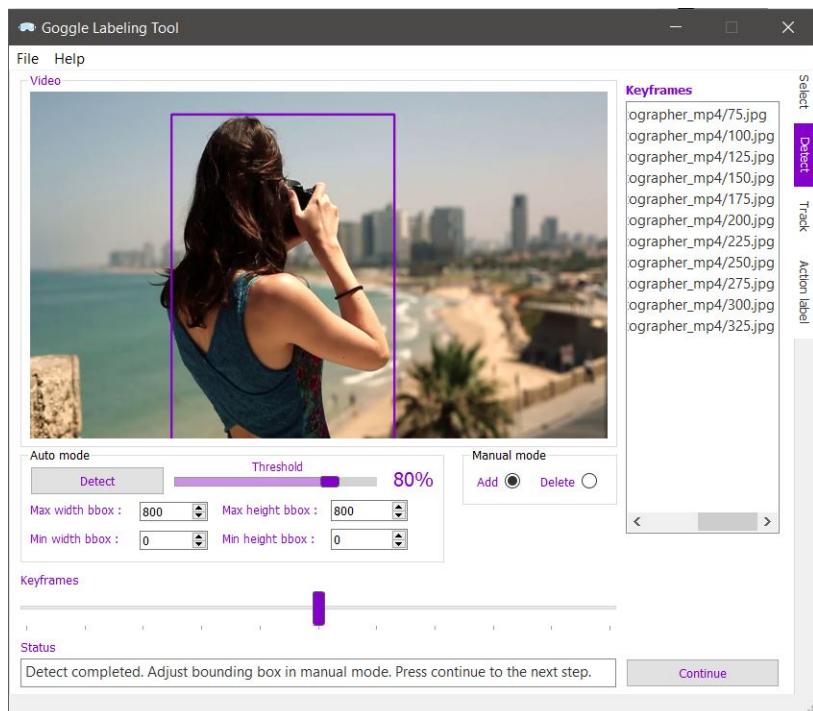
ผลลัพธ์การทำงานของหน้าต่าง Select



รูปที่ 4.5: รูปผลลัพธ์การแยกเฟรมที่มีมนุษย์อยู่ และไม่มีมนุษย์อยู่ภายในเฟรม

ขั้นตอนแรกแอปพลิเคชันจะสกัดแยกวิดีโอออกเป็นเฟรมทั้งหมด และ ทำการสัมคัญเฟรมอ กมาตามความถี่ที่ผู้ใช้งานตั้งไว้ จากนั้นแอปพลิเคชันจะนำโมเดล yolo-v3 มาตรวจสอบว่าแต่ละคัญเฟรมมีเฟรมใดบ้างที่มีมนุษย์อยู่ภายในเฟรม จากนั้นจะทำการแยกเฟรมที่มีมนุษย์อยู่ และ ไม่มีมนุษย์อยู่ ดังรูป 4.5

ผลลัพธ์การทำงานของหน้าต่าง Detect



รูปที่ 4.6: รูปคัญเฟรมที่ถูกตีกรอบสีเหลืองในส่วนที่มีมนุษย์อยู่

แอ��พลิเคชันจะนำคีย์เฟรมที่มีนุชย์ที่ได้จากหน้าต่าง Select นำมาตีกรอบสี่เหลี่ยมในส่วนของเฟรมที่มีมนุชย์อยู่โดยสามารถใช้โหมดการทำงานแบบบอตโนมัติหรือแบบแก้ไขเองก็ได้ ซึ่งผลลัพธ์ที่ได้จะได้คีย์เฟรมที่มีกรอบสี่เหลี่ยม ดังรูป 4.6 จากนั้นจะบันทึกข้อมูลในไฟล์ .txt

ผลลัพธ์การทำงานของหน้าต่าง Track



(ก) ตัวอย่างเฟรมที่ถูกตีกรอบสี่เหลี่ยม

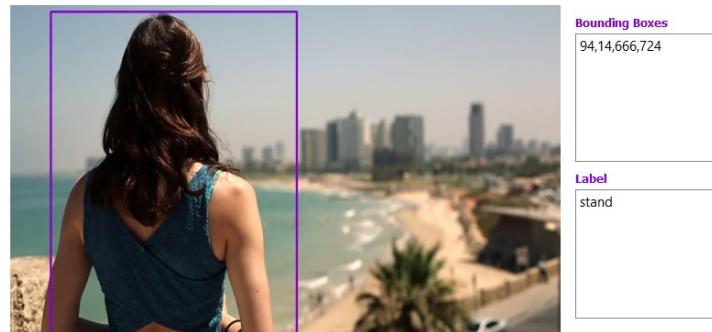
```
<?xml version="1.0"?>
- <annotation>
  <folder>D:/Goggle/Goggle_team/out/Photographer_mp4/img</folder>
  <filename>75.jpg.txt</filename>
  <path>D:/Goggle/Goggle_team/out/Photographer_mp4/img/75.jpg</path>
  - <source>
    <database>Unknown</database>
  </source>
  - <size>
    <width>1280</width>
    <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 ซึ่งผลลัพธ์ที่ได้จะได้เฟรมทุกเฟรมที่มีมนุชย์อยู่ จะถูกตีกรอบสี่เหลี่ยม ดังรูป 4.8 จากนั้นสามารถบันทึกข้อมูลออกเป็นไฟล์ XML ได้ดังรูป

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>
```

(ก) ตัวอย่างไฟล์ XML

รูปที่ 4.8: รูปผลลัพธ์การทำงานของหน้าต่าง Label

แอพพลิเคชันจะนำกรอบสีเหลืองของทุกเฟรมที่มีมนุษย์อยู่มาทำนายมนุษย์ในกรอบสีเหลืองนั้นกำลังมีการกระทำการอะไรอยู่ โดยสามารถทำงานได้ทั้งหมดอัตโนมัติหรือแบบแก้ไขเอง และสามารถบันทึกข้อมูลออกเป็นไฟล์ XML ได้ดังรูป

4.2 ผลการทดลองการตรวจจับวัตถุ

4.2.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล อ้างอิงข้อมูลจากเว็บไซต์ของ yolo

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (0.5 IOU YOLOv3 mAP)
Tiny YOLO	4	23.7
YOLOv3-tiny	4.5	33.1
SSD300	21	41.2
YOLOv3-320	22	51.5
YOLOv2 608x608	25	48.1

4.2.2 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคล

ข้อมูลผลการทำงานของโมเดลปัญญาประดิษฐ์สำหรับการทำการตรวจจับภาพบุคคลหลังจากการทดลอง

	ความเร็วต่อรูปภาพ(มิลลิวินาที)	ความแม่นยำ (0.5 IOU YOLOv3 mAP)
Tiny YOLO	X	X
YOLOv3-tiny	X	X
SSD300	X	X
YOLOv3-320	X	X
YOLOv2 608x608	X	X

4.3 ผลการทดสอบการทำนายตำแหน่งต่อไปของมนุษย์

4.4 ผลการทดสอบการระบุตัวตนของมนุษย์

4.5 ผลการทดสอบการจดจำการกระทำการของมนุษย์

4.5.1 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่าน AVA เทียบผลลัพธ์กับแหล่งอ้างอิง ได้ผลการทดลองดังตารางต่อไปนี้

	ความเร็วต่อรูปภาพ(วินาที)	ความแม่นยำ (PASCAL mAP)
แหล่งอ้างอิง	0.93	11
ผลการทดสอบของผู้วิจัย	X	X

4.5.2 ผลการทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกトレนด์ผ่าน AVA และใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์กับแหล่งอ้างอิง

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

4.5.3 ทดสอบประสิทธิภาพการทำงานของโมเดลปัญญาประดิษฐ์ที่เคยถูกเทรนด์ผ่านชุดข้อมูลสำหรับการเทรน์ที่ผู้วิจัยสร้างขึ้น และ ใช้ชุดข้อมูลที่ผู้วิจัยสร้างขึ้น ในการทดสอบและเทียบผลลัพธ์การทดสอบก่อนหน้า

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

ເອກສາຣອ້າງອີງ

- [1] Optical flow.
- [2] 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.
- [3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.
- [4] 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.
- [5] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In Scandinavian conference on Image analysis, pages 363–370. Springer, 2003.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.
- [11] 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.
- [12] 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.

ภาคผนวก